الجمـــهوريـــة الجـــزائريـة الديمقراطيـــة الشـــعبيـــة

PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA

وزارة التعـليـــم العـــالـــي والبـحث الـعلـــمي

MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

جـامعـــة عـمـــار ثـلـيـجـي بالأغواط

UNIVERSITY OF AMAR TELIDJI LAGHOUAT

كلـيـــــة العـلـــــوم

FACULTY OF SCIENCES

قسم علوم المادة

DEPARTEMENT OF MATERIAL SCIENCES

Master's dissertation

Sector: Material Sciences

Option: Materials Physics

Presented by: Ms. ROGTI Maroua

THEME

Cross-Domain Assessment of Feature Effectiveness in Functional Materials Prediction via Multiple Machine Learning Models

Board of jury:

DAHAM Taher	University of Laghouat-ALGERIA	MCA	President
HEBBOUL Zoulikha	University of Laghouat-ALGERIA	Professor	Examiner
BENGHIA Ali	University of Laghouat-ALGERIA	MCA	Supervisor
MECHRAOUI Kaima Bouchra	University of Laghouat-ALGERIA	PhD.Student	Co-supervisor

2024/2025

acknowledgment

I thank God Almighty for granting me the courage, will, and patience to complete this work.

This work was conducted in the Materials Physics Laboratory at the University of Laghouat. I would also like to express my sincere gratitude and appreciation to my supervisor, Mr. MCA Ben Ghia Ali, and my co-supervisor, Ms. Mechraoui Bouchra PHD student. I am grateful for their patience, encouragement, and assistance, as well as for their dedication and presence during the preparation of this thesis. I would also like to express my sincere thanks to Mr. MCA Daham Taher for agreeing to chair the thesis evaluation committee. I would also like to thank Professor Haboul Zlikha for agreeing to review this work and enrich it with their suggestions.

To all the professors in the Department of Materials Science, especially the physics professors. Finally, I thank everyone who helped and supported me in my work.

I dedicate this work

To my dear mother and beloved father, in gratitude for their unwavering support and sacrifices throughout my academic journey and my entire life.

To all my sisters and my sister's children—may God protect them.

to my husband, "kaouka Mustapha"

To my lifelong friends, Fatima and Dawia

List of figures

List of tables

Abbreviations list

MAE : Mean Absolute Error

MSE : Mean Squared Error

RMSE : Root Mean Squared Error

RFR : Random Forest Regression

SVR : Support Vector Regression

XGBoost : Extreme Gradient Boosting

KNN : K-Nearest Neighbors

Index

Introduction

In recent years, the application of machine learning (ML) in materials science has grown rapidly, transforming how researchers approach the design and discovery of new materials[56–65]. With the ever-expanding availability of materials data and advancements in computational power, ML has become a powerful tool to predict key properties, screen large material spaces, and uncover hidden patterns that would be difficult to detect through traditional approaches alone[66–69]. Unlike conventional methods such as density functional theory (DFT) simulations or experimental synthesis and characterization which are often time-consuming, resource-intensive, and limited in scale machine learning models offer a faster, cost-effective alternative that can significantly accelerate materials discovery. By learning from existing data, ML algorithms can quickly evaluate thousands of candidate materials, guiding researchers toward the most promising compounds for targeted applications [64, 70,71]. This shift not only reduces the time and cost associated with experimental and computational workflows but also opens new pathways for identifying functional materials in areas such as energy storage, photovoltaics, catalysis, and electronics [72-74].

Since the discovery of perovskite materials and the recognition of their exceptional physical and chemical properties, they have become central to numerous technological applications, including photovoltaics, catalysis, and thermoelectrics. Driven by the need to enhance the performance of energy-related devices, research efforts have increasingly focused on identifying and developing new perovskite compounds. In the past two decades, the design of novel functional materials with perovskite structures has reached an unprecedented level, largely due to the integration of machine learning (ML) techniques. These methods have accelerated the discovery process by enabling rapid screening and prediction of material properties. Among the most notable contributions is the work of Li et al. [75], who predicted the thermodynamic stability of over 1900 oxide perovskites using first-principles calculations. Similarly, Touati *et al.*[76] applied ML models to estimate the band gap of approximately 13,947 ABX_3 perovskite compounds, highlighting the power of data-driven approaches in materials prediction. Numerous other studies have also explored various physical properties within the perovskite family through ML-based methods.

Beyond perovskites, nonlinear optical (NLO) materials have also emerged as a class of functional materials that have transformed conventional understanding in optical physics. Although discovered relatively recently in 1961 [77]., NLO materials have demonstrated vast potential across multiple technological and biomedical domains, including light-emitting

diodes, optoelectronic devices, and biomedical imaging. Initially, the characterization of NLO properties was limited to experimental methods, which are both time-consuming and costly [78-80]. To overcome these limitations, researchers particularly in theoretical condensed matter physics began calculating NLO behavior using first-principles methods such as density functional theory (DFT) [81]. This theoretical approach laid the foundation for a faster and more systematic exploration of NLO compounds. The integration of machine learning has further revolutionized this process by dramatically reducing the computational time required to evaluate key optical properties. ML provides a powerful and efficient predictive framework that enables researchers to bypass laborious simulations and proceed directly to experimental validation, thereby optimizing resources.

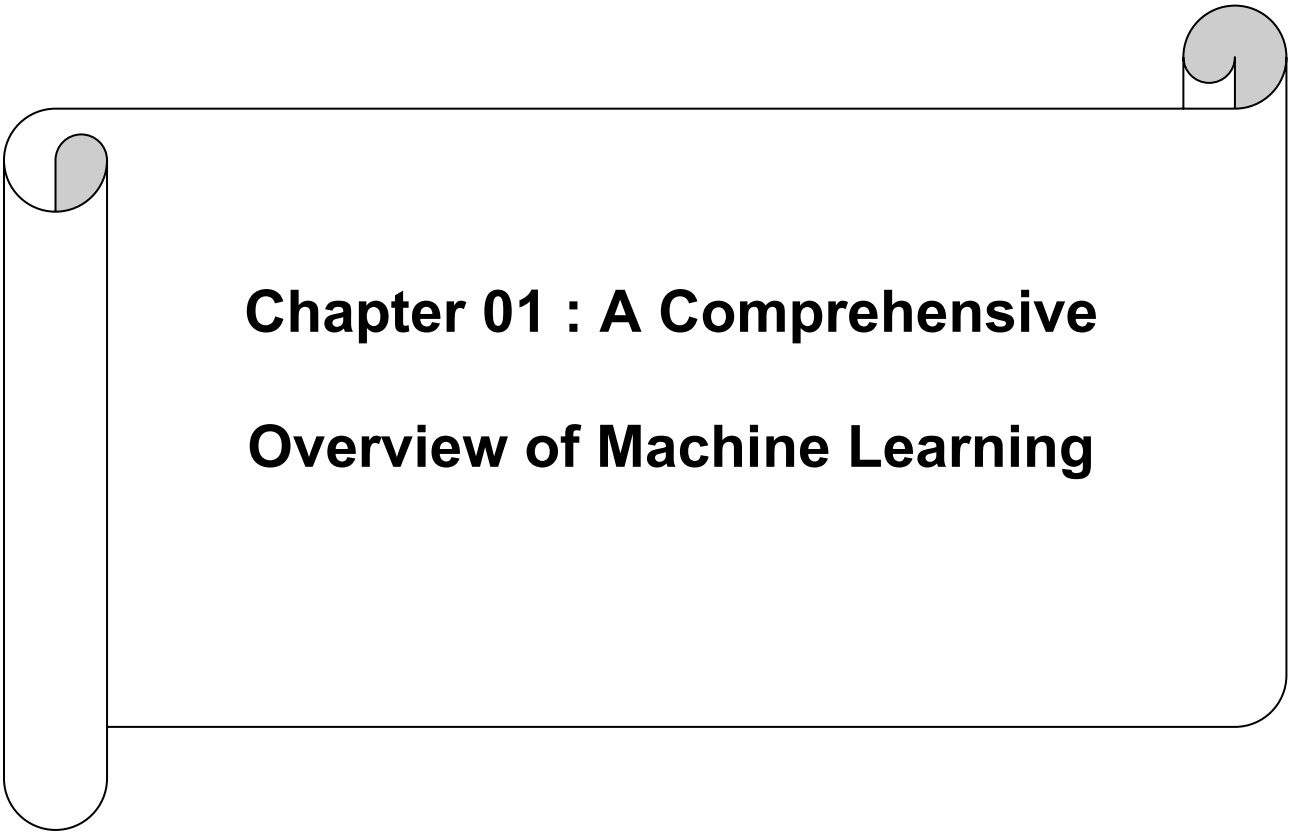
Several studies have demonstrated the efficacy of ML in predicting the behavior of NLO materials [82-85]. For instance, Zhang *et al*.[85] employed diverse ML models to predict critical properties of 411 NLO compounds. Raju *et al.* [86] introduced a novel image-processing-based ML approach to characterize NLO behavior, presenting a unique methodology in the field of materials physics. More recently, Benghia *et al*.[65] applied multiple machine learning models to predict the band gaps of various NLO materials using simplified feature sets, further illustrating the potential of data-driven discovery.

Today, machine learning stands as an indispensable tool in the discovery and design of advanced materials. While it significantly reduces the time and cost associated with traditional experimental and computational techniques, ML is not without its challenges [56, 57]. One of the most critical factors affecting model performance is feature selection. The quality and relevance of input features can substantially influence predictive accuracy [87, 88]. In this context, our study focuses on evaluating the impact of features extracted using the Pymatgen library implemented in Python on the prediction of material properties. We assess their effectiveness across multiple ML models and validate our findings using two main datasets: one comprising perovskite materials and the other containing NLO compounds.

Following this introductory section, the manuscript is organized as follows:

- **Chapter I** provides an overview of machine learning, introducing various commonly used algorithms in materials science.
- **Chapter II** offers a general overview of functional materials, focusing particularly on perovskites and NLO materials.

- **Chapter III** presents the core analysis, including the development and application of ML models to predict material properties.
- Finally, the **Conclusion** summarizes the key findings and outlines potential directions for future research in data-driven materials discovery.



Chapter 01 : A Comprehensive Overview of Machine Learning

Machine Learning (ML) represents a revolutionary area of artificial intelligence (AI) enabling systems to gain knowledge using data then refine performance absent explicit programming. It has transformed healthcare, finance, and autonomous driving now. Industries have come to see data-driven perceptions in addition to automation as a result. Authoritative references do support this paper's exploration of machine learning's fundamentals, types, applications, as well as challenges. [1]

1.1 What is Machine Learning?

Machine Learning is defined as studying algorithms allowing computers to learn with a goal to predict or decide based on data (Mitchell, 1997). ML systems can automatically recognize the trends, and they adapt along with the time, unlike the customary programming in which rules are manually coded. [1]

1.2 Types of Machine Learning

1.2.1 Supervised learning

Supervised Learning as the name indicates the presence of a supervisor or a teacher. Basically, supervised learning is a learning in which we teach or train the machine using data which is well labeled that means some data is already tagged with the correct answers. After that, the machine is provided with a new set of examples(data) so that the supervised learning algorithm analyses the training data (set of training examples) and produces a correct outcome from labeled data. The figure 1.1 below illustrates how this method works. [2]

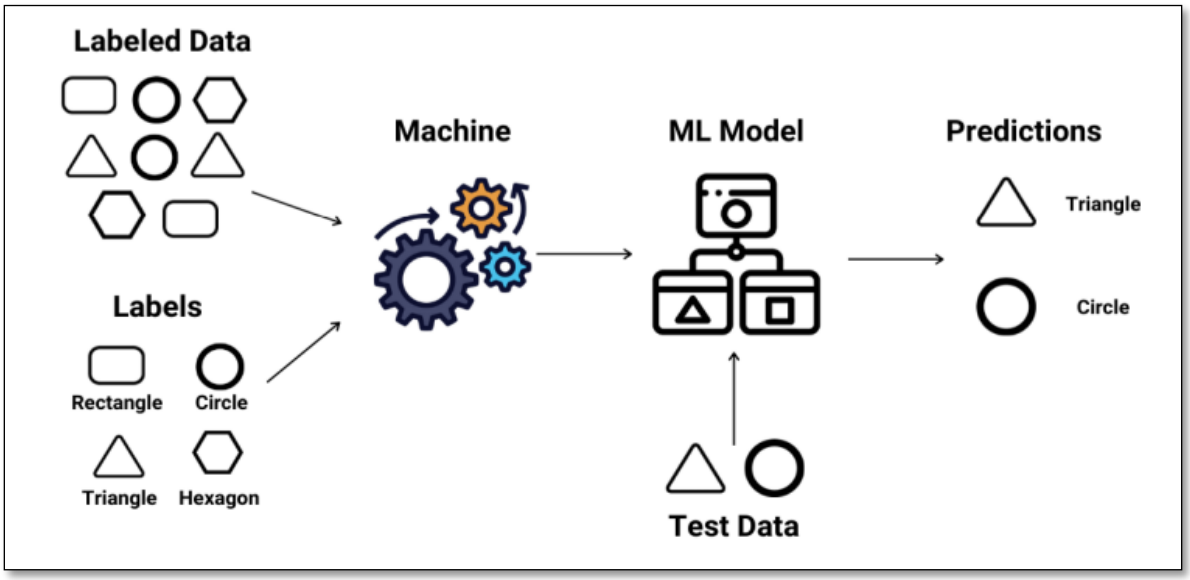


Fig1.1 An example of how the Supervised method works [2]

→ **Classification:** A classification problem is when the output variable is a category, such as “Red” or “Blue” or “Disease” or “No disease”.

→ **Regression**: A regression problem is when the output variable is a real value, such as “Dollars” or “Weight”. Fig 1.2

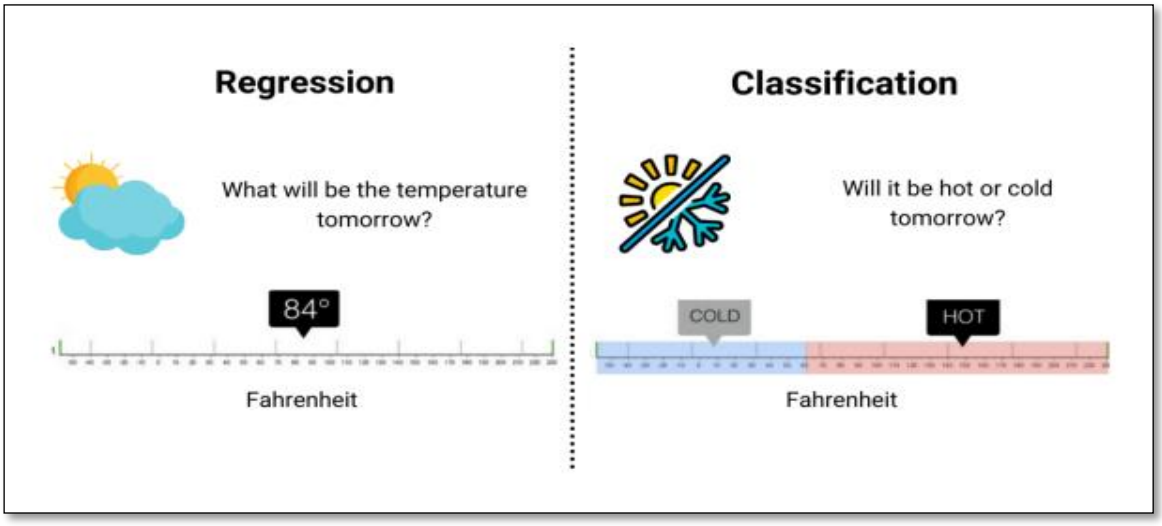


Fig 1.2The difference between Regression and Classification [2]

1.2.2 Unsupervised Learning

Unsupervised learning is a machine learning technique, where one do not need to supervise the model. Instead, you need to allow the model to work on its own to discover information. It mainly deals with the unlabeled data. Unsupervised learning algorithms allows one to perform more complex processing tasks compared to supervised learning. Although, unsupervised learning can be more unpredictable compared with other natural learning methods. [2]

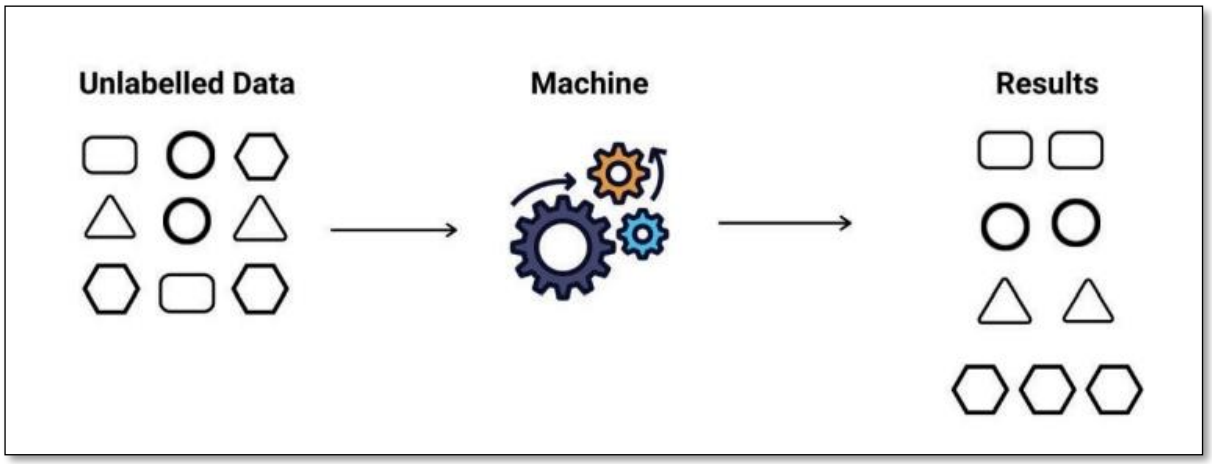


Fig1.3 An example of how the Unsupervised method works [2]

Unsupervised learning classified into two categories of algorithms:

→ **Clustering:**

Clustering is an unsupervised learning technique where a computer groups similar data points together—like organizing a messy drawer without labels. No prior training is needed, the algorithm finds patterns on its own. This technique is known for two popular algorithms: **K-Means Clustering** and **Hierarchical Clustering.** [3][4]

→ **Dimensionality Reduction:**

Dimensionality reduction simplifies complex datasets by reducing the number of variables while keeping the most important information. Think of it like summarizing a long book into key points without losing the main story. we need it to: makes data easier to visualize, speeds up machine learning algorithms, removes unnecessary noise, helps avoid overfitting. [5]

PCA (Principal Component Analysis) and t-SNE are the main techniques for dimensionality reduction. [6]

To choose the correct method, we summarize the difference between the two methods in the table below: [7]

Table 1.1: the different between PCA and t-SNE

Method	**Best For**	**Not Good For**	**Speed**
PCA	Linear patterns, speed	Complex curves	Fast
t-SNE	Seeing clusters	Large datasets (>10,000 points)	slow

1.3 Machine learning models

1.3.1 Random Forest

Random Forest is an ensemble learning method that builds multiple decision trees during training and outputs either the mode (for classification) or mean prediction (for regression) of the individual trees. It introduces randomness through bootstrap aggregation (bagging) and feature subset selection, which helps reduce overfitting and improve generalization. Key advantages include its ability to handle high-dimensional data, robustness to outliers, and provision of feature importance scores. However, it can be computationally expensive for large

datasets and is less interpretable than single decision trees. Common applications include fraud detection, medical diagnosis, and customer churn prediction. [8][9][13]

1.3.2. Support Vector Regression (SVR)

SVR adapts the principles of Support Vector Machines (SVMs) to regression tasks by identifying a hyperplane that minimizes prediction error within a specified tolerance margin (ε-insensitive tube). It leverages kernel functions (e.g., RBF, polynomial) to capture nonlinear relationships effectively. Key strengths include its performance in high-dimensional spaces, robustness to outliers, and flexibility through kernel selection. However, it is sensitive to hyperparameter tuning (e.g., kernel type, regularization parameter C, and ε) and can be computationally demanding for large datasets. SVR is commonly applied in energy consumption forecasting and financial time-series prediction.[9][11][14]

1.3.3. XGBoost (Extreme Gradient Boosting)

XGBoost is a scalable and efficient gradient-boosting framework that optimizes loss functions using L1/L2 regularization and advanced tree-pruning techniques. By leveraging parallel processing and native handling of missing values, it surpasses traditional Gradient Boosted Decision Trees (GBDT) in both speed and accuracy. Its key strengths include exceptional performance on structured data and built-in regularization to reduce overfitting. However, it requires careful hyperparameter tuning and can be prone to overfitting on small datasets. XGBoost has been widely successful in machine learning competitions, serving as the winning solution in many Kaggle challenges, and is commonly used in applications such as recommendation systems. [8][13]

1.3.4. Lasso Regression

Lasso (Least Absolute Shrinkage and Selection Operator) is a linear regression technique that applies L1 regularization to shrink coefficients and perform feature selection by driving some weights to exactly zero. It optimizes a loss function combining the sum of squared residuals and a penalty on absolute coefficient values

Key advantages include reducing multicollinearity and improving model interpretability through automatic feature selection. However, it may struggle with highly

correlated features and is less effective for nonlinear relationships. Lasso is widely used in genomics for identifying significant biomarkers and in sparse signal recovery applications. [12][14]

1.3.5.K-Nearest Neighbors (KNN)

KNN is a simple, non-parametric algorithm that makes predictions by finding the K most similar training examples (using distance metrics like Euclidean or Manhattan) and averaging their values (for regression) or taking a majority vote (for classification). While easy to implement and adapt to new data without retraining, KNN becomes computationally expensive with large datasets and suffers from the "curse of dimensionality" as it's sensitive to irrelevant features. This method proves particularly useful in recommender systems (finding similar user preferences) and medical diagnosis (matching patient cases to known outcomes).[8][10][14]

Table 1.2 : A comparison between the algorithms used in the machine learning model

Algorithm	Type	Key Strengths	Key Weaknesses	Best Use Cases
Random Forest	Ensemble	Robust, handles non-linearity	High memory usage	High-dimensional data
Support Vector Regression (SVR)	Kernel-based	Effective for non-linear data	Slow on large datasets	Small-to-medium datasets
XGBoost (Extreme Gradient Boosting)	Boosting	High accuracy, regularization	Hyperparameter-sensitive	Structured data competitions
Lasso Regression	Linear	Feature selection, interpretability	Poor for non-linear relationships	Sparse data
K-Nearest Neighbors (KNN)	Instance-based	No training, adaptable	Scalability issues	Low-dimensional data

In regression analysis, **MAR** (mean absolute residual), **MSE** (mean squared error), and **RMSE** (root mean squared error) measure prediction accuracy. MAR gives the average absolute difference, while MSE and RMSE highlight larger errors by squaring the deviations. These values are calculated the same way for all models according to the following equations:

1.Mean Absolute Residual (MAR) (Also called MAE - Mean Absolute Error): [15]

$$MAR = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (01)$$

2.Mean Squared Error (MSE):[16]

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (02)$$

3.Root Mean Squared Error (RMSE):[17]

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (03)$$

Where: y_i= True value, $\hat{y}_i$= Predicted value by the model, n = Number of observations

The difference between the methods lies in how each model generates predictions $\hat{y}_i$

Table 1.3 : variation of $\hat{y}_i$ in different models

Algorithm	$\hat{y}_i$
RFR	$\hat{y}_i = \frac{1}{T}\sum_{t=1}^{T} f_t(x_i)$, where: T = number of trees, f_t = prediction of the t-th tree
SVR	$\hat{y}_i = \langle \omega, x_i \rangle + b$, where: w = weight vector, b = bias term, optimized with an ϵ-insensitive loss
XGBoost Regression	$\hat{y}_i = \sum_{k-1}^{K} f_k(x_i)$, where f_k = k- th weak learner, optimized using gradient boosting

Lasso Regression	$\hat{y}_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$, where β_j are coefficients, optimized with L1 penalty: $\lambda \sum_{j=1}^{p} \lvert\beta_j\rvert$
(KNN) Regression	$\hat{y}_i = \frac{1}{K} \sum_{j \in N_K(x_i)} y_i$, where $N_k(x_i)$ = set of K nearest neighbors of x_i

Machine Learning relies on diverse data sources (databases) that must meet the **5 V's** of Big Data—Volume, Velocity, Variety, Veracity, and Value—to ensure high-quality, scalable, and actionable model training and predictions. The following table outlines the most important database site:

Table 1.4: database sources contains several materials properties

Database Name	Description	Materials Count	Direct Link
OQMD (Open Quantum Materials Database)	DFT-calculated thermodynamic/stability data for inorganic materials. Focuses on phase diagrams and ground-state predictions.	~1,000,000+	http://oqmd.org/
Materials Project	High-throughput DFT calculations for batteries, catalysts, and structural materials. Includes elastic tensors, band structures, and ionic diffusivity.	~150,000+	https://materialsproject.org/
AFLOW (Automatic FLOW for Materials Discovery)	Combinatorial database with calculated properties (electronic, mechanical, thermal) using high-throughput *ab initio* methods.	~4,000,000+	http://aflow.org/

NOMAD (Novel Materials Discovery)	Archive for *ab initio* calculations across multiple codes (VASP, Quantum ESPRESSO, etc.).Offers AI/ML tools for data mining.	~100,000,000 + entries	https://nomad-lab.eu/
ICSD (Inorganic Crystal Structure Database)	**Experimental + theoretical** crystal structures from journals and DFT studies. Paid access (academic licenses available).	~250,000 +	https://icsd.products.fiz-karlsruhe.de/
MPDS (Materials Platform for Data Science)	Curated experimental data (phase diagrams, crystallography, thermochemistry) linked to computational datasets.	~500,000 +	https://mpds.io/
JARVIS (Joint Automated Repository for Various Integrated Simulations)	DFT + ML models for materials properties (2D materials, topological insulators, mechanical properties).	~50,000+	https://jarvis.nist.gov/
Citrination (by Citrine Informatics)	AI-driven platform combining experimental and computational data for materials design.	~10,000,000 + data points	https://citrination.com/

1.4 Pymatgen libery

Advancements in materials science increasingly rely on computational methods to predict and analyze materials properties. Pymatgen has emerged as a cornerstone in this computational ecosystem, providing researchers with an intuitive, extensible framework to carry out materials data analysis, structure prediction, and database interfacing. Developed and maintained by the Materials Project team, Pymatgen serves as a key component in modern materials informatics workflows.

1.4.1 .Library Architecture and Design

Pymatgen is built in modular fashion with key components including: - **Core Structure Module:** Handles crystal structures, lattices, and atomic coordinates. - **IO Module:** Provides readers and writers for common materials file formats (e.g., CIF, POSCAR). – **Symmetry Module:** Interfaces with spglib for space group analysis and symmetry operations. - **Electronic Structure Module:** Facilitates the analysis of band structures, density of states (DOS), and electronic properties. – **Thermodynamics Module:** Calculates formation energies, phase diagrams, and stability.

1.4.2. Applications

Pymatgen supports a wide range of applications: - **Crystallography:** Analyze and visualize crystal structures. - **Electronic Properties:** Process outputs from DFT codes like VASP. – **Database Interfacing:** Query and retrieve materials data from the Materials Project API. - **High-throughput Workflows:** Automate structure generation, defect modeling, and input preparation. - **Integration with Matminer:** Enable machine learning applications through feature extraction.

1.4.3 .Data-Driven Materials Discovery

By integrating with Matminer and Scikit-learn, Pymatgen facilitates the construction of feature-rich datasets for training predictive models, accelerating the discovery of new materials with tailored properties.

Pymatgen has become an indispensable tool in computational materials science. Its comprehensive feature set, community support, and integration with major databases and simulation codes make it a foundational library for both academic and industrial materials research. Future development is expected to further enhance its scalability, performance, and AI-driven materials design capabilities.

The Pymatgen library is a powerful tool for machine learning in materials science, enabling efficient analysis of crystal structures, automated feature extraction, and seamless dataset generation—key steps in training accurate ML models for predicting material properties and accelerating discovery. [18][19]

Chapter 02 : Function materials

Perovskite and NLO

Functional materials are engineered to exhibit tailored responses (e.g., photovoltaic, electro-optic) for advanced applications. Perovskites (e.g., $CH_3NH_3PbI_3$) exemplify this through high-efficiency solar cells (>25%) and tunable optoelectronic properties. In nonlinear optics (NLO), materials like $BaTiO_3$ enable frequency conversion via large nonlinear coefficients (d_{ij}) and phase-matching, with emerging perovskite variants bridging both fields despite stability challenges.[20]

2.1 Perovskite Materials

Perovskites are a class of crystalline materials with the general chemical formula ABX_3, where:

- ✓ A is a monovalent cation (e.g., methylammonium, formamidinium, or cesium).
- ✓ B is a divalent metal cation (e.g., lead, tin).
- ✓ X is a halide anion (e.g., iodine, bromine)[21][22]

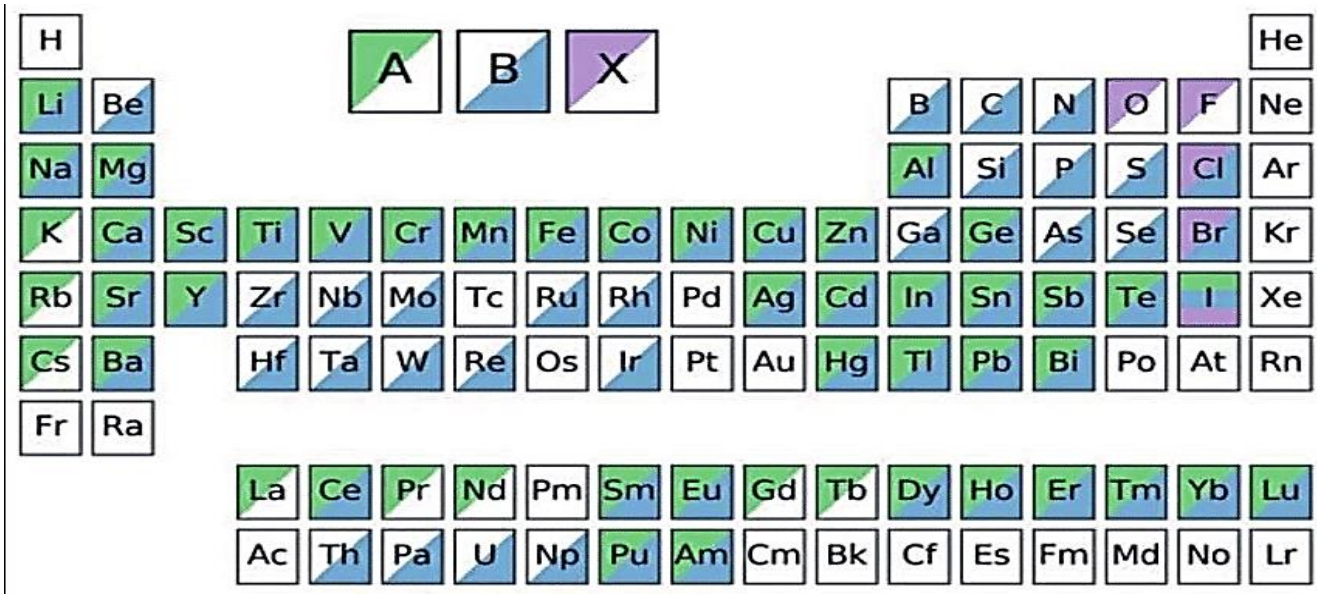


Fig 2.1 A map of the elements in the periodic table which can occupy the A, B, and/or X sites

2.2 . Ideal perovskite structure

The ideal perovskite features a simple cubic lattice in space group Pm-3m (space group number 221 in the international crystallographic tables). The crystallographic motif of perovskite is ABX3 and contains 5 atoms.

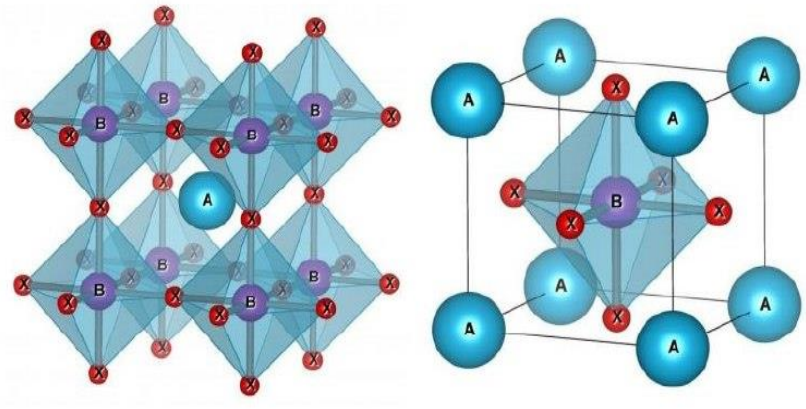


fig 2.2 Representation of ideal perovskite structure

there are two ways to describe the structure depending on the choice of origin:

- **In the first:** A in the position (0,0,0), B in the center of the cubic in the position (1/2, 1/2, 1/2) and X are in the middle of each face in the position (0, 1/2, 1/2)
- **In the second:** the origine is moved by the victor (1/2,1/2,1/2), which brings A to occupy position (1/2, 1/2, 1/2) , B to position (0,0,0), and the X atoms are located at the middle of each edge, in position (0,1/2,1/2).[23]

2.3. Different families of perovskites

Perovskites exhibit remarkable structural and compositional diversity, leading to several distinct families with unique properties. Below is a detailed classification of major perovskite families, their characteristics:

2.3.1. Halide Perovskites (ABX_3, X = Cl^-, Br^-, I^-) : Table 2.1

summarizes halide perovskites and their properties:[24][25]

Table 2 .1 : halide perovskites and their properties

Composition	Properties
A: Organic (MA^+, FA^+) or inorganic (Cs^+) cations. B : Pb^{2+}, Sn^{2+}, Ge^{2+}. X: Halides (I^-, Br^-, Cl^- or mixed).	→ Optoelectronic Excellence: High absorption coefficients (~10^5 cm^{-1}) and tunable bandgaps (1.2–3.0 eV).

2.3.2. Chalcogenide Perovskites (ABS_3, $ABSe_3$) : Table 2.2 summarizes the chalcogenide perovskites and their properties.[26][27]

Table 2.2 : chalcogenide perovskites and their properties

Composition	Properties
X: S^{2-}, Se^{2-} (instead of O^{2-} or halides). Example: $BaZrS_3$	→ Thermal Stability: Ideal for high-temperature applications. → Earth-Abundant Materials: Non-toxic, low-cost alternatives.

2.3.3. Oxide Perovskites (ABO_3, O^{2-}) :

Table 2.3 summarizes the perovskite oxides and their properties.[28][29]

Table 2.3: the perovskite oxides and their properties

Composition	Properties
A: Alkali/rare-earth metals (La^{3+}, Sr^{2+}, Ca^{2+}). B: Transition metals (Ti^{4+}, Fe^{3+}, Mn^{3+}).	→ Ferroelectricity & Magnetism: Used in capacitors, sensors, spintronics. → High-Tc Superconductors: e.g., $YBa_2Cu_3O_7$ (YBCO).

2.4. Structural stability condition of perovskite:

The stability of the peroveskite structure depends on two factor:

2.4.1. The tolerance factore:

Gold Schmidt defined a dimensional criterion, called tolerance factor, which takes into account the size of the ions to characterize the different structures derived from the perovskite structure, this tolerance factor uses the ionic radii of the atoms in the perovskite structure and can be used to indicate the stability of a certain perovskite composition. It is defined as:

$$t = \frac{r_A + r_x}{\sqrt{2}(r_B + r_x)}$$

With tolerance factore (t), r_A and r_B the ionic radii of the A and B cations respectively and r_x the ionic radius of the anion.[30]

Table 2 .4 : the variation of the tolerance factore and their corresponding structure

t	symmetry observed
$t < 0.85$	fluorite or hexagonal
$0.85 < t < 0.9$	orthohombic
$0.9 < t < 1$	rohombohedral
$t = 1$	cubic
$1 < t < 1.06$	hexagonal

2.4.2. the octahedral factor:

the stability of the perovskite is also determined by the octahedral factor **μ**, which is defined as:

$$\boldsymbol{\mu = \frac{r_B}{r_X}}$$

and should lie between 0.44 and 0.90 for stable perovskite structures.

2.5. The application of perovskite

Perovskites are transforming optoelectronics with their high efficiency and versatility. Their **solar cells** exceed 26% efficiency at low cost but need better durability. For displays, **perovskite LEDs** offer vibrant, tunable colors but shorter lifespans than OLEDs. They enable sensitive **photodetectors** for medical imaging and low-dose X-ray systems, though stability remains an issue. In **lasers**, they produce precise, energy-efficient beams but degrade over time. These materials also show promise for brain-like **computing and clean hydrogen fuel production** through water splitting, while their **thermoelectric properties** could recover waste heat. Despite their potential, stability and efficiency challenges must be solved for widespread commercial use across these groundbreaking applications.[31 − 41]

2.6. Non-linear Optical Materials

Non-linear optical materials interact with light in a non-linear manner, producing effects that are not observed in linear optics. These materials are crucial for modern optical

technologies, including telecommunications, laser systems, and quantum computing . The non-linear response arises from high-intensity electromagnetic fields altering the material's polarization. [42]

2.6.1 Fundamental Principles

The non-linear optical response is described by the polarization (P) of the material as a function of the electric field (E):

$$P = \chi^{(1)} E + \chi^{(2)} E^{2} + \chi^{(3)} E^{(3)} + \ldots\ldots$$

where:

→ $\chi(1)$ is the linear susceptibility,

→ $\chi(2)$ and $\chi(3)$ are second- and third-order non-linear susceptibilities, respectively.

Key NLO effects include:

→ **Second-harmonic generation (SHG)**: Doubling the frequency of incident light. [43]

→ **Optical Kerr effect**: Intensity-dependent refractive index change .[42]

→ **Four-wave mixing**: A third-order process used in wavelength conversion .[44]

2.6.2 Classification of NLO Materials

2.6.2.1 Inorganic NLO Materials

- **Crystals**: $LiNbO_3$, KDP (KH_2PO_4), BBO (β-BaB_2O_4) are widely used due to their high damage thresholds and efficient frequency conversion [45]
- **Semiconductors**: GaAs and ZnSe are employed in ultrafast optical applications due to their strong third-order non-linearity [46]

2.6.2.2 Organic NLO Materials

- **π-Conjugated molecules**: Exhibit high $\chi(2)$ due to electron delocalization, making them suitable for electro-optic modulators [47]
- **Poled polymers**: Used in flexible and lightweight photonic devices [47]

2.6.2.3 Hybrid and Metamaterials

- **Metal-organic frameworks (MOFs)**: Offer tunable NLO properties through structural modifications [48]
- **Plasmonic nanostructures**: Enhance non-linear effects via localized surface plasmon resonance. [49]

2.6.3 Applications

- **Laser technology**: Frequency doubling in green lasers (e.g., KDP crystals)[45]
- **Telecommunications**: All-optical signal processing using χ (3) materials.[46]
- **Biophotonics**: NLO microscopy for deep-tissue imaging [42]

NLO materials continue to evolve, with research focusing on enhancing efficiency, stability, and integration into nanophotonic devices. Future directions include quantum NLO materials and topological photonics.

Chapter 03 : Results and discussion

Two distinct datasets are utilized to investigate the influence of feature characteristics and selection on regression performance. The first dataset comprises chemically homogeneous compounds, specifically various types of perovskites with the general formula ABX_3, while the second encompasses materials with diverse chemical compositions for the application of NLO. The primary objective is to predict the **band gap** values of the materials in both datasets an essential parameter that plays a crucial role in determining material properties and subsequently guiding their suitability for specific applications.

3.1. Employing Pymatgen features

In this work, we utilized the pymatgen library implemented in python programming language [51,52] to extract a comprehensive set of compositional descriptors based on elemental properties, commonly known as **Magpie features**. These features are statistically derived (e.g., mean, maximum, range) from periodic table properties of the constituent elements in each compound, and are widely used in materials informatics to characterize chemical composition. A total of **168 features** were generated for each compound, capturing various physicochemical attributes such as atomic number, atomic weight, Mendeleev number, valence electron configuration, electronegativity, and structural parameters including space group and band gap. These descriptors form a high-dimensional representation of materials that is well-suited for machine learning applications. **Table 1** summarizes the properties used and the statistical operations applied to construct the final feature set.

Table 3.1 Extracted Magpie Features Using pymatgen

Property Name (1)	Property Name (2)	Statistical Operations Applied
Atomic Number	Mendeleev Number	min, max, range, mean, avg_dev, mode
Atomic Weight	Melting Temperature	min, max, range, mean, avg_dev, mode
Column (Periodic Group)	Row (Periodic Period)	min, max, range, mean, avg_dev, mode
Covalent Radius	Electronegativity	min, max, range, mean, avg_dev, mode
Ns Valence Electrons	Np Valence Electrons	min, max, range, mean, avg_dev, mode
Nd Valence Electrons	Nf Valence Electrons	min, max, range, mean, avg_dev, mode
Total Valence Electrons	Ns Unfilled Orbitals	min, max, range, mean, avg_dev, mode

Np Unfilled Orbitals	Nd Unfilled Orbitals	min, max, range, mean, avg_dev, mode
Nf Unfilled Orbitals	Total Unfilled Orbitals	min, max, range, mean, avg_dev, mode
GS Volume per atom (GSvolume_pa)	GS Band Gap (GSbandgap)	min, max, range, mean, avg_dev, mode
GS Magnetic Moment (GSmagmom)	Space Group Number	min, max, range, mean, avg_dev, mode

To aid interpretation, it is important to clarify a few of the terms used in the feature set. The prefixes **Ns**, **Np**, **Nd**, and **Nf** refer to the number of electrons in the **s**, **p**, **d**, and **f** valence orbitals, respectively, which are critical for understanding the electronic structure and bonding behavior of materials. The term **avg_dev** stands for **average deviation**, representing the average absolute difference of elemental values from the mean, and is useful for capturing the diversity in chemical composition. The term **mode** refers to the most frequently occurring value among the constituent elements for a given property. The prefix **GS** stands for **ground state**, indicating calculated properties such as **GSvolume_pa** (ground-state atomic volume), **GSbandgap** (electronic band gap in the ground state), and **GSmagmom** (ground-state magnetic moment). These ground-state properties are typically obtained from first-principles calculations and are essential for describing the structural and electronic behavior of materials.

3.1.1. Materials for nonlinear optical application

This class of materials was selected due to their chemically inhomogeneous compositions, based on the dataset previously compiled by *Benghia et al.* [53]from experimental reports available in the literature. Which contain 86 ternary compound and 138 quaternary compound.

The first step is to preform the prediction, the features extracted from the library. The machine learning process is presented as follow:

a. Train and test the dataset

The dataset was divided into two subsets, with 80% allocated for training and 20% for testing. This partitioning was carried out automatically using the scikit-learn library implemented in the Python programming environment. The initial stage of the analysis involved evaluating the relative importance of the input features using various regression models, including Random Forest Regression, Support Vector Regression, XGBoost, Lasso Regression, and k-Nearest Neighbors. The resulting feature importance rankings are illustrated in **Figures 3. 1, 3.2, 3.3** where the top 10 most influential features are presented. It is noteworthy that

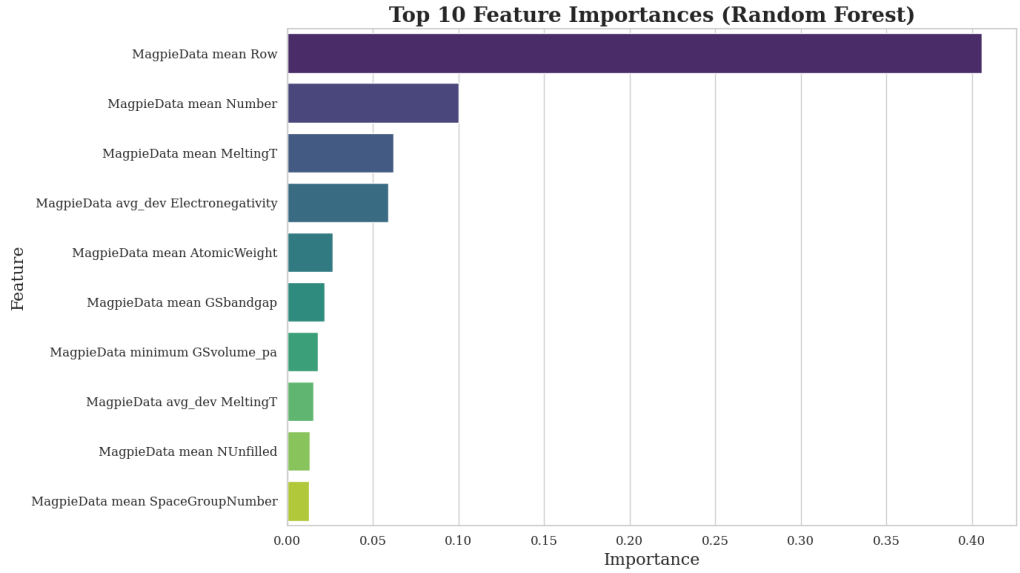


Figure 3.1: the importance of the features via random forest regression applied in the NLO dataset

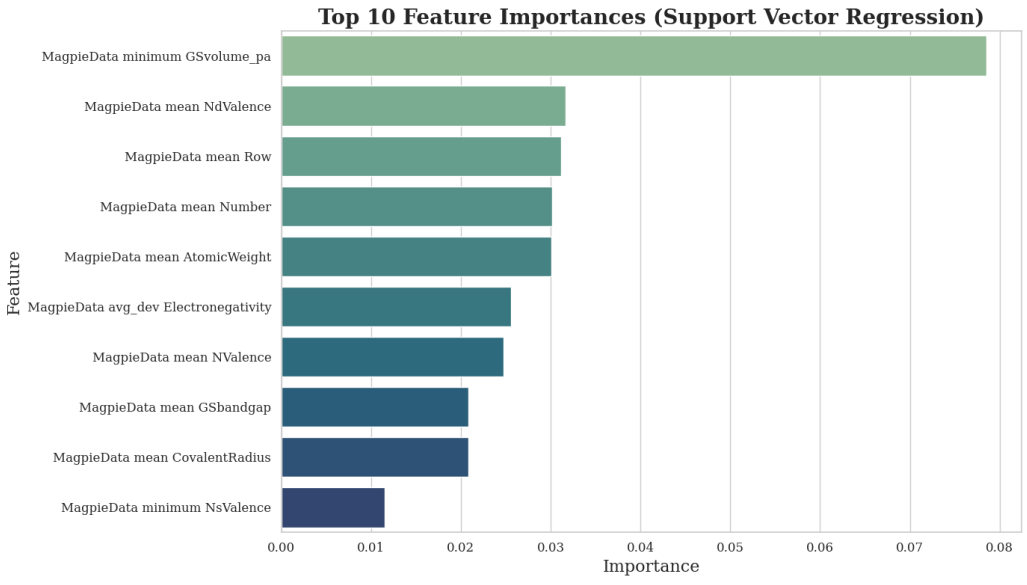


Figure 3.2: the importance of the features via support vector regression applied in the NLO dataset

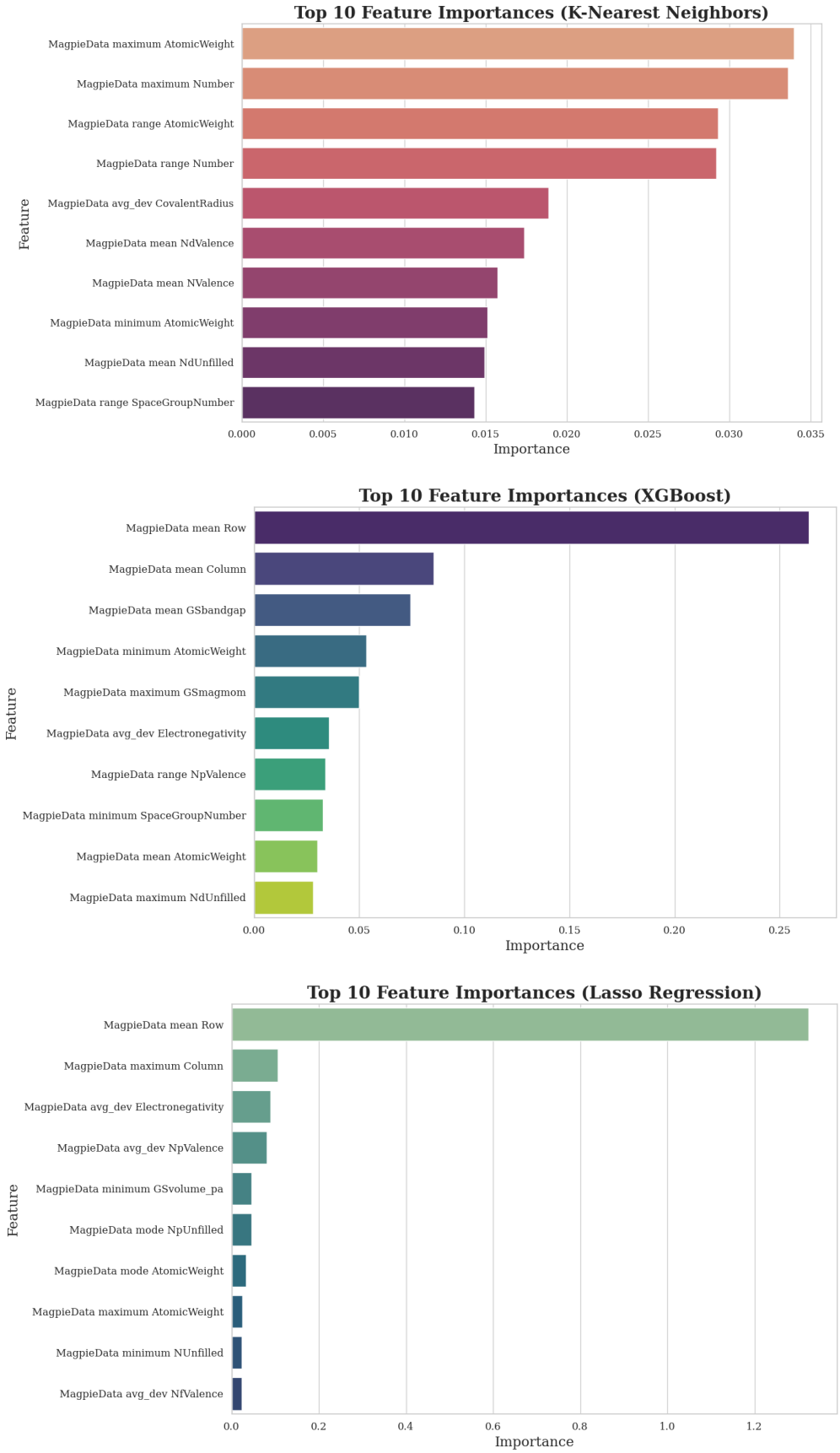


Figure3.3: the importance of the features via KNN, XGboost, and Lasso models in the NLO dataset.

each regression model produced a distinct ranking of feature importance, reflecting differences in how each algorithm interprets the contribution of features to the target property. In the Random Forest Regressor (RFR), the features *mean row* and *mean number* exhibit the highest influence on the prediction outcomes. In contrast, the Support Vector Regressor (SVR) model identifies the *minimum GS volume* as the most significant predictor. For the K-Nearest Neighbors (KNN) algorithm, the four leading features contribute equally to the model's performance, while the remaining six features show relatively uniform and lower levels of influence. The Extreme Gradient Boosting (XGBoost) model highlights the *mean row* as the most impactful parameter. Lastly, in the Lasso regression model, the *mean row* stands out as the only feature with substantial predictive power, whereas the other variables exhibit negligible influence. We remark that the mean row and the coloumn is the most impacting parameter in the prediction of this dataset

To account for variability in the train/test data splitting, the **random state** parameter was varied systematically from 10 to 50 in increments of 5. This analysis was performed using both the top 10 selected features and the full feature set. The corresponding prediction performance and the fluctuation of evaluation metrics **Mean Absolute Error (MAE), Mean Squared Error (MSE)**, and **Root Mean Squared Error (RMSE)** are illustrated in the **Figure 3.4** and **3.5**. We extract the following points:

- The Random Forest Regressor showed consistently good performance across different random states. When using all features, it achieved a minimum MAE of around **0.33 eV** and RMSE of **0.40**, indicating accurate and stable predictions. Interestingly, when limited to the top 10 features, the model still maintained strong performance, with only a slight increase in error. In some cases, such as random state 45, the top 10 features even performed comparably to the full feature set. This suggests that the model is relatively robust and can achieve high accuracy even with a reduced number of input features, which is valuable for simplifying models and improving interpretability.
- SVR delivered the **best overall performance** among all models when using the full feature set. Specifically, with random state 30, it achieved a minimum MAE of **0.29 eV** and RMSE of **0.37**, outperforming other configurations. Using only the top 10 features resulted in slightly higher error rates, but the model still maintained competitive accuracy. SVR appears to benefit from the richness of the full dataset, and while feature reduction introduces some degradation, the model remains fairly resilient and reliable across most tested scenarios.

- The performance of KNN was more variable compared to other models. With the full set of features, it achieved decent results, with MAE as low as **0.34 eV** and RMSE around **0.45**. However, when limited to the top 10 features, both MAE and RMSE increased significantly, and the model became less stable across different random states. This behavior suggests that KNN is more sensitive to feature reduction and likely relies on the full set of descriptors to maintain spatial resolution in feature space for effective predictions.
- XGBoost performed very well and demonstrated consistent accuracy across both the full feature set and the reduced top 10. With all features, it achieved MAE as low as **0.31eV** and RMSE of **0.41**, comparable to SVR. The top 10 feature model also performed competitively, maintaining stability across random states. This indicates that XGBoost is both accurate and flexible, making it suitable for complex datasets where some degree of feature selection or dimensionality reduction is desired without significant performance loss.
- Lasso regression yielded **highly unstable results** when using all features. In several random states, it produced extremely large error values (e.g., RMSE > 100), indicating possible over-regularization or numerical instability. In contrast, when limited to the top 10 features, Lasso's performance became much more stable, with MAE values ranging from **0.36 to 0.47 eV** and RMSE around **0.46 to 0.58**. While not the most accurate model overall, this behavior suggests that Lasso benefits greatly from feature selection and should be applied cautiously when working with high-dimensional data.

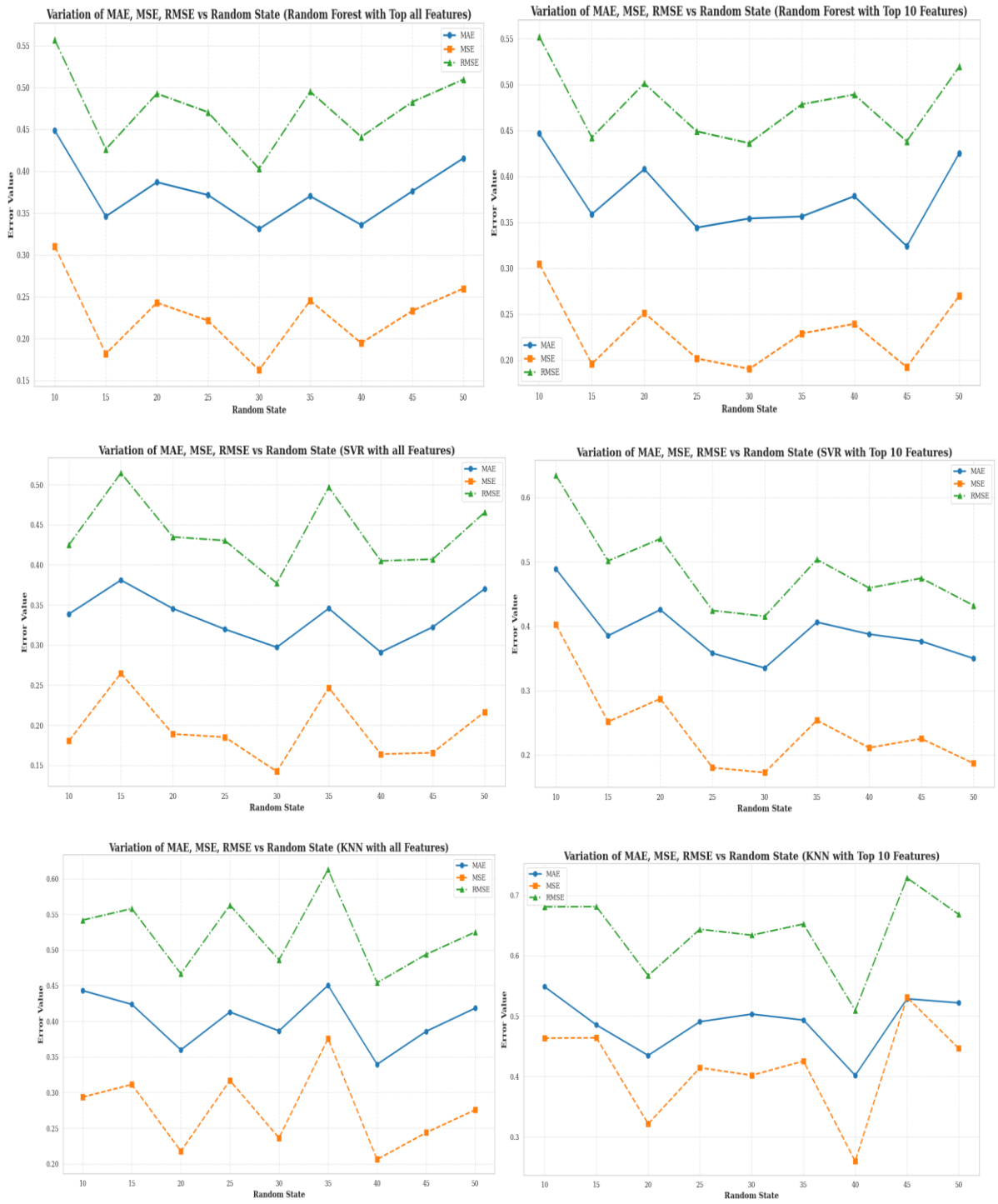


Figure3.4: the variation of the MAE, MSE, and RMSE at different random states with all the features and the top 10 features.

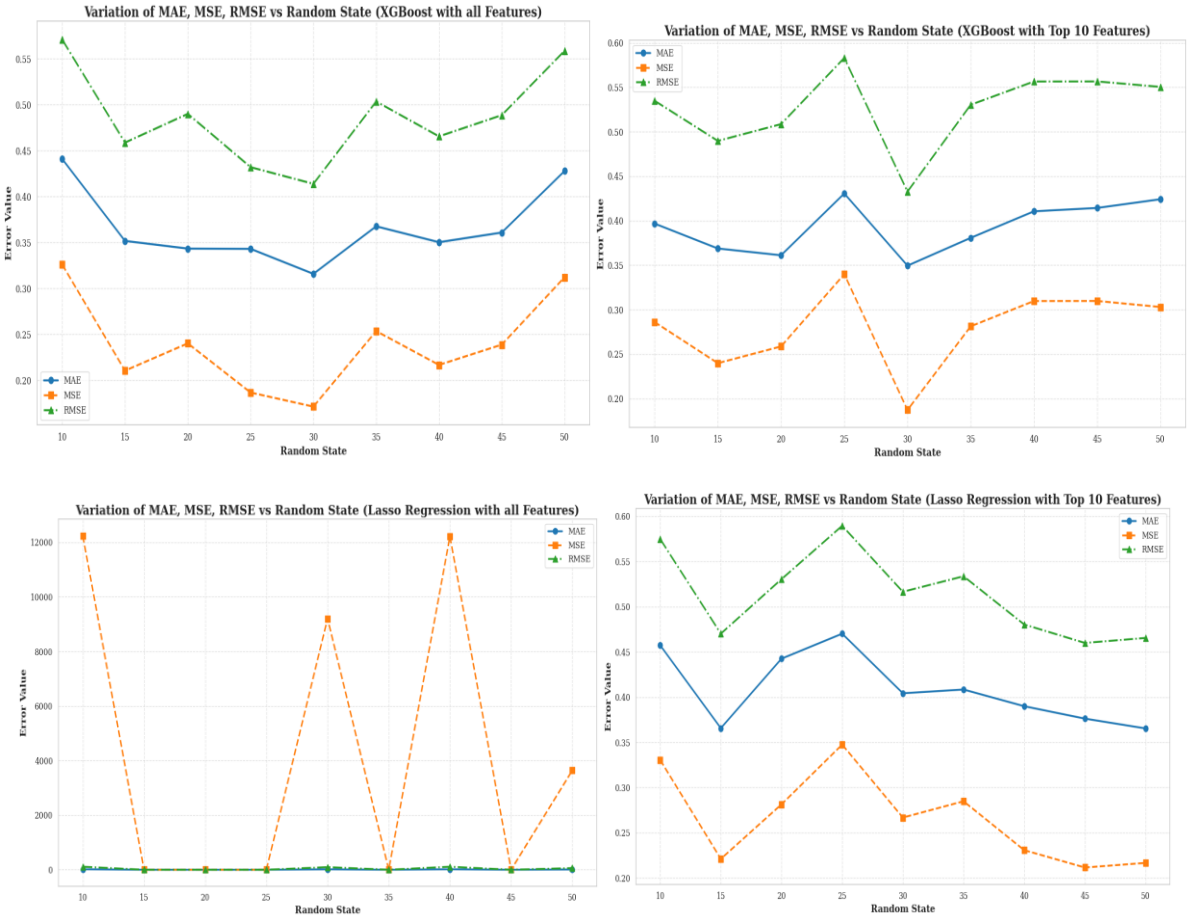


Figure3.5: the variation of the MAE, MSE, and RMSE at different random states with all the features and the top 10 features.

3.2. perovskite materials

Herein, we selected a dataset of perovskite materials sharing a common formula, comprising 643 compounds that harvested from OQMD website[54,55]. Among them, 414 are oxide-based compounds, 87 are halide-based compounds (F, Cl, Br, and I), and 142 are chalcogenides (S, Se). We followed the same preprocessing and modeling steps as those applied to the nonlinear optical (NLO) materials dataset. The feature importance derived from each machine-learning model is illustrated in **Figure 3.6** and **3.7** and **3.8**. In the Random Forest Regressor (RFR), the feature *average deviation of electronegativity* demonstrates the greatest influence on prediction accuracy. In the case of the Support Vector Regressor (SVR), ten features collectively contribute most significantly to the model's predictive performance. The K-Nearest Neighbors (KNN) algorithm assigns an approximately equal importance to all input features. For the Extreme Gradient Boosting (XGBoost) model, both the *average deviation of the column* and *average deviation of electronegativity* are identified as the most impactful variables. Finally, within the Lasso regression model, the *average deviation of electronegativity*

emerges as the dominant predictor, while the remaining features exert minimal influence. These observations collectively suggest that electronegativity plays a consistently critical role across all models in this dataset.

To avoid redundancy, it is worth noting that the influence of data partitioning on model performance has been thoroughly examined through a systematic variation of the random state parameter. As previously detailed, this evaluation conducted using both the top 10 features and the complete feature set is visually summarized in **Figure 3.9 and 3.10**, which highlights the variability of key performance indicators including MAE, MSE, and RMSE.

- The analysis reveals that Random Forest Regressor (RFR) with all features yielded the lowest MAE 0.50 eV and lowest RMSE 0.69 at random states 15 and 35, respectively, highlighting its robustness in capturing the complex nonlinearities in the data. RFR with top 10 features also demonstrated competitive performance, with an MAE of 0.49 eV and RMSE of 0.70, suggesting that dimensionality reduction does not significantly impair the model.
- For XGBoost, the best performance using all features was achieved at random state 15 with MAE = 0.52 eV and RMSE = 0.75, outperforming its top-10-feature counterpart whose best MAE 0.56 eV and RMSE 0.82 were slightly higher. This suggests that XGBoost benefits more from the complete feature set than from a reduced one.
- The Support Vector Regressor (SVR) model reached its lowest RMSE of 0.76 using all features at random state 15, compared to a much higher minimum RMSE of 0.8821 when only the top 10 features were used, confirming that SVR is sensitive to feature selection.
- K-Nearest Neighbors (KNN) performed well with both configurations. Using all features, the best RMSE achieved was 0.8040, and with top 10 features, 0.7440 indicating that KNN can maintain performance with fewer features, although results slightly fluctuate depending on the random split.
- Lasso regression, on the other hand, showed the weakest performance overall, with the best RMSE being 0.8989 using all features and 0.9503 with the reduced feature set, confirming its limitation in capturing nonlinear dependencies within the data.

In conclusion, the Random Forest Regressor outperformed all other models, especially when using the top 10 features, demonstrating both accuracy and stability. This insight supports the use of ensemble tree-based models in this context.

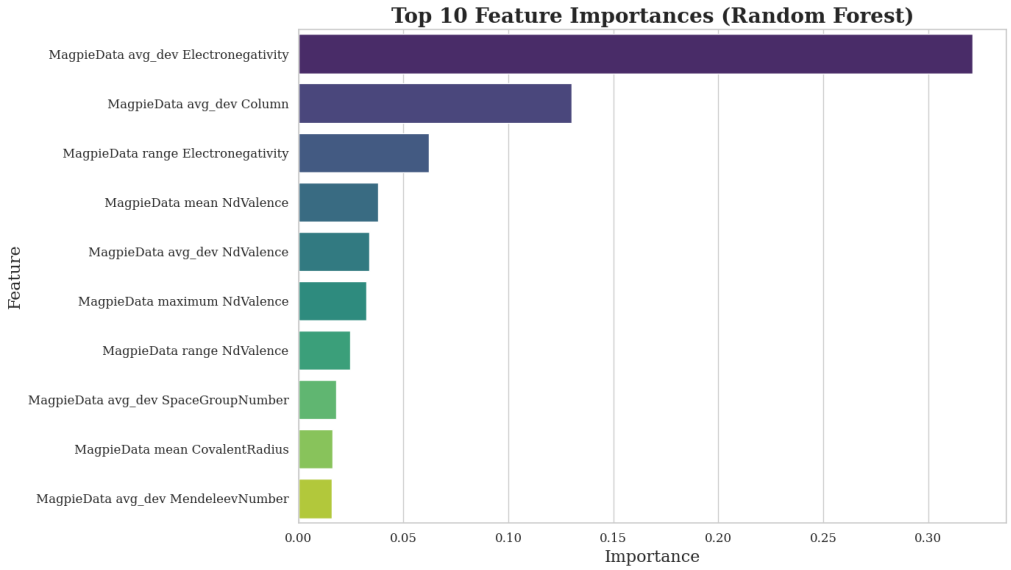


Figure 3.6: the importance of the features via RFR applied in the perovskites dataset.

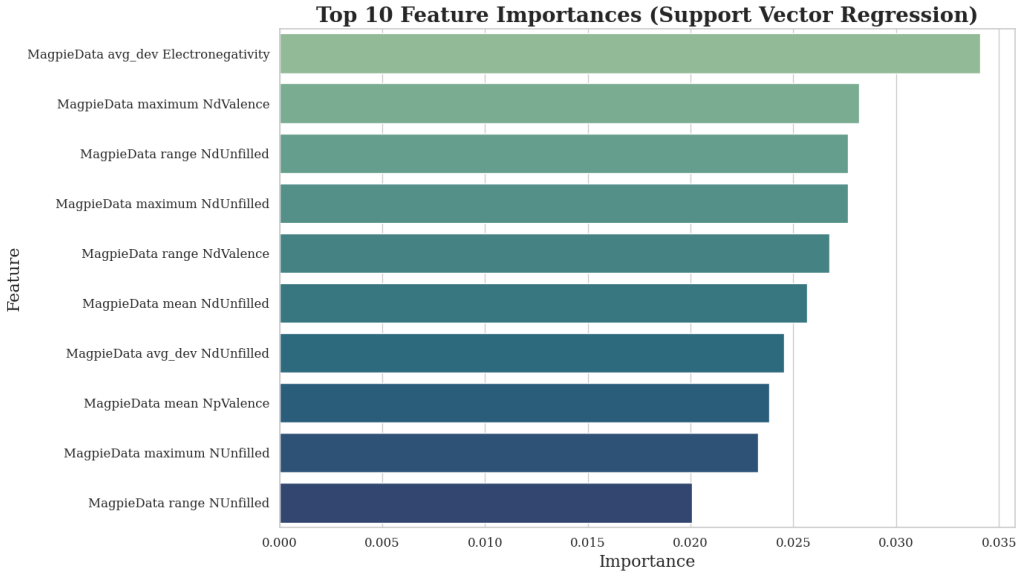


Figure 3.7: the importance of the features via SVR applied in the perovskites dataset.

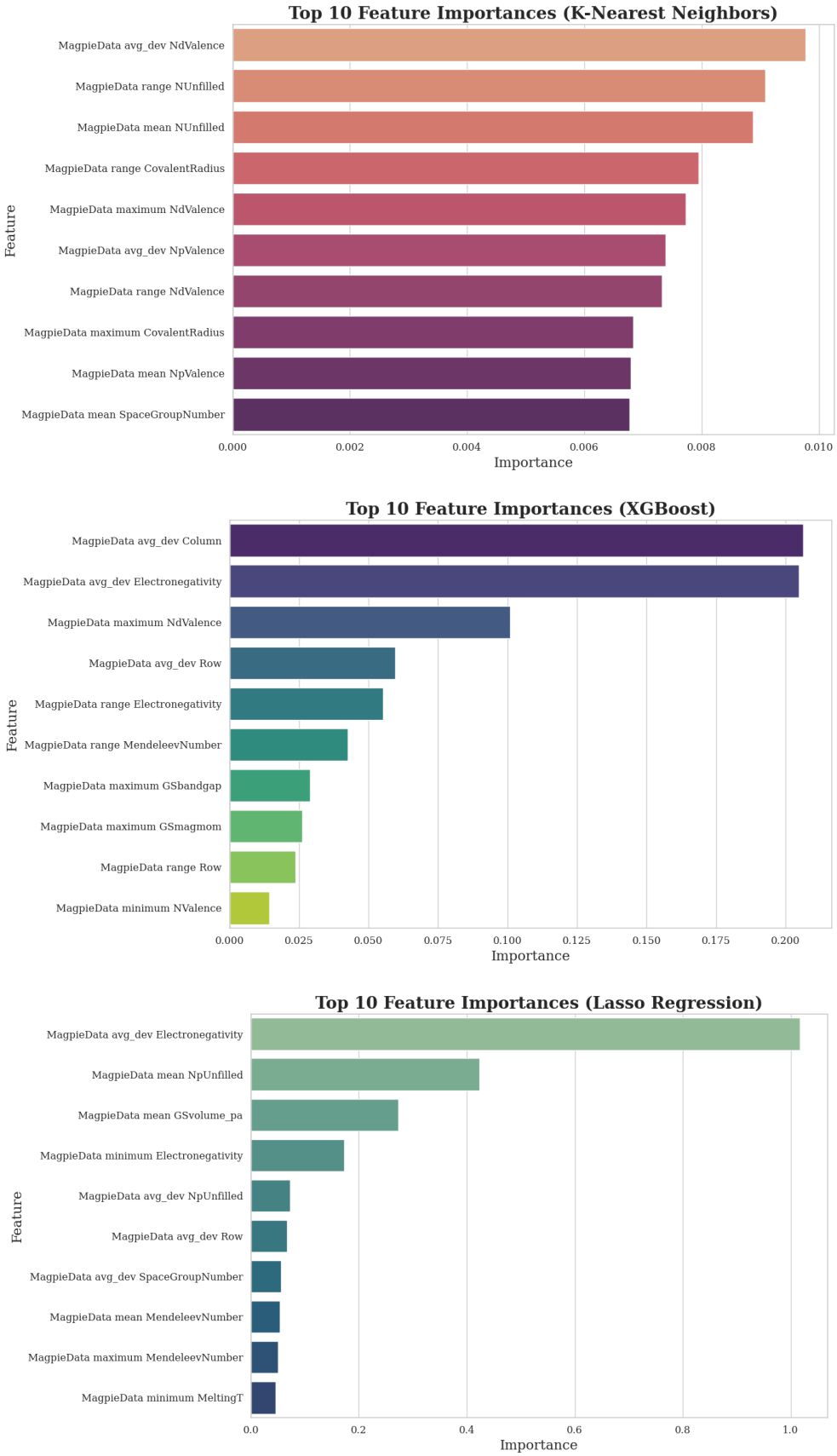


Figure3.8: the importance of the features via KNN, XGboost, and Lasso models in the perovskites dataset.

Figure3.9: the variation of the MAE, MSE, and RMSE at different random states with all the features and the top 10 features.

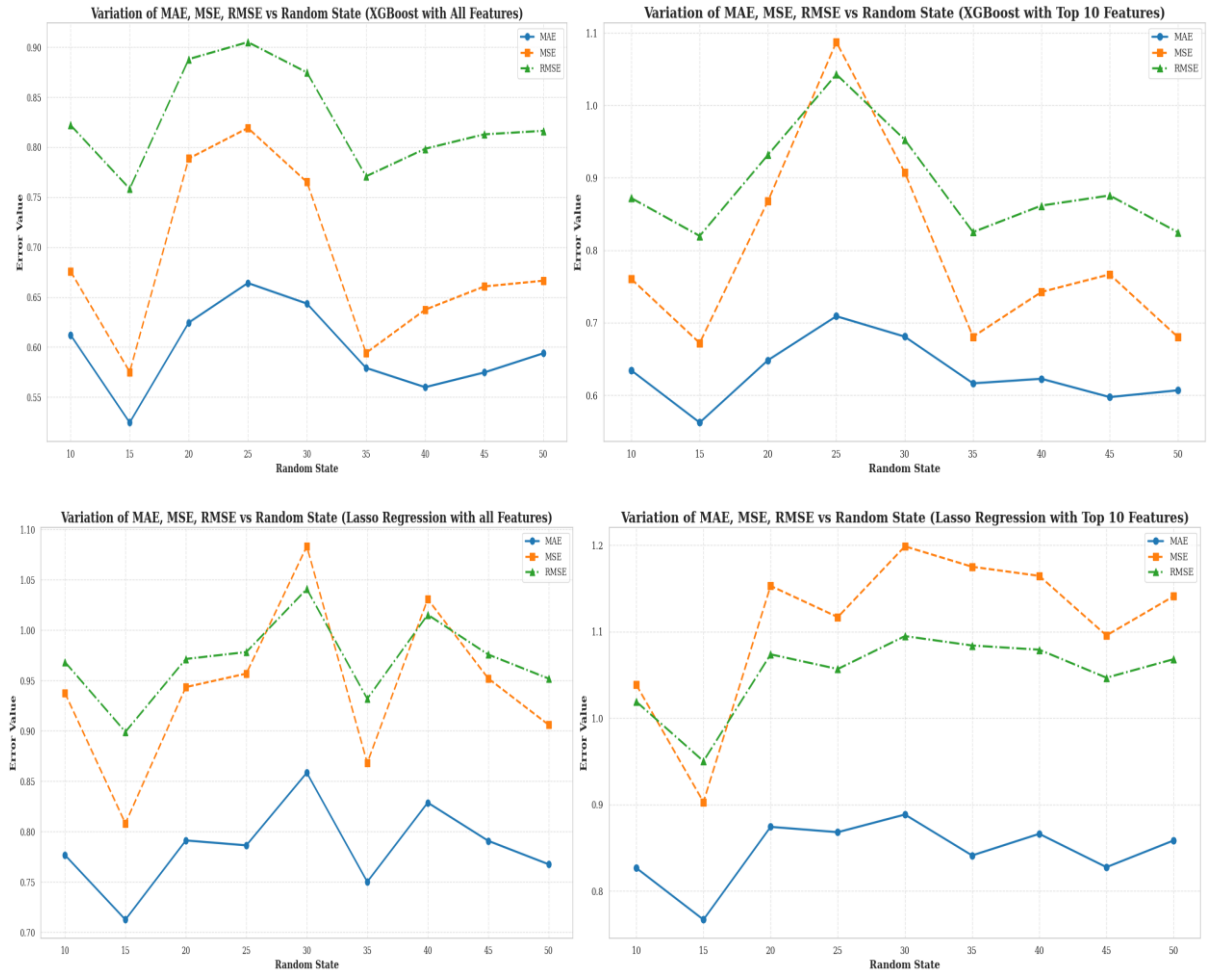


Figure3.10: the variation of the MAE, MSE, and RMSE at different random states with all the features and the top 10 features.

3.3. Applying dimensionality reduction

To enhance the conciseness and interpretability of the study, dimensionality reduction techniques were applied to both datasets to assess whether prediction accuracy is affected by a reduced feature space. Principal Component Analysis (PCA) was employed to transform the original features into a set of uncorrelated components while preserving most of the dataset's variance. **Figures 3.11** and **3.12** illustrate the explained variance, indicating the proportion of total information retained within each principal component (PC). In the NLO dataset, the first principal component (PC1) captures 18.08% of the variance, while PC2 accounts for 13.88%. Collectively, the first ten principal components explain 78.24% of the dataset's total variance. In contrast, for the perovskite dataset, PC1 and PC2 account for 22.95% and 13.88% of the variance, respectively, with the first ten components cumulatively explaining approximately 89.72% a higher retained variance compared to the NLO dataset. This suggests

that the perovskite dataset may be more amenable to dimensionality reduction without significant loss of information.

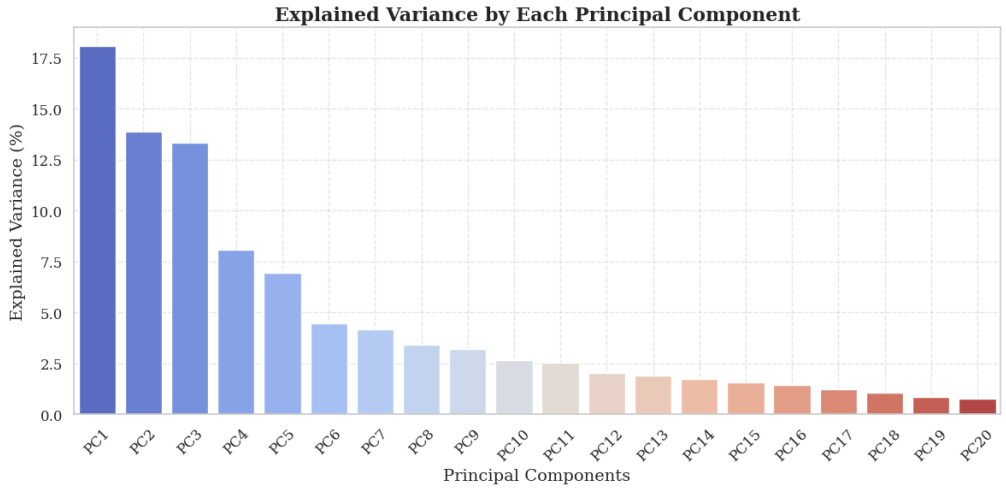


Figure 3.11: the explained variance of the reducing dimensions in the NLO dataset

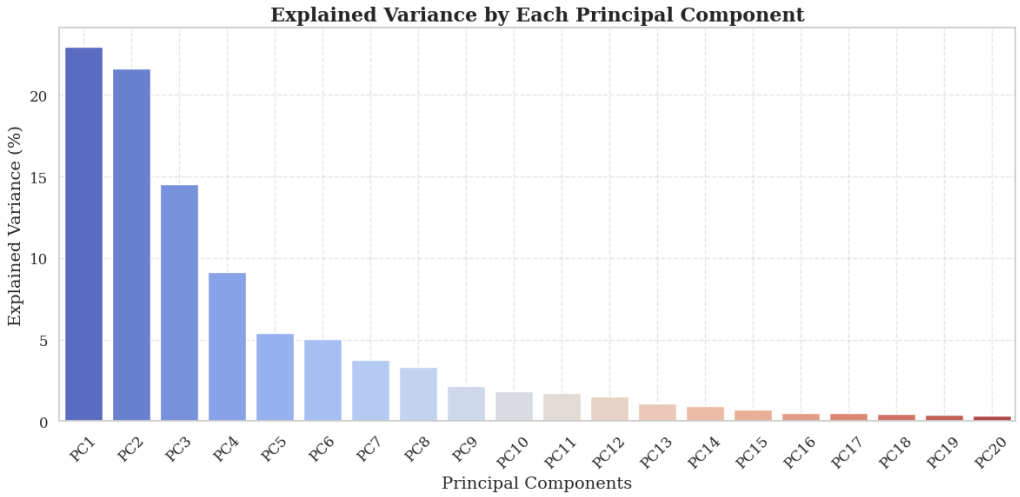


Figure 3.12: the explained variance of the reducing dimensions in the Perovskite dataset

The variability introduced by changing the random state parameter was carefully examined. As illustrated in **Figure3.13** and **3.14**, this analysis reveals fluctuations in the primary evaluation metrics: MAE, MSE, and RMSE. From this, we extract the following observations.

- Among the regression models applied to the ABX dataset, the Random Forest Regressor (RFR) consistently delivers the lowest MAE values, ranging approximately from 0.54 to 0.66 eV, indicating relatively better prediction accuracy. The XGBoost model follows closely, with MAE values between 0.59 and 0.72 eV. The Support Vector Regressor (SVR) and K-Nearest Neighbors (KNN) models show moderate performance, with MAE values generally between 0.58 and 0.70 eV for SVR, and 0.64 to 0.78 eV for

KNN. The Lasso regression performs the worst among all, with MAE values consistently above 0.83 eV, suggesting lower predictive power for this dataset

- For the NLO dataset, the Random Forest Regressor (RFR) again shows superior performance with MAE values notably lower than in ABX, ranging from approximately 0.39 to 0.49 eV. KNN and SVR models follow, with MAE values spanning roughly 0.33 to 0.56 eV and 0.33 to 0.47 eV respectively, both outperforming their ABX counterparts. XGBoost also performs well with MAE between 0.40 and 0.52 eV. Lasso

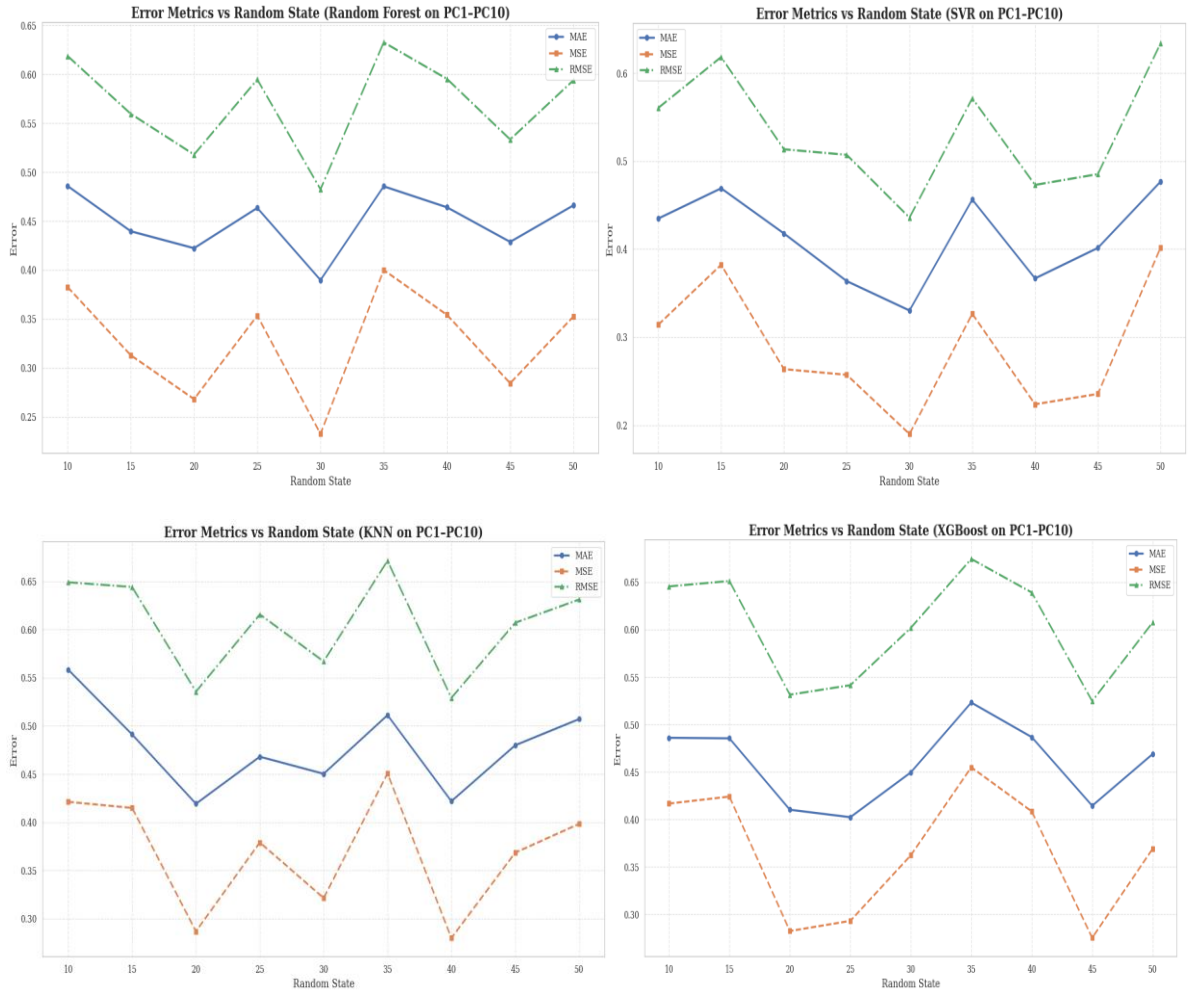


Figure3.13: the variation of the MAE, MSE, and RMSE taking in consideration the PCA features in the NLO dataset.

regression, while less accurate than the other models on NLO, still achieves MAE values lower than those observed for ABX, between 0.44 and 0.74 eV.

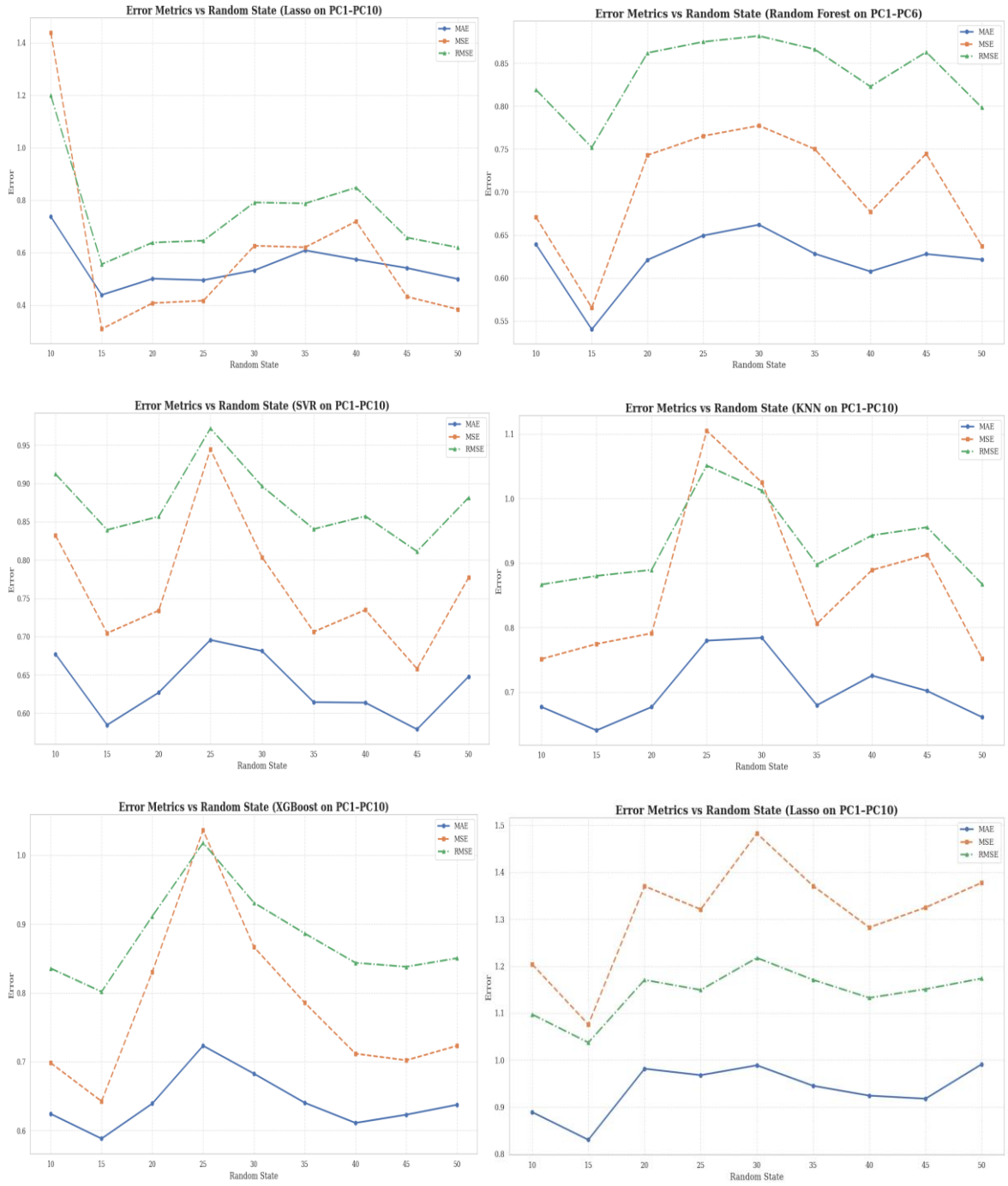


Figure 3.14: presents the changes in MAE, MSE, and RMSE values when PCA-derived features are used for the Perovskites dataset. (Note: The first graph on the left displays the Lasso regression results obtained from the NLO dataset.)

Conclusion

In summary, we conducted an in-depth study highlighting the significance of features extracted from the Pymatgen library for predicting properties across two distinct dataset types: one comprising different formula materials and the other focused on a single formula type. The NLO materials dataset includes 224 compounds, consisting of 86 ternary and 138 quaternary compounds. Additionally, the perovskite dataset features 637 compounds with the general formula ABX_3, where the X site is occupied by elements such as O, S, Se, Te, F, I, Cl, and Br each contributing a specific percentage to the dataset (percentages to be added). From this analysis, we derived the following key insights:

- In the ABX dataset, Random Forest Regressor (RFR) performs best across all settings, with lowest MAE = 0.4941 (top 10 features) and 0.5044 (all features); PCA performance is slightly lower (MAE = 0.54–0.66).
- In the NLO dataset, RFR also dominates, achieving the best MAE = 0.3309 (all features) and 0.3238 (top 10); PCA remains effective (MAE = 0.39–0.49), showing robustness to feature reduction.
- XGBoost performs better with all features in both datasets (MAE = 0.5245 in ABX, 0.3159 in NLO), with moderate accuracy loss when using top 10 features or PCA.
- SVR excels in NLO using all features (MAE = 0.2972), but shows greater sensitivity to feature reduction and PCA, especially in ABX (PCA MAE > 0.58).
- KNN benefits from the full feature set in both datasets (MAE = 0.4183 in ABX, 0.3395 in NLO), but suffers when using top 10 or PCA, especially in ABX.
- Lasso regression is unstable with all features (ABX: RMSE > 100 in some states), and only shows reasonable performance when reduced to top 10 features (MAE ~0.36–0.47); PCA helps slightly but is less effective than manual feature selection.

To further enhance model accuracy and scientific insight, future work should focus on custom feature engineering grounded in physical and chemical principles, electronic structure symmetry metrics. Incorporating graph-based descriptors that capture atomic connectivity and coordination environments, as well as dynamical and thermodynamic descriptors, can provide richer representations. Additionally, features derived from density functional theory (DFT) calculations like charge densities, partial density of states, or Bader charges alongside multi-scale descriptors combining atomic and bulk material properties, offer promising avenues.

These advanced features can help develop more interpretable models rooted in fundamental physics, enabling deeper scientific understanding and guiding the discovery of novel materials.

Reference

1. Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.
2. Alam, A. (2023). *What is Machine Learning?* Preprint. Zenodo.
3. MacQueen, J. (1967). *"Some Methods for Classification and Analysis of Multivariate Observations."*
4. Johnson, S. C. (1967). *"Hierarchical Clustering Schemes."*
5. Bellman, R. (1961). *Adaptive Control Processes.* Princeton University Press.
6. Jolliffe, I. (2002). *Principal Component Analysis.* Springer.
7. McInnes, L. (2018). *UMAP: Uniform Manifold Approximation and Projection.* arXiv.
8. Medium: "10 Popular ML Algorithms for Classification"
9. Medium: "Comparison of Logistic Regression, Decision Trees, SVM, Random Forest, XGBoost" .
10. SmartCore Documentation: KNN Implementation .
11. Nature: "Evaluating ML Algorithms for Energy Prediction" .
12. PMC: "Brain Age Prediction with Regularized Linear Models" .
13. Scikit-learn: Ensemble Methods .
14. Springboard: "Regression vs. Classification"
15. Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), 679–688
16. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.
17. Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? *Geoscientific Model Development, 7*(1), 1247–1250.
18. Ong, S. P., Richards, W. D., Jain, A., et al. (2013). Python Materials Genomics (pymatgen): Arobust, open-source python library for materials analysis. *Computational Materials Science*, 68, 314-319.
19. Jain, A., Ong, S. P., Hautier, G., et al. (2013). Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1), 011002.

20. Green et al., *Nat. Photon.* (2014) and Chen et al., *Chem. Rev.* (2021)

21. Green, M. A., et al. (2022). *Perovskite Solar Cells: Materials and Devices*. Springer.

22. Snaith, H. J. (2018). "Perovskite Solar Cells: The Future of Photovoltaics?" *Science*, 358(6364), 739-744.

23. Arar, R. *Study of the Structural, Electronic, and Mechanical Properties of Fluoro-Perovskite Compounds Based on Sodium $NaXF_3$ (X=Mg, Zn) Using the FP-LAPW Method.* [Sidi Bel Abbès]: Djilali Liabes University; 2017.

24. Jung, H. S., & Park, N. G. (2021). *"Perovskite Solar Cells: From Materials to Devices." Small*, 17(20), 2005247.

25. National Renewable Energy Laboratory (NREL). (2023). *"Best Research-Cell Efficiencies."*

26. Sun, Y. Y., et al. (2016). *"Chalcogenide Perovskites for Photovoltaics." Nano Letters*, 16(5), 3343-3348.

27. Kumar, M. H., et al. (2019). *"Chalcogenide Perovskites: Emerging Candidates for Thin-Film Photovoltaics." ACS Energy Letters*, **4**(2), 538-544.

28. Rödel, J., et al. (2015). *"Transferring Lead-Free Piezoelectric Ceramics into Application." Journal of the European Ceramic Society*, 35(6), 1659-1681.

29. Dagotto, E. (2013). *"Complexity in Strongly Correlated Electronic Systems." Science*, 309(5732), 257-262.

30. V. M. Goldschmidt, Geochemistry, Ed. Ely House, London: Oxford University Press (1958).

31. NREL (2023). Best Research-Cell Efficiency Chart. National Renewable Energy Laboratory.

32. Snaith, H.J. (2018). "Perovskite Solar Cells: The Future of Photovoltaics?" *Nature Materials* 17, 372-376.

33. Tan, Z.K. et al. (2014). "Bright Light-Emitting Diodes Based on Organometal Halide Perovskite" *Nature Nanotechnology* 9, 687-692.

34. Lin, K. et al. (2016). "Perovskite Light-Emitting Diodes with External Quantum Efficiency Exceeding 20%" *Advanced Materials* 28(32), 6687-6694.

35. Wei, H. et al. (2016). "Sensitive X-ray Detectors Made of Methylammonium Lead Tribromide Perovskite Single Crystals" *Nature Photonics* 10, 333-339.

36. Xing, G. et al. (2014). "Low-temperature Solution-processed Wavelength-tunable Perovskite Lasers" *Nature Materials* 13, 476-480.
37. Wang, Z. et al. (2020). "Memristors with Diffusive Dynamics as Synaptic Emulators for Neuromorphic Computing" *Science Advances* 6(11), eaaz5223.
38. Chen, X. et al. (2015). "Semiconductor-based Photocatalytic Hydrogen Generation" *Chemical Reviews* 115(23), 12888-12935.
39. He, J. & Tritt, T.M. (2017). "Advances in Thermoelectric Materials Research" *Science* 357(6358), eaak9997.
40. Miyasaka, T. (2018). *Perovskite Photovoltaics and Optoelectronics*. Wiley-VCH.
41. Würfel, P. (2016). *The Physics of Solar Cells: Perovskites*. Jenny Stanford Publishing.
42. Boyd, R. W. (2020). *Nonlinear Optics* (4th ed.). Academic Press.
43. Shen, Y. R. (2002). *The Principles of Nonlinear Optics*. Wiley.
44. Dmitriev, V. G., Gurzadyan, G. G., & Nikogosyan, D. N. (1999). *Handbook of Nonlinear Optical Crystals*. Springer.
45. Sutherland, R. L. (1996). *Handbook of Nonlinear Optics*. Marcel Dekker.
46. Peyghambarian, N., Koch, S. W., & Mysyrowicz, A. (1993). *Introduction to Semiconductor Optics*. Prentice Hall.
47. Prasad, P. N., & Williams, D. J. (1991). *Introduction to Nonlinear Optical Effects in Molecules and Polymers*. Wiley.
48. Zhang, W., Xiong, R., & Huang, S. (2018). "Metal-Organic Frameworks for Nonlinear Optics". *Chemical Reviews, 118*(8), 4169-4218.
49. Butet, J., Brevet, P. F., & Martin, O. J. F. (2015). "Optical Second Harmonic Generation in Plasmonic Nanostructures". *ACS Nano, 9*(11), 10545-10562.
50. Kauranen, M., & Zayats, A. V. (2012). "Nonlinear Plasmonics". *Nature Photonics, 6*(11), 737-748.
51. Russell, R., *Machine Learning: Step-By-Step Guide to Implement Machine Learning Algorithms with Python*2020: (Knxb).
52. Pilgrim, M. and S. Willison, *Dive into python 3*. Vol. 2. 2009: Springer.
53. Benghia, A., et al., *Data driven enhancement of mid-infrared non-linear optical properties of quaternary and ternary chalcogenides.* Optik, 2023. 293: p. 171432.

54. Kirklin, S., et al., *The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies.* npj Computational Materials, 2015. 1(1): p. 1-15.

55. Saal, J.E., et al., *Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD).* Jom, 2013. 65: p. 1501-1509.

56. Wei, J., et al., *Machine learning in materials science.* InfoMat, 2019. **1**(3): p. 338-358.

57. Morgan, D. and R. Jacobs, *Opportunities and challenges for machine learning in materials science.* Annual Review of Materials Research, 2020. **50**(1): p. 71-103.

58. Mueller, T., A.G. Kusne, and R. Ramprasad, *Machine learning in materials science: Recent progress and emerging applications.* Reviews in computational chemistry, 2016. **29**: p. 186-273.

59. Rodrigues, J.F., et al., *Big data and machine learning for materials science.* Discover Materials, 2021. **1**: p. 1-27.

60. Gao, C., et al., *Innovative materials science via machine learning.* Advanced Functional Materials, 2022. **32**(1): p. 2108044.

61. Chong, S.S., et al., *Advances of machine learning in materials science: Ideas and techniques.* Frontiers of Physics, 2024. **19**(1): p. 13501.

62. Nematov, D. and M. Hojamberdiev, *Machine Learning-Driven Materials Discovery: Unlocking Next-Generation Functional Materials--A minireview.* arXiv preprint arXiv:2503.18975, 2025.

63. Mortazavi, B., *Recent Advances in Machine Learning-Assisted Multiscale Design of Energy Materials.* Advanced Energy Materials, 2025. **15**(9): p. 2403876.

64. Mechraoui, B., et al., *From data to discovery: Exploring novel photovoltaic perovskite chalcogenide crystals MgMS3 (M: Zr, Ti, Hf) via machine learning and density functional theory calculations.* Materials Today Communications, 2025. **42**: p. 111517.

65. Benghia, A., et al., *Data driven enhancement of mid-infrared non-linear optical properties of quaternary and ternary chalcogenides.* Optik, 2023. **293**: p. 171432.

66. Pederson, R., B. Kalita, and K. Burke, *Machine learning and density functional theory.* Nature Reviews Physics, 2022. **4**(6): p. 357-358.

67. Ryczko, K., D.A. Strubbe, and I. Tamblyn, *Deep learning and density-functional theory.* Physical Review A, 2019. **100**(2): p. 022512.

68. del Rio, B.G., B. Phan, and R. Ramprasad, *A deep learning framework to emulate density functional theory.* npj Computational Materials, 2023. **9**(1): p. 158.

69. Xu, P., et al., *Small data machine learning in materials science.* npj Computational Materials, 2023. **9**(1): p. 42.

70. Liu, Y., et al., *Data quantity governance for machine learning in materials science.* National Science Review, 2023. **10**(7): p. nwad125.

71. Liu, Y., et al., *Generative artificial intelligence and its applications in materials science: Current situation and future perspectives.* Journal of Materiomics, 2023. **9**(4): p. 798-816.

72. Roy, P., et al., *A review on perovskite solar cells: Evolution of architecture, fabrication techniques, commercialization issues and status.* Solar energy, 2020. **198**: p. 665-688.

73. Prabhu, P. and J.-M. Lee, *Metallenes as functional materials in electrocatalysis.* Chemical Society Reviews, 2021. **50**(12): p. 6700-6719.

74. Yan, X., Y. Zhou, and S. Wang, *Nano-High Entropy Materials in Electrocatalysis.* Advanced Functional Materials, 2025. **35**(2): p. 2413115.

75. Li, W., R. Jacobs, and D. Morgan, *Predicting the thermodynamic stability of perovskite oxides using machine learning models.* Computational materials science, 2018. **150**: p. 454-463.

76. Touati, S., et al., *Predictive machine learning approaches for perovskites properties using their chemical formula: towards the discovery of stable solar cells materials.* Neural Computing and Applications, 2024: p. 1-11.

77. Franken, P., et al., *Generation of optical harmonics.* Physical Review Letters, 1961. **7**(4): p. 118.

78. Zhou, W., et al., *AAg 2 PS 4 (A= K, Na/K): the first-type of noncentrosymmetric alkali metal Ag-based thiophosphates exhibiting excellent second-order nonlinear optical performances.* Inorganic Chemistry Frontiers, 2022. **9**(19): p. 4990-4998.

79. Yang, Z. and S. Pan, *Computationally assisted multistage design and prediction driving the discovery of deep-ultraviolet nonlinear optical materials.* Materials Chemistry Frontiers, 2021. **5**(9): p. 3507-3523.

80. Sun, Y., et al., *α-Ca2CdP2 and β-Ca2CdP2: Two Polymorphic Phosphide-Based Infrared Nonlinear Crystals with Distorted NLO-Active Tetrahedral Motifs Realizing Large Second Harmonic Generation Effects and Suitable Band Gaps.* Inorganic Chemistry, 2021. **60**(10): p. 7553-7560.

81. Wang, H., et al., *Computer-Aided Development of New Nonlinear Optical Materials.* Angewandte Chemie, 2025. **137**(6): p. e202420526.

82. Cai, W., et al., *Toward the rational design of mid-infrared nonlinear optical materials with targeted properties via a multi-level data-driven approach.* Advanced Functional Materials, 2022. **32**(23): p. 2200231.

83. Wang, R., F. Liang, and Z. Lin, *Data-driven prediction of diamond-like infrared nonlinear optical crystals with targeting performances.* Scientific Reports, 2020. **10**(1): p. 1-8.

84. An, R., et al., *New Ways to Discover Novel Nonlinear Optical Materials: Scaling Machine Learning with Chemical Descriptors Information.* Small, 2025: p. 2500540.

85. Zhang, Z.-Y., et al., *Machine learning with multilevel descriptors for screening of inorganic nonlinear optical crystals.* The Journal of Physical Chemistry C, 2021. **125**(45): p. 25175-25188.

86. Raju, L., et al., *Maximized frequency doubling through the inverse design of nonlinear metamaterials.* ACS nano, 2022. **16**(3): p. 3926-3933.

87. Xu, Y., L. Jiang, and X. Qi, *Machine learning in thermoelectric materials identification: Feature selection and analysis.* Computational materials science, 2021. **197**: p. 110625.

88. Guo, J., et al. *A genetic algorithm-based artificial network method for material feature recombination*. in *2021 IEEE 6th International Conference on Smart Cloud (SmartCloud)*. 2021. IEEE.

ملخــص

تتناول هذه الدراسة تأثير اختيار الخصائص (features) على دقة التنبؤ بفجوة الطاقة (band gap) في مجموعتين مختلفتين من المواد: مركبات البصريات غير الخطية (NLO) ومواد البيروفسكايت من نوع ABX_3. تم استخراج الخصائص من مكتبة Pymatgen باستخدام لغة البرمجة Python. تضمنت قاعدة بيانات NLO عدد 224 مركبًا كيميائيًا متنوعًا تم جمعها من الأدبيات العلمية، بينما شملت قاعدة بيانات البيروفسكايت 643 مركبًا من نوع ABX_3 حيث X يمثل أحد العناصر التالية: O، S، Se، F، Cl، Br، أو I ، وتم الحصول عليها من قاعدة بيانات OQMD. تم استخدام عدة نماذج تعلم آلي، وأظهر نموذج الانحدار بالغابة العشوائية (RFR) أفضل أداء بين جميع النماذج، إذ بلغت قيمة متوسط الخطأ المطلق (MAE) في قاعدة بيانات ABX_3 حوالي 0.49 إلكترون فولت باستخدام أفضل 10 ميزات، و0.50 إلكترون فولت عند استخدام جميع الميزات، بينما حقق في قاعدة بيانات NLO قيمة MAE تبلغ 0.33 و0.32 إلكترون فولت على التوالي. ولم تسهم تقنيات تقليل الأبعاد مثل التحليل بالمكونات الرئيسية (PCA) في تحسين الأداء مقارنة بالميزات الأصلية أو المختارة يدويًا. تؤكد النتائج على أهمية اختيار الخصائص بعناية، خاصة عند استخدام موصّفات مُولّدة تلقائيًا مثل تلك المستخرجة منPymatgen ، لتعزيز كفاءة ودقة تطبيقات الذكاء الاصطناعي في اكتشاف المواد الجديدة.

الكلمات المفتاحية: مكتبةPymatgen ، المواد الوظيفية، التعلم الآلي، نماذج الانحدار، تحليل المكونات الرئيسية.

Abstract

This study explores the impact of feature selection on the accuracy of band gap prediction across two distinct material datasets: nonlinear optical (NLO) compounds and perovskite-structured ABX_3 materials. Features were extracted from the Pymatgen library in Python. The NLO dataset includes 224 chemically diverse compounds collected from literature, while the perovskite dataset consists of 643 ABX_3 materials with X being O, S, Se, F, Cl, Br, or I, sourced from the OQMD database. Several machine learning models were employed, with the Random Forest Regressor (RFR) consistently achieving the best performance. For the ABX_3 dataset, RFR yielded a mean absolute error (MAE) of 0.4941 eV using the top 10 features and 0.5044 eV with all features; for the NLO dataset, the model achieved 0.3309 eV and 0.3238 eV, respectively. Dimensionality reduction through Principal Component Analysis (PCA) did not enhance the results compared to manually selected or full feature sets. These findings highlight the significance of proper feature engineering, especially when relying on descriptors generated by automated tools like Pymatgen, to improve the efficiency and reliability of machine learning applications in materials discovery.

Key words: Pymatgen library, functional materials, Machine learning, regression models, Principal component analysis.