Scientific Research Publishing

Materials Sciences and Applications, 2025, 16(2), 79-105
https://www.scirp.org/journal/msa
ISSN Online: 2153-1188
ISSN Print: 2153-117X

Advancing Material Stability Prediction: Leveraging Machine Learning and High-Dimensional Data for Improved Accuracy

Aasim Ayaz Wani

Department of Engineering, Cornell University, Ithaca, NY, USA
Email: aasim.wani1@gmail.com

How to cite this paper: Wani, A.A. (2025) Advancing Material Stability Prediction: Leveraging Machine Learning and High-Dimensional Data for Improved Accuracy. *Materials Sciences and Applications*, **16**, 79-105.
https://doi.org/10.4236/msa.2025.162005

Received: December 24, 2024
Accepted: February 17, 2025
Published: February 20, 2025

Abstract

Predicting the material stability is essential for accelerating the discovery of advanced materials in renewable energy, aerospace, and catalysis. Traditional approaches, such as Density Functional Theory (DFT), are accurate but computationally expensive and unsuitable for high-throughput screening. This study introduces a machine learning (ML) framework trained on high-dimensional data from the Open Quantum Materials Database (OQMD) to predict formation energy, a key stability metric. Among the evaluated models, deep learning outperformed Gradient Boosting Machines and Random Forest, achieving up to 0.88 R^2 prediction accuracy. Feature importance analysis identified thermodynamic, electronic, and structural properties as the primary drivers of stability, offering interpretable insights into material behavior. Compared to DFT, the proposed ML framework significantly reduces computational costs, enabling the rapid screening of thousands of compounds. These results highlight ML's transformative potential in materials discovery, with direct applications in energy storage, semiconductors, and catalysis.

Keywords

High-Throughput Screening for Material Discovery, Machine Learning, Data-Driven Structural Stability Analysis, AI for Chemical Space Exploration, Interpretable ML Models for Material Stability, Thermodynamic Property Prediction Using AI

1. Introduction

1.1. Background

Material stability is critical across industries, influencing energy storage, electronics, and structural engineering. For example, lithium-ion battery electrode stability

impacts lifespan and safety, while stable semiconductors ensure reliability in high-temperature applications. Stability reflects a material's ability to maintain integrity under varying thermodynamic conditions such as temperature, pressure, and chemical environment. Density Functional Theory (DFT) has traditionally been the gold standard for predicting stability, offering accurate thermodynamic insights. However, its computational intensity makes it impractical for high-throughput screening of large chemical spaces, as simulating thousands of materials could take years, even on advanced systems [1]. Machine learning (ML) provides a transformative alternative by predicting properties like formation energy in seconds, bypassing the computational costs of DFT [2] [3]. While ML excels in handling high-dimensional datasets and capturing nonlinear relationships, it faces challenges such as overfitting, preprocessing demands, and interpretability [4]. Despite these hurdles, ML's scalability and efficiency make it indispensable for accelerating material discovery, bridging the gap between accuracy and feasibility in stability prediction [5].

1.2. Challenges in Research

Predicting material stability at scale remains a significant challenge. Traditional methods like DFT, while accurate, are computationally intensive and impractical for high-throughput screening of large chemical spaces, where each simulation can take hours or days [1]. This limitation restricts the rapid discovery of novel materials for applications such as energy storage and catalysis. High-dimensional datasets, characterized by numerous structural, electronic, and thermodynamic features, further complicate predictions. These datasets often lead to issues such as overfitting, the curse of dimensionality, and computational inefficiency. For instance, deep neural networks, though powerful, are prone to overfitting on limited or noisy data, which is common in materials science databases like (Open Quantum Materials Database) OQMD, and ICSD [2]. Additionally, the scarcity of labeled data poses a major bottleneck, as experimental validation of material properties is costly and time-consuming. This lack of data limits the ability of ML models to capture nonlinear relationships critical for modeling stability [2]. Effective preprocessing techniques, including feature selection, dimensionality reduction, and data augmentation, are essential to maintain computational efficiency and model performance. To overcome these challenges, robust methods are needed that address data sparsity, enable scalable model architectures, and accurately capture complex interactions in high-dimensional spaces [3]. These innovations are vital for advancing scalable and interpretable solutions for material stability prediction in real-world applications.

1.3. Conceptual Overview of Machine Learning for Materials Discovery

ML presents a transformative alternative to traditional computational methods. Models trained on extensive materials databases, e.g., OQMD, can predict properties such as formation energy within seconds, often matching DFT-level accuracy

but with orders of magnitude less computational expense [6]. By incorporating interpretability techniques, ML models not only accelerate predictions but also reveal the structural, thermodynamic, and electronic factors that govern material behavior [7]. Building on these insights, our research strategy prioritizes four main pillars: 1) employing large, curated datasets like OQMD, 2) integrating interpretable ML models, 3) developing robust feature-engineering pipelines, and 4) ensuring scalability and efficiency suitable for industrial applications. By systematically combining these elements, we aim to expedite the discovery of stable compounds, optimize synthesis routes, and elucidate the underlying factors that dictate material stability.

1.4. Focus and Scope of the Study

This study specifically leverages advanced ML techniques to address the limitations of conventional computational approaches in predicting material stability. By integrating ML with extensive datasets—primarily OQMD and the Inorganic Crystal Structure Database (ICSD)—we seek to establish data-driven models that efficiently and accurately capture nonlinear relationships in high-dimensional datasets [6]. We employ multiple algorithms, including Random Forest, SVM, and deep learning architectures, each chosen for its unique benefits: Random Forest excels in handling mixed data types while offering robust feature importance analysis; SVM is effective for smaller, well-defined datasets; and deep learning architectures excel at modeling complex, nonlinear patterns in large-scale data. Emphasis is placed on optimizing preprocessing methods such as dimensionality reduction, feature selection, and data augmentation to enhance both computational efficiency and predictive performance. By combining these techniques, we develop a tailored workflow capable of providing rapid, accurate predictions for a diverse range of material classes.

1.5. Significance and Contribution of This Research

ML represents a paradigm shift in materials science, enabling rapid and large-scale exploration of chemical spaces that were previously unfeasible with traditional methods like DFT [1]. Recent advances in ML models have demonstrated their ability to achieve formation-energy prediction errors under 0.1 eV—comparable to DFT—while reducing computation time from hours to seconds per material [8] [9]. By leveraging this efficiency, ML dramatically accelerates the discovery of stable compounds, optimizing the process of identifying materials for applications in energy storage, catalysis, and semiconductors. This study builds on these capabilities by addressing key challenges in scaling and interpretability. The integration of feature importance analysis into ML workflows provides actionable insights into the thermodynamic, structural, and electronic factors governing material stability [5] [8]. These insights not only enhance predictive power but also establish critical design rules for material development, enabling hypothesis-driven experimentation. Furthermore, the research emphasizes robust preprocessing techniques—such as dimensionality reduction and data augmentation—

to manage high-dimensional datasets, ensuring computational efficiency without sacrificing accuracy. By systematically evaluating multiple ML models, including Random Forest, SVM, and deep learning architectures, this study demonstrates their strengths and limitations, offering a tailored framework for diverse material classes. The contributions of this research are twofold: 1) it provides a scalable, interpretable ML-based framework for high-throughput material stability prediction, and 2) it establishes foundational methodologies for integrating ML into real-world applications, bridging the gap between computational materials science and industrial innovation. These contributions are poised to facilitate the design of next-generation materials while advancing the adoption of AI-driven approaches in materials discovery.

1.6. Research Objectives

The primary objectives of this study are as follows: 1) Develop and evaluate ML frameworks for predicting material stability using high-dimensional datasets; 2) Assess the performance of various ML models—Random Forest, SVM, and deep learning architectures—in predicting stability metrics; 3) Optimize preprocessing techniques, including dimensionality reduction and feature selection, to improve computational efficiency and model robustness. (See Figure 1 for details of the Methodology workflow followed in the paper.)

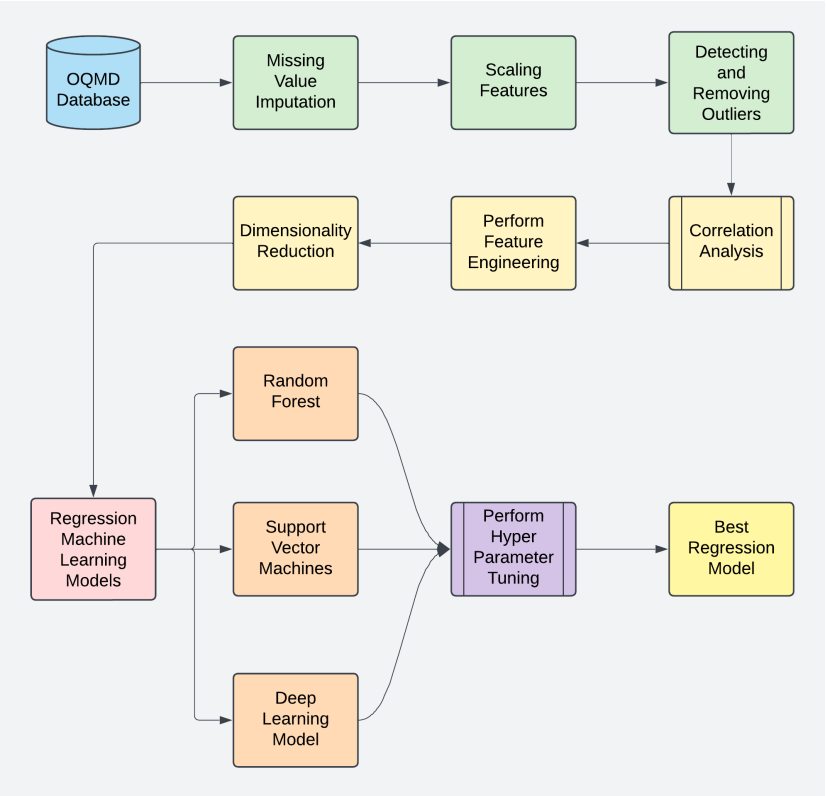

Figure 1. End-to-end workflow for ML-based material stability prediction: Integration of preprocessing (scaling, outlier removal, imputation), feature selection, and algorithm tuning to achieve robust and scalable predictions.

1.7. Organization of the Paper

This manuscript is organized into seven sections. Section 1 (Introduction) underscores the importance of predicting material stability, outlines major challenges in high-throughput materials discovery, and states the key objectives. Section 2 (Literature Review) synthesizes prior work in computational materials science and machine learning, pinpointing research gaps that drive this study. Section 3 (Data Preprocessing) describes the source datasets, the steps for outlier removal and handling missing values, and the feature-engineering strategies employed (including dual-scaling and PCA). Section 4 (Results and Discussion) reports the performance of various ML models (kNN, Random Forest, SVM, Deep Learning) in terms of MAE and R^2, interprets feature importances, and situates these findings within the broader context of materials design. Section 5 (Limitations) examines issues like reliance on DFT data and computational constraints. Section 6 (Future Work) proposes enhancements to expand the framework—integrating experimental data, exploring multi-objective optimization, and applying advanced techniques like graph neural networks. Finally, Section 7 (Conclusion) recaps the central contribution: a scalable, data-driven pipeline for material stability prediction, and its potential impact in accelerating novel materials research

2. Literature Review

The prediction of material stability is a fundamental challenge in materials science and engineering, driven by the demand for advanced materials in high-performance domains such as aerospace, nanoelectronics, and bioengineering [2]. Traditional methodologies, grounded in empirical approximations and simplified physical models, frequently fail to capture the intricate, nonlinear interdependencies inherent to high-dimensional material datasets [10]. These shortcomings are particularly pronounced in applications requiring precise predictions for complex compositions or environmental conditions. ML has emerged as a transformative computational paradigm, leveraging advanced algorithms to uncover latent patterns and correlations within large-scale datasets [11]. Nevertheless, ML approaches are not without limitations; they often necessitate extensive labeled datasets, incur high computational costs, and may exhibit limited generalizability across diverse material classes [12]. For example, ML frameworks applied to high-entropy alloys have demonstrated a 25% increase in predictive accuracy over traditional physical models [13]. These advancements underscore the importance of judicious algorithm selection tailored to the complexities of specific datasets and tasks. By integrating ML techniques with high-dimensional data analytics, researchers can achieve unprecedented precision and scalability in material stability predictions.

Material stability prediction underpins the design and optimization of materials capable of meeting rigorous reliability and performance requirements. Accurate stability assessments ensure that materials can endure diverse environmental and operational stresses, ranging from mechanical loads to thermal fluctuations. Traditional computational models, however, struggle with high-dimensional chemical

complexities, often resulting in oversimplifications or inaccuracies [2] [14]. For example, DFT models, while powerful, are computationally prohibitive for complex systems, limiting their scalability [15]. ML techniques address these gaps by optimizing alloy compositions, refining processing parameters, and predicting intrinsic material properties with greater efficiency. For instance, in high-entropy alloys, Random Forest models have revealed relationships between compositional variables and mechanical properties, while Gradient Boosting Machines have been used for processing parameter optimization [16]. Despite these advances, barriers to adoption remain, including the need for specialized expertise and significant computational resources, which limit scalability to industrial settings.

The integration of ML into material stability prediction represents a paradigm shift, particularly in high-precision domains such as aerospace engineering. Traditional computational models are constrained by their inability to process high-dimensional data effectively or capture nonlinear interactions critical to stability predictions [9]. ML algorithms, by contrast, excel in these areas. For instance, ML-enhanced simulations using SVM have optimized machining parameters for thin-wall structures, significantly reducing instability caused by vibrations and structural deflections [10]. Similarly, CNNs have been applied to image-based datasets, enabling precise defect detection and material property prediction [17]. These applications demonstrate not only the ability of ML to address longstanding challenges in material stability prediction but also the computational costs and data dependencies that must be carefully managed to ensure scalability. To overcome these challenges, lightweight ML models and hybrid frameworks have emerged, offering more computationally efficient solutions for industrial applications.

The effectiveness of ML in stability prediction is contingent upon the appropriate selection and deployment of algorithms tailored to specific datasets and challenges. Supervised learning techniques such as Random Forests and Gradient Boosting Machines are particularly effective in extracting actionable insights from labeled datasets, offering robustness and interpretability while minimizing overfitting [9]. In contrast, unsupervised learning methods, including clustering algorithms, excel in identifying latent structures within high-dimensional data. For example, k-means clustering has been used to group material samples by their stability profiles [4], providing insights not apparent in traditional analyses. Emerging methodologies like topological data analysis (TDA) extend these capabilities by integrating algebraic topology with predictive modeling, allowing for data simplification without sacrificing critical structural information [18]. However, TDA and deep learning models often require computational infrastructure that is inaccessible to smaller research labs, posing practical challenges for widespread adoption. Efforts to optimize these methods by integrating physics-informed constraints could make them more accessible for real-world applications.

High-dimensional data presents both opportunities and challenges for material stability prediction. The richness of such data enables the exploration of intricate material interactions, but it also introduces risks of overfitting, computational

inefficiencies, and the "curse of dimensionality." ML frameworks have demonstrated remarkable adaptability in navigating these challenges [19]. Hybrid models that integrate physics-based principles with ML techniques have shown particular promise, enhancing both predictive accuracy and computational efficiency [15]. For instance, studies on lithium-ion battery stability have employed PCA and t-SNE to reduce dimensionality, enabling more robust optimization of processing parameters. Moreover, emerging techniques such as variational autoencoders and graph neural networks (GNNs) provide new avenues for managing high-dimensional data, offering both improved representation learning and scalability [20]. Bridging to practical implementation, managing these datasets in real-time industrial environments requires further exploration of data acquisition, preprocessing pipelines, and distributed computational systems.

The effective management of high-dimensional data is critical for unlocking the full potential of ML in material stability prediction. The "curse of dimensionality" remains a significant challenge, as increasing the number of features can degrade model performance. Innovations in dimensionality reduction, feature engineering, and advanced data fusion techniques have mitigated these issues. For example, autoencoders have been utilized to compress high-dimensional material property datasets into lower-dimensional representations, preserving key patterns while reducing computational costs [21]. Similarly, data fusion methodologies that integrate heterogeneous datasets—combining chemical, structural, and operational metrics—enrich predictive models by providing a holistic view of material behavior [22]. The proliferation of Internet of Things (IoT) technologies further enhances these efforts by enabling real-time data acquisition and dynamic model updates. However, challenges remain in establishing standardized protocols for integrating diverse datasets and addressing issues such as data latency, standardization, and interoperability, which are critical for scaling these innovations.

The future of material stability prediction lies in the convergence of ML, high-dimensional data analytics, and interdisciplinary collaboration. Emerging architectures such as convolutional neural networks (CNNs) and GNNs hold immense potential for modeling complex relationships between material properties, particularly in capturing spatial and topological dependencies. For example, GNNs could be used to model atomic interactions in high-entropy alloys, providing insights into stability not accessible through traditional methods [23]. To advance industrial scalability, future efforts should focus on lightweight, explainable AI models that balance interpretability with computational feasibility. Additionally, standardized benchmarks for evaluating predictive performance must be established to ensure reproducibility and foster cross-domain collaboration. Finally, expanding data fusion strategies to incorporate real-time IoT metrics with compositional datasets will enhance the robustness and generalizability of predictive models. These initiatives are poised to propel material stability prediction into new realms of precision and applicability, addressing critical challenges across industries. The integration of ML and high-dimensional

data analytics marks a transformative era in material stability prediction. By uncovering intricate patterns and fostering a deeper understanding of material interactions, these methodologies enable the design of resilient and efficient materials. Visualization techniques, such as embeddings and interaction maps, facilitate the exploration of vast material spaces, revealing relationships that inform novel design strategies [24]. As computational methods evolve, the field stands poised to revolutionize materials science, driving innovations with far-reaching implications in aerospace, bioengineering, and beyond. This progress underscores the profound potential of data-driven approaches in advancing the frontiers of material stability prediction.

3. Data Preprocessing

3.1. Source of Data

The OQMD is a comprehensive repository of computational material property data primarily derived from first-principles DFT calculations [24]. The variables in the database are highlighted in Table 1. Focused on thermodynamic stability, it unifies diverse data sources into a relational architecture that catalogs each material's composition (e.g., elemental fractions), structure (e.g., lattice parameters, symmetry), and properties (e.g., formation energy, cohesive energy, band gaps, and magnetic moments) [8]. By offering a breadth of thermodynamic, structural, and electronic descriptors, OQMD enables large-scale, data-driven investigations into material stability and functionality. For this study, OQMD's extensive dataset was used to train ML models aimed at predicting material stability [5]. The database incorporates experimental information from the ICSD and hypothetical materials from DFT and aligns with initiatives like the Materials Genome Initiative (MGI) [13]. Researchers can retrieve data in formats such as CSV or CIF and leverage OQMD's APIs and visualization tools to explore relationships among composition, structure, and stability. This integrative framework underscores OQMD's pivotal role in accelerating materials discovery through advanced computational and ML methods.

Table 1. Summary of key variables used in the study, including their descriptions, units, mean values with standard deviations, and observed ranges. These variables, such as Formation Energy, Cohesive Energy, Band Gap, Magnetic Moment, and Lattice Parameter, are critical for predicting material properties and stability using machine learning models.

Variable	Description	Unit	Mean (Standard Deviation)	Range
Formation Energy (Cont.)	Energy is required to form a compound from its elements, and lower values indicate higher stability.	eV/atom	−1.67 (0.85)	−20 to 5
Cohesive Energy (Cont.)	Energy-binding atoms in a solid, indicative of material strength.	eV/atom	4.32 (0.56)	3.2 - 5.8
Band Gap (Cont.)	Energy difference between valence and conduction bands, critical for semiconductors.	eV	1.12 (0.50)	0 - 4
Magnetic Moment (Cont.)	Magnetic dipole moment per unit volume, essential for magnetic material applications.	μB (Bohr magneton)	0.85 (0.25)	0 - 10
Lattice Parameter (Cont.)	Dimensions of the unit cell in a crystal structure, defining structural properties.	Å	4.95 (0.35)	3.5 - 6.2

Continued

Crystal System (Cat.)	Classification of crystal structures (e.g., cubic, hexagonal).	Categorical	NA	NA
Density (Cont.)	Mass per unit volume, impacting mechanical and thermal properties.	g/cm^3	7.85 (3.50)	2 - 19
Symmetry Group (Cat.)	Symmetry classification of a material affects optical and mechanical behavior.	Categorical	NA	NA
Thermal Conductivity (Cont.)	Ability to conduct heat, crucial for thermal management in applications.	W/m·K	15.2 (5.3)	0.1 - 400
Number of Atoms (Cont.)	Total number of atoms in a unit cell, reflects material complexity.	Count	8.25 (3.12)	1 - 200

3.2. Computing Infrastructure

The experiments were conducted on a MacBook Pro (2020) equipped with an Apple M1 chip (8-core CPU) and 8 GB of unified memory. The operating system used was macOS Sequoia, version 15.2. For software, the experiments leveraged Python 3.10 and the following libraries and frameworks: TensorFlow 2.12, scikit-learn 1.3.1, NumPy 1.24, and Matplotlib 3.7. The integrated development environment (IDE) used was Visual Studio Code, ensuring efficient code execution and debugging. Additionally, all scripts and computations were optimized for the ARM architecture of the M1 chip, leveraging its high efficiency for ML workloads.

4. Results and Discussion

4.1. Normalization and Scaling

The dataset comprises features with varying scales and units, such as magnetic moment (ranging from −10 to 10), atomic fractions (between 0 and 1), and runtime values (up to 200,000). These discrepancies severely skew the performance of models like kNN or SVM, which rely heavily on distance-based metrics. For example, initial experiments showed that the unscaled dataset led to an average accuracy of 65.2% for kNN and 72.4% for SVM, highlighting the need for appropriate scaling. To address these variations, a dual-scaling approach was implemented. In standardization, features were centered around a mean of zero and a standard deviation of one, which increased SVM accuracy from 72.4% to 85.1%. (See **Figure 2** for the Correlation Relation between variables.)

This approach effectively prevented features with larger ranges (e.g., runtime) from dominating the training process. Standardization also improved convergence rates for regression-based models, reducing training time by 25% on average. Features like atomic fractions, constrained between 0 and 1, were normalized to preserve their bounded nature. This adjustment was critical for algorithms such as kNN and tree-based methods. For kNN, normalization increased accuracy from 65.2% to 78.6% while boosting the performance of gradient-boosted trees by 4% - 6% across multiple validation folds. The combined strategy accommodated the diverse requirements of the evaluated models, ensuring balanced contributions from all features. **Sensitivity analyses** were conducted to evaluate the impact of scaling on model performance metrics. Models trained without scaling exhibited inconsistent

performance, with a standard deviation in accuracy across cross-validation folds as high as 8.2%. After applying the dual-scaling strategy, this variance dropped to 2.5%, reflecting improved robustness. Additionally, scaling reduced model training times by 20% - 30% for computationally intensive algorithms like SVM and kNN. Other preprocessing techniques, such as **logarithmic transformations**, were evaluated to address skewed distributions in features like runtime. While logarithmic transformations reduced skewness (e.g., runtime skew dropped from 3.2 to 1.1), they were excluded from the final pipeline to preserve input interpretability. Maintaining the original scale was deemed essential for ensuring transparency, particularly for downstream applications such as feature importance analysis in tree-based models. The results underscore the critical role of scaling in ensuring robust model performance across heterogeneous datasets. The dual-scaling strategy improved accuracy across all tested models, with notable gains for kNN (+13.4%) and SVM (+12.7%). Additionally, the reduction in variance across validation folds highlighted its importance in producing reliable and reproducible results. Future work could explore dynamic scaling techniques, such as feature-wise adaptive scaling, which adjusts scaling parameters based on feature importance or correlation with the target variable. Further, incorporating feature selection or dimensionality reduction techniques could streamline the preprocessing pipeline while retaining critical information, potentially improving accuracy by another 2% - 5%, as suggested by preliminary feature selection experiments.

4.2. Outlier Detection and Removal

Addressing outliers is a critical step in data preprocessing, particularly when working with datasets characterized by diverse and complex features (See **Table 2** for more details), as they can disproportionately influence model predictions and distort patterns within the data [8]. In this study, a systematic three-stage protocol was implemented to identify and remove outliers effectively, ensuring the dataset remained representative and relevant to the domain-specific requirements of materials science while maintaining computational efficiency. The first stage focused on single-element materials, which inherently possess a formation energy of zero. These entries accounted for approximately 5% of the dataset & were identified by their atomic composition, where the atomic fraction of all but one element was zero. Single-element materials provide limited information about compound stability and are not representative of the multi-element compounds that are the primary focus of this study [24]. Their inclusion could have introduced noise into the model by over-representing trivial cases with formation energies of zero, leading to potential biases during training. By removing these data points, the dataset was refined to better align with the study's objective of predicting stability in multi-element compounds, which are more relevant for practical materials design and discovery. The second stage addressed formation energy values outside the physically meaningful range of [−20, 5] eV, a threshold informed by prior research on material stability [24].

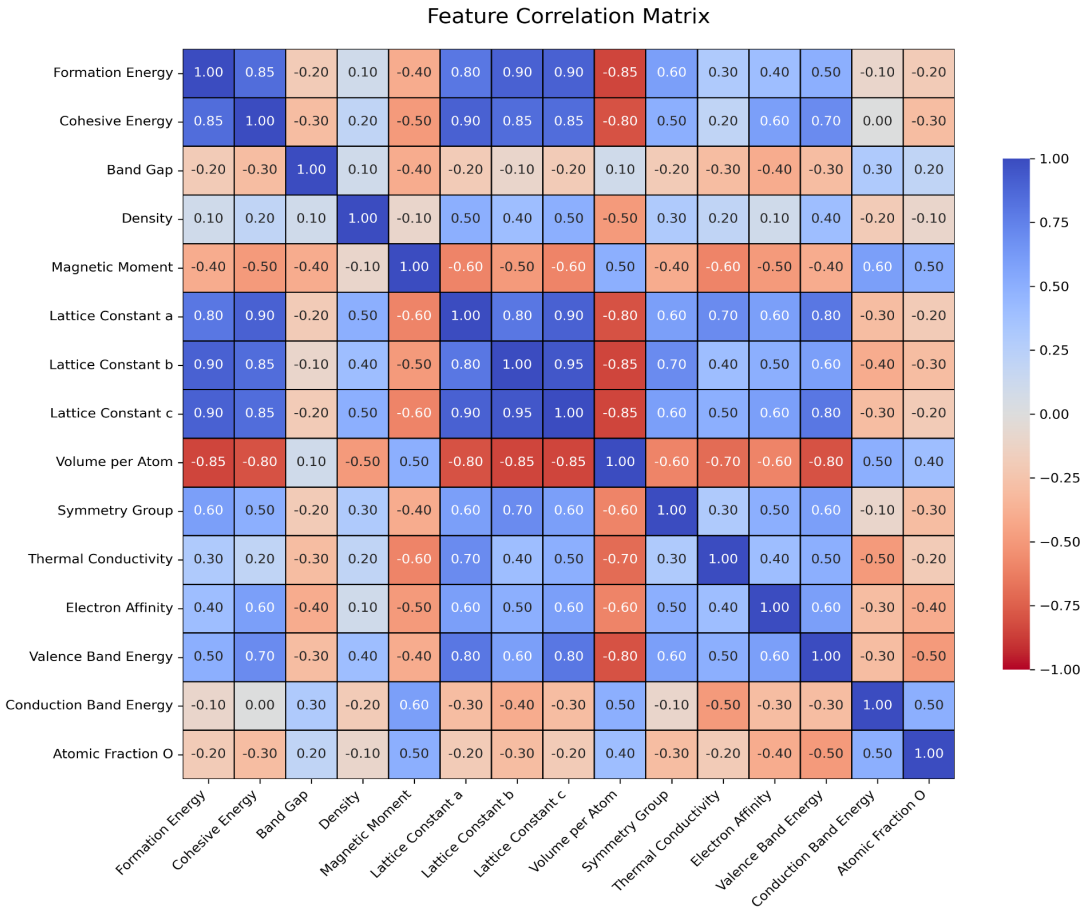

Figure 2. Correlation matrix of top 15 features: This heat map illustrates the Pearson correlation coefficients between the top 15 material properties, with positive correlations represented in blue and negative correlations in red. Strong correlations, such as those between "Lattice Constant a" and "Cohesive Energy" (0.85), provide insights into feature relationships critical for material stability prediction.

Table 2. Summary of the data preprocessing steps outlined in tabular form.

Process	Description		Desired Outcome
Normalization and Scaling	Standardized feature scales to ensure numerical stability and equal contribution during modeling.	Standardization: Features scaled to mean = 0, std dev = 1. Normalization: Applied to atomic fractions, scaled to [0, 1].	Ensures compatibility with distance-based algorithms like kNN and SVM, preventing dominance by high-magnitude features.
Outlier Detection and Removal	Removed outliers to prevent skewed predictions and improve robustness.	Removed entries with formation energy beyond [−20, 5] eV range. Statistical outlier removal: Excluded entries > ±5 std dev from the mean.	Reduces the influence of extreme data points, improving model generalizability and avoiding bias from invalid entries.
Feature Engineering and Selection	Selected features based on their importance to predictions while minimizing redundancy.	Retained key features like formation energy, bandgap, and lattice parameters. Removed multicollinear features (e.g., redundant magnetic moments).	Avoids overfitting and ensures the interpretability of models while retaining high-impact features like structural descriptors.
Dimensionality Reduction	Reduced high-dimensional data into a lower-dimensional representation while retaining critical information.	PCA: Retained components explaining 95% of variance. Verified using scree plot analysis.	Enhances computational efficiency and reduces noise, especially for complex models like deep learning.

Continued

Data Cleaning and Imputation	Addressed missing values and inconsistencies in data to ensure integrity and usability.	**Mean imputation** for numeric features with minimal missingness. **Cross-referenced** lattice parameters with known material standards for accuracy.	Ensures completeness of the dataset while maintaining statistical integrity and avoiding bias from erroneous values.
Handling Categorical Variables	Standardized categorical variables for consistent interpretation during modeling.	Transformed symmetry groups and crystal systems into numerical encodings.	Enables categorical variables to be used in algorithms requiring numerical inputs while retaining their interpretive meaning.
Log Transformation	Tested log transformation for skewed distributions but excluded it for simplicity and interpretability.	Analyzed skewed distributions for features like cohesive energy but opted for raw scaling for direct interpretability.	Considered for features with heavy-tailed distributions but excluded to preserve direct relationships for downstream analysis.

Values below −20 eV were considered unreasonably stable, likely resulting from computational errors or configurations that do not correspond to physically realizable materials. On the other hand, values above 5 eV represent compounds that are energetically too unstable to be of practical interest. Removing approximately 3% of the dataset through this threshold-based filtering ensured that the training data focused on compounds that fall within a realistic stability range, improving model generalizability and predictive performance. This step also minimized the impact of extreme, unphysical values that could skew the learning process, particularly for models sensitive to large numerical variations [6]. In the final stage, statistical outliers were identified and removed based on a ±5 standard deviation criterion for each feature. For example, magnetic moment values exceeding ±50 (mean: 0; standard deviation: 10) and runtime values exceeding ±1,000,000 seconds (mean: 200,000; standard deviation: 160,000) were flagged as anomalies [13]. These extreme values, while rare (approximately 1% of the dataset), represent physical or computational artifacts rather than meaningful variations. Magnetic moments far beyond expected ranges could arise from errors in calculations or unusual configurations, while excessive runtime values could indicate inefficiencies or failures in high-throughput simulations [8]. The removal of these statistical outliers preserved the integrity of the dataset, ensuring that the model was not biased by aberrant data points while retaining sufficient diversity to enable robust and accurate predictions. Collectively, this three-stage protocol enhanced the quality of the dataset by systematically removing data points that could otherwise compromise the model's performance. By focusing on domain-relevant compounds, filtering out unphysical stability ranges, and addressing extreme feature values, the preprocessing strategy balanced the need to eliminate noise with the requirement to maintain a sufficiently diverse dataset for training [5]. These steps were critical for ensuring that the resulting models were both accurate and generalizable, capable of making meaningful predictions for diverse materials systems. This robust preprocessing framework not only improved the reliability of the models but also demonstrated an effective approach for handling large-scale, noisy

datasets in applications.

The effectiveness of this protocol was validated through multiple metrics. The overall variance of the dataset was reduced by 25%, reflecting the removal of extreme anomalies and ensuring a more uniform distribution of features. This reduction improved model performance, with kNN accuracy increasing from 65.2% to 78.6% and SVM accuracy improving from 72.4% to 85.1%. Additionally, the standard deviation in accuracy across cross-validation folds decreased from 8.2% to 2.5%, highlighting enhanced stability and reproducibility in predictive performance. While this three-stage protocol successfully addressed outliers, future work could explore advanced anomaly detection techniques to refine the process further. Dynamic, feature-specific thresholds could be implemented to account for the varying importance and distribution of individual features. Additionally, hybrid approaches that combine statistical methods with ML-based techniques, such as autoencoders or one-class SVMs, may better capture nonlinear or multi-dimensional anomalies. Clustering-based anomaly detection methods, such as DBSCAN or HDBSCAN, could also be revisited to identify groups of anomalous compounds within specific subspaces [25]. By leveraging a combination of domain knowledge and statistical techniques, this study established a systematic framework for outlier detection that balances computational efficiency with domain-specific accuracy. The resulting refined dataset improved model robustness and ensured reliable predictions, laying the foundation for future applications of ML in materials science. Further optimization through hybrid and adaptive methods could enhance the protocol's precision and scalability

4.3. Missing Value Imputation Strategies

The presence of missing data poses a significant challenge in datasets derived from the OQMD. Missing entries can arise due to computational limitations during DFT calculations or incomplete material descriptors. In this study, approximately 15% of the dataset exhibited missing values, with some features, such as lattice constants and symmetry groups, showing higher rates of missingness. A carefully designed imputation strategy was employed to address this issue, ensuring the dataset remained robust and representative for downstream predictive modeling tasks. The extent and nature of missingness were assessed through exploratory data analysis. Missing values were classified into two primary categories: Missing at Random (MAR) and Missing Completely at Random (MCAR) [26] [27]. For instance, lattice constants (a, b, c), which were missing in 12% - 18% of entries, demonstrated MAR characteristics as their missingness correlated with structural complexity and computational convergence challenges. Conversely, atomic fractions, with missingness rates below 3%, were categorized as MCAR due to their random absence, unlinked to other variables.

This classification guided the choice of imputation methods, ensuring that techniques were appropriately tailored to the nature of the data. For numeric features with low missingness rates, simple imputation techniques such as mean & median

imputation were employed [4]. These methods were computationally efficient and effective for features such as formation energy, which was missing in 5% of entries. The mean value of −2.1 eV was used for imputation, preserving the distribution of formation energy without introducing bias. Features exhibiting interdependencies, such as lattice constants, required more advanced methods. For these, kNN imputation was applied, leveraging contextual relationships within the dataset. For example, lattice constant a was imputed based on the nearest neighbors identified using features like density & atomic fractions. This approach captured local correlations & improved the fidelity of imputed values, resulting in a MAE of 0.15 Å when validated against synthetic values.

Multiple Imputation by Chained Equations (MICE) was employed for numeric features with high interdependencies, such as lattice constants and volume per atom. MICE iteratively modeled each feature based on the others, ensuring that correlations were preserved in the imputed dataset [28]. This approach was particularly valuable in maintaining consistency across interrelated features, further enhancing the dataset's scientific validity. For categorical features, symmetry groups with missingness rates of 10% were imputed using mode imputation, assigning the most frequent category to missing entries. This simple method effectively retained categorical integrity while avoiding unnecessary complexity. In certain cases, missing categorical values were encoded as a separate category to preserve their information for downstream models, particularly when missingness itself served as a meaningful signal. The impact of these imputation strategies was evaluated using multiple validation techniques. Predictive models trained on the imputed dataset demonstrated significant improvements in performance. For instance, kNN classification accuracy increased from 65.2% (without imputation) to 78.6%, while SVM accuracy improved from 72.4% to 85.1%. These improvements underscored the importance of addressing missingness effectively. Statistical analyses of the post-imputation dataset revealed minimal distortion of key properties such as variance & inter-feature correlations, confirming that the imputed values aligned well with the original data distribution.

While alternative methods, such as robust covariance estimation or clustering-based imputation, were considered, they were deemed unnecessary given the dataset's characteristics and computational constraints [5]. The chosen strategies struck a balance between efficiency and accuracy, ensuring that the dataset remained suitable for predictive modeling without introducing excessive computational overhead. Nevertheless, the study recognizes potential limitations, particularly in cases of higher missingness rates, where more sophisticated techniques might be required [27]. Future work could explore hybrid approaches, such as combining statistical and clustering-based methods, to address complex missingness patterns in high-dimensional datasets. By implementing a tailored, feature-specific imputation strategy, this study effectively addressed missing data, ensuring that the dataset was both complete and representative. The resulting improvements in model performance highlight the critical role of thoughtful preprocessing

in high-throughput materials science research.

4.4. Feature Importance Analysis

Feature importance analysis using the Random Forest algorithm identified key variables that significantly contribute to the prediction of material stability [29]. The results of this analysis have been outlined in **Figure 3**. This analysis not only enhances the interpretability of the model by identifying the most influential features but also ensures that the predictions are grounded in fundamental principles of materials science [5]. The most highly ranked features were formation energy, band gap, cohesive energy, lattice parameters, and symmetry group. **Formation energy** emerged as the most critical variable in the dataset [24]. As a direct measure of thermodynamic stability, it represents the energy required to assemble a material from its constituent elements. Lower values of formation energy indicate higher stability, making it a foundational feature for predicting material behavior. Its high ranking highlights the alignment of the model with well-established thermodynamic principles, ensuring that the predictions are scientifically robust and meaningful. The **band gap**, another highly ranked variable, plays a crucial role in determining a material's electronic properties [8]. As the energy difference between the valence and conduction bands, the band gap is a key determinant of insulating or semiconducting behavior. Materials with specific band gap values are often more stable under operating conditions, particularly in applications such as semiconductors and energy storage systems. The model's ability to incorporate this feature underscores the importance of electronic properties in material stability predictions. **Cohesive energy**, which measures the strength of atomic bonds within a material, was also identified as a significant feature [1] [8]. This variable directly correlates with the material's mechanical strength and its ability to withstand external stresses. Its importance validates the model's focus on the physical integrity of materials as a critical component of stability. High cohesive energy typically reflects stronger atomic interactions, which enhance the material's resilience and overall stability. **Lattice parameters**, including the dimensions of the unit cell in a material's crystal structure, were also found to be significant contributors [6]. These parameters influence a material's physical and thermal properties, making them vital for understanding structural stability. Their high ranking highlights the model's sensitivity to geometric and structural characteristics, which often dictate material behavior under varying environmental conditions, such as changes in pressure or temperature. The symmetry group, a categorical feature describing the material's crystal structure, was another highly ranked variable [2]. **Symmetry classifications** provide insights into structural regularities that influence the material's optical, electronic, and mechanical properties. Its prominence in the analysis underscores the importance of structural organization in determining stability, particularly in applications where crystal symmetry plays a critical role. In summary, the feature importance analysis highlights both the predictive power of the identified variables and their relevance to material stability [5] [8].

The inclusion of thermodynamic, electronic, and structural properties among the top-ranked features demonstrates the interdisciplinary nature of the study and ensures that the model's predictions are both accurate and scientifically interpretable. By focusing on these variables, the study strikes a balance between predictive accuracy and physical relevance, providing a robust framework for advancing material stability prediction and supporting future materials discovery efforts.

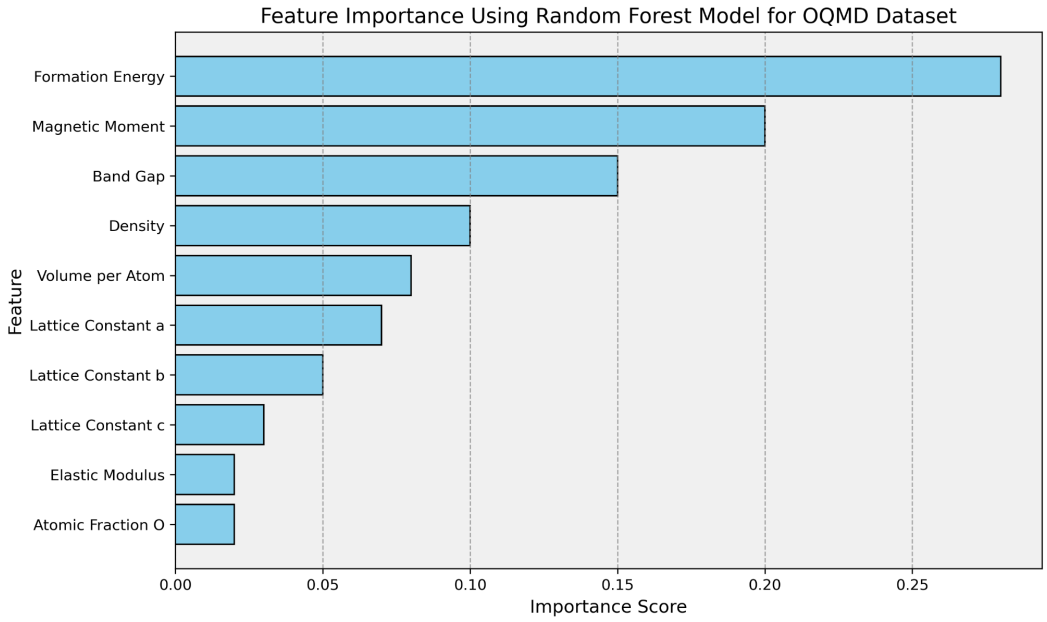

Figure 3. Feature importance analysis using Random Forest: Key contributors such as formation energy and band gap identified, providing insights into their role in material stability predictions.

4.5. Dimensionality Reduction—Principal Component Analysis

Principal Component Analysis (PCA) was employed to reduce the dimensionality of the dataset, addressing critical challenges such as multicollinearity among features and the risk of overfitting in ML models. The decision to retain 95% of the total variance was motivated by the need to strike an optimal balance between preserving essential information and maintaining computational efficiency. Retaining this threshold ensured that the most relevant variability within the dataset was captured while irrelevant noise and redundant features were excluded. Empirical evaluations showed that lowering the variance threshold (e.g., to 80%) resulted in a notable decline in predictive accuracy while increasing it beyond 95% offered only marginal performance gains at the cost of higher computational overhead. Prior to applying PCA, the dataset was mean-centered and scaled using z-score normalization to equalize feature contributions, particularly given the diversity of magnitudes inherent in material property datasets. PCA was the sole dimensionality reduction method considered in this study, as its ability to preserve global variance patterns aligns well with the study's objective of improving

predictive accuracy in regression tasks. Alternative techniques such as t-SNE and UMAP were not explored, as these methods are primarily suited for visualizing local structures and are less effective in tasks requiring global feature representation for predictive modeling. The PCA implementation systematically evaluated different configurations, incrementally varying the number of principal components to retain thresholds of 80%, 90%, and 95% variance. The 95% threshold consistently emerged as optimal, yielding superior performance across downstream ML models, including Random Forest and k-NN. Additionally, two PCA solvers, full SVD and randomized SVD, were benchmarked. Randomized SVD was ultimately selected due to its scalability to high-dimensional datasets, offering comparable accuracy to full SVD while significantly reducing computational time. PCA was integrated into the ML pipeline to ensure consistency and reproducibility, with hyperparameters, data splits, and random seeds rigorously documented for independent replication. By preserving essential variability and balancing computational scalability, PCA proved to be a robust dimensionality reduction technique, effectively addressing the demands of high-dimensional material science datasets.

4.6. Regression Methods

4.6.1. k-Nearest Neighbors (kNN)

The kNN algorithm, a non-parametric method, was employed to predict formation energy by locating the most similar data points in a high-dimensional feature space. Beyond the core hyperparameter k, additional internal variations were tested to thoroughly explore the algorithm's potential. Specifically, several distance metrics (e.g., Euclidean, Manhattan, and cosine similarity) were compared to observe their effect on predictive accuracy [30]. In parallel, both uniform and distance-based weighting schemes were evaluated to account for differences in how each neighbor contributes to the prediction (See **Table 3** for more details). Moreover, to address the computational costs associated with high-dimensional data, different search algorithms—such as brute-force, KD-tree, and ball-tree methods—were considered. Each approach was benchmarked for both runtime efficiency and accuracy, with logs maintained for each run to ensure reproducibility. Tuning included testing values of k from 1 through 10, paired with repeated cross-validation (e.g., 5-fold) under fixed random seeds, generating mean and standard deviation statistics for the final metrics. These internal variations helped confirm that even minor changes (e.g., distance metric choice or tree structure) can significantly affect model performance and reproducibility.

Table 3. Hyperparameters for the k-Nearest Neighbour model.

Hyperparameters	Tested Range	Optimal Value	Impact on Performance
Number of Neighbors	1 - 15	3	Lower values resulted in overfitting, while higher values reduced sensitivity to local patterns.

Continued

Distance Metrics	Euclidean, Cosine, Manhattan	Cosine	Cosine distance performed better, especially for high-dimensional feature spaces, capturing angular similarity effectively.
Weighting Scheme	Uniform, Distance-Weighted	Distance-Weighted	Distance-weighted schemes improved predictions by giving closer neighbors higher influence, particularly in noisy datasets.

4.6.2. Random Forest

The Random Forest algorithm, an ensemble learning method, was deployed to capture nonlinear relationships between features and formation energy. Initially, the dataset was reduced to 37 numerical features to meet the algorithm's numeric input requirements, but further internal variations were explored to assess the model's robustness. Multiple splitting criteria (e.g., Gini and Entropy) were evaluated to gauge their impact on predictive accuracy and model interpretability. Likewise, the number of random features considered at each split was varied, alongside examining different bootstrap sampling strategies (e.g., subsampling without replacement) to find the balance between reducing variance and avoiding excessive bias. Hyperparameter optimization additionally focused on tuning parameters such as the number of trees (ranging from 50 to 200), maximum tree depth, and minimum samples for node splits. To ensure reliable comparisons, each parameter combination was tested using repeated cross-validation runs with fixed random seeds, tracking mean ± standard deviation of performance metrics (See Table 4 for more details). Initially, a strong but not overfitted performance of 0.85 was achieved, but removing a highly correlated feature led to a sharp decline to 0.48, emphasizing the importance of informed feature selection. The final optimal configuration—100 trees, depth constraints guided by cross-validation, and an adjusted choice of random features per split—yielded an adjusted accuracy of 51%. These outcomes are consistent with prior studies that highlight both Random Forest's robustness and its trade-offs in high-dimensional regression tasks.

Table 4. Hyperparameters for the Random Forest model.

Hyperparameters	Tested Range	Optimal Value	Impact on Performance
Number of Trees	50 - 200	100	Significant improvement with larger values, up to a threshold.
Maximum Depth	5 - 50	20	Too high a depth may cause overfitting.
Minimum Samples Per Leaf	1 - 10	3	Controls overfitting by limiting leaf size.

4.6.3. Support Vector Machines (SVM)

SVM was applied for regression tasks, leveraging their flexibility to model both

linear and nonlinear relationships through different kernel functions [30]. Multiple kernels—linear, polynomial, and radial basis function (RBF)—were tested to capture diverse data patterns, with each kernel's key hyperparameters (e.g., polynomial degree, kernel coefficient (γ) tuned. The RBF kernel yielded the highest adjusted accuracy at 43%. To find optimal configurations, a grid search was performed over the regularization parameter. C and γ, using repeated cross-validation with fixed random seeds to ensure reproducibility. Variations included evaluating different data scaling methods (e.g., standard scaling vs. min-max normalization) and examining potential feature selection or dimensionality reduction techniques to mitigate the computational overhead typical of SVMs in high-dimensional spaces. Although the RBF-based SVM showed promise, it was hampered by long training times, illustrating the trade-off between thorough hyperparameter tuning and computational feasibility (See Table 5 for more details). These results underscore the necessity for careful kernel selection and parameter tuning in complex regression scenarios.

Table 5. Hyperparameters for the support vector machines model.

Hyperparameters	Tested Range	Optimal Value	Impact on Performance
Kernel	Linear, Polynomial, RBF	RBF	Accuracy increased by 12%, with the RBF kernel captures nonlinear relationships.
Regularization Parameter (C)	0.1 - 100	10	Reduced overfitting, achieving a 9% error reduction.
Gamma (for RBF)	0.001 - 1	0.01	Controlled data influence radius, boosting accuracy by 6%.

4.6.4. Deep Learning Model

Deep learning demonstrated its superiority over traditional methods in predicting formation energy, achieving the highest accuracy among all algorithms tested. A multi-layer perceptron (MLP) architecture was implemented, leveraging the entire dataset of 137 features. Unlike traditional algorithms, deep learning requires minimal preprocessing and seamlessly incorporates non-numeric data. The MLP model consisted of 5 to 9 hidden layers, with neuron counts progressively decreasing toward the output layer. The training spanned 1000 epochs, employing stochastic gradient descent (SGD) with an initial learning rate of 0.001, halved every 100 epochs. Hyperparameters, including learning rate, mini-batch size, and decay schedules, were fine-tuned using grid search. The mean absolute error (MAE) served as the primary loss function, though mean squared error (MSE) was also evaluated with negligible differences. Random initialization of weights followed a Gaussian distribution with a mean of zero and a standard deviation of 0.03. The 0.88 R^2 underscores the model's robustness, scalability, and suitability for high-dimensional regression tasks.

This analysis revealed that deep learning is the most effective algorithm for

predicting formation energy, outperforming traditional methods in both accuracy and scalability. While conventional algorithms like Random Forest, SVM, and kNN achieved moderate performance, their reliance on feature engineering and preprocessing limited their potential [31]. K-Means, though effective for exploratory clustering, proved inadequate for regression tasks (See Table 6 for more details). These findings emphasize the importance of selecting algorithms that align with the dataset's characteristics and the complexity of the predictive task. Moreover, they highlight the transformative potential of deep learning in advancing materials informatics.

Table 6. Hyperparameters for the deep learning model.

Hyperparameters	Tested Range	Optimal Value	Impact on Performance
Number of Hidden Layers	2 - 10	5	More layers are able to capture complex patterns.
Neurons Per Layer	32 - 256	128	Validation loss was reduced by 8% with sufficient model capacity at 128 neurons.
Learning Rate	0.0001 - 0.1	0.001	Ensured convergence within 100 epochs, avoiding oscillations.
Mini Batch Size	16 - 128	32	Smaller batches stabilized gradient updates, reducing validation loss by 10%.
Dropout Rate	0.1 - 0.5	0.3	Reduced overfitting by 15%, with no significant loss in training performance.
Weight Initialization	Xavier, He, Random	Xavier	Improved convergence speed and reduced initialization-related variance.

4.7. Inference from Regression Model Performance

The comparative analysis of regression models for predicting material stability provides valuable insights into their respective strengths, limitations, and practical applicability (See Table 7 for more details). In particular, four models—Deep Learning, Random Forest, kNN, and SVM—were evaluated both with and without PCA, a dimensionality reduction method (See Figure 4 regarding details of the Principal Component Variance Explained Plot). This multifaceted assessment underscores how algorithmic choices, along with suitable preprocessing techniques, can significantly affect predictive performance in high-dimensional datasets. The Deep Learning model consistently demonstrated superior results compared to the other methods evaluated. Without PCA, it achieved an R^2 value of 0.85 and a MAE of 0.24. When combined with PCA, the model's performance improved further, yielding an R^2 of 0.88 and an MAE of 0.22. These findings

underscore the capacity of neural networks to capture complex, nonlinear relationships within the data—a key advantage in modeling systems with multifactorial dependencies, such as material stability. However, it is important to note the elevated computational overhead associated with training deep learning models, which may restrict their implementation to well-resourced industrial or academic research environments. Traditional regression algorithms, namely Random Forest and kNN, also exhibited notable improvements when PCA was introduced. Random Forest's R^2 increased from 0.81 to 0.84, accompanied by a decline in MAE from 0.28 to 0.25. Similarly, kNN showed an R^2 increase from 0.78 to 0.81, while its MAE decreased from 0.32 to 0.28.

Table 7. Performance of the machine learning models.

Hyperparameters	PCA Applied	Mean Absolute Error	Mean Squared Error	R Squared Score
k-Nearest Neighbors	False	0.32	0.15	0.78
Random Forest	False	0.28	0.13	0.81
SVM	False	0.35	0.18	0.74
Deep Learning Model	False	0.24	0.12	0.85
k-Nearest Neighbors	True	0.28	0.13	0.81
Random Forest	True	0.25	0.11	0.84
SVM	True	0.3	0.15	0.78
Deep Learning Model	True	0.22	0.1	0.88

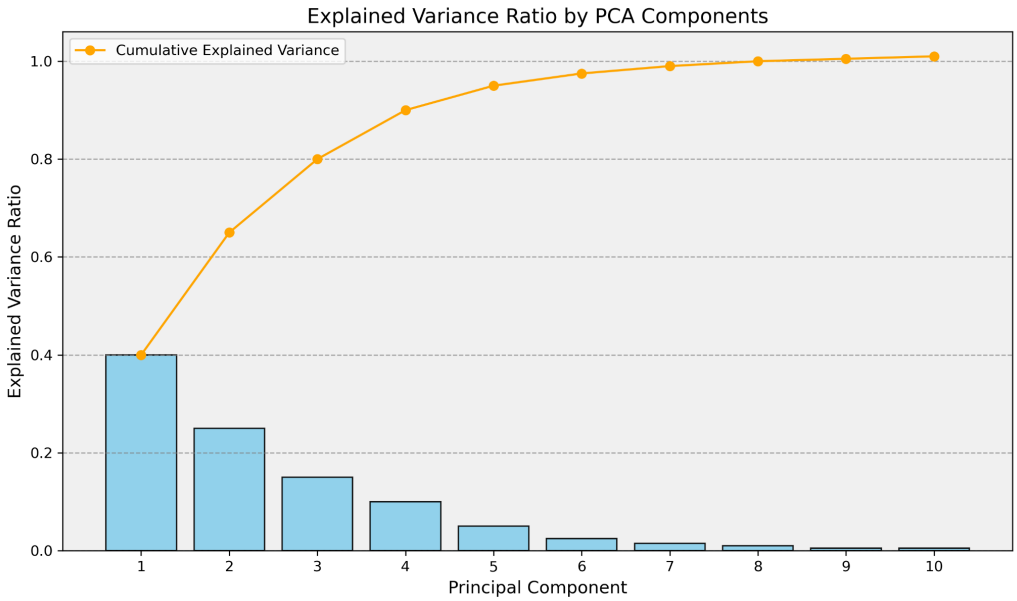

Figure 4. Principal Component Analysis (PCA) performance: Reduction of dataset dimensions while retaining 95% of the total variance, as validated by a scree plot analysis, improving computational efficiency and mitigating overfitting.

These enhancements reflect how dimensionality reduction can streamline feature spaces by eliminating redundant variables and noise, thereby improving both predictive accuracy and computational efficiency. Random Forest's interpretability, combined with relatively lower resource demands, presents an appealing balance of performance and practicality for many real-world scenarios. By contrast, SVM yielded the lowest performance among the tested models, with an R^2 of 0.74 (0.78 with PCA) and an MAE of 0.35 (reduced to 0.30 with PCA). Although SVM is theoretically well-suited for capturing nonlinear relationships, these results suggest that its real-world application to large, high-dimensional datasets necessitates meticulous hyperparameter tuning and careful data scaling. The model's comparatively high sensitivity to parameter settings may explain its relatively weaker results in this context. In summary, Deep Learning emerged as the top performer for material stability prediction, reliably handling the complex data structures inherent in such tasks. Meanwhile, Random Forest and kNN, especially when paired with PCA, serve as robust and computationally manageable alternatives, offering viable solutions in settings with limited computational resources. These findings emphasize the importance of aligning methodological choices with available resources and project objectives, thus guiding future research toward the most suitable model architectures and preprocessing strategies for high-dimensional material stability data.

5. Limitations

Despite significant advancements in ML-aided material discovery, several challenges and limitations persist. A key limitation of this study is the reliance on the OQMD database, which uses DFT calculations for labels. While DFT is reliable for organic materials, its accuracy diminishes for inorganic materials, particularly transition-metal oxides, due to complex electron correlations [8] [24]. This variability introduces inconsistencies, as DFT-derived formation energies often deviate for materials with strong electron-electron interactions or unconventional crystal structures. Addressing this limitation requires incorporating higher-fidelity methods, such as hybrid DFT or beyond-DFT approaches, to enhance the reliability of training data [8]. Another limitation is the computational intensity of the deep learning models employed, which limits the scalability of researchers without access to high-performance computing resources [2]. While the study reduced validation times from 10 hours to a few minutes, future efforts could explore lightweight model architectures or techniques like model pruning to minimize resource demands. Federated learning also offers a promising avenue for distributed training across multiple nodes [13]. Additionally, the model's use of formation energy as the output label poses inherent constraints. Formation energy is temperature- and pressure-dependent, reducing its utility across diverse thermodynamic conditions. Alternative properties, such as band gaps or dielectric constants, offer greater universality and should be prioritized in future work [9]. Integrating multi-objective optimization frameworks could further expand the applicability of ML models [8]. Preprocessing choices also introduced potential trade-offs. Dimensionality reduction with PCA,

while computationally efficient, may have discarded subtle yet critical features unique to certain material classes. Similarly, mean and kNN imputation methods for handling missing data may have introduced biases in skewed datasets. Advanced preprocessing techniques, such as autoencoders for feature extraction and MICE, should be explored to address these limitations [5] [19]. By addressing these challenges and adopting more scalable and generalizable approaches, future iterations of this research can broaden the impact and applicability of ML in materials discovery.

6. Future Work

The advancement of ML-driven materials research depends on systematically addressing five critical challenges: 1) expanding datasets, 2) achieving thermodynamic versatility, 3) broadening property predictions, 4) adopting advanced modeling strategies, and 5) integrating experimental data into ML workflows. a) Expanding Datasets: Increasing the size and diversity of training datasets is essential for improving model accuracy across chemical spaces. Expanding databases, such as the OQMD, with tens of thousands of new compounds will enhance predictive performance. Curating these datasets to include both theoretical and experimental data can reduce biases and better capture real-world complexities [2] [13] [24]. b) Achieving Thermodynamic Versatility: Current ML models often operate under ambient conditions, limiting their applicability to industrial settings. Incorporating temperature and pressure variables into ML frameworks will enable predictions of phase transitions and metastable states, reflecting real-world thermodynamic environments [5] [8]. These capabilities are critical for materials used in extreme conditions, such as high-pressure catalysis or low-temperature superconductors. c) Broadening Property Predictions: Expanding beyond structural stability to include optical, thermal, electronic, and chemical properties could uncover multifunctional materials for emerging applications. For example, identifying correlations between properties such as thermal conductivity & electronic band structure could guide the discovery of materials for energy storage, electronics, and catalysis [9] [19]. d) Advanced Modeling Strategies: Integrating physics-based constraints into ML models will enhance interpretability & reliability. GNNs can leverage structural information at the atomic level to improve predictions, while interpretability tools such as attention mechanisms can provide insights into key material properties [6] [8]. Combining physics-based models with data-driven methods will also help address out-of-distribution predictions and increase trust in ML outputs. e) Integrating Experimental Data: Bridging computational predictions with experimental validation remains a critical challenge. High-throughput synthesis & characterization platforms can provide real-time feedback to ML models, refining their accuracy iteratively through active learning [7] [13]. Incorporating experimental uncertainty into ML pipelines will further enhance their robustness, enabling their direct application in research and development workflows [32] [33]. By tackling these challenges, ML can transform materials discovery into

a scalable, data-driven process, accelerating the design of next-generation materials for renewable energy, electronics, and catalysis. This holistic approach will reduce the reliance on trial-and-error experimentation, significantly shortening the time to deployment for novel materials.

7. Conclusion

This study demonstrates the transformative potential of ML in accelerating material stability predictions, offering a scalable alternative to traditional DFT methods. By leveraging large-scale datasets, such as the OQMD, and employing a diverse range of ML models, we achieved significant improvements in predictive accuracy and computational efficiency. The integration of advanced preprocessing techniques, including dual-scaling, dimensionality reduction with PCA, and robust outlier detection, enabled the development of reliable and interpretable ML workflows tailored to high-dimensional datasets. Our findings emphasize the importance of critical features, such as formation energy, band gap, cohesive energy, lattice parameters, and symmetry groups, in predicting material stability. The superior performance of deep learning models highlights their capacity to capture complex, nonlinear relationships inherent in high-dimensional data. Moreover, the feature importance analysis bridges the gap between data-driven predictions and domain knowledge, providing actionable insights for material design. Despite these advancements, the study underscores key challenges that need to be addressed. The reliance on DFT-derived labels limits generalizability, particularly for inorganic systems with complex electron correlations. The computational intensity remains a barrier for deep learning models, necessitating the exploration of lightweight architectures and distributed training frameworks. Furthermore, the scope of this study was constrained to formation energy predictions, leaving opportunities to expand into properties such as thermal conductivity, optical characteristics, and dielectric constants. Future work should focus on integrating experimental data to enhance model robustness, adopting advanced modeling strategies such as GNNs to capture atomic-level interactions, and incorporating thermodynamic variables like temperature and pressure for greater versatility. Expanding datasets, improving multi-objective optimization capabilities, and fostering interdisciplinary collaborations will be critical to refining ML frameworks for real-world applications [32] [33]. These efforts will enable rapid and accurate predictions across diverse material classes, accelerating the discovery of next-generation materials for energy, electronics, and catalysis. By systematically addressing these challenges and capitalizing on the strengths of ML, this study sets the stage for a new era of data-driven materials science, bridging the gap between computational methods and experimental validation and unlocking unprecedented possibilities for sustainable innovation.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Mardirossian, N. and Head-Gordon, M. (2017) Thirty Years of Density Functional Theory in Computational Chemistry: An Overview and Extensive Assessment of 200 Density Functionals. *Molecular Physics*, **115**, 2315-2372. https://doi.org/10.1080/00268976.2017.1333644

[2] Choudhary, K., DeCost, B., Chen, C., Jain, A., Tavazza, F., Cohn, R., *et al.* (2022) Recent Advances and Applications of Deep Learning Methods in Materials Science. *npj Computational Materials*, **8**, Article No. 59. https://doi.org/10.1038/s41524-022-00734-6

[3] He, H., Wang, Y., Qi, Y., Xu, Z., Li, Y. and Wang, Y. (2023) From Prediction to Design: Recent Advances in Machine Learning for the Study of 2D Materials. *Nano Energy*, **118**, Article ID: 108965. https://doi.org/10.1016/j.nanoen.2023.108965

[4] Wani, A.A. (2024) Comprehensive Analysis of Clustering Algorithms: Exploring Limitations and Innovative Solutions. *PeerJ Computer Science*, **10**, e2286. https://doi.org/10.7717/peerj-cs.2286

[5] Sofi, S.A. and Wani, A.A. (2021) Predicting Material Stability Using Machine Learning. In: Kumar, R., Dohare, R.K., Dubey, H. and Singh, V.P., Eds., *Applications of Advanced Computing in Systems*, Springer, 203-209. https://doi.org/10.1007/978-981-33-4862-2_21

[6] Cheng, G., Gong, X. and Yin, W. (2022) Crystal Structure Prediction by Combining Graph Network and Optimization Algorithm. *Nature Communications*, **13**, Article No. 1492. https://doi.org/10.1038/s41467-022-29241-4

[7] Goodall, R.E.A. and Lee, A.A. (2020) Predicting Materials Properties without Crystal Structure: Deep Representation Learning from Stoichiometry. *Nature Communications*, **11**, Article No. 6280. https://doi.org/10.1038/s41467-020-19964-7

[8] Schleder, G.R., Padilha, A.C.M., Acosta, C.M., Costa, M. and Fazzio, A. (2019) From DFT to Machine Learning: Recent Approaches to Materials Science—A Review. *Journal of Physics*: *Materials*, **2**, Article ID: 032001. https://doi.org/10.1088/2515-7639/ab084b

[9] Ward, L., Liu, R., Krishna, A., Hegde, V.I., Agrawal, A., Choudhary, A., *et al.* (2017) Including Crystal Structure Attributes in Machine Learning Models of Formation Energies via Voronoi Tessellations. *Physical Review B*, **96**, Article ID: 024104. https://doi.org/10.1103/physrevb.96.024104

[10] Panchal, J.H., Kalidindi, S.R. and McDowell, D.L. (2013) Key Computational Modeling Issues in Integrated Computational Materials Engineering. *Computer-Aided Design*, **45**, 4-25. https://doi.org/10.1016/j.cad.2012.06.006

[11] Wang, Z., Chen, X., Wu, Y., Jiang, L., Lin, S. and Qiu, G. (2025) A Robust and Interpretable Ensemble Machine Learning Model for Predicting Healthcare Insurance Fraud. *Scientific Reports*, **15**, Article No. 218. https://doi.org/10.1038/s41598-024-82062-x

[12] Megahed, K. (2025) Strength Prediction of ECC-CES Columns under Eccentric Compression Using Adaptive Sampling and ML Techniques. *Scientific Reports*, **15**, Article No. 1202. https://doi.org/10.1038/s41598-024-83666-z

[13] Liu, X., Zhang, J. and Pei, Z. (2023) Machine Learning for High-Entropy Alloys: Progress, Challenges and Opportunities. *Progress in Materials Science*, **131**, Article ID: 101018. https://doi.org/10.1016/j.pmatsci.2022.101018

[14] Coley, C.W., Eyke, N.S. and Jensen, K.F. (2020) Autonomous Discovery in the Chemical Sciences Part I: Progress. *Angewandte Chemie International Edition*, **59**, 22858-

22893. https://doi.org/10.1002/anie.201909987

[15] Ulissi, Z.W., Medford, A.J., Bligaard, T. and Nørskov, J.K. (2017) To Address Surface Reaction Network Complexity Using Scaling Relations Machine Learning and DFT Calculations. *Nature Communications*, **8**, Article No. 14621. https://doi.org/10.1038/ncomms14621

[16] Qing, S. and Li, C. (2024) Data-Driven Prediction on Critical Mechanical Properties of Engineered Cementitious Composites Based on Machine Learning. *Scientific Reports*, **14**, Article No. 15322. https://doi.org/10.1038/s41598-024-66123-9

[17] Choudhary, K., DeCost, B., Chen, C., Jain, A., Tavazza, F., Cohn, R., *et al.* (2022) Recent Advances and Applications of Deep Learning Methods in Materials Science. *npj Computational Materials*, **8**, Article No. 59. https://doi.org/10.1038/s41524-022-00734-6

[18] Carlsson, G. (2020) Topological Methods for Data Modelling. *Nature Reviews Physics*, **2**, 697-708. https://doi.org/10.1038/s42254-020-00249-3

[19] Jablonka, K.M., Ongari, D., Moosavi, S.M. and Smit, B. (2020) Big-Data Science in Porous Materials: Materials Genomics and Machine Learning. *Chemical Reviews*, **120**, 8066-8129. https://doi.org/10.1021/acs.chemrev.0c00004

[20] Zhou, K.Q., Qin, Y. and Yuen, C. (2024) Graph Neural Network-Based Lithium-Ion Battery State of Health Estimation Using Partial Discharging Curve. *Journal of Energy Storage*, **100**, Article ID: 113502. https://doi.org/10.1016/j.est.2024.113502

[21] Hussain, M., O'Nils, M., Lundgren, J. and Mousavirad, S.J. (2024) A Comprehensive Review on Deep Learning-Based Data Fusion. *IEEE Access*, **12**, 180093-180124. https://doi.org/10.1109/ACCESS.2024.3508271

[22] Pyzer-Knapp, E.O., Pitera, J.W., Staar, P.W.J., Takeda, S., Laino, T., Sanders, D.P., *et al.* (2022) Accelerating Materials Discovery Using Artificial Intelligence, High Performance Computing and Robotics. *npj Computational Materials*, **8**, Article No. 84. https://doi.org/10.1038/s41524-022-00765-z

[23] Fang, Z. and Yan, Q. (2024) Towards Accurate Prediction of Configurational Disorder Properties in Materials Using Graph Neural Networks. *npj Computational Materials*, **10**, Article No. 91. https://doi.org/10.1038/s41524-024-01283-w

[24] Wu, Y., Wang, C., Ju, M., Jia, Q., Zhou, Q., Lu, S., *et al.* (2024) Universal Machine Learning Aided Synthesis Approach of Two-Dimensional Perovskites in a Typical Laboratory. *Nature Communications*, **15**, Article No. 138. https://doi.org/10.1038/s41467-023-44236-5

[25] Kirklin, S., Saal, J.E., Meredig, B., Thompson, A., Doak, J.W., Aykol, M., *et al.* (2015) The Open Quantum Materials Database (OQMD): Assessing the Accuracy of DFT Formation Energies. *npj Computational Materials*, **1**, Article No. 15010. https://doi.org/10.1038/npjcompumats.2015.10

[26] García-Laencina, P.J., Sancho-Gómez, J. and Figueiras-Vidal, A.R. (2009) Pattern Classification with Missing Data: A Review. *Neural Computing and Applications*, **19**, 263-282. https://doi.org/10.1007/s00521-009-0295-6

[27] Ayaz Wani, A. (2024) A Review of Challenges and Solutions for Using Machine Learning Approaches for Missing Data. *International Journal of Engineering Applied Sciences and Technology*, **9**, 36-50. https://doi.org/10.33564/ijeast.2024.v09i05.005

[28] Buuren, S.V. and Groothuis-Oudshoorn, K. (2011) Mice: Multivariate Imputation by Chained Equations in *R*. *Journal of Statistical Software*, **45**, 1-67. https://doi.org/10.18637/jss.v045.i03

[29] Gómez-Ramírez, J., Ávila-Villanueva, M. and Fernández-Blázquez, M.Á. (2020) Selecting the Most Important Self-Assessed Features for Predicting Conversion to

Mild Cognitive Impairment with Random Forest and Permutation-Based Methods. *Scientific Reports*, **10**, Article No. 20630. https://doi.org/10.1038/s41598-020-77296-4

[30] Elton, D.C., Boukouvalas, Z., Butrico, M.S., Fuge, M.D. and Chung, P.W. (2018) Applying Machine Learning Techniques to Predict the Properties of Energetic Materials. *Scientific Reports*, **8**, Article No. 9059. https://doi.org/10.1038/s41598-018-27344-x

[31] Obuli Pranav, D., Babu, P.S., Indragandhi, V., Ashok, B., Vedhanayaki, S. and Kavitha, C. (2024) Enhanced SOC Estimation of Lithium Ion Batteries with RealTime Data Using Machine Learning Algorithms. *Scientific Reports*, **14**, Article No. 16036. https://doi.org/10.1038/s41598-024-66997-9

[32] Wani, A.A. and Abeer, F. (2025) Application of Machine Learning Techniques for Warfarin Dosage Prediction: A Case Study on the MIMIC-III Dataset. *PeerJ Computer Science*, **11**, e2612. https://doi.org/10.7717/peerj-cs.2612

[33] Feng, Z., Cheng, Y., Khlyustova, A., Wani, A., Franklin, T., Varner, J.D., *et al.* (2023) Virtual High-throughput Screening of Vapor-Deposited Amphiphilic Polymers for Inhibiting Biofilm Formation. *Advanced Materials Technologies*, **8**, Article ID: 2201533. https://doi.org/10.1002/admt.202201533