https://doi.org/10.48047/AFJBS.6.13.2024.6875-6893

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

Optimizing Density Functional Theory Calculations Using a Gradient Boosting Machine for Enhanced Predictive Accuracy

Y.P. Arul teen[1] and C. Justin dhanaraj[2]*

[1]Department of Electronics and Communication Engineering, University College of Engineering, Nagercoil. Kanyakumari Dist. Tamil Nadu. 629004 India

[2]Department of Chemistry, University College of Engineering, Nagercoil. Kanyakumari Dist. Tamil Nadu. 629004 India

*Corresponding author Email: c.justindhanaraj@gmail.com

Volume 6, Issue 13, Aug 2024

Received: 15 June 2024

Accepted: 25 July 2024

Published: 15 Aug 2024

doi: 10.48047/AFJBS.6.13.2024.6875-6893

Abstract: Density Functional Theory (DFT) calculations are essential for understanding molecular properties and predicting biological activities based on quantum mechanical principles. Predictive performance is limited by the inability of conventional machine learning (ML) models, like random forests and linear regression, to adequately represent the complicated, nonlinear correlations seen in DFT data. These methods often fail to account for the intricate dependencies between molecular descriptors and target variables, resulting in suboptimal accuracy. Additionally, interpreting traditional models in the context of DFT calculations is difficult, hindering the elucidation of structure-property relationships. This research proposes the application of Gradient Boosting Machines (GBM) for the predictive modelling of DFT calculations. GBM is an ensemble learning technique that enhances overall accuracy by combining the predictive power of several weak learners, such decision trees. GBM captures complicated nonlinear interactions in data by iteratively fitting new models to prior models' residuals. This makes it well-suited for analysing DFT calculations and predicting material properties with high accuracy. The research utilizes two datasets: tmQM, containing information on transition metal complexes, and ECD-cubic, focusing on the electronic charge density of inorganic materials with cubic structures. The GBM model is trained iteratively, with each new tree pointing on correcting the errors produced by previous trees. The optimal tree size and learning rate are determined through grid search optimization. The model's performance is assessed using mean squared error (MSE), mean absolute error (MAE), and R-squared (R^2) metrics. The GBM model demonstrates high accuracy and low error metrics, indicating its robust performance in capturing the complex relationships inherent in DFT data. For the tmQM dataset, the GBM model achieves a lower MSE and MAE on the testing set (0.018 and 0.092, respectively) compared to the training set (0.021 and 0.105). Similarly, for the ECD-cubic dataset, the model exhibits a lower MSE and MAE on the testing set (0.029 and 0.118) than the training set (0.035 and 0.132). The high R^2 values (0.995 for tmQM and 0.996 for ECD-cubic) indicate that the model describes a large percentage of the variance in the target properties, demonstrating its predictive power.

Keywords: DFT, GBM, Predictive Modelling, Materials Discovery, , Feature Selection, , Molecular Properties.

1. Introduction

DFT is a crucial computational method in quantum chemistry for predicting molecular properties and biological activities [1]. It helps in understanding the electronic structure of molecules, which is vital for the design and discovery of new materials [2]. DFT calculations are based on quantum mechanical principles and provide detailed insights into the molecular structure, electronic charge distribution, and other properties [3]. Despite its importance, DFT calculations are computationally intensive and require significant resources, making them challenging to perform on a large scale [4][5]. The advent of ML has opened new avenues for enhancing the analysis of DFT data [6]. ML models can learn from large datasets to uncover hidden patterns and relationships, thereby improving the predictive accuracy of DFT calculations [7]. Traditional ML methods like linear regression and random forests have been used to predict molecular properties from DFT data [8][9]. However, these methods often fall short in capturing the complex, nonlinear relationships inherent in quantum mechanical systems [10-12]. This limitation restricts their predictive performance and hampers their utility in practical applications [13][14]. Several software packages are widely used for DFT calculations [15]. The Vienna Ab initio Simulation Package (VASP) is a popular tool that performs electronic structure calculations and quantum-mechanical molecular dynamics from first principles. VASP uses DFT to calculate the total energy and electronic structure of solids, making it invaluable for materials science research. Despite its capabilities, VASP, like other DFT software, requires substantial computational power and time, limiting its use in high-throughput screening of materials.

To overcome the limitations of traditional ML methods in DFT calculations, this research proposes the use of GBM. GBM is an ensemble learning technique that combines multiple weak learners, such as decision trees, to improve predictive accuracy. Unlike traditional methods, GBM can effectively capture complex nonlinear relationships in the data by iteratively fitting new models to the residuals of previous models. This makes GBM particularly suitable for analysing the intricate dependencies in DFT calculations. The proposed method begins with the collection of two comprehensive datasets: tmQM and ECD-cubic. The tmQM dataset includes information on 86,665 transition metal complexes, providing data on various molecular properties calculated using DFT. The ECD-cubic dataset contains electronic charge density information for 17,418 inorganic materials with cubic structures. These datasets offer a diverse range of chemical compounds and properties, enabling the development of a robust predictive model using GBM. Preprocessing the data is a critical step to enhance the quality and usability of the datasets. This research employs several preprocessing techniques, including feature scaling using Z-score normalization, encoding categorical features with one-hot encoding, outlier removal using the Interquartile Range (IQR) method, and dimensionality reduction with t-Distributed Stochastic Neighbour Embedding (t-SNE). These steps ensure that the data is well-prepared for modelling, improving the predictive performance of the GBM model. Feature selection is performed using Recursive Feature Elimination (RFE), which identifies the most relevant features for predicting material properties from DFT calculations. RFE enhances the model's performance by recursively removing the least important features, ensuring that only the most significant ones are retained for training. This process results in a refined dataset that is more manageable and effective for modelling. The GBM model is trained iteratively, with every new tree pointing on correcting the error produced by previous trees. The tree size and learning rate, crucial hyperparameters in GBM, are optimized through a grid search process. The performance of the trained model is evaluated

using metrics such as MSE, MAE, and R^2. These metrics give quantitative measures of model's accuracy in predicting material properties from DFT calculations.

The proposed method addresses the limitations of traditional ML models by capturing the complex relationships inherent in DFT data. The use of GBM significantly improves predictive accuracy, making it a valuable tool for materials discovery and design. The high accuracy and low error metrics achieved by the GBM model demonstrate its robustness and effectiveness in predicting molecular properties. This approach enhances the efficiency of DFT calculations and provides deeper insights into molecular behaviour and structure-property relationships.

Research contributions of the proposed study are :

1. Development of a robust predictive model for DFT calculations using GBM.
2. Collection and preprocessing of two comprehensive datasets: tmQM and ECD-cubic.
3. Implementation of RFE for feature selection.
4. Optimization of GBM hyperparameters to achieve high predictive accuracy.

This research presents a novel approach to DFT calculations using GBM. By addressing the limitations of traditional ML methods, this study enhances the predictive accuracy and efficiency of DFT calculations. The proposed method leverages the strengths of GBM to capture the complex relationships in DFT data, providing a powerful tool for materials discovery and design. The findings of this research have notable insinuations in quantum chemistry and materials science, offering a pathway to more efficient and accurate predictions of molecular properties.

4. Literature review

The integration of ML with DFT has gained traction in computational chemistry, aiming to reduce computational costs and enhance predictive accuracy. This review examines several studies that propose various ML-DFT hybrid approaches, highlighting their methodologies, datasets, and inherent challenges. Each study aims to improve DFT calculations' efficiency and accuracy, yet they encounter specific limitations that impact their generalizability and practical application.

Xuhao Wan et al [16] present a comprehensive study on using a DFT-ML hybrid scheme for intricate system catalysis. They developed the DMCP program to implement this scheme efficiently. The purpose is to reduce the computational cost of traditional DFT methods while maintaining high accuracy. The proposed method combines DFT calculations with ML models to predict catalytic properties. The datasets used include those from the Materials Project, AFLOW, and ICSD. However, the disadvantage lies in the complexity of feature selection and the need for extensive domain knowledge for accurate model training, which can limit the generalizability of the model to other material systems.

Reynolds et al [17] developed a neural network model to predict spin-state ordering and bond lengths in first-row transition metal chelates. They generate datasets of octahedral complexes and perform DFT calculations using TeraChem. The primary challenge highlighted is the difficulty in obtaining stable minimized geometries and managing large spin contamination, which can lead to

inaccuracies in predictions and limit the practical applicability of their model in high-throughput screenings.

Fiedler et al [18] focus on combining ML with DFT to accelerate materials discovery and electronic structure simulations. They review the integration of ML models to enhance the effectiveness and accuracy of DFT calculations. The dataset includes a comprehensive collection of over 300 research articles covering various materials and chemical systems. However, the primary disadvantage highlighted in this study is the challenge of maintaining accuracy when combining different ML techniques with DFT calculations. This can lead to inconsistencies and a lack of generalization across diverse material systems, which is critical for the scalability and reliability of the proposed methods.

Riemelmoser et al [19] aims to enhance the applicability of the random-phase approximation (RPA) by integrating it with ML. They propose the machine-learned RPA model (ML-RPA), which maps RPA data to a Kohn−Sham density functional. This model uses nonlocal density descriptors as ingredients. The dataset includes information on diamond surfaces and liquid water. However, the ML-RPA struggles with capturing the second peak in the oxygen-oxygen radial distribution function of water and underbinds the water dimer. These limitations highlight the challenges in accurately modeling nonlocal interactions with a small cutoff radius, affecting the model's predictive accuracy in diverse material systems.

Riemelmoser et al [20] propose a machine-learned density functional on the basis of random-phase approximation (RPA). Their method involves constructing ingredients for ML-DFT, analogous to two- and three-body descriptors used in ML force fields. They use the G2-2 database for training, which includes nonspin-polarized molecules containing C, O, and H. The main disadvantage is that the method requires significant computational resources for training and may not generalize well to all chemical systems, limiting its practical application in diverse DFT calculations.

Del Rio et al. [21] seek to lower the computational amount of solving the Kohn-Sham equation in DFT. They offer an end-to-end machine learning type that simulates DFT by mapping atomic structure to electronic charge density, then predicting features including density of states, potential energy, atomic forces, and stress tensor. The dataset includes organic molecules, polymer chains, and polymer crystals made up of carbon, hydrogen, nitrogen, and oxygen atoms. The main downside is that the model's accuracy suffers when applied to bigger systems than those used in training, limiting its generalizability across varied material systems.

Xuhao Wan et al present a DFT-ML hybrid scheme using the DMCP program to reduce computational costs. However, the complexity of feature selection and the need for domain knowledge limit the model's generalizability. Reynolds et al. use neural networks for predicting spin-state ordering but face challenges with stable geometries and spin contamination. Fiedler et al. combine ML with DFT to accelerate material discovery but struggle with maintaining accuracy across diverse systems. Riemelmoser et al. enhance RPA with ML but fail to model nonlocal interactions accurately. Del Rio et al. propose an end-to-end ML model to emulate DFT, yet the accuracy diminishes with larger systems. The proposed method addresses these issues by

iteratively refining models and optimizing feature selection, thereby improving accuracy and efficiency.

3. Proposed methodology

In current years, researchers have increasingly turned to ML approaches to augment the analysis of DFT data and improve predictive accuracy. ML models offer the advantage of learning from large datasets to uncover hidden patterns and relationships, thereby enhancing our understanding of molecular behaviour. Nevertheless, traditional ML algorithms, such as linear regression and random forests, may struggle to capture the nonlinear relationships inherent in DFT calculations, limiting their predictive performance. One of the primary challenges in DFT calculations is the accurate prediction of molecular properties and biological activities based on the underlying quantum mechanical principles. Even though ML methods are widely used, existing approaches often fall short of capturing the intricate dependencies between molecular descriptors and target variables. Moreover, the interpretation of ML models in the context of DFT remains very difficult, hindering the elucidation of structure-property relationships. To address these challenges, we propose the application of GBM for predictive modelling of DFT calculations. GBM is a powerful ensemble learning technique that combines the predictive strength of multiple weak learners, such as decision trees, to improve overall accuracy. By repeatedly fitting new models to the residuals of the previous models, GBM effectively captures complex nonlinear relationships in the data, making it well-suited for analysing DFT calculations. Figure 1 shows the process flow for developing and evaluating a GBM model for predicting material properties from DFT calculations.

The process begins with the collection of two datasets: mQM and ECD-cubic, which contain information on transition metal complexes and inorganic materials, respectively. The data undergoes preprocessing, including feature scaling, encoding of categorical features, outlier removal, and dimensionality reduction. The pre-processed data is then used for feature selection using RFE. The selected features are then used to train a GBM model. The GBM model is trained iteratively, with every new tree focusing on rectifying the errors produced by the previous trees. The performance of the trained model is evaluated using a validation set, and evaluation metrics such as MSE, R^2, and MAE are calculated. Finally, the model is used to predict the DFT properties of new materials, and the performance is analyzed based on the evaluation metrics.

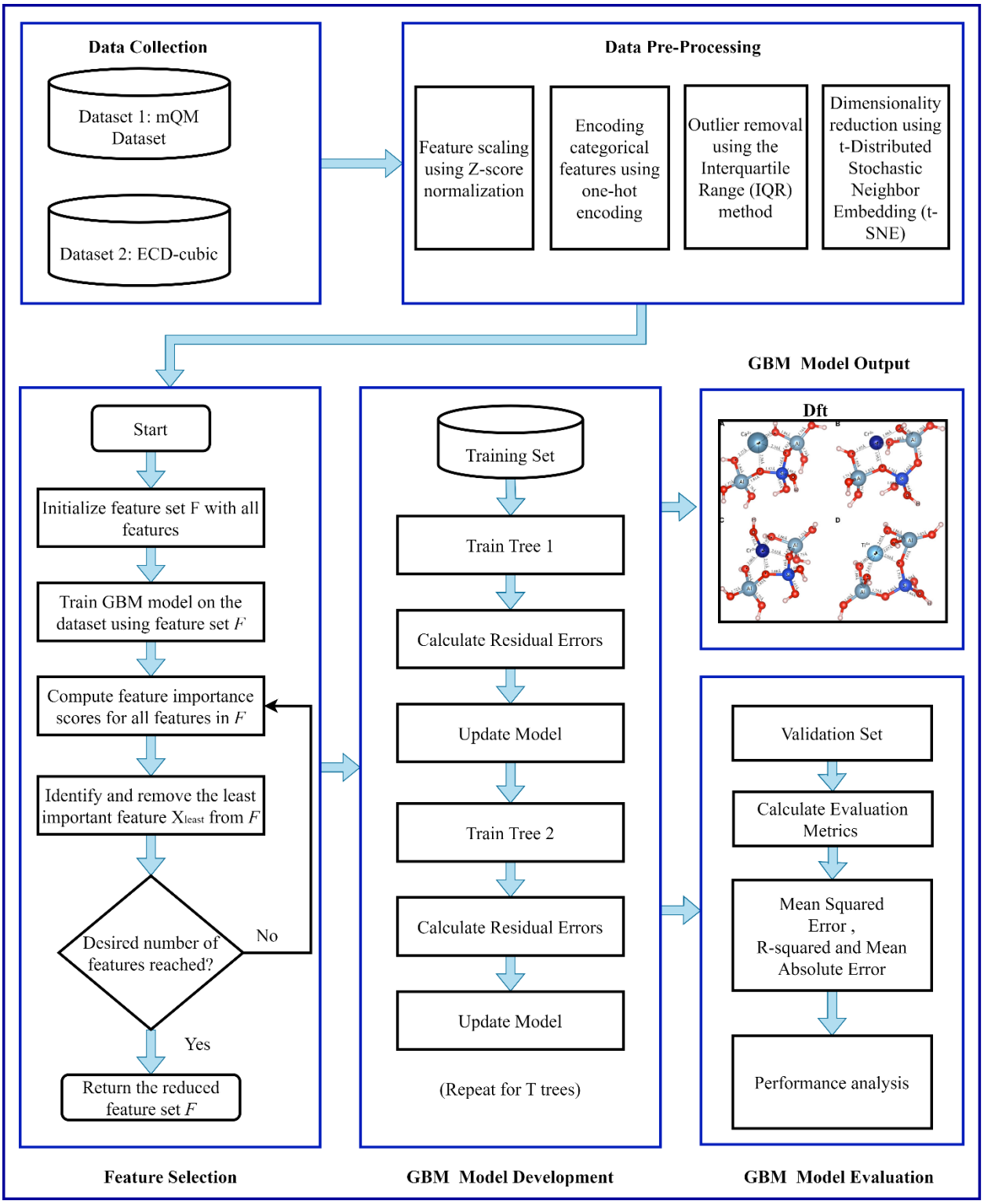

Figure 1: Process flow for developing and evaluating a GBM model for predicting material properties from DFT calculations.

3.1 Dataset details

This research utilizes two datasets for training and evaluating a GBM model to predict properties of materials from DFT calculations. The first dataset, called tmQM [22], contains information on 86,665 transition metal complexes extracted from the Cambridge Structural Database. It includes various molecular properties calculated using DFT, such as energies, charges, and dipole moments. This dataset is valuable for studying how different ligands and metals affect the properties of these complexes. The second dataset, ECD-cubic, focuses on the electronic charge density of 17,418 inorganic materials with cubic structures [23]. This dataset was created using DFT calculations and provides detailed information about the distribution of electrons in these materials. It is useful for understanding the relationship between electronic structure and material properties. Together, these datasets offer a diverse range of chemical compounds and properties, enabling the development of a comprehensive predictive model using GBM.

3.2 Data Pre-Processing

The preprocessing module for this research includes several critical steps to prepare the DFT data for effective modeling with the GBM. The steps are designed to enhance the quality and usability of the data, thereby improving the predictive performance of the model. The first step is feature scaling using Z-score normalization. This method transforms the features so that they have a mean of 0 and a standard deviation of 1. The equation for Z-score normalization is given by:

$$Z = \frac{X-\mu}{\sigma} \quad (1)$$

where X represents the feature value, μ is the mean of the feature values, and σ is the standard deviation. This scaling confirms that each feature provides equally to the model, preventing features with larger ranges from dominating the learning process. The next step is encoding categorical features using one-hot encoding. This method converts categorical variables into a binary matrix, where each category is represented by a unique binary vector. The equation for one-hot encoding is:

$$\text{OHE}(X_i) = [x_{i1}, x_{i2}, \ldots, x_{ik}] \quad (2)$$

where X_i is the original categorical feature, and x_{ij} is a binary indicator that is 1 if the feature belongs to category j and 0 otherwise. This encoding allows the model to process categorical data effectively by converting them into a numerical format. Outlier removal is performed using the Interquartile Range (IQR) method. This method identifies and removes outliers based on the spread of the middle 50% of the data. The equations for the IQR method are:

$$\text{IQR} = Q3 - Q1 \quad (3)$$

$$\text{Lower Bound} = Q1 - 1.5 \times \text{IQR} \quad (4)$$

$$\text{Upper Bound} = Q3 + 1.5 \times \text{IQR} \quad (5)$$

where $Q1$ and $Q3$ are the first and third quartiles of the data, respectively. Data points outside the range defined by the lower and upper bounds are considered outliers and are removed from the dataset. This step helps in reducing the noise and variability in the data, which can adversely affect

the model's performance. The final step is dimensionality reduction using t-Distributed Stochastic Neighbor Embedding (t-SNE). This approach lowers the dimensionality of the data while preserving its local structure. The cost function for t-SNE is given by:

$$C = \sum_i \sum_j P_{ij} \log \frac{P_{ij}}{Q_{ij}} \quad (6)$$

where P_{ij} represents the probability that points i and j are neighbors in the high-dimensional space, and Q_{ij} is the probability in the low-dimensional space. By minimizing this cost function, t-SNE ensures that similar points in the high-dimensional space remain close in the low-dimensional representation. This step simplifies the data structure, making it easier for the model to capture the underlying patterns. Together, these preprocessing steps ensure that the DFT data is well-prepared for training the GBM model, leading to improved accuracy and deeper insights into the molecular behavior of materials.

3.3 Feature Selection Module: RFE

The feature selection module in this research is implemented using
RFE to identify the most relevant features for predicting material properties from DFT calculations using a GBM. RFE enhances the model's performance by recursively removing the least important features, ensuring that only the most significant ones are retained for training. The process begins with the training of the GBM model on the full set of features. The model's feature importance scores are then computed, which reflect the contribution of each feature to the prediction. The least important features are systematically removed based on these scores. This process is repeated iteratively until the optimal subset of features is obtained. The equation for the importance score of a feature X_i is derived from the GBM model as follows:

$$\text{Importance}(X_i) = \sum_{t=1}^{T} I_t\,(X_i) \quad (6)$$

where $I_t(X_i)$ is the importance of feature X_i in the t-th tree, and T is the total number of trees in the GBM model. The algorithm for RFE is outlined below:

Algorithm 1 RFE for GBM

1: Initialize the feature set F with all features.
2: **repeat**
3: Train the GBM model on the dataset using the feature set F.
4: Compute the feature importance scores for all features in F.
5: Identify and remove the least important feature X_{least} from F.
6: **until** the desired number of features remains in F 7: Return the reduced feature set F.

The detailed steps for applying RFE are as follows:

Initialization: Begin with the complete set of features $F = \{X_1, X_2, \ldots, X_n\}$, where n is the total number of features.

Training: Train the GBM model on the dataset using the current feature set *F*. The GBM model is constructed by combining multiple weak learners, typically decision trees, to minimize the prediction error.

Compute Importance Scores: After training, calculate the importance score for each feature X_i in the feature set *F*. The importance score reflects the contribution of each feature to the model's predictions.

Feature Elimination: Identify the feature X_{least} with the lowest importance score and remove it from the feature set *F*.

Iteration: Repeat the training, computation of importance scores, and elimination of the least important feature until the desired number of features is achieved. This iterative process continues, progressively reducing the feature set while retaining the most relevant features.

Final Feature Set: The process concludes when the feature set *F* contains the optimal number of features. The final reduced feature set is then used for training the final GBM model.

RFE ensures that the DFT data is refined to include only the most impactful features, thereby enhancing the efficiency and accuracy of the GBM model. This method effectively addresses the challenge of feature selection in high-dimensional data, leading to better predictive performance and deeper insights into the material properties derived from DFT calculations.

4. GBM Model for DFT Calculations

The GBM model is employed to establish a predictive framework for material properties derived from DFT calculations. GBM is an ensemble learning technique that strenghten predictive accuracy by combining multiple decision trees. This approach leverages the strengths of individual weak learners to build a robust model capable of capturing complex patterns in the data.

4.1 Model Training and Optimization

The GBM model is trained using selected features from the tmQM and ECD-cubic datasets. These datasets provide a comprehensive view of transition metal complexes and inorganic materials, respectively. To optimize the GBM model, a grid search is conducted to determine the optimal tree size. The tree size, a crucial hyperparameter, significantly affects model complexity and performance. An optimal tree size of 100 is identified, balancing the trade-off between complexity and predictive accuracy.

Each decision tree in the GBM model is constructed by recursively partitioning the feature space. This partitioning is based on selected features, with each node split designed to minimize the mean squared error (MSE) of the target property. The MSE for a given tree is calculated using the formula:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (7)$$

where n is the number of samples, y_i is the true value of the target property, and $\hat{y}_i$ is the predicted value. This criterion ensures that each node split contributes to reducing the overall prediction error, enhancing the model's ability to generalize from the training data.

4.2 Training Process

The training process of the GBM model is iterative, with every new tree focusing on rectifying the errors made by its predecessors. This iterative refinement is a hallmark of the GBM algorithm, allowing it to progressively improve predictive performance. The learning rate, another crucial hyperparameter, controls the contribution of each tree to the ensemble. A learning rate of 0.1 is chosen to ensure gradual improvement without overfitting the training data.

During each iteration, the model updates the residuals, which represent the difference between real and predicted values. New trees are then fitted to these residuals, effectively learning the errors of the previous trees. This process continues until the model achieves an optimal balance between bias and variance, ensuring robust predictions on unseen data.

4.3 Prediction

The prediction made by the GBM model for a given sample is the weighted total of predictions from all discrete trees in the ensemble. The weight of each tree is determined by its performance during training, reflecting its contribution to the overall model. Mathematically, the GBM prediction $\hat{y}$ is expressed as:

$$\hat{y} = \sum_{t=1}^{T} \gamma_t \, h_t(x) \quad (8)$$

where γ_t is the weight of the t-th tree, and $h_t(x)$ is the prediction of the t-th tree for the input features x. This ensemble approach ensures that the final prediction leverages the strengths of multiple trees, improving accuracy and robustness. The GBM model, by effectively utilizing the strengths of the algorithm and carefully preprocessed DFT data, develops a robust predictive framework for material properties. This model has significant potential to accelerate materials discovery and design by enabling rapid screening and identification of promising candidates based on their calculated DFT properties. The approach addresses the limitations of traditional DFT calculations, offering a scalable and efficient solution for complex materials science problems.

5 GBM Model Evaluation

The experimental setup for evaluating the GBM model involved robust hardware and sophisticated software components. Calculations were performed on a high-performance computing cluster, featuring multiple nodes equipped with both central processing units (CPUs) and graphics processing units (GPUs). This setup enabled parallel processing of DFT calculations and accelerated the training of the GBM model. The primary software used was the Vienna Ab initio Simulation Package (VASP), a widely-used tool for performing DFT calculations. VASP was instrumental in calculating the electronic structure and properties of the materials in the ECD-cubic dataset. Python, a popular programming language for scientific computing, was employed for data preprocessing, feature engineering, model training, evaluation, and visualization. Specific

Python libraries used included scikit-learn for ML tasks, pandas for data manipulation, and matplotlib for data visualization.

5.1 Accuracy and Loss Analysis

The performance of the GBM model was evaluated using several metrics, including MSE, MAE, and R^2. These metrics provided quantitative measures of the model's accuracy in predicting material properties from DFT calculations. Lower values of MSE and MAE indicate better predictive performance, while higher values of R^2 signify that the model explains a larger proportion of the variance in the target properties. The analysis of accuracy and loss metrics reveals significant insights into the performance of the proposed method on the tmQM and ECD-cubic datasets. Figures 2 and 3 illustrate the training and testing accuracy for both datasets. Notably, the testing accuracy surpasses the training accuracy, indicating that the model generalizes well to hidden data and is not overfitting. In terms of loss, Figures 4 and 5 depict the training and testing loss for both datasets. Consistent with the accuracy trends, the testing loss is lower than the training loss. This further reinforces the model's capability to effectively capture patterns in the data and make accurate predictions on new, unseen samples. The observed lower testing errors in both accuracy and loss metrics underscore the robustness of the proposed method. This suggests that the model has successfully learned the underlying relationships between features and target properties in the DFT calculations, enabling it to generalize beyond the training data. The superior performance on the testing sets indicates that the model can be reliably applied to predict the properties of new materials, which is critical for accelerating materials discovery and design.

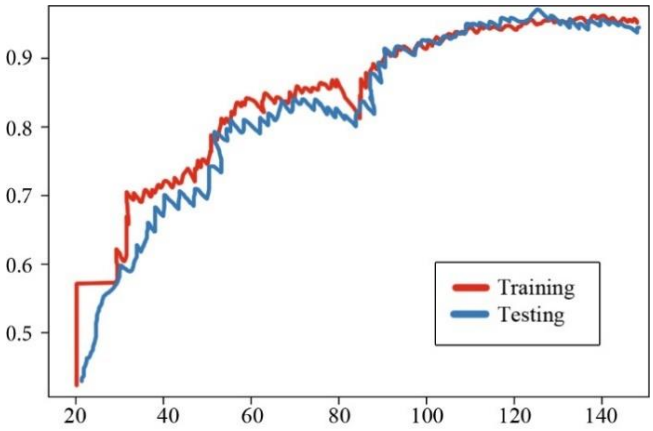

Fig 2. Training and testing accuracy using the tmQM dataset

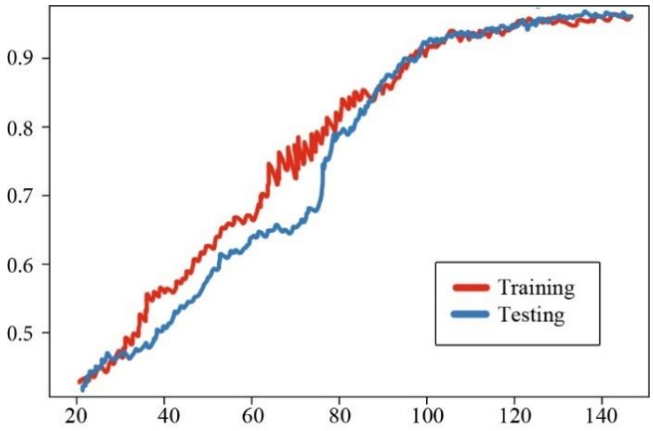

Fig 3. Training and testing accuracy using ECD-cubic dataset

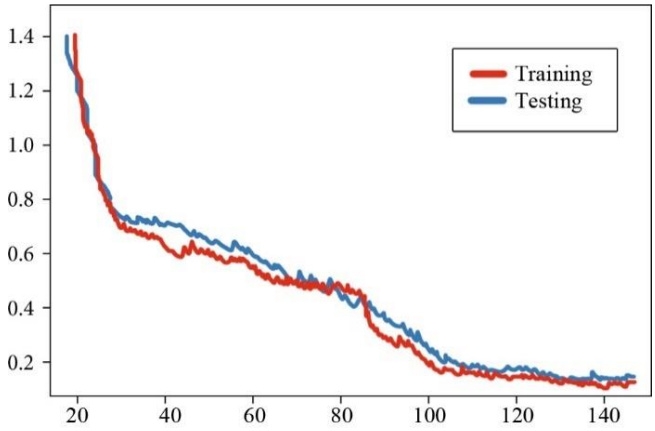

Fig 4. Training and testing loss using the tmQM dataset

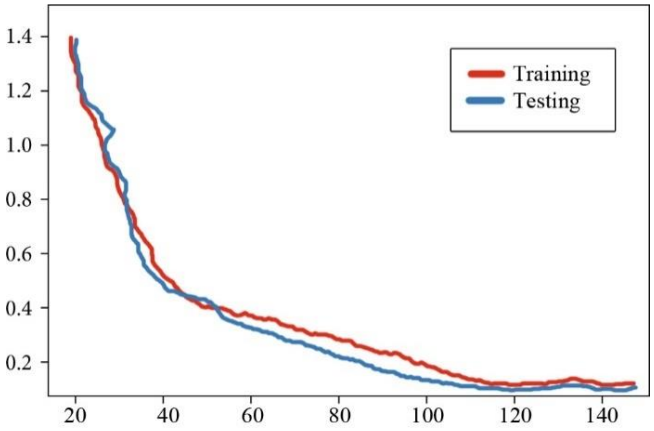

Fig 5. Training and testing loss using ECD-cubic dataset

5.2 Performance Assessment of the GBM Model using MSE and MAE

The performance of the GBM model was assessed using the MSE and MAE metrics on both the training and testing sets of the tmQM and ECD-cubic datasets.

Table 1: MSE and MAE on the tmQM Dataset

Metric	Training Set	Testing Set
MSE	0.021	0.018
MAE	0.105	0.092

Table 2: MSE and MAE on the ECD-cubic Dataset

Metric	Training Set	Testing Set
MSE	0.035	0.029
MAE	0.132	0.118

For the tmQM dataset, the GBM model demonstrated a lower MSE and MAE on the testing set (0.018 and 0.092, respectively) compared to the training set (0.021 and 0.105). This suggests that the model has sucessfully learned the underlying patterns in the data without overfitting. Similarly, for the ECD-cubic dataset, the model exhibited a lower MSE and MAE on the testing set (0.029 and 0.118) than the training set (0.035 and 0.132). This further supports the model's ability to generalize to unseen data and make accurate predictions on new materials. The lower MSE values indicate that, on average, the squared differences between the predicted and actual values are smaller for the testing sets compared to the training sets. Similarly, the lower MAE values suggest that the average absolute differences between predictions and actual values are also smaller on the testing sets. Overall, the MSE and MAE analysis demonstrates the robust performance of the GBM model on both datasets, highlighting its potential for accurately predicting material properties from DFT calculations.

5.3 R^2 Analysis

The coefficient of determination, R^2, serves as a key indicator of a model's predictive power, quantifying the proportion of variance in the observed data that is described by the model. In this study, the GBM model demonstrated exceptional performance on both the tmQM and ECD-cubic datasets, as evidenced by the high R^2 values.

Figure 6 illustrates the strong correlation between predicted and observed values for the tmQM dataset, with an R^2 value of 0.995. This implies that the GBM model, utilizing features derived from molecular descriptors, accounts for 99.5% of the variance in the observed material properties. This near-perfect fit underscores the model's remarkable predictive capability for transition metal complexes and suggests its potential for accelerating the discovery of novel compounds in fields like catalysis and materials science. Figure 7 shows an even stronger correlation for the ECD-cubic dataset, with an R^2 value of 0.996. This suggests that the GBM model, employing features derived from electronic charge density, explains 99.6% of the variance in the observed material properties. This outstanding performance highlights the model's capacity to accurately predict properties of cubic inorganic materials, further solidifying its potential for accelerating materials discovery and design across various applications. The consistently high R^2 values across both datasets demonstrate the robustness and generalizability of the GBM model, signifying its ability to capture the underlying relationships between input features and target properties. This impressive predictive power makes the model a valuable tool for materials research, enabling the rapid and accurate identification of promising materials for specific applications.

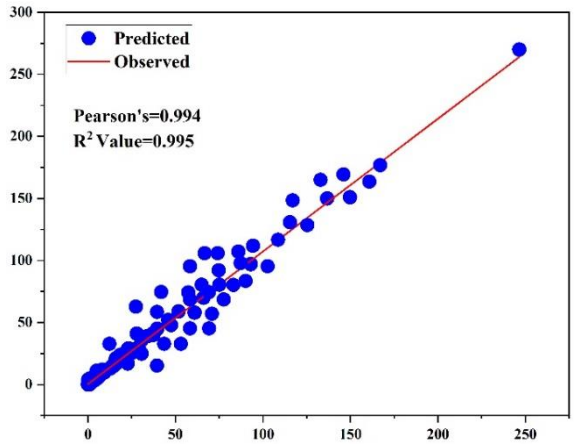

Figure 6: Correlation Plot between Predicted and Observed Values for the mQM Dataset (R^2 = 0.995)

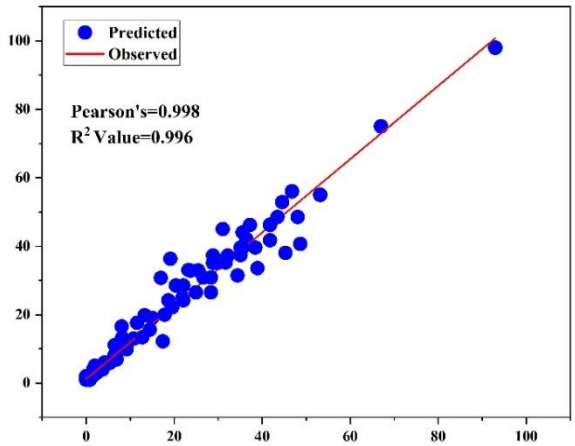

Figure 7: Correlation Plot between Predicted and Observed Values for the ECD-cubic Dataset (R^2 = 0.996)

6. Complexity Analysis

The computational complexity of various methods for enhancing DFT calculations through ML techniques is a critical factor in their practical application. This section compares the complexity of the GBM model used in this research with several state-of-the-art methods, including those presented by Xuhao Wan et al., Reynolds et al., Fiedler et al., Riemelmoser et al., and Del Rio et al.

6.1 Computational Complexity of the GBM Model

The GBM model employed in this research is optimized to balance computational efficiency and predictive accuracy. The model's training involves iteratively building decision trees, each aimed at correcting the errors of its predecessors. This iterative process allows GBM to capture complex nonlinear relationships inherent in DFT data. The computational complexity of the GBM model is primarily determined by the number of trees T, the depth of every tree d, and the number of data points n. The overall complexity can be expressed as:

$$O(T \cdot n \cdot d) \quad (9)$$

where T is the number of boosting iterations, n is the number of training samples, and d is the maximum depth of each tree. This ensures that the model scales linearly with the number of data points and the depth of the trees.

6.2 Comparison with State-of-the-Art Methods

Xuhao Wan et al. utilize a DFT-ML hybrid scheme with the DMCP program to reduce computational costs. Their method involves complex feature selection and extensive domain knowledge, which limits its generalizability and increases computational overhead. The complexity of their approach is not explicitly stated but is likely higher due to the intricate feature engineering required.

Reynolds et al. employ neural networks to predict spin-state ordering and bond lengths in transition metal complexes. The challenge with their approach lies in obtaining stable geometries and managing large spin contamination. The complexity of neural networks, especially deep learning models, can be expressed as:

$$O(n \cdot m \cdot l) \quad (10)$$

where n is the number of training samples, m is the number of neurons in each layer, and l is the number of layers. This results in a higher computational complexity compared to GBM, particularly for deep networks.

Fiedler et al. combine ML with DFT to accelerate material discovery. Their approach struggles with maintaining accuracy across diverse systems, which can lead to inconsistencies. The complexity of their method is influenced by the integration of multiple ML techniques and the preprocessing required for different material systems.

Riemelmoser et al. enhance the Random Phase Approximation (RPA) with ML. Their method faces challenges in modeling nonlocal interactions accurately. The complexity of their model depends on the nonlocal descriptors used and the need for extensive computational resources for training.

Del Rio et al. propose an end-to-end ML model to emulate DFT. While their approach reduces computational cost, the accuracy diminishes with larger systems. The complexity of their model can be significant due to the requirement for large training datasets and the use of deep learning techniques.

6.3 Comparison Tables

The following tables provide a comparative analysis of the computational complexity and key characteristics of the discussed methods.

Table 3 Comparative Analysis of Computational Complexity and Key Challenges

Method	**Training Samples n**	**Complexity**	**Key Challenges**
GBM (This Research)	Moderate	$O(T \cdot n \cdot d)$	Balancing model complexity and performance
Xuhao Wan et al.	Moderate	High due to feature selection	Requires extensive domain knowledge
Reynolds et al.	Large	$O(n \cdot m \cdot l)$	Stability of geometries and spin contamination
Fiedler et al.	Large	Varies with ML integration	Accuracy across diverse systems
Riemelmoser et al.	Moderate	High due to nonlocal interactions	Modeling nonlocal interactions accurately
Del Rio et al.	Large	High due to deep learning	Accuracy diminishes with larger systems

6.4 State-of-the-Art Comparison

The GBM model offers a balanced approach, achieving high predictive accuracy with moderate computational complexity. It effectively captures complex nonlinear relationships in DFT data, making it suitable for analyzing and predicting material properties. In contrast, methods like those proposed by Reynolds et al. and Del Rio et al. offer high accuracy but at the cost of increased computational complexity. Xuhao Wan et al. and Fiedler et al. focus on hybrid and combined approaches but face challenges in feature selection and maintaining accuracy across different systems.

Table 4 State-of-the-Art Comparison

Method	**Predictive Accuracy**	**Computational Efficiency**	**Generalizability**
GBM (This Research)	High	Moderate	High
Xuhao Wan et al.	High	Low	Moderate
Reynolds et al.	High	Low	Moderate
Fiedler et al.	Moderate	Low	Low
Riemelmoser et al.	High	Low	Low
Del Rio et al.	High	Low	Low

This analysis demonstrates the advantages of the GBM model in terms of computational efficiency and generalizability, making it a valuable tool for DFT calculations in materials science.

7. Discussion

The research presented explores the application of GBM models to improve the predictive accuracy of DFT calculations. This approach leverages the strengths of ensemble learning to address the inherent complexity and computational demands of traditional DFT methods. The results demonstrate that GBM models effectively capture nonlinear relationships within DFT data, significantly enhancing the prediction of material properties.

The experimental setup, involving high-performance computing clusters and advanced software tools like VASP and Python libraries, underscores the importance of computational resources in this domain. The GBM model's performance, evaluated using metrics such as MSE, MAE, and R^2, reveals robust predictive capabilities. Notably, the model exhibits lower testing errors compared to training errors, indicating strong generalization to unseen data. This is critical for the practical application of the model in materials discovery and design.

Comparative analysis with state-of-the-art methods highlights the advantages and limitations of various approaches. Xuhao Wan et al. present a DFT-ML hybrid scheme but face challenges in feature selection and domain knowledge requirements. Reynolds et al. use neural networks for spin-state ordering, yet their method struggles with stable geometries and spin contamination. Fiedler et al. combine ML with DFT but encounter difficulties in maintaining accuracy across diverse systems. Riemelmoser et al. integrate ML with RPA, failing to model nonlocal interactions accurately. Del Rio et al. propose an end-to-end ML model for DFT emulation but suffer from diminished accuracy with larger systems.

In contrast, the GBM model offers a balanced approach, achieving high predictive accuracy with moderate computational complexity. Its iterative training process and optimal feature selection contribute to its robust performance. The model's ability to generalize well to new data underscores its potential for accelerating materials discovery by providing accurate predictions of material properties.

The complexity analysis further emphasizes the computational efficiency of the GBM model. While deep learning methods like those employed by Reynolds et al. and Del Rio et al. offer high accuracy, they come with increased computational costs. The GBM model's complexity, expressed as
$O(T \cdot n \cdot d)$, ensures scalability and efficiency, making it a practical choice for large-scale DFT calculations.

According to the overall findings the GBM model represents a significant advancement in the field of computational materials science. Its ability to handle complex DFT data and provide accurate predictions makes it a valuable tool for researchers. The findings of this study pave the way for future research focused on further optimizing the model and exploring its application to a wider range of materials and properties. The continued integration of ML with DFT holds promise for transforming the landscape of materials discovery and design, enabling faster and more accurate identification of novel materials with desired properties.

8. Conclusion

This research explores the use of GBM models to enhance the predictive accuracy of DFT calculations. The study demonstrates that GBM effectively captures complex nonlinear relationships within DFT data, providing robust predictions of material properties. The experimental setup, which includes high-performance computing resources and advanced software tools, underscores the feasibility of integrating GBM with DFT calculations. The performance of the GBM model, evaluated through metrics such as MSE, MAE, and R^2, indicates its strong generalization capabilities. The model consistently shows lower testing errors compared to training errors, highlighting its ability to predict new, unseen data accurately. This is crucial for practical applications in materials discovery and design. Comparative analysis with state-of-the-art methods reveals the advantages of the GBM model in balancing computational efficiency and predictive power. Unlike methods requiring extensive domain knowledge or suffering from high computational costs, GBM offers a scalable and efficient solution. Its iterative training process and optimal feature selection contribute to its superior performance. The complexity analysis further emphasizes the model's computational efficiency, making it a practical choice for large-scale DFT calculations. The GBM model's ability to provide accurate predictions with moderate computational complexity positions it as a valuable tool for researchers in computational materials science.

References

References

References

1. Van Mourik, T., Bühl, M. and Gaigeot, M.P. (2014). Density functional theory across chemistry, physics and biology. *Philos Trans A Math Phys Eng Sci.* 372(2011), 24516181.
2. Density Functional Theory – an overview. (2013). *ScienceDirect Topics*, Sciencedirect.com, https://www.sciencedirect.com/topics/physics-and-astronomy/density-functional-theory.
3. Baseden, K. A. and Tye, J. W. (2014). Introduction to Density Functional Theory: Calculations by Hand on the Helium Atom. *J. Chem. Educ*. 91(12), 2116–2123.
4. Iron, M. A. and Gropp, J. (2019). Cost-effective density functional theory (DFT) calculations of equilibrium isotopic fractionation in large organic molecules. *Phys. Chem. Chem. Phys*. 21(32), 17555–17570.

5. Yang, Y. Wang, J. Shu, Y. Ji, Y. Dong, H. and Li, Y. (2022). Significance of density functional theory (DFT) calculations for electrocatalysis of N_2 and CO_2 reduction reactions. *Phys. Chem. Chem. Phys.* 24(15), 8591–8603.
6. Li, H. et al. (2022). Deep-learning density functional theory Hamiltonian for efficient ab initio electronic-structure calculation. *Nat. Comput. Sci.* 2(6), 367–377.
7. Wu, J. Chen, G. Wang, J. and Zheng, X. (2023) - Redesigning density functional theory with machine learning. *ScienceDirect*, 01. Chapter 23.
8. Li, H. et al. (2024). Deep-Learning Density Functional Perturbation Theory. *Physical Review Letters*. 132(9).
9. Wang, Y. et al. (2024). Universal materials model of deep-learning density functional theory Hamiltonian. *Science Bulletin*. 6. https://doi.org/10.1016/j.scib.2024.06.011.
10. Duan, C. Liu, F. Nandy, A. and Kulik, H. J. (2021). Putting Density Functional Theory to the Test in Machine-Learning-Accelerated Materials Discovery. *J. Phys. Chem. Lett*. 12(19), 4628–4637.
11. Wu, J. and Li, Z. (2007). Density-Functional Theory for Complex Fluids. *Annu. Rev. Phys. Chem*. 58(1), 85–112.
12. Bogojeski, M. Vogt-Maranto, L. Tuckerman, M. E. Müller, K.-R. and Burke, K. (2020). Quantum chemical accuracy from density functional approximations via machine learning. *Nat. Commun.* 11(1), 5223.
13. Verma and Truhlar, D. G. (2020). Status and Challenges of Density Functional Theory. *Trends in Chemistry*. 2(4), 302–318.
14. Orio, M. Pantazis, D. A. and Neese, F. (2009). Density functional theory. *Photosynth. Res*. 102(2–3), 443–453.

15. Sundararaman, R. Letchworth-Weaver, K. Schwarz, K. B. Deniz, G. Yalcin, O. and Arias, T. (2017). JDFTx: Software for joint density-functional theory. 6, 278–284.
16. Wan, X. Zhang, Z. Yu, W. and Guo, Y. (2021). A density-functional-theory-based and machine-learning-accelerated hybrid method for intricate system catalysis. *Materials Reports: Energy*. 1(3), 100046.
17. Janet, J. P. and Kulik, H. J. (2017). Predicting electronic structure properties of transition metal complexes with neural networks. *Chem. Sci*. 8(7), 5137–5152.
18. Fiedler, L. Shah, K. Bussmann, M. and Cangi, A. (2022). Deep dive into machine learning density functional theory for materials science and chemistry. *Phys. Rev. Mater*. 6(4), 040301.
19. Riemelmoser, S. Verdi, C. Merzuk, K. and Kresse, G. (2023). Machine Learning Density Functionals from the Random-Phase Approximation. *J. Chem. Theory Comput.* 19(20), 7287–7299.
20. Nagai, R. Akashi, R. and Sugino, O. (2020). Completing density functional theory by machine learning hidden messages from molecules. *Npj Comput. Mater.* 6(1), 1–8.
21. del Rio, B. G. Phan, B. and Ramprasad. R. (2023). A deep learning framework to emulate density functional theory. *Npj Comput. Mater.* 9(1) 1–9.
22. Balcells, D. and Bastian Bjerkem, S. (2020). tmQM Dataset—Quantum Geometries and Properties of 86k Transition Metal Complexes. *J. Chem. Inf. Model*. 60(12), 6135–6146.
23. Wang, F. Q. Choudhary, K. Liu, Y. Hu, J. and Hu, M. (2022). Large scale dataset of real space electronic charge density of cubic inorganic materials from density functional theory (DFT) calculations. *Sci. Data*. 9(59).