IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 28, Issue 1, Ser. 1 (Jan. – Feb. 2026), PP 96-106
www.iosrjournals.org

Physics-Guided Nonlinear Machine Learning for Accurate Band Gap Prediction from Composition-Derived Electronic Descriptors

Priti Goyal[1]* and Shaivi Goyal[2]
[1]*Acharya Narendra Dev College, Department of Physics, University of Delhi, INDIA*
[2]*Bharati Vidyapeeth's College of Engineering, GGSIPU, Delhi, INDIA*

Abstract: *The logical design of functional oxides for applications in energy and optoelectronics requires a knowledge of the compositional parameters that controls optical band gaps in oxide materials. Although band gaps can be accurately predicted using machine learning techniques, many of the existing models rely on intricate structural descriptors or black box architecture that restrict their physical interpretability. In this work, we present a physics guided machine learning framework that uses a small set of physically justified descriptors to directly relate optical band gaps of oxide materials to chemical composition. Composition based quantities such as average d-electron count, iconicity, metallic fraction, mean electronegativity, electronegativity difference, average atomic number and valence electron count are employed to capture key aspects of chemical bonding and electronic structure. To assess predictive performance and robustness linear and nonlinear regression models are trained and evaluated using cross-validation. Analysis of model coefficients and feature importance shows distinct, physically meaningful trends linking compositional parameters to band gap behavior in oxides. From the optimized Random Forest model, $R^2$of 0.7902, MAE of 0.461eV and RMSE of 0.694 eV were obtained. Correlation analysisshowsthat band gap is moderately correlated with the average d electron count (-0.45) and metallic fraction (-0.36).These findings provide a conceptually effective and transparent approach to materials screening and understanding.This shows that interpretable and composition only machine learning models can provide valuable insights into band gap physics even in limited data regimes.*

Keywords: *Physics guided machine learning, Band gap, Linear and nonlinear regression models, Random Forest, XGboost*

I. Introduction

Band gap formation in crystalline inorganic oxides results from complex interactions between lattice structure, bonding properties, and electrical configuration. Traditional theoretical methods are computationally demanding for extensive screening. As a result, data driven machine learning frameworks have become powerful tools for facilitating interpretable structure- property analysis and accelerating band gap prediction.

Band gaps in oxides result from the interaction of orbital hybridization, metal-oxygen bonding and electronic structure, with contributions from atomic number, valence electron configuration and electronegativity difference [1, 2]. Density functional theory (DFT) and other first-principles electronic structure techniques have provided deep insight into these mechanism and are indispensable for quantitative analysis [3]. However, the scalability of DFT for large scale materials screening is limited as it often requires computationally expensive methods beyond the standard exchange- correlation functional for accurate band gap prediction [4].

Machine learning (ML) has become powerful complementary approach for predicting material properties, including electronic and optical band gaps[5]. A vast array of ML models have been developed, ranging from kernel methods and tree- based algorithms to deep neural networks and graph-based architectures [6]. Many of these studies use extensive sets of engineered descriptors or precise structural information to achieve high predictive accuracy. Despite their effectiveness, these methods frequently function as black boxes and provide little information about physical factors controlling the target property.

Interpretability and physical understanding are just as crucial to condensed matter physics and materials research as predictive performance. Models that mask the connection between input descriptors and physical mechanisms provide limited guidance for materials design and hypothesis generation. This motivates the development of physics- guided machine learning techniques that emphasize interpretability above algorithm complexity and specifically include physically meaningful descriptors.

Several studies have successfully applied machine learning to predict electronic and optical properties of materials. For example, machine learning model has been employed to predict band gap of crystalline materials using chemical composition [7]. ML regression methods such as support vector regression, random forests, and gradient-boosted trees have been used to predict band gaps in transition metal compounds and doped semiconductors with good accuracy [8, 9]. Graph neural networks have also been developed to capture

local chemical environments for optical band gap prediction in doped carbon nitrides [10]. Larger frameworks employing elemental and structural descriptors have been applied to complex oxides and perovskites to model band gaps and other material properties [11, 12]. Previousresearch demonstrate the feasibility of data-driven prediction of band gap values by using complex structural descriptors or deep learning architectures to maximize predictive accuracy.This motivates the present study to focus oninterpretable, composition-based descriptors to uncover fundamental compositional trends in oxide band gaps.

To capture key aspects of bonding and electronic structure without depending on crystallographic information, we restrict the input features to a limited number of physically motivated composition-based descriptors, such as mean electronegativity, electronegativity difference, average atomic number and valence electron count. Linear and nonlinear regression models are employed to evaluate the extent to which these descriptors identify band gap patterns in a wide range of oxides.

The present work emphasizes a physics-guided and interpretable approach, concentrating solely on oxide materials and employing a small set of composition-based descriptors that have distinct physical meanings. In addition to predicting optical band gaps, this study seeks to clarify the underlying compositional trends governing electronic structure in oxides by restricting the model inputs to chemically motivated parameters. This method facilitates rapid screening without relying on structural information or first principles calculations and provides meaningful insight even in limited data regimes.

II. Methodology

Dataset and features

Dataset of crystalline oxides was obtained from the Materials Project [13]. It offers band gaps of structurally confirmed inorganic oxides, estimated by density functional theory (DFT) and experimentally reported. Noble gases and outliers were excluded in the dataset. Initially, the raw dataset contained only the chemical formula and the associated band gap values. Materials with band gap values ≤ 0eV were eliminated. The band gap range was restricted to 0-6eV to limit the analysis to realistic semiconducting and insulating compounds. In addition, only oxide materials were retained based on chemical composition filtering.Only stoichiometric oxides containing oxygen as the anionic component were retained, ensuring chemical consistency and avoiding mixed anion systems such as nitrides, sulfides, or halides.

The chemical formula was used to compute physically interpretable compositional descriptors without using external database searches or structural information. For a compound consisting of i elements with atomic fraction x_i, the average atomic number (Z_{avg}), average atomic mass(M_{avg}), average electronegativity (χ_{avg}), and average valence electron count (V_{avg}) were calculated as weighted averages over the constituent elements. The difference between maximum and minimum Pauling electronegativity values was defined as electronegativity difference ($\Delta\chi$). These descriptors identify important chemical trends pertaining to bonding, ionicity and electronic structure while being low dimensional and interpretable.

Python with locally available elemental properties was used for all descriptor computations. The need for external programming interfaces was eliminated and it ensured complete reproducibility.

2.1 Feature Engineering

Descriptors were obtained using weighted average schemes from the elemental properties and stoichiometric composition.

- Average d-electron count (d electron avg)
- Average atomic number (Z avg)
- Average electronegativity (χ avg)
- Ionicity index
- Fraction of metals / nonmetals
- Fraction of elements like Cu, Mn, P, O
- Structural density
- Unit cell volume
- Polarizability average
- O/metal ratio
- Average valence electron count

These features does not require crystal optimization and can encode electronic structure tendencies, bonding character and compositional effects.

2.2 Model Development

Regression models with increasing order of complexity were usedfor prediction of band gap. Ridge regression, a linear baseline model, was selected to determine if changes in crystalline oxides can be represented

by linear compositional trends. Ensemble tree based methods like Random Forest [14]and Extreme Gradient Boosting (XGBoost) [15] were applied to account for nonlinear interactions among electronic and compositional descriptors. More complex deep architectures were intentionally avoided to prevent overfitting and to preserve physical transparency.

Standard regression metrics were applied for quantitative assessment of model performance. The coefficient of determination (R^2), Mean absolute error (MAE) and root mean square error (RMSE) were calculated using following formula

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}$$

$$MAE = \frac{1}{n}\sum|y_i - \hat{y}_i|$$

$$RMSE = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2}$$

Where y_i represents the actual band gap, $\hat{y}_i$ the predicted value, and $\bar{y}_i$the mean of observed band gaps. All these metrics together provide complementary evaluation of model accuracy, robustness and error distribution.

III. Results and Discussion

3.1 Dataset Characteristics

The compositional bias was reduced and generalizability improved by taking the final dataset containing N oxide compositions from transition metal oxides, main group oxides, rare-earth oxides and mixed metal oxides. This oxide only dataset provided a chemically coherent and electronically diverse platform for studying the structure-property relationships governing band gap formation in crystalline oxides.

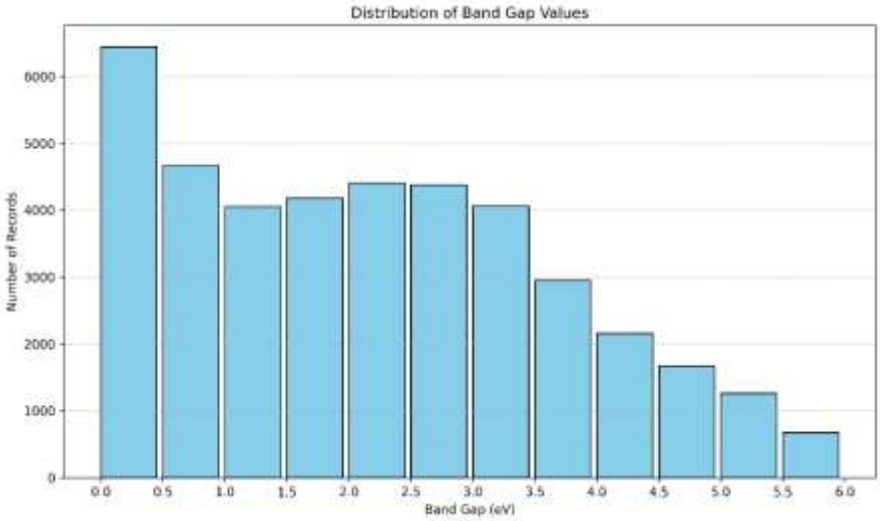

Fig.1 Distribution of Band Gap Values

The fig.1 shows the histogram of band gap in the range of 0-6 eV. As can be seen from the figure, there is high concentration of compounds in low band gap region (0-1 eV). With the increase in band gap, the frequency of materials decreases. Small percentage of compounds are observed with band gaps above 4.5 eV. This results in a right-skewed distribution. It is specified by a long tail that extends toward higher energy values. Statistically, the dataset contains large number of small-gap materials, fewer mid-gap semiconductors and very small proportion of wide-gap insulating oxides.

As can be observed from the figure, the distribution is non-Gaussian.There is an increase in variance in the higher band gap region and the mean has surpassed the median as a result of the long right tail. This disparity suggests that while predictive models may perform well in the low- to mid- gap range but their accuracy may suffer in high-gap materials due to data sparsity.

A set of compositional descriptors like Electronic structure descriptors (e.g., average d-electron count, valence electron count), Atomic descriptors (e.g. average atomic number, mean atomic weight), Bonding descriptors (e.g., electronegativity difference, ionicity), Structural proxies (e.g., density, atomic volume), Compositional fractions (e.g., oxygen fraction, metal fraction, element-specific fractions) were generated to describe each oxide composition. All descriptors were computed from elemental properties using composition-weighted averaging. Features were standardized and screened for missing values to ensure numerical stability prior to model training.

3.2 Descriptor Correlation Analysis

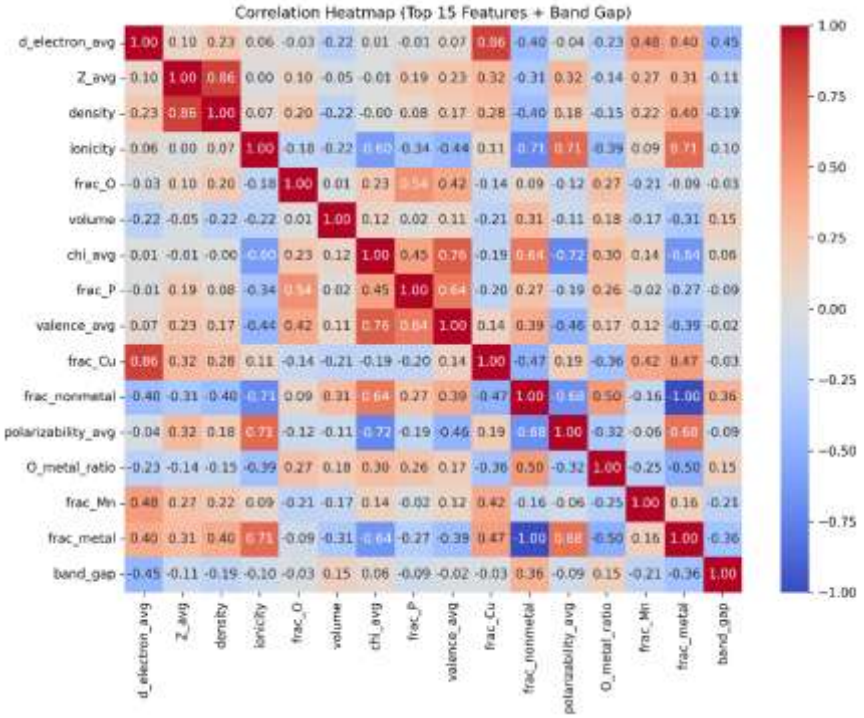

Figure 2. Pearson correlation heatmap of the top 15 descriptors along with band gap for crystalline oxide compounds.

Inter-feature relationship and their direct linear association with the target property are revealed by the heatmap. The average d-electron count and the band gap exhibits a moderate negative correlation (r = -0.45), confirming that an increased d-electron occupancy of transition metals generally leads to reduced band separation. This finding is consistent with enhanced metal-oxygen hybridization and increased electronic delocalization in d-rich oxides.

Additionally, a weak negative correlation is also observed between band gap and average atomic number (Zavg), suggesting that narrower band gaps are typically produced by heavier transition metals. Descriptors like polarizability and ionicity, on the other hand, exhibit a moderately positive connection with band gap, which is consistent with wider gaps in more ionic and less covalent oxides.

Strong inter-feature correlations are evident among certain structural and compositional descriptors. Among few structural and compositional descriptors, strong inter-feature correlations are evident.Zavg and density show strong positive correlation, χavg and frac P display strong coupling and polarizability avg and frac nonmetal exhibit negative correlation. The strong negative correlations is observed between some descriptors like metal fraction and non-metal fraction (-1.00). It indicates that certain features contain overlapping information. This result supports the need for feature ranking and dimensionality reduction to remove the redundancy and to improve model reliability. As can be seen from fig, no descriptor exhibits a strong correlation (r> 0.5) with band gap, indicating that the prediction problem is not linear and requires nonlinear ensemble learning approaches like Random Forest and XGBoost.

3.3 Model Development and Performance Evaluation

3.3.1 Predicted vs Actual Analysis

Figure 3 showsActual versus predicted band gap values for the three models

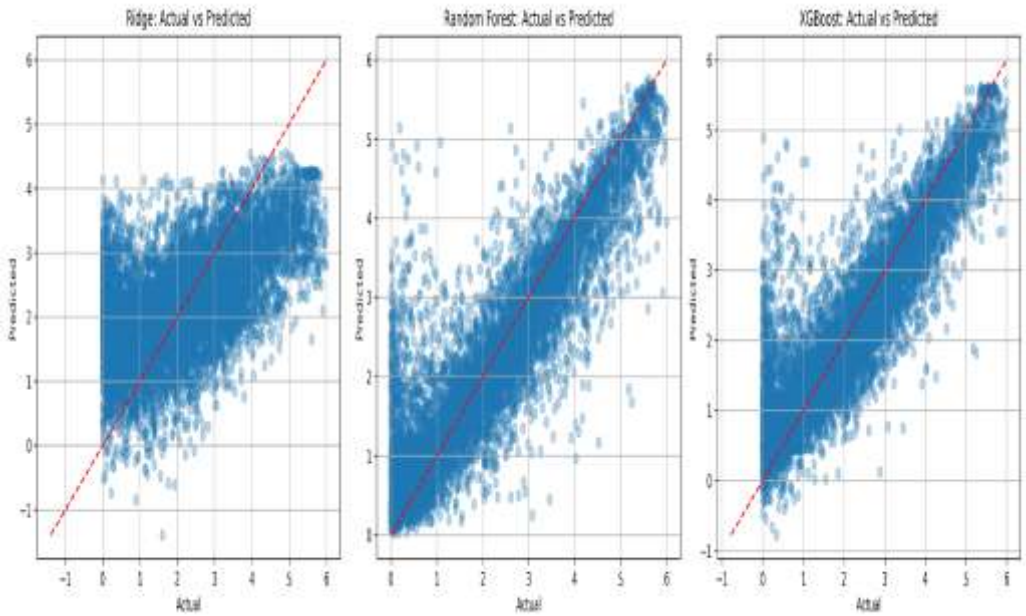

Figure 3. Actual versus predicted band gap values for (a) Ridge Regression, (b) Random Forest, and (c) XGBoost models. The red dashed line represents perfect agreement (y = x).

The Ridge Regression model fig 3(a) shows significant dispersion around the parity line, with predictions grouped together in a narrow band regardless of the magnitude of the band gap. Strong underfitting, in which linear model is unable to account for the nonlinear compositional interactions governing oxide band gaps, is indicated by this compression effect. Large deviations observed especially for high band gap materials confirms the limited ability of linear regression to model metal-oxygen electronic hybridization.

The Random Forest model in fig 3(b) exhibits strong alignment with the parity line throughout the full band gap range. Predictions closely resemble the ideal trend, with fairly even distribution along the diagonal. The improved agreement suggests the ability of ensemble tree method to capture nonlinear descriptor interactions and threshold behaviors. For wide band gap oxides, minor underestimation is observed, which implies there is increased prediction uncertainty in wide-gap regimes.

XGBoost, fig 3(c) exhibits tight clustering along the parity line. Its performance is similar to Random Forest. In the intermediate band gap ranges, there is slight increase in scatter compared to Random Forest. The model effectively captures nonlinear patterns, but in densely populated compositional regions, it shows low variance.

Thus for the ensemble models, a slight increase in dispersion is seen at higher band gap values, indicating heteroscedastic uncertainty, where orbital localization and dielectric screening effects become more pronounced. Residuals are symmetrically distributed around the parity line. This implies there is no systematic over or under estimation. Hence imbalance in parity alignment between the linear and ensemble models offers visual proof thatnonlinear modeling capable of capturing coupled electronic and compositional interactions is required for accurate oxide band gap prediction.

3.3.2 Quantitative Model Performance Comparison

Table 1 summarizes the quantitative performance of the three models.

Model	R^2	MAE(eV)	RMSE(eV)
Random Forest	0.7902	0.461	0.694
XG Boost	0.7872	0.4882	0.699
Ridge Regression	0.3462	0.9891	1.2251

Table 1. Performance metrics for regression models.

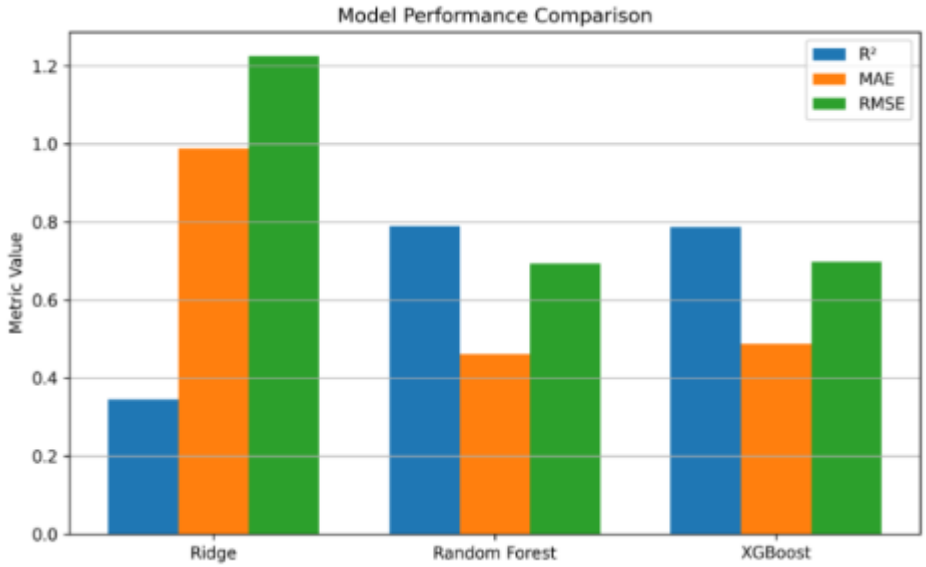

Fig. 4 Model performance comparison

The Random Forest Model attains the maximum predictive accuracy with R^2 of 0.7902, MAE of 0.461 and RMSE of 0.694 eV. These values are seen to be comparable to XGBoost.The close performance of the two ensemble models suggests that in oxide materials, the formation of band baps is by nonlinear interactions.With R^2 of 0.3462 and RMSE of 1.2251 eV, Ridge Regression exhibits poor performance. The notable performance gap between tree based and linear models demonstrates that for oxide band gap prediction linear combinations of compositional descriptors are insufficient.

3.4 Feature Importance and Descriptor Dominance

Figure 5 shows the top twenty most influential descriptors suggested by the Random Forest model for oxide band gap prediction. A highly skewed importance distribution is observed. As can be seen from figure, the average d-electron count (d electron avg) is dominating all other features.

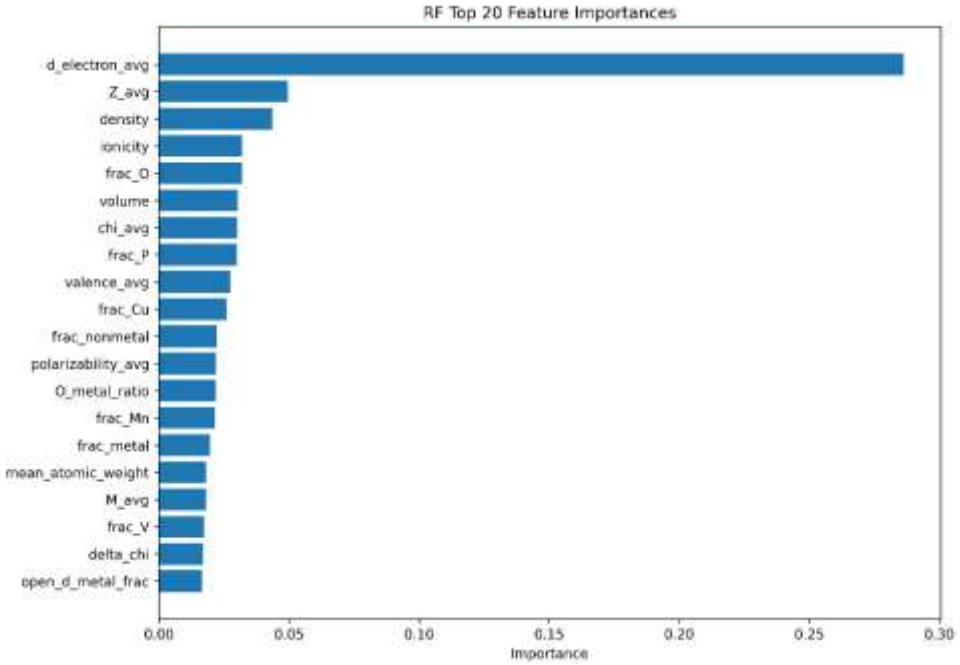

Figure 5. Top 20 feature importance rankings obtained from Random Forest model.

As observed from the dominance of d electron avg, the main factor that influences the band gap behavior is the electronic configuration of transition metals. They account for approximately 28-30% of the total model importance. They directly participate in metal-oxygen hybridization and controls the position of

conduction and valence band. Zavg, density, ionicity, oxygen fraction, volume and χavg are secondary contributors. They collectively represent atomic packing effects, band polarity, compositional stoichiometry and structural-electronic coupling. Metal/nonmetal ratios and specific fractions (Cu, Mn, V) show substantial but less impact, indicating compositional fine tuning affects band gap only after the primary electronic configuration is taken into consideration. The cumulative contribution exceeds to ~ 50% of the predictive power when combined with next four descriptors and the top 8-10 descriptors collectively account for ~ 80-85% of the model importance. Beyond these, additional features contribute to only marginal gains. This behavior implies that the predictive structure of the model is driven by compact set of physically meaningful electronic descriptors rather than diffuse combination of weak predictors.

3.5 Feature Reduction and Model Stability

A reduced feature model was evaluated to examine redundancy as a result of sharp drop in feature relevance that occurred after the top ten descriptors. Only the top ten descriptors were used to train the Random Forest model, which resulted in negligible decrease in performance ($\Delta R^2 < 0.01$), indicating that lower ranked features have a limited impact on prediction accuracy. This result illustrates the robustness of electronic structure driven learning, the lack of considerable overfitting and the descriptor redundancy among lower-ranked features. The interpretability and generalizability of the model are improved by such compact descriptor dependence.

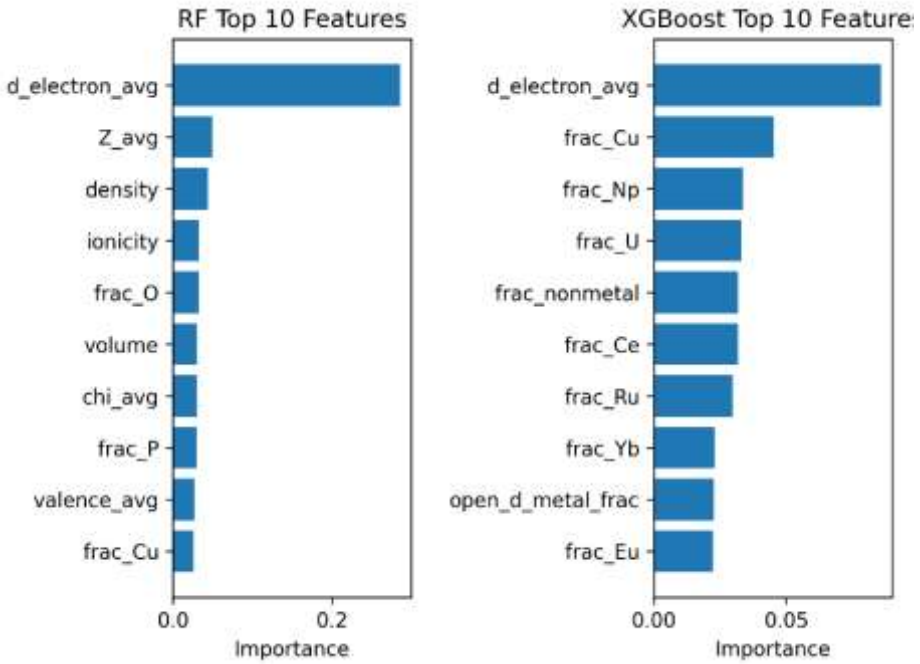

Figure 6. Top 10 feature importance rankings for (a) Random Forest and (b) XGBoost models in oxide band gap prediction.

Fig 6 shows that the average d-electron count (d electron avg) is the most important predictor of oxide band gap in both models. This finding is of physical importance as the transition metal d-orbital occupancy directly affects metal-oxygen hybridization and band separation. In general, electronic delocalization is improved by higher d-electron density which results in smaller band gaps. The average atomic number, density, ionicity, oxygen proportion and average electronegativity are other significant parameters for Random Forest model. These features collectively capture compositional factors, bonding polarity and atomic packing that affect electronic bandwidth and localization. Similarly XGBoost prioritizes d-electron count and gives emphasis on specific elemental fractions (e.g. Cu, U, Ce, Ru, and Yb) and nonmetal fraction. This suggests that gradient boost captures localized compositional influences and element-specific threshold effects. Thus both models identify similar dominant descriptors. Random Forest gives emphasis on global electronic trends whereas XGBoost show finer element specific contributions.

Top N Features	20	15	10	5	3	1
R2	0.7495	0.7376	0.7171	0.6376	0.4796	0.3163
MAE(eV)	0.5081	0.5244	0.5498	0.6362	0.7961	0.9981
RMSE(eV)	0.7583	0.7761	0.8058	0.9120	1.0930	1.2528
NRMSE	0.1264	0.1294	0.1344	0.1521	0.1822	0.2089

Table 2. Model performance as a function of the number of top-ranked physics informed descriptors.

A progressive feature reduction analysis was performed to evaluate the robustness of the model and determine the minimum number of descriptors required for reliable prediction. Starting from the full ranked feature set, the model was trained using only the top N (= 20, 15, 10, 5, 3, 1) most important features. The resulting performance metrics are summarized in Table 2.

To evaluate the robustness of the model and determine the minimum number of descriptors required for reliable prediction, a progressive feature reduction analysis was performed. The results are summarized in Table 2. The model was trained using only the top N (= 20, 15, 10, 5, 3, 1) important features. The results show deterioration in predictive performance as the number of retained features decreases. As can be seen from the table. Using top 20 features give R^2 as 0.7495, MAE as 0.5081 eV and RMSE of 0.7583 eV. Reducing the feature set to15 and then to 10 results in only minor performance loss. However, elimination of more features degrades the performance significantly. As can be seen from the table, with the reduction in features from five and then to three, R^2 further decreases. A single feature model performs poorly (R^2 = 0.3163, RMSE = 1.2528 eV). The normalized RMSE follows the same trend, confirming loss of predictive precision.

The analysis of successive feature reduction shows that even though the model maintains a significant predictive ability with ten descriptors, aggressive reduction considerably weakens the performance. This confirms that rather than being controlled by a single dominant feature, the desired property results from multivariate interactions across electronic, structural and compositional descriptors. An optimal balance between dimensionality and accuracy is achieved using the top-ranked subset of features.

3.6 Predicted vs Actual Parity Analysis

Predicted vs real parity plots with statistical error bands were examined for all regression models in order to assess predictive fidelity and show model agreement across the entire band gap range.

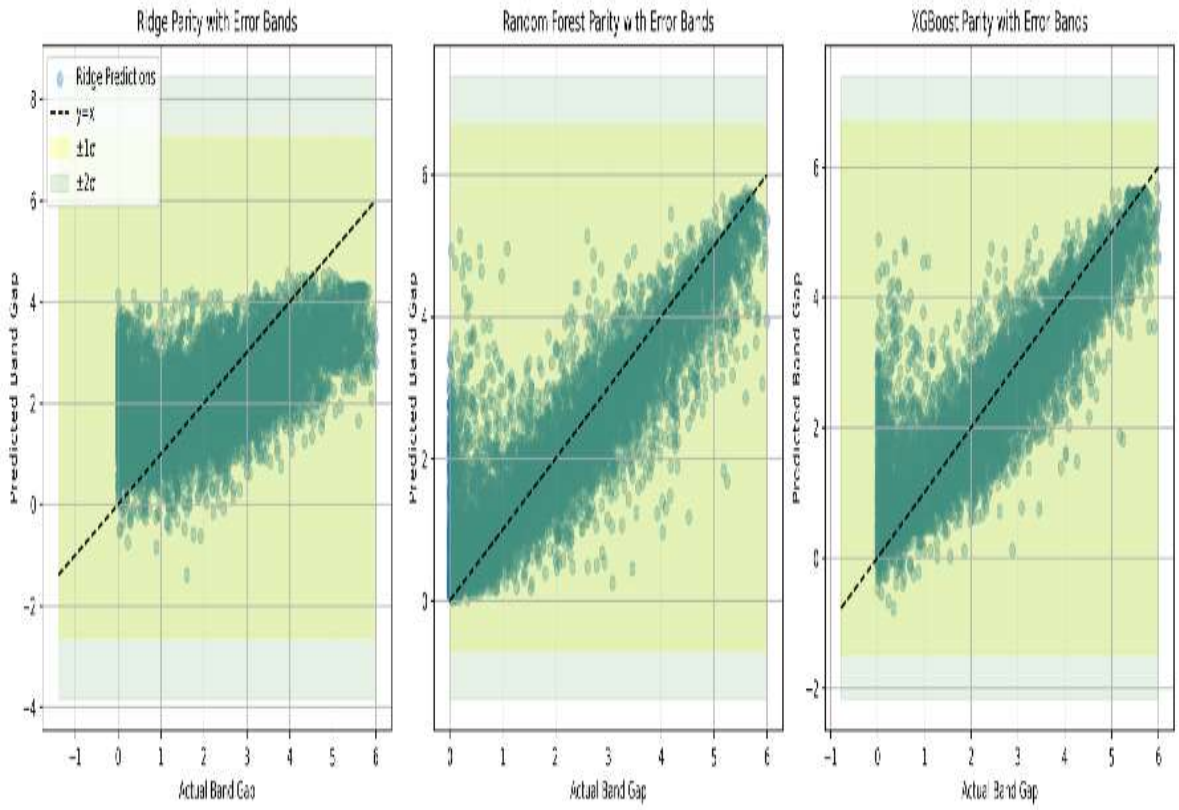

Figure 7. Parity plots with ±1σ and ±2σ error bands for Ridge, Random Forest, and XGBoost models, comparing predicted and actual band gap values.

The parity plots (fig 7) show the comparative predictive performance of the three models against the ideal y=x agreement line. At higher band gap values, the Ridge Model shows significant dispersion and deviation suggesting underfitting and limited ability to capture nonlinear electronic effects. Random Forest shows markedly improved alignment with the parity line and reduced error spread, on the other hand, the XGBoost further reduces the prediction error, particularly in the mid- to high- band gap regime. Hence, oxide band gap prediction is intrinsically nonlinear as evidenced by the progressively narrower dispersion from Ridge to tree based ensemble models.

3.7 Residual Diagnostics

The residuals (Predicted-Actual band gap) were examined as a function of dominant descriptors in order to further assess the robustness of model and identify systematic biases.

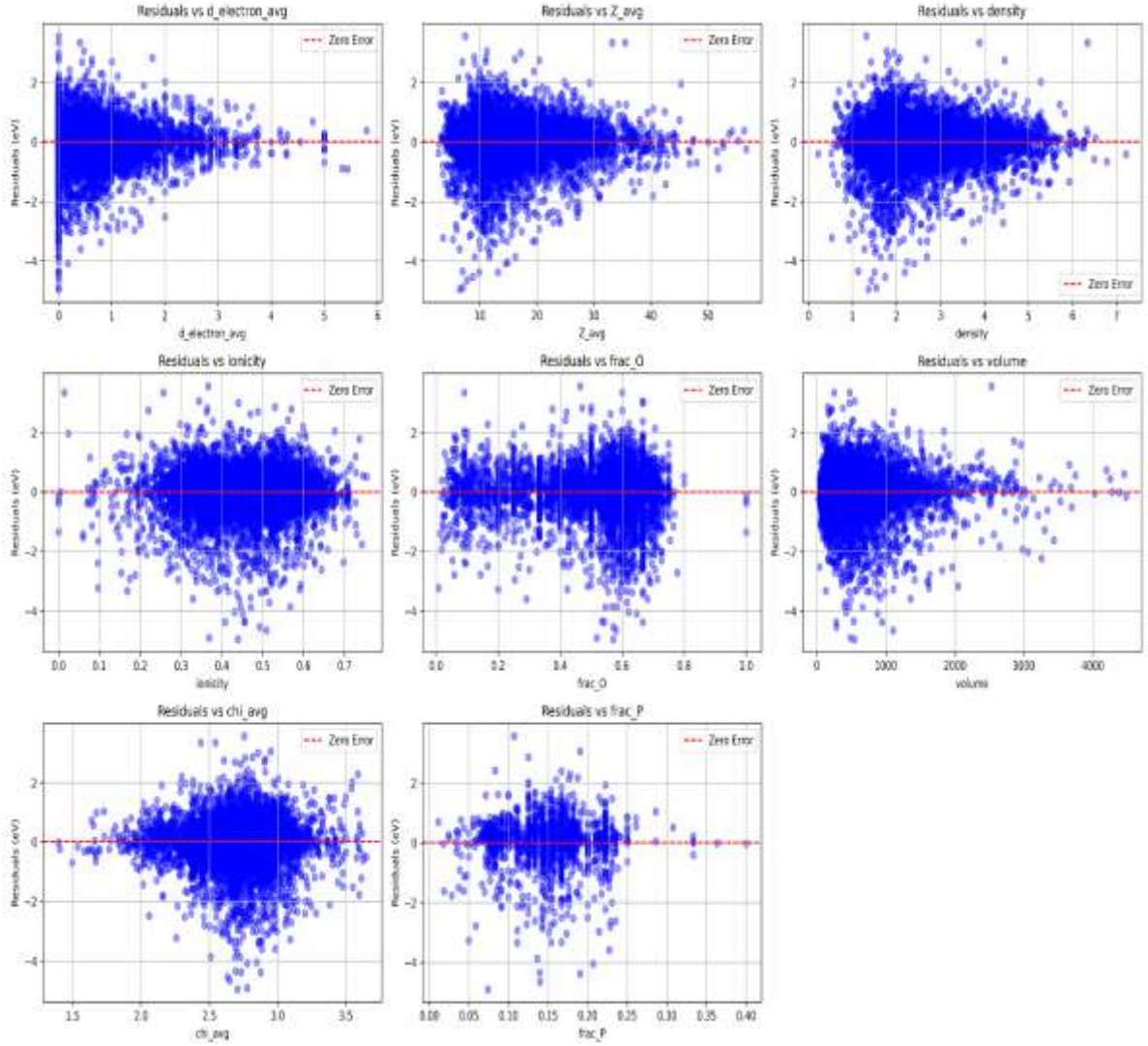

Figure 8 – Residual Distribution across Descriptors

As can be seen from fig 8, across all the descriptors, residuals are distributed symmetrically around zero. It indicates the absence of systematic overestimation or underestimation trends. The distribution is uniform across the descriptor range, which confirms that the Random Forest does not exhibit strong heteroscedastic behavior.

For the d electron avg and Zavg at lower descriptor values, where larger error variance appears, a slight funnel shaped distribution is observed. This is expected due to increased electronic diversity in early transition metal oxides and mixed valance systems. However, monotonic residual trends are not evident confirming that the model adequately captures nonlinear electronic structure effects. Residuals vs density and volume show no systematic slope, which indicates that structural proxies are properly incorporated without inducing bias. Similarly, ionicity, oxygen fraction and average electronegativity show centered residual distribution.

Importantly, no descriptor exhibits a consistent positive or negative residual gradient confirming the absence of descriptor dependent bias, good generalization across compositional space and stability of nonlinear mapping.

The residual diagnostic show that the model is robust, physically consistent and statistically unbiased. Strong generalization over compositional space is confirmed by the lack of descriptor-dependent trends. Reliable nonlinear mapping free from systematic bias is indicated by the stable error distribution. In general, predictive deviations arise from intrinsic material complexities rather than model deficiency.

IV. Conclusion

In this study, machine learning framework has been developed for accurate prediction of band gaps in crystalline inorganic oxides using physically interpretable compositional descriptors. Chemical homogeneity was ensured by restricting the dataset exclusively to oxide compounds. It enabled meaningful interpretation of structure property relationships governing electronic behavior. Random Forest and XGBoost, significantly outperformed linear regression, confirming that oxide band gap formation is governed by complex nonlinear electronic structure effects rather than linear trends.

Moderate linear relationships between band gap and key electronic descriptors was revealed by correlation analysis. Feature importance analysis showed a pronounced dominance of average d-electron count, highlighting the key role of transition metal electronic configuration in determining band gap. Contributions from atomic number, ionicity, density and oxygen fraction further highlight the connection between electronic occupancy, bonding polarity and lattice packing.Robustness and interpretability of models was supported by cumulative performance analysis. It showed that a small and compact subset of descriptors can predict more accurately.Residual diagnostics and Parity plots confirmed the absence of systematic bias across electronic and compositional regimes.

This work establishes that oxide band gap prediction can be achieved with high accuracy using physically grounded and interpretable descriptor set. The strong alignment between machine learning, importance rankings and well-known sold-state physics principles enhances the mechanistic credibility of the model. The proposed framework provides a reliable tool for accelerated screening and rational design of functional oxide materials for electronic, optical and energy applications.

Acknowledgements

The author would like to thank the principal, Acharya Narendra Dev College, University of Delhi, for the infrastructural assistance and direction.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Funding sources

The author declares that no funds, grants, or other support were received during the preparation of this manuscript.

References

[1]. Jin Suntivich, W. Hong, Yueh-Lin Lee, J. Rondinelli, Wanli Yang, J. Goodenough, B. Dabrowski, J. Freeland, Y. Shao-horn Estimating Hybridization of Transition Metal and Oxygen States in Perovskites from O K-edge X-ray Absorption Spectroscopy J. Phys. Chem. C 2014, 118, 4, 1856–1863 https://doi.org/10.1021/jp410644j

[2]. Samadhan Kapse, Maria Voccia, Francesc Viñes, Francesc Illas, Chemical bonding and electronic properties along Group 13 metal oxides, Journal of Molecular Modeling 2024 30:161https://doi.org/10.1007/s00894-024-05957-6

[3]. Jiang, H., & Zhang, M. (2020). Density-functional theory methods for electronic band structure properties of materials. SCIENTIA SINICA Chimica.Volume 50, Issue 10: 1344 - 1362 2020 https://doi.org/10.1360/ssc-2020-0142

[4]. Carl Belle, V. Aksakalli, S. Russo, A machine learning platform for the discovery of materials, Journal of Cheminformatics, 2021 13:42 https://doi.org/10.1186/s13321-021-00518-y

[5]. Zhuo, Y., Mansouri Tehrani, A., & Brgoch, J. Predicting the Band Gaps of Inorganic Solids by Machine Learning. The journal of physical chemistry letters, 2018, 9 7, 1668-1673, https://doi.org/10.1021/acs.jpclett.8b00124

[6]. Satveer Kaur, Evaluating Machine Learning Models in Various Domains: An Extensive Analysis and Comparison, Journal of Software Engineering and Simulation 10 (11) 2024 60-63, https://doi.org/10.35629/3795-10116063

[7]. Andrew Ma, Owen Dugan, and Marin Soljaci, Predicting band gap from chemical composition: A simple learned model for a material property with atypical statistics,cond-mat.mtrl-sci ArXiv 2025https://doi.org/10.48550/arXiv.2501.02932

[8]. Machine learning prediction of band gap in transition metal trihalides. Journal of Nepal Physical Society 2023. 9(2):34-41, DOI:10.3126/jnphyssoc.v9i2.62287

[9]. Yuqi Tang, Haiyuan Chen, Jianwei Wang and Xiaobin Niu Machine learning-aided band gap prediction of semiconductors with low concentration doping. Phys. Chem. Chem. Phys. 2023, 25, 18086-18094https://doi.org/10.1039/D3CP02431H

[10]. Chen Chen, Enze Xu, Defu Yang, Haibing Yin, Tao Wei, Hanning Chen, Yong Wei, Minghan, Chemical Environment Adaptive Learning for Optical Band Gap Prediction of Doped g-C_3N_4 Nano sheets. ArXiv (2023). https://doi.org/10.48550/arXiv.2302.09539

[11]. Chan Gao, Xiaoyong Yang, Ming Jiang, Lixin Chen, Zhiwen Chen and Chandra Veer Singh, Machine learning-enabled band gap prediction of TMD alloys.Phys. Chem. Chem. Phys. 2022, 24, 4653-4665https://doi.org/10.1039/D1CP05847A
[12]. Wen Yao, Wanli Jia, Ruofan Shen, Jiayao Wang, Lin Zhang, Xinmei Wang, ML prediction of band gap and formation energy in 2D metal oxides. Physica B (2025) Condensed Matter Volume 717, 2025, 417821https://doi.org/10.1016/j.physb.2025.417821
[13]. Dunn, A., Wang, Q., Ganose, A., Dopp, D., & Jain, A. Matbench mp gap dataset, ColabFit/Materials Project. DOI:10.60732/fb4d895d
[14]. L. Breiman, Mach. Learn., 2001, 45, 5-32, doi: 10.1023/A: 1010933404324.
[15]. T. Chen and C. Guestrin, Proceedings of the 22nd acmsigkdd international conference on knowledge discovery and data mining 2016, 785-794, doi: 10.1145/2939672.2939785.