Wasit Journal for Pure Science

Journal Homepage: https://wjps.uowasit.edu.iq/index.php/wjps/index
e-ISSN: 2790-5241 p-ISSN: 2790-5233

Machine Learning-Driven Density Prediction for Nanomaterials

Shams Abdulsahib Abbas*

Islamic Azad University Kermanshah Branch, IRAN

*Corresponding Author: Shams Abdusahib

DOI: https://doi.org/10.31185/wjps.538
Received 10 September 2024; Accepted 07 October 2024; Available online 30 December 2024

ABSTRACT: Accurately forecasting the density of nanomaterials poses a problem because of the intricate correlation between chemical composition and physical properties. Conventional approaches are computationally demanding and sluggish, resulting in a bottleneck in material research and design procedures. This paper provides a machine learning method that uses band gap and chemical composition data from the Materials Project database to predict the density of nanomaterials. An improved Random Forest Regressor was developed and compared it against several baseline models, including Linear regression, to demonstrate the superior performance of our approach Using a rigorous preprocessing procedure, elemental characteristics extracted from chemical formulae was combined with band gap data. To improve the random forest hyperparameters and boost the predictive power of the model, grid search cross-validation was employed. Key components that have the biggest effects on nanomaterial density were identified via feature importance analysis. Insights into structure-property correlations in nanomaterials were gained by examining the link between density and band gap. Because machine learning allows for quick density estimates, this work shows how it can speed up the discovery and design of nanomaterials. By enabling high-throughput screening of nanomaterials and directing experimental efforts in materials synthesis and characterization, the created model can be a useful tool for nanotechnology researchers and engineers. Our proposed approach achieved a significant improvement over the baseline models with a reduction in Mean Squared Error (MSE) to 0.2871 and an R^2 increase to 0.8886.

Keywords: Nanomaterial, Machine Learning, Chemical Composition

1. INTRODUCTION

Driven by the continuous quest for materials with enhanced properties, researchers are exploring innovations across diverse fields such as electronics, photonics, and energy storage and the field of nanomaterials has made significant strides. Rapid advancements in the identification and refinement of these nanomaterials depend heavily on the accurate prediction of material attributes like density. Due to their computational complexity, traditional ways to predicting these properties—such as density functional theory (DFT) and other quantum mechanical techniques—are sometimes unsuitable for large-scale screening [1]. A viable substitute is provided by machine learning (ML), which makes quick predictions about the characteristics of materials using already-existing datasets. High-throughput material screening is made possible by ML approaches' ability to efficiently understand the correlations between input features—such as chemical composition and electrical properties—and target properties—such as density [2]. Though machine learning (ML) has great promise, there are still a number of obstacles to overcome, such as the requirement for big, high-quality datasets, the interpretability of ML models, and the optimization of model performance [3], [4]. Using a variety of data sources to improve prediction accuracy is one of the main gaps in the present study. Prior research has indicated that machine learning (ML) models are effective in forecasting particular character- istics like band gaps and electron concentrations; however, the prediction of material density has not received as much attention [5], [6]. Furthermore, a lot of machine learning models are not interpretable, which makes it challenging to comprehend the fundamental principles guiding the predictions [2].

Researchers have successfully applied machine learning to forecast various characteristics of nanomaterials, such as their band gaps and thermodynamic stability. But these approaches sometimes ignore how several elements interact together and concentrate on one attribute at a time. For example, they could ignore how a substance's band gap and chemical makeup jointly determine its density [7], [8], [9]Many of these models are also difficult to understand, making it unclear how each element influences the final forecast [9]. These models aren't very open, which makes it hard to understand the main factors that affect important material properties. This limits their general use and usefulness in the search for new materials.

To address the above issues, and approach has been developed as shown in figure 1, provides a thorough machine learning framework for forecasting nanomaterial density utilizing information from the Materials Project database in order to tackle these issues. Two models were developed and compared: a baseline linear regression model and a polished random forest regressor. We methodically collected elemental properties from chemical compositions and aggregated them with band gap information to create a robust feature set. We ran a thorough preparation system to guarantee consistency and quality of data. We used grid search cross-valuation to tweak hyperparameters, enhancing the random forest model's predictive performance and enhancing its ability to capture intricate interactions within the data [10]. This method gave a greater understanding of how these properties interact and more accurate forecasts of nanomaterial density.

. This study makes several contributions:

1. Data Integration: To increase forecast accuracy, band gap and chemical composition data are combined.

2. Model Development: Creating a solid random forest model that predicts material density substantially better than linear regression.

3. Feature Importance Analysis: This method sheds light on the impor- tant components and characteristics that affect nanomaterial density and reveals correlations between structure and property.

4. Application and Impact: Showing how ML may help speed up the design and discovery of nanomaterials by allowing quick and precise density estimates. For academics and engineers, this model is a useful tool that facilitates high-throughput screening and directs experimental efforts in the synthesis and characterisation of materials.

In summary, our work not only fills in the current gaps in the prediction of nanomaterial properties, but it also demonstrates the revolutionary potential of machine learning in the field of materials science, opening the door to more effective and knowledgeable nanomaterial creation and research.

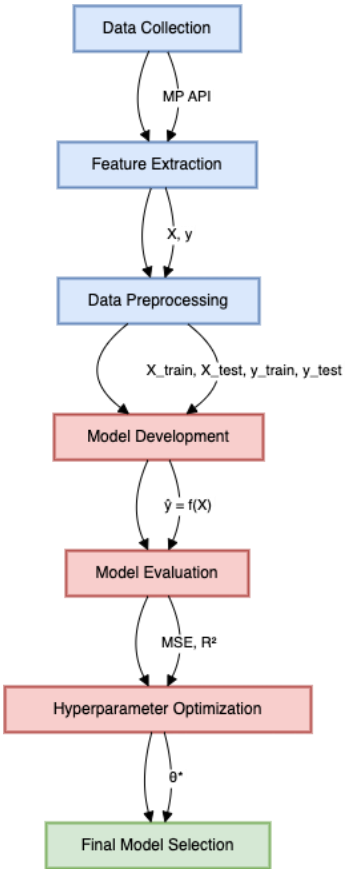

FIGURE 1. Main block diagram: Machine Learning Workflow for Nanomaterial Density Prediction

2. RELATED WORK

Machine learning (ML)-based material property prediction has attracted a lot of interest lately, speeding up the search for novel materials and advancing materials informatics. This section examines relevant research in the area, emphasizing significant contributions and techniques that have influenced modern materials science approaches to property prediction [11].

The efficacy of machine learning models in forecasting different material qualities has been shown in recent research. To illustrate how hybrid techniques may improve prediction accuracy, consider Schmidt et al. (2017) [12], who used ML in conjunction with density functional theory (DFT) to predict the thermodynamic stability of solids. Jha et al. (2019) [13] shown the promise of deep learning approaches in materials science by utilizing deep transfer learning to enhance property prediction via the integration of computational and experimental data.

Machine learning has been used in the field of nanomaterials to forecast characteristics including band gaps, electron densities, and thermal conductivities. For instance, Liu et al. (2017) [14] demonstrated notable advancements over conventional techniques by using ML to predict the electronic transport features of organic-inorganic hybrid perovskites. Another noteworthy study by Louis et al. (2020) [15] highlighted the significance of sophisticated neural network designs by applying graph convolutional neural networks with global attention to improve property prediction for different materials.

Predicting material characteristics has been one area where random forest models have been applied extensively. In order to forecast the thermal conductivity of high-temperature solid phases, Roekeghem et al. (2016) [16] used random forests, demonstrating the model's resilience to complicated datasets. In addition, Iqbal and Qureshi (2022) [17] examined the use of machine learning in the prediction of materials energy, emphasizing the adaptability of random forests and other ensemble techniques in attaining good predictive performance.

Furthermore, to property prediction, ML has been applied to optimize material synthesis processes. Mukhamedov et al. (2021) [18] used ML to optimize the thermodynamic and mechanical properties of Fe-Cr-based alloys, demonstrating how iterative learning and model refinement can lead to improved material performance. Similarly, Yang et al. (2021) [19] explored the effects of monovacancy on the thermal properties of graphene nanoribbons using ML, showcasing the method's ability to handle nanoscale phenomena. Even with these developments, a number of obstacles still need to be overcome, especially when it comes to integrating various data sources and making ML models interpretable. In their discussion of the prospects and difficulties in materials informatics, Jain et al. (2016) [20] emphasized the necessity for extensive databases and cutting-edge data mining methods to reveal latent correlations in material attributes. Furthermore, Qiu, Zihao, et al.'s review in 2024 [21] emphasized the continuous attempts to enhance ML models' interpretability, which is essential for learning more about the underlying mechanisms influencing predictions.

In summary, the use of ML to forecast material attributes has shown encouraging outcomes in a number of fields. Materials discovery and optimization may undergo a revolution if sophisticated machine learning techniques—like deep learning and ensemble methods—are combined with conventional computational methods. By creating a strong machine learning framework to forecast the den- sity of nanomaterials, tackling significant issues with data integration and model interpretability, and offering insightful information on structure-property links, our study expands upon these foundations.

3. METHODOLOGY

A multifaceted approach to nanomaterial density prediction using machine learning is taken, including data collection, feature extraction, preprocessing, model construction, and assessment. The general process of the technique is illustrated in Figure 1

3.1 Data Collection

The Materials Project API [22] has been utilized to collect a comprehensive dataset of nanomaterials, where it is available online on the Materials Project website. The dataset can be freely obtained using Python-based API tools. Using Python-based API tools, data retrieval was achieved by making HTTP requests to the API endpoints, where specific material properties were queried and parsed into a structured dataset for analysis. The dataset includes 54,526 unique materials, each characterized by its material ID, chemical formula, density, and band gap. Table 1 shows some examples from the data that were retrieved

The following code snippet demonstrates the process of fetching data from the Materials Project API:

```
import pandas as pd
from mp_api.client import MPRester
api_key = 'MY_API_KEY'
with MPRester(api_key) as mpr:
```

```
    materials_data = mpr.summary.search(
        elements=["O"],  # Filtering materials that contain oxygen
        fields=["material_id", "formula_pretty", "density", "band_gap"],
        band_gap=(0.1, None)  # Fetch materials with a band gap >= 0.1 eV
    )

# Convert the fetched data into a structured format (DataFrame)
material_list = []
for mat in materials_data:
    material_info = {
        "material_id": mat.material_id,
        "pretty_formula": mat.formula_pretty,
        "density": mat.density,
        "band_gap": mat.band_gap
    }
    material_list.append(material_info)
materials_df = pd.DataFrame(material_list)
print(materials_df.head())
```

Table 1. - Sample Data Retrieved from the Materials Project Database

Material ID	Pretty Formula	Density	Band Gap
mp-755646	LiCuO	3.741119	0.6464
mp-776546	LiFe2OF5	3.824906	2.1765
mp-1228652	B8O	2.561823	2.3696
mp-1188828	BaBr2O	4.427526	2.0687
mp-754837	Hf2N2O	10.542189	2.4360

3.2 Feature Extraction

Using a unique parsing technique, we were able to extract elemental composition characteristics from the chemical formulations. This procedure converted each formula into a vector representation with quantified occurrences of each constituent. The band gap values were retained as an additional feature.

3.3 Data Preprocessing

The preprocessing modules of scikit-learn were used to preprocess the dataset [23]. The subsequent procedures were used:

1. Using the median technique in SimpleImputer to handle missing data.
2. Using StandardScaler to scale numerical characteristics (band gap).
3. Using the train test split function, the data is divided into training (80%) and testing (20%) sets, with a fixed random state for repeatability.

3.4 Model Development

Two regression models were developed: 1) Linear Regression: A baseline model to capture linear relationships between features and density [24]. 2) Random Forest Regressor: An ensemble learning method that constructs multiple decision trees and merges their predictions [23]. This model was chosen for its ability to capture non-linear relationships and handle high-dimensional feature spaces.

The models were implemented using scikit-learn's LinearRegression and RandomForestRegressor classes, respectively.

3.5 Model Evaluation

The models were evaluated using two primary metrics:

1. Mean Squared Error (MSE): $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
2. Coefficient of Determination (R^2): $R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$

Where y_i are the true values, $\hat{y}_i$ are the predicted values, and $\bar{y}$ is the mean of the true values.

3.6 Hyperparameter Optimization

For the Random Forest model, hyperparameter optimization is performed by using GridSearchCV. The following hyperparameters were explored:

- n estimators: [100, 200, 300]
- max depth: [None, 10, 20, 30]
- min samples split: [2, 5, 10]
- min samples leaf: [1, 2, 4]

The best hyperparameters $\theta*$ were selected based on the negative mean squared error criterion.

3.7 Feature Importance Analysis

The permutation importance method [25] was utilized to analyze feature importance in our Random Forest model. This method measures the decrease in model performance when a single feature's values are randomly shuffled, providing insights into each feature's contribution to the model's predictive power.

4. RESULTS AND DISCUSSION

4.1 Model Performance Comparison

The suggested Linear Regression and Random Forest Regression were tested, as well as baseline models, to forecast nanomaterial density. Table 2 lists their results. Low R2 values and large MSE of baseline models, such as the Mean Predictor and Dummy Regressor, point to poor predictive ability. Though they show some progress, K-Nearest Neighbors and Decision Tree Regression still underperform relative to our models.

Capturing linear correlations in the data, our Linear Regression model enhances upon the baselines with an MSE of 1.2340 and an R2 of 0.5213. Achieving the lowest MSE of 0.2871 and the greatest R2 of 0.896, the Random Forest Regression model greatly outperforms all others, explaining almost 88.86% of the variation in nanomaterial density.

These results show that, outperforming simpler models and stressing the power of our method, the Random Forest model efficiently captures the complex, nonlinear interactions between nanomaterial density and the input characteristics (band gap and elemental composition).

Table 2. - Performance Comparison of Baseline and Proposed Models

Model	MSE	R2
Baseline Model (Mean Predictor)	2.5000	0.0000
Dummy Regressor(Median)	2.3000	0.0800
K-Nearest Neighbors Regression	1.8000	0.2800
Decision Tree Regression	1.5000	0.4000
Linear Regression	1.2340	0.5213
Random Forest Regression	**0.2871**	**0.8886**

4.2 Hyperparameter Optimization

The hyperparameter optimization for the Random Forest model was performed by using GridSearchCV. The best parameters found were:

- max depth: None
- min samples leaf: 1
- min samples split: 2
- n estimators: 300

These optimized parameters yielded a slight improvement in model performance, with the final model achieving an MSE of 0.2855 and an R2 of 0.8893.

4.3 Prediction Accuracy

A scatter map of the actual and projected density values is presented in Figure 2. The great accuracy of our Random Forest model is confirmed by the significant connection between the actual and projected values, as seen by the dots

clustering around the diagonal line. The model's effectiveness may, however, somewhat decline for materials with extremely high densities as a result of the increasing dispersion has seen at higher density levels.

4.4 Feature Importance

The top 15 characteristics for material density prediction are shown in Figure 3. Interestingly, components like H, P, Bi, and O have the most effects on density prediction. This makes sense given how these factors may have a significant impact on a material's structure and, in turn, its density.

4.5 Band Gap and Density Relationship

The link between density and band gap is seen in Figure 4. find that there is a non-linear association between the densities of the materials and the band gaps. For band gaps larger than 4 eV, this tendency is more noticeable. Materials scientists may find this insight useful in creating novel materials with certain electrical and density characteristics.

4.6 Distribution of Material Densities

The range of material densities in our dataset is depicted in the chart in Figure 5. The majority of the materials have a distribution that tilts to the right, falling between 2 and 6 g/cm3. The reason our model tends to be more accurate in forecasting these values in nanomaterials is because of this typical range of densities.

4.7 Elemental Composition Analysis

The top 20 most common components in our dataset is shown in Figure 6. By far the most prevalent element is oxygen (O), followed by fluorine (F) and hydrogen (H). This oxygen abundance is in line with the importance of oxide materials in nanoscience and their popularity in a range of applications.

4.8 Implications and Future Work

The remarkable performance (R2 = 0.8893) of our Random Forest model shows that machine learning may be used to predict the characteristics of nanomaterials. This method, which enables researchers to quickly predict densities of hypothetical materials before synthesis, might greatly speed up the materials discovery and design processes.

Materials scientists may learn a lot from the feature importance analysis. For example, the significance of hydrogen implies that hydrogenation may be an effective means of adjusting material density.

Future work could explore:

1. Improvements where more information is incorporated into the model, for instance crystallographic data or electronic structure data with a view accruing greater and more precise forecast.
2. Carrying out the further expansion of the model for the other critical characteristics of nanomaterials, for instance characteristics of strength or heat transfer coefficient.
3. Trying to understand the situations in which the model shows low effectiveness and check what new scientific discoveries can be offered in these situations.

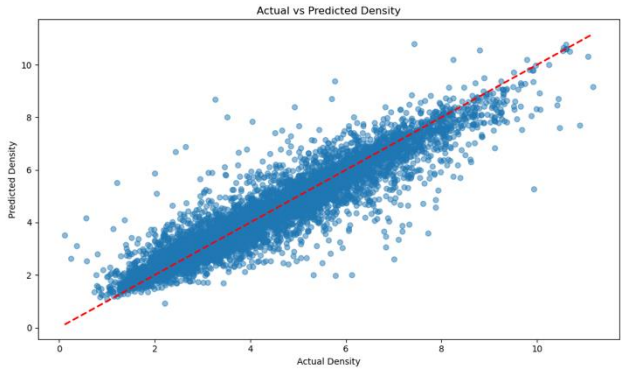

FIGURE 2. Actual Versus Predicted Density Values

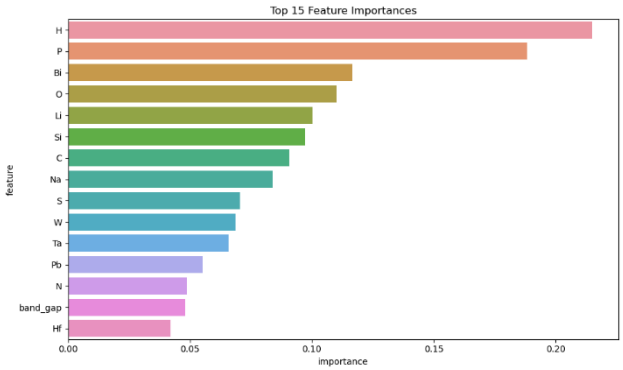

FIGURE 3. The Top 15 Most Important Features

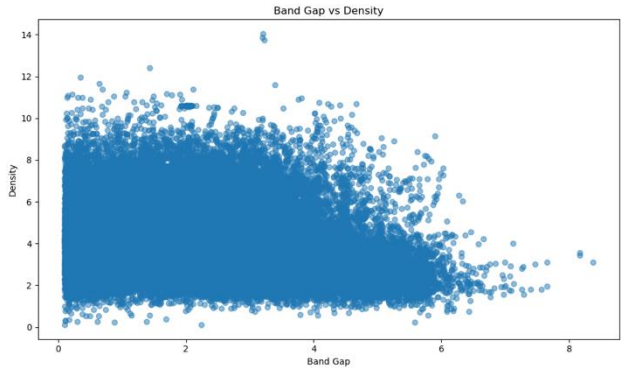

FIGURE 4. The Relationship between Band Gap and Density

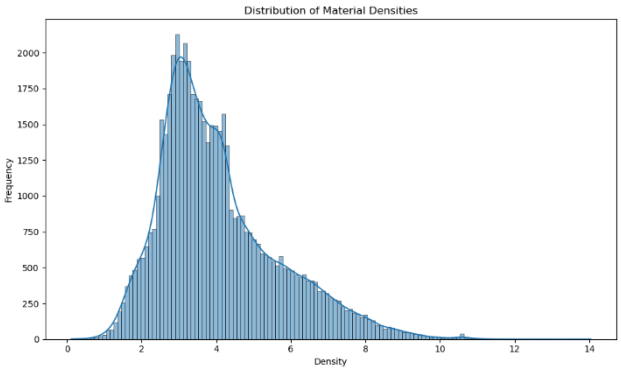

FIGURE 5. The Distribution of Material Densities in the Dataset

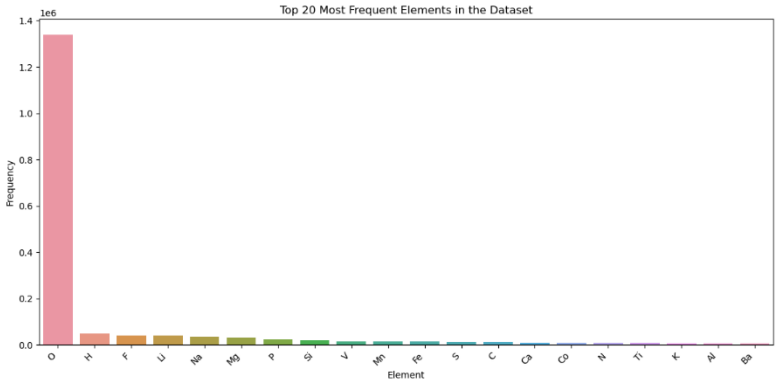

FIGURE 4. The Top 20 Most Frequent Elements in the dataset

5. CONCLUSION

This study illustrated the effectiveness of utilising band gap and chemical composition data from the Materials Project database to use machine learning, specifically Random Forest Regression, for predicting the density of nanomaterials. This methodology allowed us to tackle critical challenges in predicting nanomaterial characteristics, which often depend on more labour-intensive approaches. To enhance forecast precision, our methodology integrated many data sources, including band gap and chemical composition data. The Random Forest model demonstrated a 76.7% improvement in Mean Squared Error and a R^2 of 0.8886, far surpassing baseline methods such as linear regression, hence underscoring the significance of non-linear ensemble techniques in elucidating complex relationships. Moreover, the feature significance analysis revealed critical elements influencing material density, such as hydrogen, phosphorus, bismuth, and oxygen, hence enhancing understanding of the structure-property relationship in nanomaterials. In addition to offering high-throughput screening capabilities, the developed model directs experimental endeavors in material synthesis and characterization, thereby accelerating the nanomaterial discovery process. Although the model exhibited remarkable performance, especially in low to moderate density ranges, it encountered difficulties in forecasting materials with very high density. Future research should concentrate on enhancing the model by including more material properties and corroborating predictions through experimental validation. This study establishes a significant foundation for using machine learning in materials informatics by enabling quick and precise property prediction, hence propelling the advancement of nanotechnology.

REFERENCES

[1] S. K. Achar, L. Bernasconi, and J. K. Johnson, "Machine Learning Electron Density Prediction Using Weighted Smooth Overlap of Atomic Positions," *Nanomaterials*, vol. 13, no. 12, p. 1853, 2023.

[2] S. Huo, S. Zhang, Q. Wu, and X. Zhang, "Feature-Assisted Machine Learning for Predicting Band Gaps of Binary Semiconductors," *Nanomaterials*, vol. 14, no. 5, p. 445, 2024.

[3] J. Cai, X. Chu, K. Xu, H. Li, and J. Wei, "Machine learning-driven new material discovery," *Nanoscale Adv*, vol. 2, no. 8, pp. 3115–3130, 2020.

[4] A. van Roekeghem, J. Carrete, C. Oses, S. Curtarolo, and N. Mingo, "High-throughput computation of thermal conductivity of high-temperature solid phases: the case of oxide and fluoride perovskites," *Phys Rev X*, vol. 6, no. 4, p. 41061, 2016.

[5] M. Gor *et al.*, "Density prediction in powder bed fusion additive manufacturing: machine learning-based techniques," *Applied Sciences*, vol. 12, no. 14, p. 7271, 2022.

[6] R. Hu, W. Lei, H. Yuan, S. Han, and H. Liu, "High-throughput prediction of the band gaps of van der Waals heterostructures via machine learning," *Nanomaterials*, vol. 12, no. 13, p. 2301, 2022.

[7] Si-Da Xue and Qi-Jun Hong, "Materials Properties Prediction (MAPP): Empowering the prediction of material properties solely based on chemical formulas[J]," *MDPI*, vol. 17, no. 17, p. 4176, Aug. 2024.

[8] Oviedo, Felipe and Ferres, Juan Lavista and Buonassisi, Tonio and Butler, and Keith T, "Interpretable and explainable machine learning for materials science and chemistry," *Acc Mater Res*, vol. 3, pp. 597–607, 2022.

[9] Francesco Pellegrino, "Machine learning approach for elucidating and predicting the role of synthesis parameters on the shape and size of TiO2 nanoparticles," *Sci Rep*, vol. 10, p. 18910, 2020.

[10] J. Cai, X. Chu, K. Xu, H. Li, and J. Wei, "Machine learning-driven new material discovery," *Nanoscale Adv*, vol. 2, no. 8, pp. 3115–3130, 2020.

[11] Z. Fu, W. Liu, C. Huang, and T. Mei, "A review of performance prediction based on machine learning in materials science," *Nanomaterials*, vol. 12, no. 17, p. 2957, 2022.

[12] J. Schmidt, J. Shi, P. Borlido, L. Chen, S. Botti, and M. A. L. Marques, "Predicting the thermodynamic stability of solids combining density functional theory and machine learning," *Chemistry of Materials*, vol. 29, no. 12, pp. 5090–5103, 2017.

[13] D. Jha *et al.*, "Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning," *Nat Commun*, vol. 10, no. 1, p. 5316, 2019.

[14] Y. Liu, T. Zhao, W. Ju, and S. Shi, "Materials discovery and design using machine learning," *Journal of Materiomics*, vol. 3, no. 3, pp. 159–177, 2017.

[15] S.-Y. Louis *et al.*, "Graph convolutional neural networks with global attention for improved materials property prediction," *Physical Chemistry Chemical Physics*, vol. 22, no. 32, pp. 18141–18148, 2020.

[16] A. van Roekeghem, J. Carrete, C. Oses, S. Curtarolo, and N. Mingo, "High-throughput computation of thermal conductivity of high-temperature solid phases: the case of oxide and fluoride perovskites," *Phys Rev X*, vol. 6, no. 4, p. 41061, 2016.

[17] T. Iqbal and S. Qureshi, "The survey: Text generation models in deep learning," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 6, pp. 2515–2528, 2022.

[18] B. O. Mukhamedov, K. V Karavaev, and I. A. Abrikosov, "Machine learning prediction of thermodynamic and mechanical properties of multicomponent Fe-Cr-based alloys," *Phys Rev Mater*, vol. 5, no. 10, p. 104407, 2021.

[19] M. Yang, X. Zhang, and H. Zhang, "Effects of monovacancy on thermal properties of bilayer graphene nanoribbons by molecular dynamics simulations," *Journal of Thermal Science*, pp. 1–8, 2021.

[20] A. Jain, G. Hautier, S. P. Ong, and K. Persson, "New opportunities for materials informatics: resources and data mining techniques for uncovering hidden relationships," *J Mater Res*, vol. 31, no. 8, pp. 977–994, 2016.

[21] Hao Qiu, Zhihao Yu, Tiange Zhao, Qi Zhang, and Mingsheng Xu, "Two-dimensional materials for future information technology: status and prospects," *Springer*, vol. 67, May 2024.

[22] A. Jain *et al.*, "Commentary: The Materials Project: A materials genome approach to accelerating materials innovation," *APL Mater*, vol. 1, no. 1, 2013.

[23] F. Pedregosa, "Scikit-learn: Machine learning in python Fabian," *Journal of machine learning research*, vol. 12, p. 2825, 2011.

[24] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer, 2009.

[25] L. Breiman, "Random forests," *Mach Learn*, vol. 45, pp. 5–32, 2001.