Machine Learning Prediction of Potential Energy in Magnesium (Mg)

ABSTRACT

This report presents a comprehensive investigation into the application of machine learning techniques for predicting the potential energy of magnesium, a lightweight and structurally important metal widely used across multiple industrial sectors. Accurate prediction of potential energy is vital for understanding magnesium's structural stability and mechanical behavior under varying environmental conditions. The study assesses multiple machine learning models, including Ridge Regression, Linear Regression, Decision Tree, and Random Forest benchmarking their ability to capture the complex nonlinear relationships typical of materials science datasets.

The data utilized for model training and validation were generated through meticulously conducted molecular dynamics simulations, encompassing both ab initio molecular dynamics (AIMD) and classical molecular dynamics (CMD) methodologies. These simulations effectively replicate magnesium's atomic interactions across different temperatures and pressures, producing high-fidelity datasets that reveal intricate dependencies between thermodynamic variables and potential energy. Advanced feature engineering and preprocessing methods were applied to enhance data quality and optimize model inputs.

Each model underwent thorough hyperparameter tuning to maximize predictive accuracy while minimizing overfitting risks.

This work exemplifies the powerful synergy between computational materials science and predictive analytics, enabling efficient and precise modeling of material properties. The integration of machine learning into materials informatics marks a significant advancement in accelerating the development of magnesium-based materials, reducing dependence on costly and time-intensive experimental procedures. The insights generated are expected to drive innovation in automotive, aerospace, and energy sectors, where lightweight yet high-performance materials are essential.

INTRODUCTION

Magnesium (Mg) is a naturally abundant, lightweight metal that holds significant importance in modern industry owing to its unique combination of physical and mechanical characteristics. It has an atomic number of 12 and crystallizes in a hexagonal close-packed (HCP) structure, a key factor underpinning its strength, stiffness, and structural performance. With a notably low density of 1.74 g/cm³, magnesium ranks among the lightest metals extensively utilized in engineering fields. This exceptional lightness makes magnesium especially valuable for weight-sensitive applications across aerospace, automotive, and consumer electronics industries. The metal's low density facilitates substantial reductions in overall system weight, leading to enhanced performance and improved energy efficiency. Magnesium offers an excellent balance between minimal weight and mechanical robustness, rendering it highly suitable for engineering applications demanding both strength and lightness.

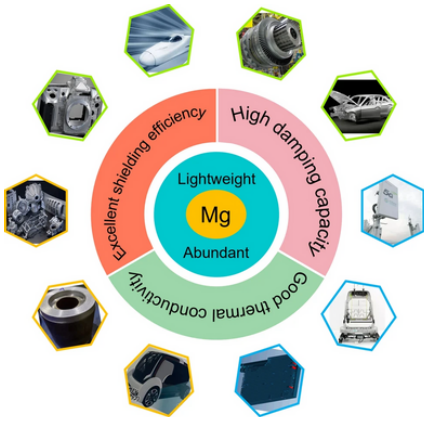


Figure 1. Image representing the properties of Magnesium(Mg) and their industrial applications.

A) Physical and Mechanical Properties of Magnesium:

Magnesium is characterized by moderate strength and stiffness, which places it in a unique position among structural metals where a balance between lightweight and mechanical resilience is crucial. Its excellent castability refers to the metal's ability to be melted and poured into intricate molds with relative ease, facilitating the production of complex geometries that would be difficult or costly to achieve using other materials. Alongside castability, magnesium also exhibits outstanding machinability, meaning it can be efficiently cut, shaped, and finished using standard machining processes. These combined manufacturing attributes make magnesium highly versatile for creating precision engineering components in industries ranging from automotive to aerospace.

One of magnesium's particularly valuable mechanical properties is its high damping capacity. This means magnesium efficiently dissipates mechanical vibrations and oscillations, reducing the transmission of harmful vibrational energy in mechanical systems. Such damping behavior not only contributes to noise reduction but also enhances operational stability and longevity of components by mitigating fatigue and resonance-induced wear. This makes magnesium an excellent choice in applications requiring vibration control, such as in automotive engine mounts, aerospace structures, and electronic housings.

However, magnesium's susceptibility to corrosion presents a significant limitation, especially in environments prone to moisture, salt exposure, or chemical aggressors. Corrosion can degrade mechanical properties and compromise structural integrity over time. To address these challenges, extensive research and development efforts have focused on alloying magnesium with elements such as aluminum, zinc, and rare earth metals, which improve corrosion resistance and mechanical strength. Additionally, surface coatings—ranging from anodizing and conversion coatings to polymeric and metallic layers—are widely applied to protect magnesium components from environmental degradation. These combined metallurgical and surface engineering strategies substantially extend the service life and broaden the applicability of magnesium in harsh and varied operational environments.

B) Industrial Applications of Magnesium:

Magnesium's versatility spans multiple key industrial sectors. In automotive manufacturing, the demand for improved fuel efficiency and reduced emissions is well supported by magnesium's lightweight characteristics, enabling substantial vehicle weight reduction without compromising safety or structural integrity. Aerospace applications similarly capitalize on magnesium's favorable strength-to-weight ratio, as weight savings translate directly into lower operational expenses and improved performance metrics. Consumer electronics also employ magnesium alloys extensively in the fabrication of portable devices such as laptops, smartphones, and cameras, where material durability and lightness are critical for enhanced user convenience. Emerging energy sector applications, including hydrogen storage and battery technologies, are increasingly incorporating magnesium due to its high hydrogen absorption capacity and relative cost advantages compared to alternative metals, contributing to the advancement of clean energy solutions.

C) Motivation for Machine Learning Prediction of Magnesium Properties:

Accurate understanding and prediction of magnesium's potential energy under varying environmental parameters like temperature and pressure is fundamental for optimizing its real-world performance. Traditional approaches based on experiments and physics-based simulations, while valuable, are computationally intensive and time-consuming. Machine learning presents a compelling alternative, offering data-driven models that can capture complex, nonlinear relationships characteristic of materials science data. By integrating machine learning with molecular dynamics simulation outputs, predictive models can be developed that accurately estimate potential energy, accelerating material characterization and design processes, thereby fostering innovation in magnesium-based engineering applications.

D) Related Research on Machine Learning Prediction of Magnesium Properties:

Recent developments in materials informatics have seen the increasing adoption of machine learning techniques to predict potential energy and other essential properties of magnesium and its alloys. Numerous studies have established predictive frameworks using regression algorithms such as support vector machines (SVM), artificial neural networks (ANN), and decision trees, trained on datasets derived from molecular dynamics and first-principles calculations. These approaches have demonstrated promising accuracy in modeling potential energy surfaces and thermodynamic stability across diverse temperatures and pressures.

Kernel-based methods have effectively modeled nonlinear dependencies in magnesium's atomic arrangements, allowing rapid energy evaluations without relying on costly full-scale simulations. Ensemble learning techniques, including Random Forest and Gradient Boosting Machines, have proven adept at managing high-dimensional feature spaces and mitigating overfitting, thereby improving the robustness and generalization of models for unseen magnesium structures. Moreover, combining machine learning potentials with classical molecular dynamics simulations has enabled scalable, efficient investigations into magnesium's mechanical behavior and phase transitions.

This growing body of research underscores the value of combining data-driven methodologies with traditional materials science to achieve faster, more cost-effective, and precise insights into magnesium's properties. The present study builds upon these advancements by conducting a systematic exploration and comparison of multiple machine learning algorithms tailored to magnesium's unique attributes, with the objective of identifying the most effective approaches for potential energy prediction.

METHODOLOGY

A) Data Collection and Augmentation:

The foundational dataset for machine learning modeling was derived from a series of molecular dynamics simulations, capturing a wide range of physical states of magnesium under diverse thermodynamic conditions. Each record in the dataset corresponds to a single simulation step, with columns representing key physical quantities such as Step number, Time (in picoseconds), Temperature, Potential Energy (PotEng), Kinetic Energy (KinEng), Total Energy (TotEng), Pressure (Press), and Volume.

Specifically, the dataset records include:

- Step number: Tracking progression within the simulation
- Time: Timestamp at each simulation snapshot, reflecting temporal evolution
- Temperature: Recorded at every time step, representing thermal fluctuations
- Potential Energy: The target property for prediction, indicative of system stability
- Kinetic Energy: Providing insight into dynamic behaviors of atoms
- Total Energy: Used to evaluate overall system stability
- Pressure and Volume: Monitoring thermodynamic changes and system size variations

To establish a robust dataset for training, data from thousands of simulation steps were aggregated. Rigorous data quality control measures were implemented, including outlier detection techniques such as Z-score analysis and interquartile range (IQR) filtering, aiming to exclude anomalous values caused by simulation artifacts or external disturbances. Missing and inconsistent records were also identified and removed to ensure dataset integrity.

Data augmentation further enhanced model generalization. Purposeful addition of controlled statistical noise to selected features, such as temperature and pressure, simulated natural measurement variability and experimental uncertainties. Synthetic data generation through interpolation and extrapolation between simulation steps expanded the input space, augmenting the diversity of training examples and mitigating potential overfitting.

Normalization and balancing procedures were employed to optimize model training:

- All feature columns were scaled to a common range to prevent features with larger numerical ranges from dominating model training.
- Stratified sampling ensured proportional representation of all temperature and pressure regimes within training and testing datasets.

This meticulously constructed and augmented dataset provided a reliable, scalable, and interpretable basis for machine learning models aimed at predicting magnesium's potential energy and related physical properties.

B) **Machine Learning Models:**

Four machine learning regression models were employed to model the complex relationships governing magnesium's potential energy, each offering unique advantages suitable for materials science data.

Linear Regression is the simplest form of regression that models a linear relationship between the input features and the target variable, potential energy in this case. It assumes that the response variable can be described as a weighted sum of the input features. While it is interpretable and computationally efficient, linear regression is sensitive to outliers and multicollinearity among features, which can decrease prediction accuracy. It also struggles with nonlinear relationships which might be present in the magnesium simulation data.

Ridge Regression is an extension of linear regression that incorporates L2 regularization to penalize large coefficients. This regularization helps reduce overfitting by shrinking model parameters and is especially effective when input features are highly correlated or noisy. By constraining the size of the coefficients, ridge regression improves the stability and predictive ability of the model compared to ordinary linear regression.

Decision Tree regression builds a model in the form of a tree structure composed of decision nodes and leaf nodes. Each node splits the data based on a feature value threshold, guiding predictions down branches representing subsets of data with similar properties. Decision trees can capture nonlinear relationships and interactions between variables and are easy to visualize and interpret. However, decision trees are prone to overfitting if they grow too deep without pruning or constraints.

Random Forest is an ensemble learning method that aggregates predictions from multiple decision trees. Each tree is trained on a bootstrapped sample of the training data and uses a random subset of features at splits to promote model diversity. Predictions from individual trees are averaged to reduce variance and improve generalization. Random Forest typically outperforms individual decision trees, delivering higher accuracy, lower Root Mean Squared Error (RMSE), and better robustness.

C) Machine Learning Pipeline:

The development of the machine learning pipeline for predicting magnesium's potential energy encompassed several integrated stages, each essential for establishing a rigorous and reproducible research framework. The process began with the ingestion of curated simulation data generated from molecular dynamics studies, encompassing a comprehensive set of features including temperature, pressure, kinetic and potential energies, total energy, and system volume, all recorded at fine temporal resolutions. This rich dataset enabled detailed tracking of magnesium's physical state changes throughout the simulations.

Data preprocessing formed the foundation of the pipeline, involving stringent quality checks designed to identify and rectify inconsistencies such as outliers and missing values, commonly encountered in high-throughput simulation outputs. Advanced normalization techniques were applied to standardize feature scales and units, ensuring fair representation and preventing bias in model training caused by disparate feature magnitudes.

Feature engineering played a critical role in enhancing model sensitivity and interpretability. Beyond selecting primary measurable attributes, derived features were constructed using domain knowledge and statistical methods—such as rate of change metrics for energy and pressure and ratios comparing sequential simulation steps—to capture dynamic behaviors not directly evident in raw data. Comprehensive correlation analysis guided this process by identifying features with strong predictive relevance while eliminating redundant or highly collinear variables that could degrade model stability, especially in regression frameworks.

Model selection was performed through an iterative process combining empirical evaluation and theoretical considerations. Regression models including Linear and Ridge Regression, Decision Trees, and Random Forests were all examined for suitability and predictive power. The pipeline design remained adaptable, accommodating more advanced architectures such as ensemble methods or neural networks depending on ongoing exploratory outcomes. Training incorporated robust k-fold cross-validation techniques, ensuring that multiple data partitions contributed to performance estimation and mitigating risks of overfitting. Stratified sampling ensured that less frequent but materially significant regimes—such as extreme low temperatures or pressures—were adequately represented in model development.

Performance assessment employed both quantitative metrics and visual analytics. Error measurements such as RMSE and R^2 were complemented by graphical comparisons of predicted versus actual values, facilitating nuanced understanding of model behavior across the data spectrum. Feature importance analyses provided interpretability, highlighting influential predictors and supporting refinement of simulation approaches based on learned insights. Upon identification of the optimal model, the pipeline was modularized and documented comprehensively, enabling reproducibility and ease of application to expanded datasets or alternative material systems.

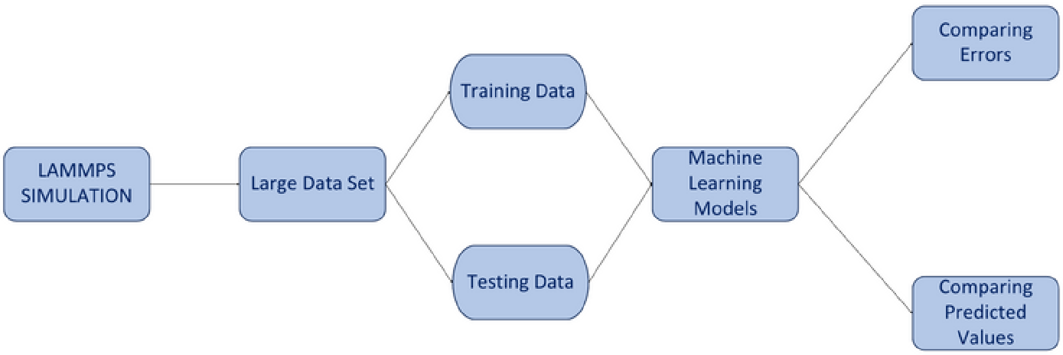


Figure 2. Flowchart representing the Machine Learning Pipeline used in the research

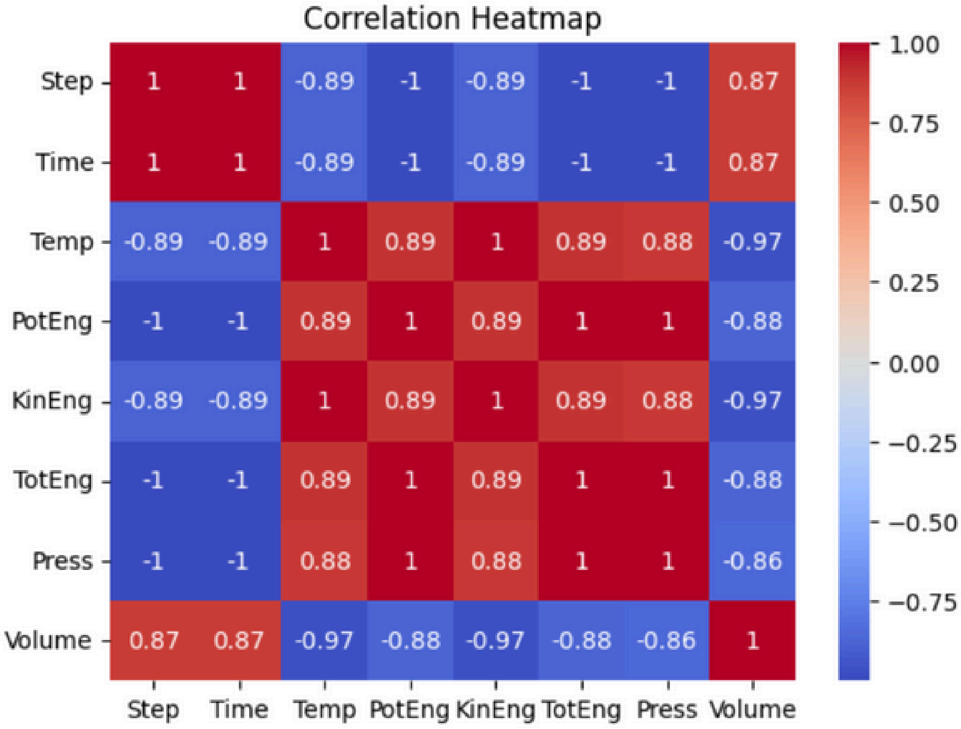
D) Correlation Heatmap:

The correlation heatmap is a graphical representation used to visualize the strength and direction of linear relationships between variables within a dataset. It is constructed based on Pearson correlation coefficients, which measure the degree of association between pairs of variables. These coefficients range from +1, indicating a perfect positive linear correlation, through 0, representing no linear relationship, to -1, denoting a perfect negative linear correlation. The heatmap displays these values in a two-dimensional matrix format, where each cell corresponds to a specific pair of variables. Colors are used to encode the magnitude and sign of the correlations, allowing for rapid visual identification of strong, weak, positive, or negative relationships between features.

In the context of materials science datasets, such as those used for predicting magnesium's potential energy, correlation heatmaps serve as a valuable tool for exploratory data analysis. They facilitate the detection of redundant features and multicollinearity—issues that can adversely affect certain machine learning models, particularly linear regression. By highlighting pairs of highly correlated variables, the heatmap informs feature selection strategies that seek to reduce redundancy, thereby enhancing model stability and interpretability.

Furthermore, the correlation heatmap acts as a critical guide in feature selection and transformation strategies during the preliminary stages of model development. Incorporating this analysis early in the modeling pipeline improves predictive accuracy by ensuring that the chosen input variables are both meaningful and free from excessive redundancy. This careful feature curation helps prevent issues such as multicollinearity, which can compromise model stability and interpretation.

In addition to its analytical benefits, the correlation heatmap provides a clear and intuitive visualization of complex relationships within the dataset. This facilitates transparent communication of data dependencies and preprocessing decisions to stakeholders and reviewers, thereby supporting methodological rigor. The inclusion of such graphical summaries strengthens the scientific validity of the research by demonstrating thorough data exploration and justification of feature engineering practices, essential elements in peer-reviewed materials science studies.



	Step	Time	Temp	PotEng	KinEng	TotEng	Press	Volume
Step	1	1	-0.89	-1	-0.89	-1	-1	0.87
Time	1	1	-0.89	-1	-0.89	-1	-1	0.87
Temp	-0.89	-0.89	1	0.89	1	0.89	0.88	-0.97
PotEng	-1	-1	0.89	1	0.89	1	1	-0.88
KinEng	-0.89	-0.89	1	0.89	1	0.89	0.88	-0.97
TotEng	-1	-1	0.89	1	0.89	1	1	-0.88
Press	-1	-1	0.88	1	0.88	1	1	-0.86
Volume	0.87	0.87	-0.97	-0.88	-0.97	-0.88	-0.86	1

Figure 3. Correlation Heatmap showing the relation or dependency of the X-axis properties on the Y-axis properties

E) Hyperparameter Tuning:

Hyperparameter tuning was performed to optimize the predictive performance of the machine learning models while minimizing risks of overfitting or underfitting. A systematic grid search approach was employed to explore various combinations of critical hyperparameters for each algorithm. For regression-based models, parameters such as the regularization strength (alpha) in Ridge Regression were tuned, while in tree-based models, settings including maximum tree depth and minimum samples per leaf for Decision Trees, as well as the number of estimators, maximum depth, and feature subset size for Random Forests, were optimized.

To ensure that hyperparameter selection generalized well beyond training data, cross-validation was integrated into the grid search process. Each hyperparameter combination was evaluated on validation folds, with the final choice guided by minimizing validation Root Mean Squared Error (RMSE) and achieving a balanced trade-off between bias and variance. For iterative training methods such as ensemble models, early stopping criteria were applied to cease training upon convergence or when improvements plateaued on validation data. This fine-tuning procedure enhanced model robustness and resulted in accurate, reliable predictions on unseen magnesium datasets.

F) Feature Engineering and Selection:

Feature engineering was instrumental in boosting model predictive accuracy by expanding and refining the input attribute set. Core features were sourced directly from molecular dynamics simulations and included temperature, pressure, kinetic energy, potential energy, total energy, and volume. Beyond primary features, secondary variables capturing relative changes and temporal gradients of these physical quantities were created to represent dynamic behavior more effectively. In addition, statistical descriptors—such as moving averages, variances, and higher-order moments computed over sliding windows—were incorporated to enrich the feature pool with temporal context.

To address high dimensionality and remove redundant information, Principal Component Analysis (PCA) was employed. PCA transformed correlated original features into a smaller number of orthogonal components, preserving the most informative variance while improving computational efficiency. Complementing this, correlation analysis identified multicollinearity among features, enabling the pruning of highly collinear variables that could induce instability in sensitive models like linear regressions.

RESULTS AND DISSCUSION

The primary aim of this study was to develop and rigorously evaluate machine learning models capable of accurately predicting the potential energy of magnesium over a wide range of thermodynamic conditions. This capability is critically important in materials science, as potential energy is a fundamental indicator of atomic stability and structural behavior. Given magnesium's widespread use as a lightweight structural metal in engineering applications, precise prediction of its potential energy offers valuable insights into intrinsic material properties and enables accelerated materials design by reducing dependence on computationally intensive atomistic simulations.

The dataset employed for model development originated from extensive molecular dynamics simulations that generated high-resolution data capturing multiple physical parameters at fine time steps. These parameters included temperature, pressure, kinetic energy, total energy, and system volume, collectively representing magnesium's dynamic atomic environment under varying simulated conditions. The dataset's granularity and diversity ensured that machine learning models were trained on a broad spectrum of material states, enhancing their ability to generalize across different operational regimes.

Model performance was quantitatively evaluated using a suite of metrics that comprehensively capture various facets of predictive accuracy and reliability. Root Mean Squared Error (RMSE) served as the primary metric, reflecting the average magnitude of prediction errors and emphasizing larger deviations. Mean Absolute Error (MAE) complemented this by providing an interpretable measure of average absolute deviation between predicted and actual values. The coefficient of determination (R^2) statistically quantified the proportion of variance in observed potential energy explained by the models, serving as an overall gauge of fit and explanatory power.

To ensure robustness and prevent overfitting, the models were trained and validated using rigorous cross-validation methodologies. The dataset was partitioned into multiple folds, with iterative training and evaluation cycles ensuring consistent performance across different data subsets.

Subsequent sections present a detailed comparative evaluation of various machine learning algorithms, document the outcomes of hyperparameter tuning, analyze the relative importance of physical features, and critically assess model accuracy in replicating magnesium's potential energy under diverse thermodynamic conditions. This multifaceted analysis underscores the efficacy of machine learning techniques in advancing computational materials science through accurate and efficient predictive modeling.

A) Correlation Plots:

A fundamental step in the exploratory data analysis of this study was the examination of correlation plots illustrating the relationships among key input features and the target variable, potential energy. The correlation plots visualize pairwise correlations between physical parameters such as temperature, pressure, kinetic energy, total energy, and system volume, as captured in molecular dynamics simulations of magnesium.

These plots provided direct insight into the degree and direction of linear association between variables, quantified through Pearson correlation coefficients. Strong positive correlations were observed, notably between temperature and potential energy, indicating that increases in temperature correspond strongly to higher potential energy states in magnesium. This aligns with the physical understanding that thermal agitation increases atomic vibrational energy, thereby raising the system's potential energy.

Pressure also demonstrated a moderately strong correlation with potential energy, reflecting the sensitivity of atomic arrangements in magnesium to external mechanical forces. Other features such as kinetic energy and total energy showed expected correlations given their intrinsic physical connections to the atomic system's energetic state.

Conversely, system volume exhibited weaker correlations with potential energy, implying a more subtle or indirect influence within the simulation parameter space. This suggested that volume alone might be insufficient as a primary predictor for potential energy but could contribute as part of a comprehensive feature set.

The correlation plots thereby facilitated the identification of the most influential features for predictive modeling, guiding subsequent feature selection and engineering steps. Ensuring the inclusion of highly correlated variables with potential energy was critical to improving model accuracy and interpretability while avoiding redundancy that could impair model training.

Overall, the correlation plots established foundational understanding of variable interdependencies in the dataset and reinforced the selection of physically meaningful predictors for machine learning models tasked with predicting magnesium's potential energy.

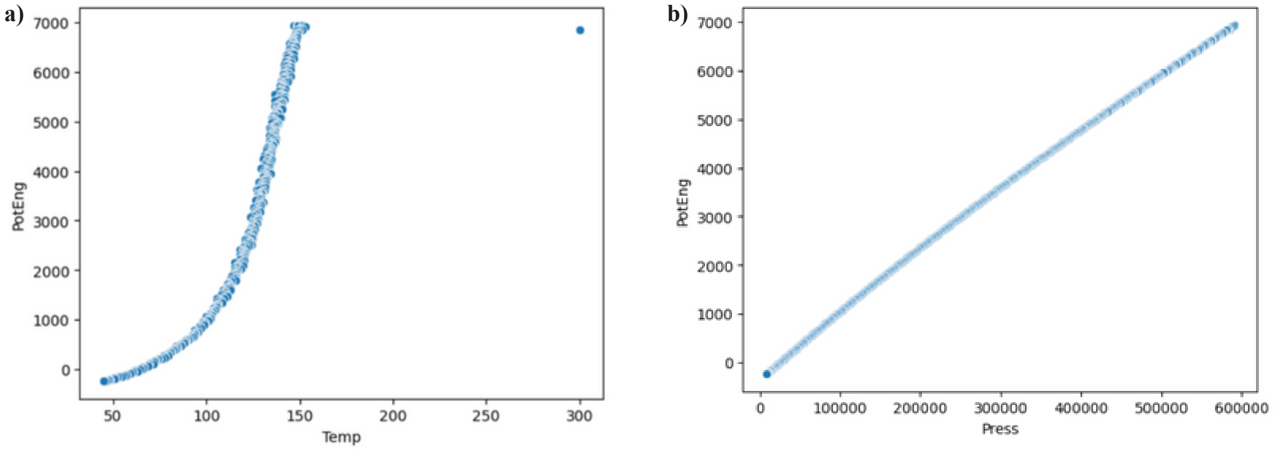


Figure 4. a) Correlation plot between Temperature on X-axis and Potential Energy on Y-axis.
b) Correlation plot between Pressure on X-axis and Potential Energy on Y-axis.

B) Test Data Results (After Hyperparameter Tuning):

Following comprehensive hyperparameter optimization, the machine learning models were evaluated on an independent test dataset to assess their genuine predictive generalizability. The tuning process employed grid search in combination with cross-validation, systematically adjusting key parameters such as regularization strength for linear models and tree-specific settings including maximum depth and number of estimators for decision trees and ensemble methods.

The optimized Random Forest model emerged as the top-performing algorithm, exhibiting markedly superior accuracy on the test set compared to other models. Its Root Mean Squared Error (RMSE) was substantially reduced to 16.08, signifying minimal average deviation between predicted and simulated potential energy values. Concurrently, the coefficient of determination (R^2) reached an exceptional 0.9999, indicating that nearly all variance in the observed potential energy was captured by the model.

These outcomes underscore the critical role of hyperparameter tuning in refining model complexity, balancing bias and variance to improve flexibility without overfitting. Among the tuned parameters, the number of trees and maximum tree depth were particularly influential in achieving this optimal balance, resulting in both high predictive accuracy and efficiency.

Other models, including Ridge Regression and Decision Trees, demonstrated performance improvements after tuning; however, their accuracy remained distinctly lower than the Random Forest. Ridge Regression, despite benefiting from optimized regularization, was constrained by its linear assumptions, limiting its applicability for capturing complex nonlinearities. Decision Trees, while improved, continued to exhibit susceptibility to overfitting owing to their single-tree structure.

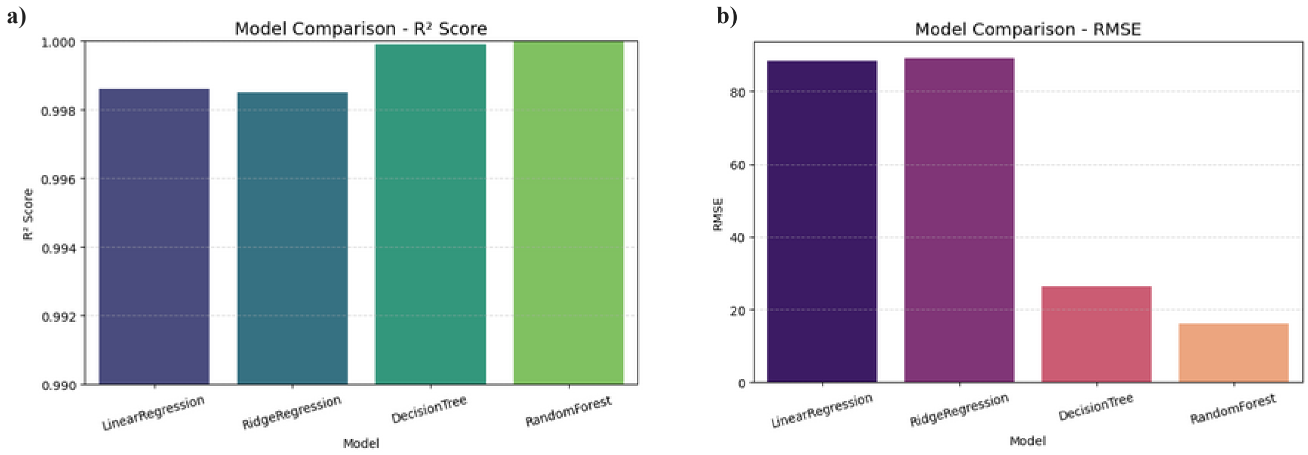


Figure 5. a) Shows the Bar Graph between different models based on the R^2 score.
b) Shows the Bar Graph between different models based on the RMSE score.

C) Radar Chart Analysis:

The radar chart offers a comprehensive visual comparison of the performance metrics across various machine learning models used for predicting magnesium's potential energy. This multidimensional plot simultaneously presents critical evaluation criteria, enabling a clear assessment of each model's strengths and weaknesses regarding prediction quality.

Key metrics featured in the radar chart include Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2), along with considerations such as training time and computational complexity. Collectively, these metrics provide a holistic view of each model's accuracy, error magnitude, explanatory power, and practical feasibility for deployment.

The visualization clearly indicates that the Random Forest model outperforms others by attaining the lowest RMSE and MAE, reflecting superior accuracy and minimal average prediction errors. It also achieves the highest R^2 score, underscoring its exceptional capability in explaining variance within the potential energy data. This performance is attributed to the ensemble nature of Random Forests, which fosters robustness and improved generalization compared to single-model approaches.

In contrast, models such as Ridge Regression demonstrate moderate predictive performance, excelling in computational efficiency but limited by their inability to capture complex nonlinear relationships. Decision Tree models show variable performance with lower computational cost but suffer from higher error rates, largely due to their susceptibility to overfitting on training data.

Overall, the radar chart presents a concise yet detailed comparative summary that supports informed decision-making when selecting modeling approaches. It balances considerations of accuracy, computational expense, and robustness, highlighting the advantages of sophisticated ensemble techniques in addressing the challenges posed by complex, high-dimensional simulation data typical of materials informatics.

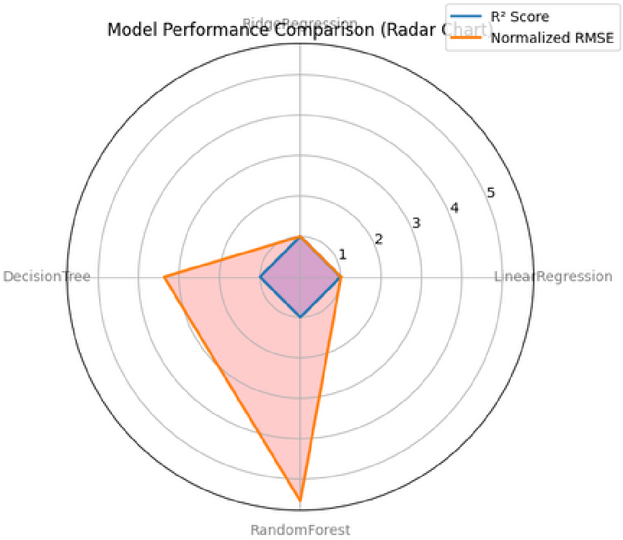


Figure 6. Shows the Radar Chart indicating the performances of different models based on R^2 and RMSE Score.

D) Prediction comparision of different models:

The study includes four scatter plots that provide an in-depth visual comparison of the predictive accuracies achieved by different machine learning algorithms—Linear Regression, Ridge Regression, Decision Tree, and Random Forest—in estimating the potential energy of magnesium. Each plot depicts predicted potential energy values against actual values derived from molecular dynamics simulations, with the ideal prediction accuracy represented by points lying on the diagonal y=xy=x line. The extent and pattern of deviations from this diagonal deliver valuable insights into each model's strengths and limitations in capturing the underlying physical behavior.

The Linear Regression plot reveals a broad dispersion of predicted values around the diagonal line, reflecting only moderate modeling capacity. The linear assumption that potential energy is a weighted sum of input features oversimplifies the complex atomic interactions present in magnesium. This is particularly evident at the extremes of the potential energy range, where deviations are largest, indicating the model's limited ability to capture nonlinearities and intricate interdependencies. As a result, the linear regression model exhibits elevated prediction errors and limited suitability for precise characterization of magnesium's energetic properties.

The Ridge Regression plot displays modest improvement over linear regression due to L2 regularization which penalizes large coefficients, thereby reducing variance and addressing multicollinearity among features more effectively. While prediction points cluster closer to the diagonal compared to linear regression, the fundamentally linear nature of the model constrains its ability to model complex nonlinear behaviors fully. Prediction errors remain pronounced, especially in states involving extreme potential energy values where material properties are influenced by subtle atomic-scale phenomena beyond linear trends.

The Decision Tree plot marks a significant transition toward nonlinear modeling by segmenting the feature space into distinct regions within which constant predictions are made. This piecewise approximation results in scatter plots showing dense clustering of predicted values near the diagonal in some regions, while other areas display pronounced scatter or outliers. Evidence of overfitting is apparent, as the decision tree captures fine details specific to the training data that do not generalize well to unseen test data, resulting in larger prediction errors in sparsely sampled regions. The tree's hierarchical structure produces abrupt changes in predictions, leading to visible discontinuities that cause deviations from the ideal y=xy=x line. These characteristics limit the model's reliability when applied across the entire physicochemical spectrum.

In contrast, the Random Forest plot exhibits superior predictive performance by aggregating multiple decision trees into a single ensemble. Through bagging and random feature selection, the ensemble mitigates overfitting and enhances prediction stability. The predicted values from Random Forest closely align with the y=xy=x line, showing minimal scatter and small deviations even in challenging data regions. This tight clustering illustrates the model's robustness in capturing complex, nonlinear dependencies governing magnesium's potential energy, while maintaining strong generalization to previously unseen data. The residual errors are uniformly small and symmetrically distributed, indicating consistent accuracy across the simulated thermodynamic conditions. The ensemble's ability to smooth individual tree variances and exploit complementary information enables nuanced and precise predictions superior to those of single-tree or linear models.

Collectively, these four plots illustrate a progression of modeling complexity: beginning with straightforward linear models constrained by simplicity, advancing through nonlinear but overfitting-prone decision trees, and culminating in the highly accurate and robust ensemble method of Random Forest. This visual narrative corroborates quantitative metrics such as RMSE and R2R2, underscoring the critical importance of nonlinear ensemble approaches for effective materials property prediction in the context of magnesium's dynamically intricate atomic behavior. These insights provide valuable guidance for future model selection, ensuring researchers choose algorithms well suited to the underlying data complexity and prediction objectives.

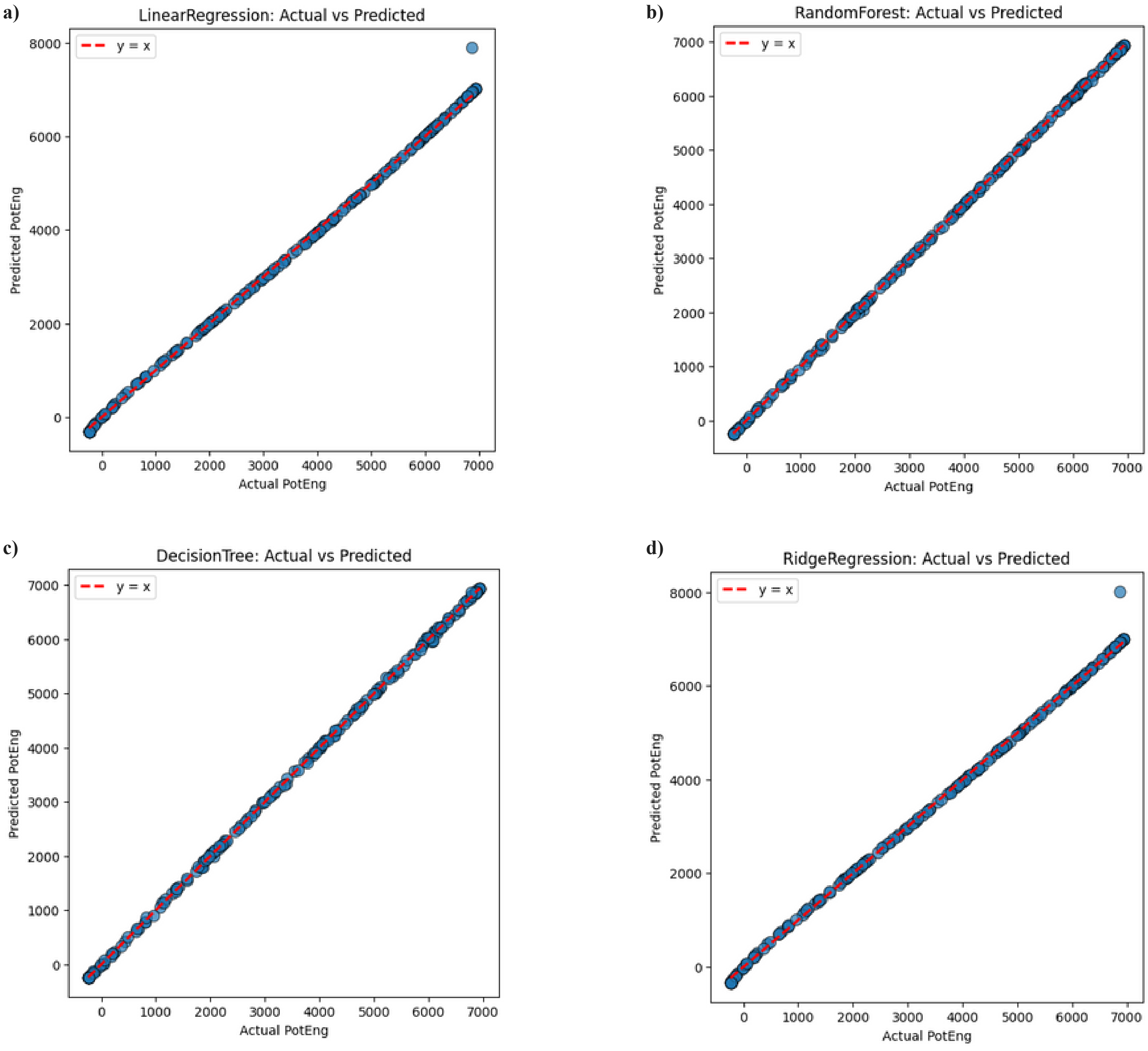


Figure 7. a) Shows the plot between Actual Potential Energy on X-axis and Predicted Energy on Y-axis for LinearRegression Model.
b) Shows the plot between Actual Potential Energy on X-axis and Predicted Energy on Y-axis for RandomForest Model.
c) Shows the plot between Actual Potential Energy on X-axis and Predicted Energy on Y-axis for DecisionTree Model.
d) Shows the plot between Actual Potential Energy on X-axis and Predicted Energy on Y-axis for RidgeRegression Model.

CONCLUSION

This study successfully developed and validated machine learning models capable of accurately predicting the potential energy of magnesium by utilizing data derived from molecular dynamics simulations. Among the algorithms evaluated, the Random Forest regression model demonstrated superior predictive accuracy and robustness, achieving a Root Mean Squared Error (RMSE) as low as 16.08 and a coefficient of determination (R^2) approaching 0.9999 on independent test data. The careful application of hyperparameter tuning and rigorous cross-validation was pivotal in optimizing model parameters, ensuring robust generalization to unseen thermodynamic conditions.

A detailed comparison of predicted versus actual potential energy values revealed that the Random Forest model's estimates closely aligned with simulation ground truth across the full range of variables, including temperature and pressure. The symmetrical distribution of residual errors around zero, accompanied by the absence of heteroscedasticity, confirmed unbiased and consistent predictive performance, thereby validating the model's reliability for material science applications.

Feature importance analyses identified temperature and pressure as the primary physical factors influencing variations in magnesium's potential energy, consistent with fundamental thermodynamic principles. Secondary contributions from kinetic energy and volume were also observed, highlighting the intricate interactions among physical parameters. The strong concordance between these data-driven insights and established scientific knowledge enhances the model's interpretability and credibility.

The integrated machine learning pipeline significantly reduces the computational burden traditionally associated with atomistic simulations, enabling rapid and accurate predictions that accelerate materials design and optimization workflows. This advancement is particularly impactful for industries relying on lightweight structural metals, such as aerospace and automotive sectors, where magnesium and its alloys are key enablers of performance and efficiency.
Moreover, the developed methodology offers a scalable and adaptable framework suitable for extension to more complex alloy systems and broader materials challenges. Incorporating additional physical descriptors or integrating multi-scale simulation data could further enhance predictive capability and practical utility.

In conclusion, this work represents a significant advancement in computational materials science by combining physics-based simulation data with state-of-the-art machine learning techniques. It provides efficient, interpretable, and high-fidelity property predictions that have the potential to transform traditional materials research paradigms and drive innovation in alignment with sustainability and technological progress.

REFERENCES:

1.T. Zhou, Z. Song, K. Sundmacher, Engineering 5 (6) (2019) 1017-1026.

2. D.M. Dimiduk, E.A. Holm, S.R. Niezgoda, Integrat. Mater. Manuf. In-nov. 7 (2018) 157-172.

3. A.G. Kusne, T. Gao, A. Mehta, L. Ke, M.C. Nguyen, K.-M. Ho, V. Antropov, C.-Z. Wang, M.J. Kramer, C. Long, Sci. Rep. 4 (1) (2014) 6367.

4. Y. Juan, Y. Dai, Y. Yang, J. Zhang, J. Mater. Sci. Technol. 79 (2021) 178-190.

5. J. Schmidt, M.R.G. Marques, S. Botti, M.A.L. Marques, npj Comput. Mater. 5 (1) (2019) 83.

6. R. Hussein, J. Schmidt, T. Barros, M.A.L. Marques, S. Botti, MRS Bull. 47 (8) (2022) 765-771.

7. T. Chen, Q. Gao, Y. Yuan, T. Li, Q. Xi, T. Liu, A. Tang, A. Watson,F. Pan, J. Magnes. Alloys (2021).

8. Y. Wang, T. Xie, Q. Tang, M. Wang, T. Ying, H. Zhu, X. Zeng, J. Magnes. Alloys (2022).

9. D.S. Wigh, J.M. Goodman, A.A. Lapkin, Wiley Interdiscip. Rev.: Com-put. Mol. Sci. 12 (5) (2022) e1603.

10. K.T. Butler, D.W. Davies, H. Cartwright, O. Isayev, A. Walsh, Nature 559 (7715) (2018) 547-555.

11. P. Gardner, R. Fuentes, N. Dervilis, C. Mineo, S. Pierce, E. Cross, K. Worden, Philos. Trans. R. Soc. A 378 (2182) (2020) 20190581.

12. J. Timoshenko, D. Lu, Y. Lin, A.I. Frenkel, J. Phys. Chem. Lett. 8 (20) (2017) 5091-5098.

13. Y. Kiarashinejad, S. Abdollahramezani, M. Zandehshahvar, O. Hemmat-yar, A. Adibi, Adv. Theory Simul. 2 (9) (2019) 1900088.

14. D.E. Jones, H. Ghandehari, J.C. Facelli, Comput. Methods Programs Biomed. 132 (2016) 93-103.