DEVELOPMENT OF ARTIFICIAL INTELLIGENCE MODELS FOR PREDICTING THE PROPERTIES OF BINARY AND TERNARY STRUCTURES.

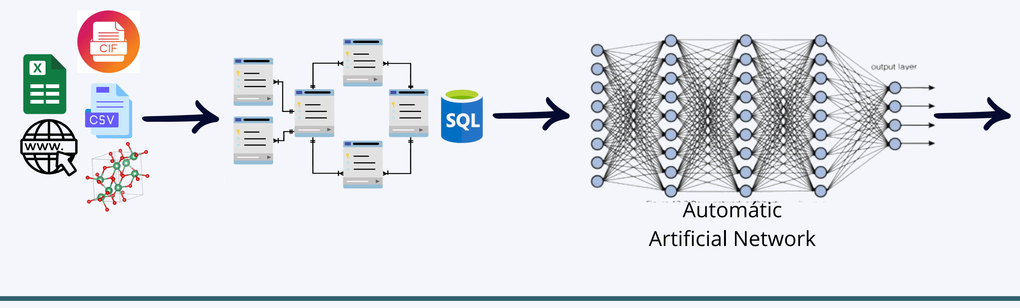
M. GABRIEL GARCIA ZAMBRANO, DR. ARTURO ISAÍAS MARTÍNEZ ENRÍQUEZ, DR. DANIEL OLGUÍN MELO, DR. FELIPE MONDACA ESPINOZA.

ABSTRACT

Knowing the band gap in crystalline structures is essential for designing materials with tailored electronic and optical properties. This study proposes an artificial intelligence framework based on a SQL database of approximately 2,700 binary and ternary structures, encompassing 273 theoretical, experimental, and DFT HSE06–computed variables. XGBoost and artificial neural network (ANN) models were optimized via Grid Search and cross-validation. Although XGBoost exhibited excellent training performance, its R^2 fell to ≈0.44 in validation. In contrast, the ANN achieved $R^2 = 0.96$ on the training set and $R^2 = 0.92$ on the validation set, faithfully reproducing the Gaussian distribution of experimental values. This methodology reduces the estimation time from days (with DFT) to minutes. It establishes a scalable, reproducible protocol adaptable to predicting other electronic and optical properties, thereby accelerating the rational design of materials.

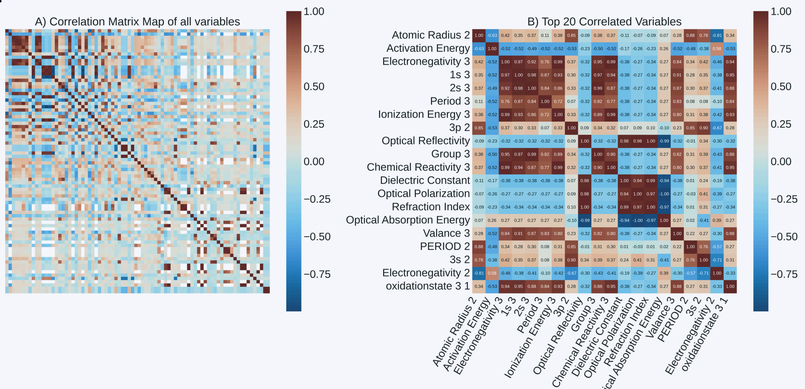
Keywords: Crystalline structures, band gap Prediction, deep learning (ANN), machine learning (XGBoost), DFT HSE06.

METHODOLOGY

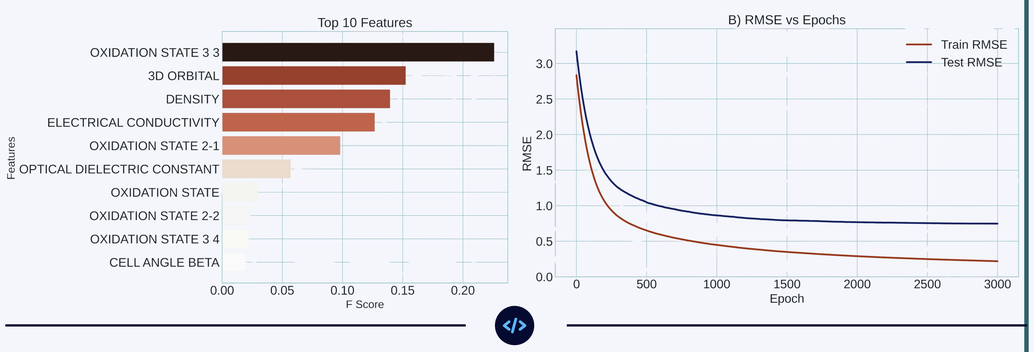


DISCUSSION

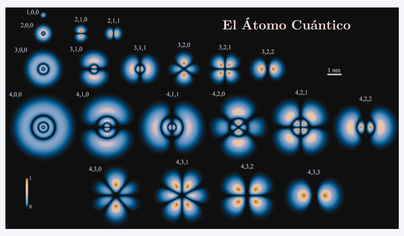
The correlation matrix displays, in a heatmap, the degree of linear association among more than 270 variables, using reddish hues for positive correlations and bluish hues for negative ones. From this overview, the 20 variables with the highest correlations—including atomic radius, activation energy, and electronegativity—are selected to focus the analysis on the most relevant attributes.

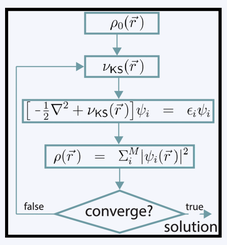


The first figure shows the ten variables with the highest importance scores (F-Scores), allowing identification of which attributes contribute most to the model's predictive power. The second figure displays the evolution of the root mean squared error (RMSE) on both the training and validation sets across epochs, demonstrating the model's convergence and providing a criterion for selecting the optimal stopping point to avoid overfitting.



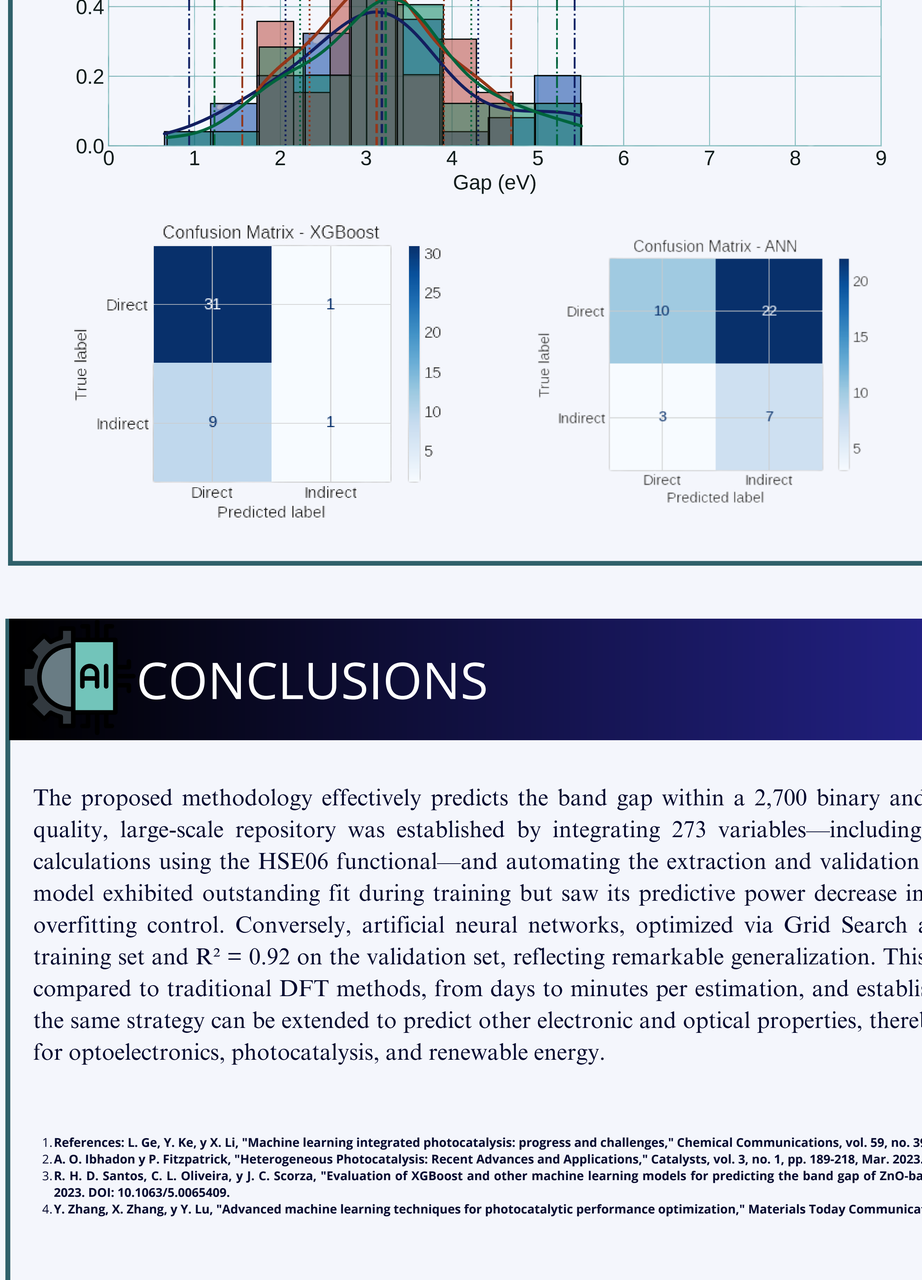
In the first figure, atomic orbitals for various quantum numbers (n, ℓ , m) are displayed, representing the electronic probability density of the stationary states of the Schrödinger equation for the atom. The second figure shows the flowchart of the Kohn–Sham self-consistent cycle in density functional theory, where from an initial density $\rho_0(r)$ the Kohn–Sham potential is computed, the Kohn–Sham equations are solved to obtain wavefunctions ψ_i and eigenvalues ε_i, the density is updated, and the process is repeated until convergence.

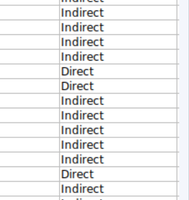




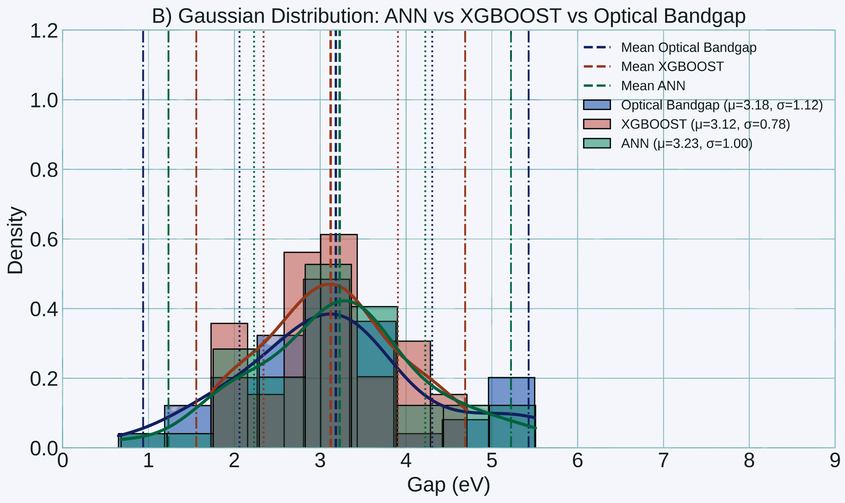
Converting thousands of crystallographic files (.cif) into a data repository enriched with physical, chemical, structural, and optical properties to train neural networks; analogous to the self-consistent DFT cycle, which achieves the minimum energy through iterative convergence, these networks seek to minimize predictive error.

RESULTS





Structure	Reported Gap (eV)	XGBOOST (eV)	ANN (eV)	Type of Prediction
$ZnCr_2O_4$	1.8	1.73	1.8	Validation
$ZnMn_2O_4$	1.86	1.97	1.97	Validation
ZnS	3.67	2.59	3.49	Validation
$ZnSnO_3$	4.21	2.65	3.7	Validation
ZnO	3.37	3.24	3.75	Validation
$ZnFe_2O_4$	2	2.01	2.22	Validation
$Mo_3Zn_2O_8$	3.14	2.65	2.1	Validation
Zn2GeO4	4.68	4.71	4.6	Validation
Zn2SnO4	2.5	2.43	2.54	Validation
ZnO3Si	2.87	1.91	2.85	Validation



STRUCTURE	REAL_LABEL	XGB_PRED	ANN_PRED
LaTiO3	Direct	Direct	Indirect
NiO	Direct	Direct	Indirect
AgNiO2	Direct	Direct	Indirect
ZnGeO3	Direct	Direct	Indirect
Zn3P2	Direct	Direct	Indirect
Zn3	Direct	Direct	Direct
V3ZnO8	Direct	Direct	Direct
Fe2ZnO4	Direct	Direct	Indirect
Sr3Ti2O7	Direct	Direct	Indirect
Mo3Zn2O8	Direct	Direct	Indirect
MnZnO3	Direct	Direct	Indirect
BaZnO2	Direct	Direct	Indirect
Mn2ZnO4	Direct	Direct	Direct
ZnO3Se	Direct	Direct	Direct
ZnSe	Direct	Direct	Indirect
Bi2MoO6	Direct	Direct	Indirect
Bi2MoO6	Direct	Direct	Indirect
ZnO3Si	Direct	Direct	Indirect
Mo3Zn2O8	Direct	Direct	Indirect
ZnO	Direct	Direct	Direct
TaNO	Direct	Direct	Indirect
ZnO	Direct	Direct	Indirect
Na2Ti3O7	Direct	Direct	Direct
TiZnO3	Direct	Direct	Direct
SnZn2O4	Direct	Direct	Indirect
ZnS	Direct	Direct	Indirect
Zn2GeO4	Direct	Direct	Direct
GeO2	Direct	Direct	Indirect
MgCO3	Direct	Indirect	Direct
Mg2B2O5	Direct	Direct	Indirect
Mg2B2O5	Direct	Direct	Direct
Zn2O4Si	Direct	Direct	Indirect
BiRhO3	Indirect	Indirect	Indirect
Cr2ZnO4	Indirect	Direct	Direct
CaMnO3	Indirect	Direct	Direct
V2Zn3O8	Indirect	Direct	Direct
WO3	Indirect	Direct	Indirect
Rh2O3	Indirect	Direct	Indirect
ZnCO3	Indirect	Direct	Indirect
Zn2I4O12	Indirect	Direct	Indirect
ZnO2	Indirect	Direct	Indirect

CONCLUSIONS

The proposed methodology effectively predicts the band gap within a 2,700 binary and ternary crystalline structures database. A high-quality, large-scale repository was established by integrating 273 variables—including experimental data, literature values, and DFT calculations using the HSE06 functional—and automating the extraction and validation of gap magnitude and band type. The XGBoost model exhibited outstanding fit during training but saw its predictive power decrease in validation (R^2≈0.44), underscoring the need for overfitting control. Conversely, artificial neural networks, optimized via Grid Search and cross-validation, achieved $R^2 = 0.96$ on the training set and $R^2 = 0.92$ on the validation set, reflecting remarkable generalization. This approach drastically reduces computational cost compared to traditional DFT methods, from days to minutes per estimation, and establishes a reproducible, scalable protocol. Moreover, the same strategy can be extended to predict other electronic and optical properties, thereby contributing to the rational design of materials for optoelectronics, photocatalysis, and renewable energy.

1. References: L. Ge, Y. Ke, y X. Li, "Machine learning integrated photocatalysis: progress and challenges," Chemical Communications, vol. 59, no. 39, pp. 5795-5806, Apr. 2023. DOI: 10.1039/D3CC00989K.
2. A. O. Ibhadon y P. Fitzpatrick, "Heterogeneous Photocatalysis: Recent Advances and Applications," Catalysts, vol. 3, no. 1, pp. 189-218, Mar. 2023. DOI: 10.3390/catal3010189.
3. R. H. D. Santos, C. L. Oliveira, y J. C. Scorza, "Evaluation of XGBoost and other machine learning models for predicting the band gap of ZnO-based photocatalysts," Journal of Applied Physics, vol. 132, no. 14, pp. 145703, Oct. 2023. DOI: 10.1063/5.0065409.
4. Y. Zhang, X. Zhang, y Y. Lu, "Advanced machine learning techniques for photocatalytic performance optimization," Materials Today Communications, vol. 31, pp. 103519, Jan. 2023. DOI: 10.1016/j.mtcomm.2022.103519.