Report on Multivariate Analysis and Modeling of the Band Gap

Sebastián Mindiola

December 7, 2025

1 Introduction

The development of functional inorganic materials is essential for a wide variety of modern technological applications, such as light-emitting diodes (LEDs) [Fasol, 1996, Schubert and Kim, 2005], transistors [Radisavljevic et al., 2011], photovoltaic cells [Polman et al., 2016, Ahn et al., 2010], and scintillators. Effective design of these devices requires meticulous knowledge of the band gap (E_g) [Zhuo et al., 2018]. Traditionally, Density Functional Theory (DFT) has been used to predict E_g *a priori*; however, this method presents significant limitations, most notably a systematic underestimation of band gap values compared to experimental values when standard exchange-correlation functionals such as PBE are employed [Perdew, 2009, Seidl et al., 1996].

Although more precise methods exist, such as hybrid functionals [Heyd and Scuseria, 2004, Garza and Scuseria, 2016] or GW-type methods [Gerosa et al., 2015], their high computational cost prevents their use in high-throughput screening of materials [Zhuo et al., 2018]. Given this challenge, machine learning has emerged as a robust alternative for estimating electronic properties based solely on material composition.

This report takes as its starting point the experimental database compiled by Zhuo et al., who demonstrated that models trained with experimental data can surpass the accuracy of DFT calculations. Nevertheless, the use of compositional descriptors entails inherent statistical challenges, primarily the high collinearity among atomic properties.

The objective of this work is to deepen the analysis of the characteristics of such compositional data and evaluate predictive models. Specifically, it seeks to: (1) implement feature engineering based on statistical aggregations and reduce the resulting dimensionality through Factor Analysis to eliminate redundancy due to collinearity, interpreting the latent constructs obtained; (2) optimize the response variable (E_g) through power transformations (Yeo-Johnson and Box-Cox) to approximate normality and stabilize error variance; and (3) evaluate and compare the performance of various regression algorithms (linear and non-linear) to determine the predictive capacity of the latent factors on the transformed data.

2 Methodology

The methodological strategy was divided into five phases: data curation, feature engineering, dimensionality reduction, preprocessing of the target variable, and predictive modeling.

2.1 Data Source and Non-Metal Filtering

The dataset provided by Zhuo et al. was used, which consolidates 3896 experimentally measured band gap values [Kiselyova et al., 2016, Strehlow and Cook, 1973]. Given that the interest of this study lies in the quantitative prediction of the magnitude of the energy gap, the subset of interest S was defined by excluding metals, such that:

$$S = \{x_i \in D \mid E_g(x_i) > 0 \text{ eV}\} \tag{1}$$

where D represents the original dataset and $E_g(x_i)$ the band gap of composition i.

2.2 Feature Engineering and Vectorization

Originally, the data are presented as stoichiometric chemical formulas. To transform this qualitative information into a numerical vector space, a descriptor generation scheme based on elemental properties was employed.

A base set of 57 physical-chemical atomic properties was selected. For each composition, four statistical aggregation functions were applied to these properties, weighted by the stoichiometric fraction: average (*avg*), sum (*sum*), variance (*var*), and range (*range*). This results in a high-dimensional feature space with a total of $m = 57 \times 4 = 228$ predictor variables per observation.

2.3 Feature Analysis and Collinearity Reduction

The generated space of 228 variables presents high multicollinearity. To mitigate redundancy, the procedure was carried out in two stages:

1. **Elimination of Extreme Redundancy:** The Pearson correlation matrix was calculated, and variables presenting a coefficient $\rho > 0.95$ with respect to another variable were eliminated, reducing the initial space and avoiding mathematical singularity problems.

2. **Exploratory Factor Analysis:** A factor model was applied to the remaining variables to extract orthogonal latent constructs.

Viability Assessment

The suitability of the correlation matrix $\mathbf{R}$ was verified using two statistics:

- **Bartlett's Test of Sphericity:** Contrasts the null hypothesis of uncorrelation ($H_0 : \mathbf{R} = \mathbf{I}$). The statistic is defined as:

$$\chi^2 = -\left(n - 1 - \frac{2p+5}{6}\right) \ln |\mathbf{R}| \tag{2}$$

- **Kaiser-Meyer-Olkin (KMO) Index:** Measures sampling adequacy. It is calculated as:

$$KMO = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} u_{ij}^2} \tag{3}$$

Factor Model and Rotation

The model assumes that the vector of observable variables $\mathbf{X}$ depends on common factors $\mathbf{F}$ and errors ϵ:

$$\mathbf{X} = \mathbf{\Lambda F} + \epsilon \tag{4}$$

To facilitate the interpretation of the resulting factors, an orthogonal **Varimax** rotation was applied, which maximizes the variance of the squared factor loadings per column, tending to polarize the loadings toward 1 or 0.

2.4 Transformation of the Response Variable (E_g)

The target variable E_g presents a positive skew. Prior to transformation, extreme value filtering was applied by restricting the data to the 99th percentile (P_{99}) to avoid distortions by outliers. Subsequently, power transformations were evaluated:

- **Box-Cox Transformation:** Defined for $y > 0$:

$$y^{(\lambda)} = \begin{cases} \dfrac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(y) & \text{if } \lambda = 0 \end{cases} \tag{5}$$

- **Yeo-Johnson Transformation:** Robust generalization for non-strictly positive values:

$$y^{(\lambda)} = \begin{cases} \dfrac{(y+1)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, y \geq 0 \\ \ln(y+1) & \text{if } \lambda = 0, y \geq 0 \\ -\dfrac{(-y+1)^{2-\lambda} - 1}{2-\lambda} & \text{if } \lambda \neq 2, y < 0 \\ -\ln(-y+1) & \text{if } \lambda = 2, y < 0 \end{cases} \tag{6}$$

The resulting normality was evaluated using the **Shapiro-Wilk** test.

2.5 Predictive Modeling and Evaluation

To establish the functional relationship between the extracted latent factors and the band gap (both original and transformed), five regression algorithms were trained and compared: Multiple Linear Regression, Ridge Regression (L_2 regularization), Random Forest, Gradient Boosting, and Support Vector Regression (SVR) with radial basis function (RBF) kernel.

Validation and Metrics

The dataset was randomly divided into a training subset (80%) and a test subset (20%). Model performance was quantified using the Coefficient of Determination (R^2) and the Root Mean Squared Error (RMSE):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \quad RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the mean of the observed values.

3 Results

3.1 Data Adequacy and Preprocessing

The initial dataset consisted of 228 compositional descriptors for 2459 non-metal observations. Given that extreme multicollinearity can destabilize factor solutions, prior filtering was performed identifying pairs of variables with a Pearson correlation coefficient greater than 0.95. This process resulted in the elimination of 66 redundant variables, reducing the feature space to a final matrix of dimensions 2459 × 162.

Upon this curated set, statistical adequacy tests were applied, yielding the following results:

- **Bartlett's Test of Sphericity:** A p-value of 0.0 was obtained, rejecting the null hypothesis of uncorrelation. This confirms that, even after eliminating extreme redundancy, a significant covariance structure exploitable by factor analysis persists.
- **Kaiser-Meyer-Olkin Index:** The overall value increased slightly to 0.874. This result is categorized as "meritorious" and suggests that the elimination of highly collinear variables improved the quality of the partial correlation structure for the analysis.

3.2 Determination of the Number of Factors

The number of factors to retain was determined using the scree plot criterion and the percentage of explained variance.

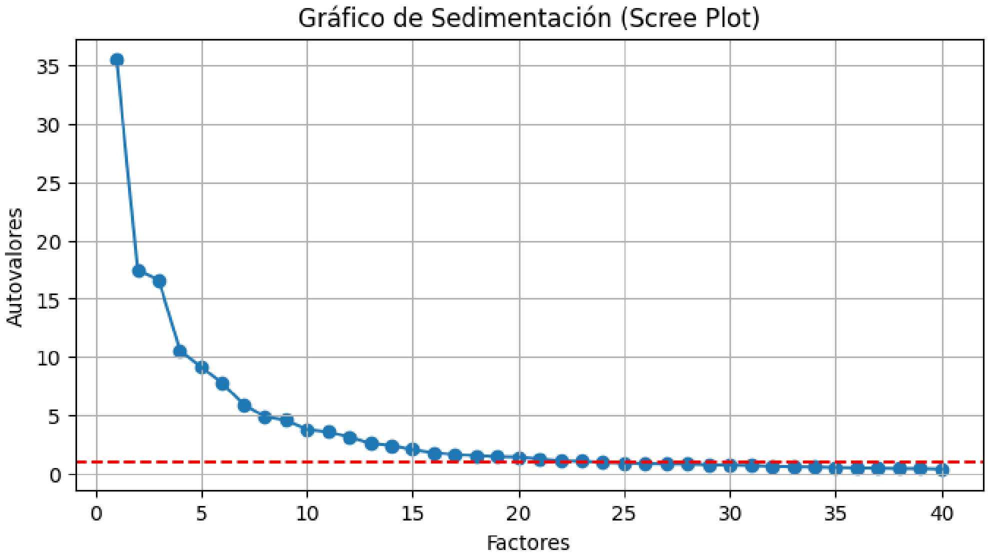

Figure 1: Scree plot. The curve suggests that a model of 20 factors is suitable for capturing the main structure of the data.

Table 1 presents the variance decomposition for the 20 selected factors. It is observed that the first factor explains 11.63% of the total variance, and the first 5 factors accumulate nearly 42%. With the retention of 20 latent factors, the model is capable of explaining 83.03% of the total variability present in the 162 variables analyzed. This result indicates highly efficient dimensionality reduction, condensing complex chemical and physical information into a manageable number of orthogonal constructs.

Table 1: Variance explained by the 20 factors retained after redundancy elimination. Sum of Squared (SS) Variance, Proportional, and Cumulative are detailed.

Factor	**1**	**2**	**3**	**4**	**5**
Variance SS	18.83	17.01	11.59	10.33	10.14
% Proportional	11.63%	10.50%	7.16%	6.38%	6.26%
% Cumulative	11.63%	22.13%	29.28%	35.66%	41.92%
Factor	**6**	**7**	**8**	**9**	**10**
Variance SS	8.17	7.22	6.38	6.33	5.92
% Proportional	5.04%	4.46%	3.94%	3.91%	3.66%
% Cumulative	46.96%	51.41%	55.35%	59.26%	62.92%
Factor	**11**	**12**	**13**	**14**	**15**
Variance SS	5.06	4.89	4.28	4.13	3.84
% Proportional	3.12%	3.02%	2.64%	2.55%	2.37%
% Cumulative	66.04%	69.06%	71.70%	74.25%	76.62%
Factor	**16**	**17**	**18**	**19**	**20**
Variance SS	2.73	2.08	1.93	1.91	1.74
% Proportional	1.68%	1.28%	1.19%	1.18%	1.07%
% Final Cumulative	**78.30%**	**79.59%**	**80.78%**	**81.96%**	**83.03%**

3.3 Interpretation of Latent Factors

Once the reduced dimensional structure was established, the physical meaning of the factors was interpreted through the weights of the variables representing them. Figures 2 and 3 illustrate the original variables with the highest factor loading in the definition of the main constructs.

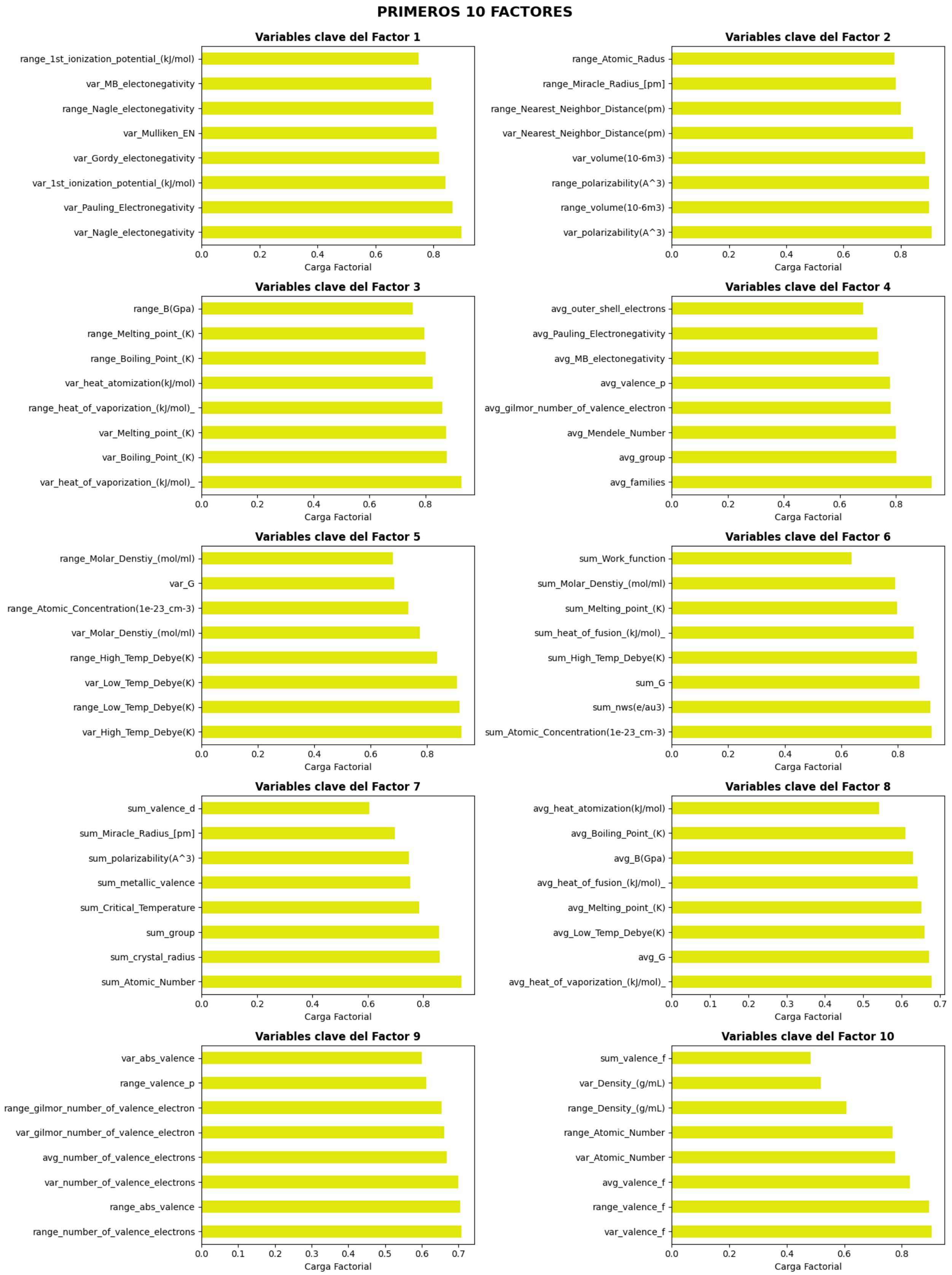

Figure 2: Most influential variables of the first 10 factors. A predominance of dispersion measures (var, range) is observed in the first three components.

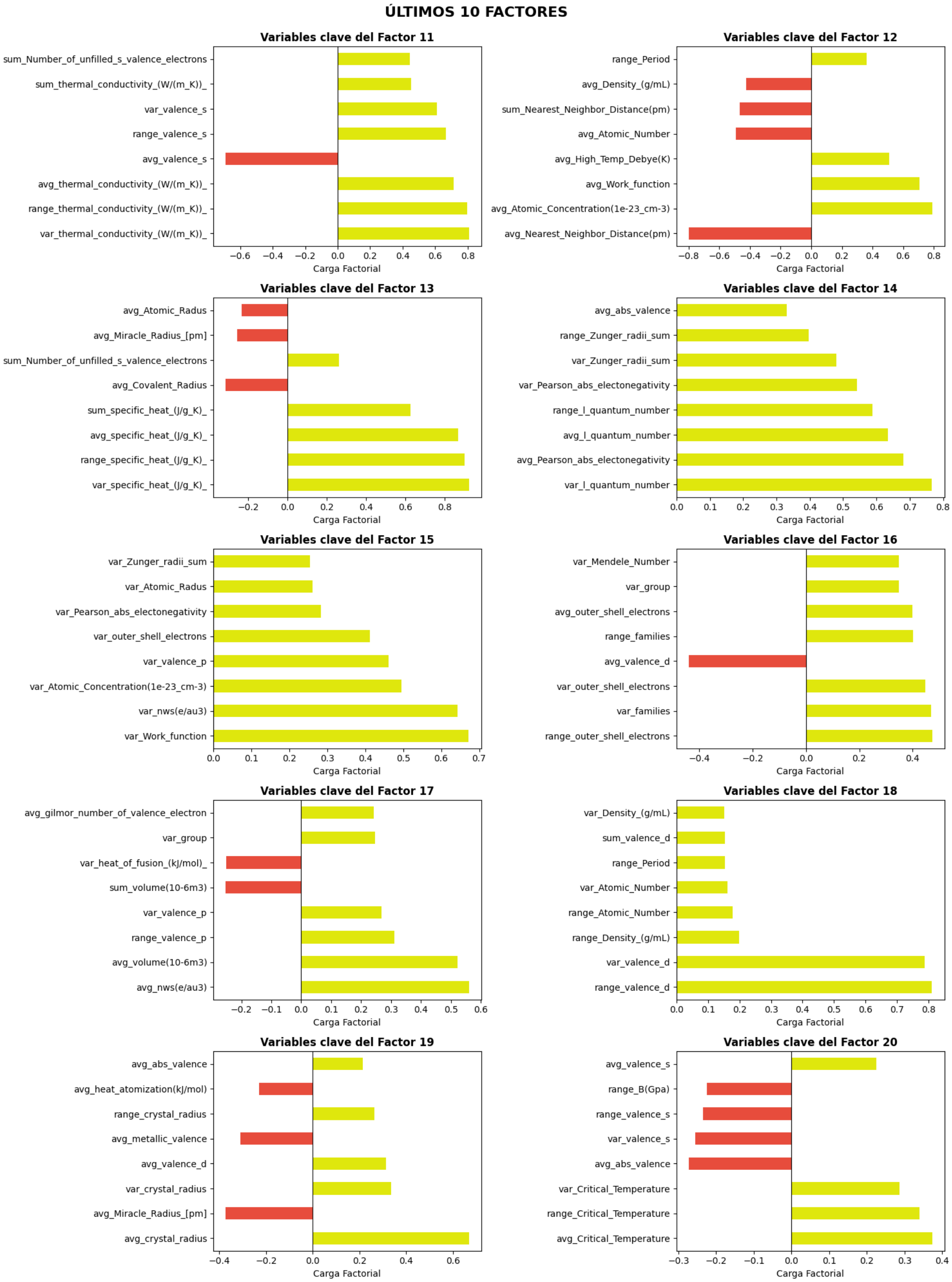

Figure 3: Most influential variables of factors 11-20, highlighting specific properties such as heat capacity and transition orbitals.

Analysis of the factorial structure reveals a fundamental physical pattern visible in Figure 2: the most determining factors are dominated by dispersion measures (*variance* and *range*), while averages (*avg*) and sums (*sum*) appear mostly in subsequent factors. This suggests that the magnitude of E_g in non-metals critically depends on the **asymmetry** or **contrast** among the elements forming the compound.

Below, the physical-chemical interpretation of the identified factor groups is detailed:

Group 1: Bond Disparity and Structure Factors

As evidenced in the first bars of Figure 2:

- **Factor 1 (Ionicity and Bond Polarity):** Is defined almost exclusively by the *variance* and *range* of electronegativity and ionization potential. A high variance in these loadings indicates the simultaneous presence of strong anions and cations, capturing the transition from covalent to ionic bonding, which tends to open the band gap.

- **Factor 2 (Steric Mismatch):** Groups the variance of polarizability and atomic radii. Represents structural tension (”size mismatch”) and crystal lattice distortion caused by combining atoms of very disparate sizes.

- **Factor 3 (Cohesion Disparity):** Dominated by the variance in melting points and enthalpies. Reflects the mixture of refractory elements with volatile ones, indicating the stability of the potential well.

Group 2: Chemical Identity and Periodicity Factors

- **Factor 4 (Average Periodic Position):** Unlike the previous ones, this factor loads on the *averages* (*avg*) of Group and Family. It locates the material in the periodic table (e.g., differentiating III-V from II-VI semiconductors), defining its global electronic configuration.

- **Factor 12 (Packing Density):** Positively relates atomic concentration and work function, describing lattice compactness and surface electronic density.

Group 3: Lattice Dynamics and Specific Orbitals

Figure 3 shows how certain factors isolate specific quantum characteristics:

- **Factor 5 (Lattice Rigidity):** Defined by the variance of the Debye Temperature, linked to phonon vibration modes and bond rigidity.

- **Factors 10 and 18 (f and d Orbitals):** Factor 10 isolates the variance of f electrons (Rare Earths) and Factor 18 (visible in the second graph) isolates d electrons (Transition Metals). These are critical as they form narrow bands that can drastically alter E_g.

- **Factor 11 (Thermal Transport):** Groups thermal conductivity with s orbital valence, consistent with the delocalized nature of these electrons facilitating transport.

Finally, the remaining factors (such as 6, 7, and 13) gather extensive and additive properties (sums of atomic number, specific heat capacity), acting as adjustment components that complete the thermodynamic description of the system without dominating the principal variance.

3.4 Transformation and Normalization of the Response Variable

The original target variable E_g presents a positively skewed distribution, typical of energy properties that are lower-bounded by zero but can extend to high values. Prior to evaluating transformations, extreme value filtering was applied, restricting the analysis to observations located below the 99th percentile. This preliminary measure was adopted to mitigate the influence of severe ”outliers” that could distort the estimation of normalization parameters.

To stabilize variance and improve the performance of linear regression and support vector-based models, three transformation techniques were evaluated on this curated set: Logarithmic ($\log(1 + y)$), Yeo-Johnson, and Box-Cox.

Table 2 summarizes the results of the Shapiro-Wilk test for each scenario.

Table 2: Results of the Shapiro-Wilk normality test ($N = 2459$). A W value close to 1 indicates greater proximity to normality.

Transformation	**Statistic** W	**Value** p
None (Original)	0.9405	2.33×10^{-30}
Logarithmic ($\log(1 + y)$)	0.9855	3.76×10^{-15}
Yeo-Johnson	0.9878	1.16×10^{-13}
Box-Cox	**0.9917**	1.10×10^{-10}

As observed in the p values of Table 2, no transformation managed to satisfy the null hypothesis of strict normality ($p > 0.05$). This result is expected given the sample size ($n = 2459$), as the Shapiro-Wilk test is extremely sensitive to minor deviations in large datasets.

However, the W statistic offers a comparative measure of the distribution shape. The original variable ($W = 0.9405$) shows considerable deviation. The application of the Box-Cox transformation maximized the W statistic to 0.9917, indicating the best possible approximation to a Gaussian distribution within the tested families of transformations.

This improvement is corroborated visually in Figure 4.

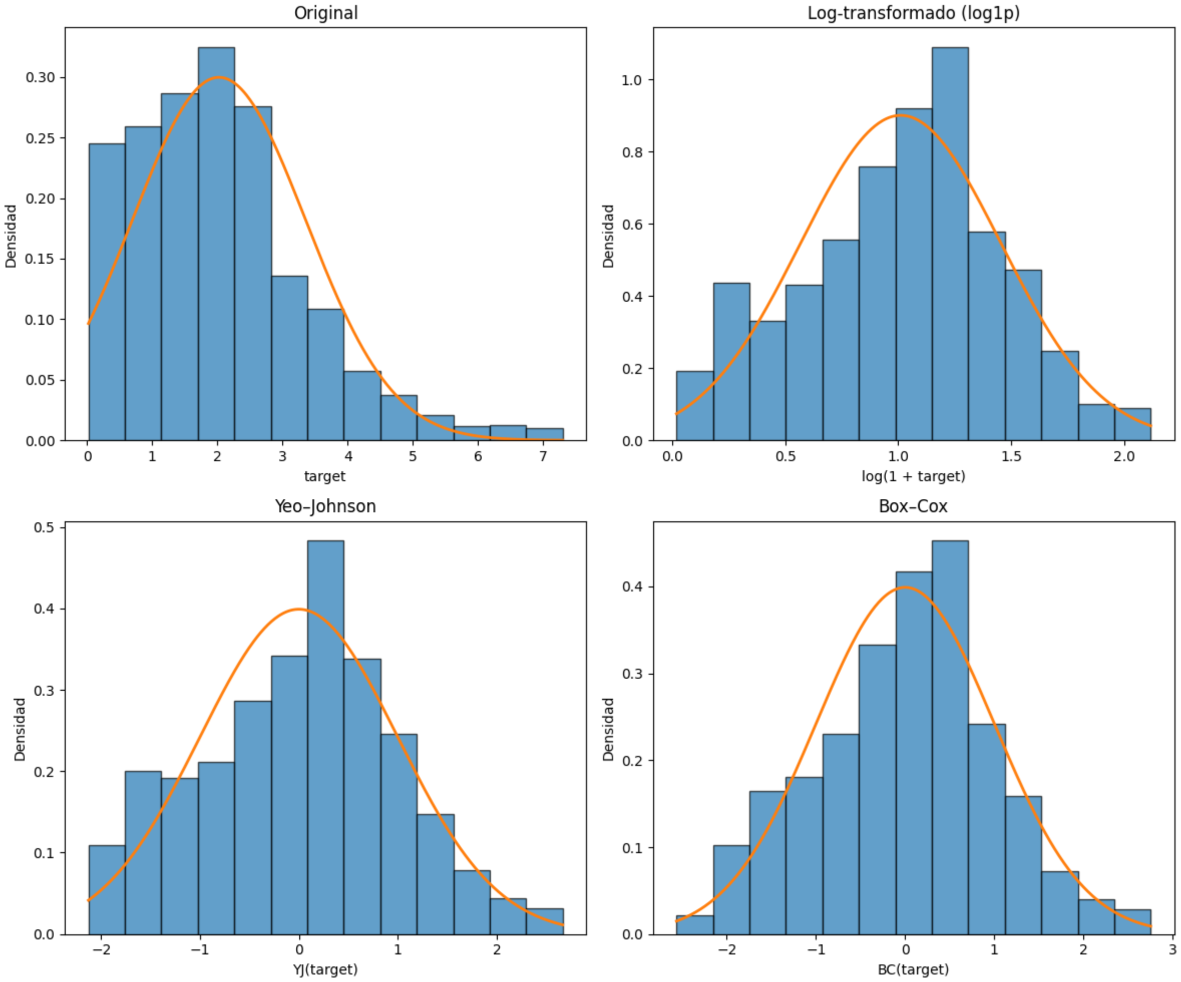

Figure 4: Comparison of probability densities. The ”Original” graph shows a clear skew to the right. The ”Box-Cox” transformation achieves the most effective symmetry, aligning the histogram (blue) almost perfectly with the theoretical normal curve (orange).

Analyzing the histograms:

- **Original:** Presents a long tail to the right (positive kurtosis), which could bias models towards the lower and more frequent band gap values.
- **Log-transformed:** Although it reduces skewness, it tends to slightly ”over-correct,” generating an accumulation in the central-left part.
- **Box-Cox:** Generates the most symmetric and centered distribution. The fit of the density curve (orange line) over the histogram is superior to that of Yeo-Johnson and Logarithm, significantly reducing the impact of outliers at the extremes of the distribution.

Consequently, the variable transformed by Box-Cox was selected as the definitive target (in addition to the original untransformed variable) for the training of subsequent predictive models.

3.5 Evaluation of Predictive Models

To determine the predictive capacity of the compositional descriptors, an exhaustive comparative study was carried out evaluating five distinct regression algorithms: Linear Regression, Ridge, Gradient Boosting, Random Forest, and SVR. The experimental design contemplated four scenarios resulting from the combination of two input types (original variables vs. latent factors) and two transformations of the target variable (original y_{fill} vs. Box-Cox y_{bc}).

Table 3 synthesizes the global performance of the models. The results evidence a clear superiority of ensemble methods, specifically Gradient Boosting and Random Forest, when trained on the latent factor space, reaching coefficients of determination (R^2) close to 0.80. This demonstrates that dimensionality reduction via factor analysis was effective in filtering noise and redundancy, providing robust orthogonal features. On the other hand, the SVR model showed the best performance when using the original variables ($R^2 = 0.76$), but its predictive capacity decreased drastically when applied to the latent factors without specific hyperparameter retuning.

Additionally, it was confirmed that the Box-Cox transformation is critical for precision. Although the R^2 coefficient remained stable, the RMSE experienced a substantial improvement in non-linear models. The Random Forest algorithm trained on latent factors achieved the minimum absolute error of the study (RMSE = 0.4627), validating that the normalization of the response variable effectively minimizes the magnitude of residuals.

Table 3: Comparative summary of performance. Results using Latent Factors vs. Original Variables are contrasted for both target variable transformations.

		R^2 Score (Higher is better)		**RMSE (Lower is better)**	
Input	**Model**	**Box-Cox (y_{bc})**	**Original (y_{fill})**	**Box-Cox (y_{bc})**	**Original (y_{fill})**
Latent Factors	**Gradient Boosting**	0.7771	**0.7991**	0.4683	0.5889
	Random Forest	**0.7824**	0.7830	**0.4627**	0.6120
	Ridge	0.7148	0.7312	0.5297	0.6812
	Regresión Lineal	0.7051	0.7235	0.5386	0.6909
	SVR (RBF)	0.1889	0.1953	0.8934	1.1786
Original Variables	**SVR (RBF)**	**0.7353**	**0.7626**	**0.5103**	**0.6402**
	Random Forest	0.7206	0.7314	0.5243	0.6809
	Gradient Boosting	0.6843	0.6889	0.5573	0.7329
	Ridge	0.5897	0.6063	0.6354	0.8244
	Regresión Lineal	0.5896	0.6062	0.6354	0.8245

Figure 5 presents a comparative panel with the four analyzed scenarios, allowing visualization of the dispersion of predictions versus actual values:

- **Upper Panel (Original Variables):** For both the original variable (left) and the transformed one (right), considerable dispersion is observed, especially in high band gap values. Although SVR achieves a good global fit ($R^2 \approx 0.76$), points distant from the ideal diagonal exist.

- **Lower Panel (Latent Factors):** Here, the improvement in robustness is evidenced. The Random Forest model (right, with Box-Cox) shows the most compact point cloud aligned with the red diagonal, significantly reducing residual variance compared to the upper models. This visually confirms that the combination of dimensionality reduction (Factors) and normalization (Box-Cox) produces the most stable and precise predictions.

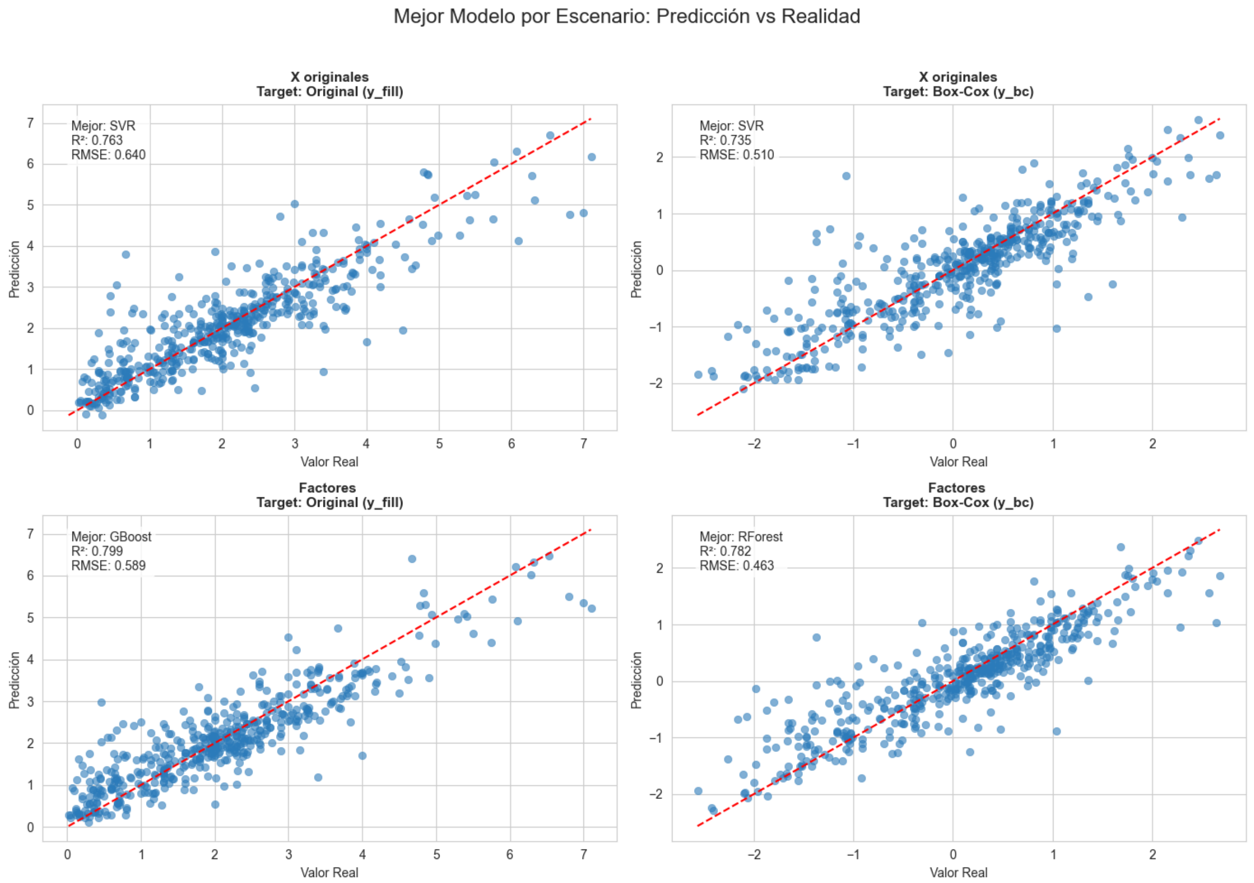

Figure 5: Performance comparison of the best models in the four scenarios. Top: Models trained with original variables (SVR). Bottom: Models trained with latent factors (Random Forest). The Box-Cox transformation (right column) and the use of factors (bottom row) generate the most compact fit (least dispersion).

Finally, to understand the topology of the prediction surface and the sufficiency of the main factors, a simplified model was trained using exclusively the first two latent factors: F1 (Electronic Disparity) and F2 (Steric Mismatch). Figure 6 presents the resulting contour curves. The complexity of the contours evidences the highly non-linear nature of the response function. However, the low R^2 (0.3902) of this reduced model confirms that the band gap is a multidimensional phenomenon requiring the integration of the 20 latent factors to be predicted with precision.

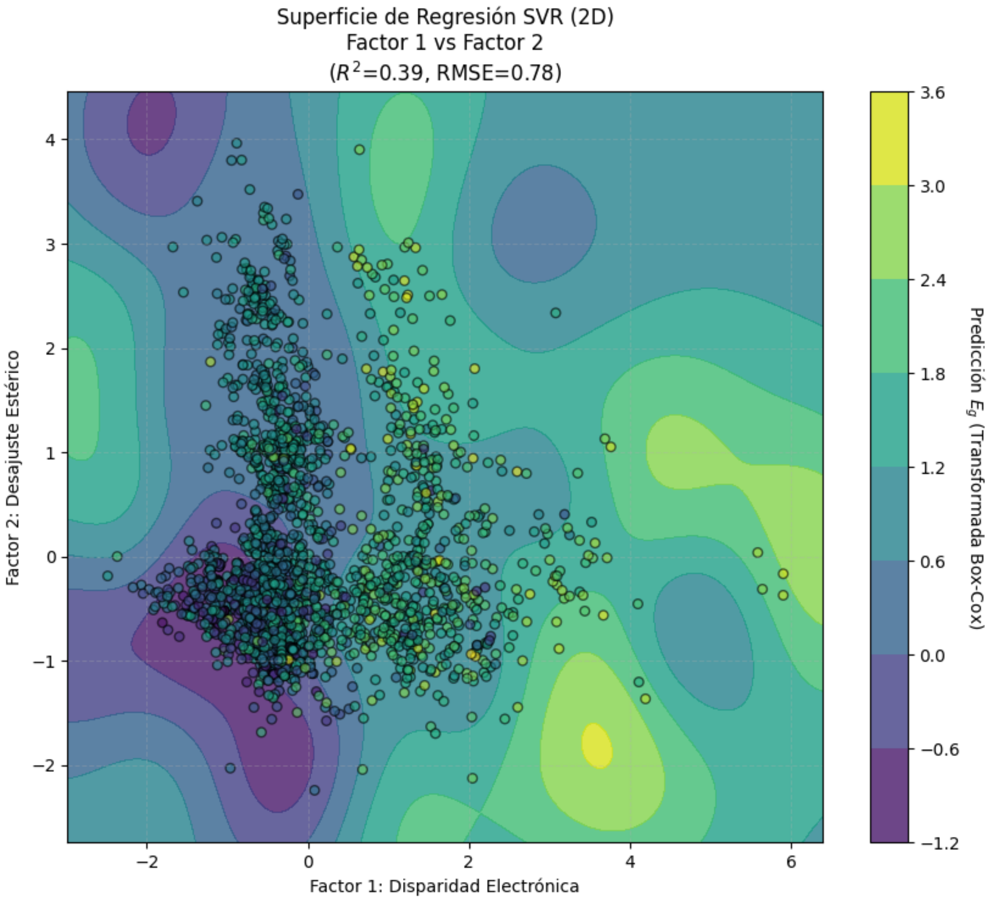

Figure 6: Decision surface of the simplified model using only Factors 1 and 2. The complex topology demonstrates the non-linearity of the problem, while the low explained variance underscores the need to integrate all latent factors.

4 Conclusions

The statistical exploration carried out on the compositional descriptors allowed evidencing that a significant redundant information structure exists in atomic properties, which can be efficiently condensed. Through factor analysis, it was possible to reduce the original 228 variables to 20 latent factors, retaining most of the system's variance and providing a robust feature space for machine learning model training.

This reduction process revealed a relevant structural pattern: the factors explaining the greatest variability are defined by dispersion measures (*variance* and *range*) and not by averages. This suggests that, in the context of these data, the differences between the properties of constituent atoms—such as disparity in electronegativity or size—provide more statistical information than their mean values to describe variations in the band gap.

Regarding model fitting, it was observed that the use of latent factors together with ensemble algorithms (Random Forest and Gradient Boosting) generated the best results, even surpassing the performance obtained with original variables in several scenarios. However, it is important to highlight that no model managed to exceed an R^2 of 0.80. This threshold suggests that there is a remnant variability in the band gap that cannot be explained exclusively through chemical composition; it is likely that the incorporation of structural or crystallographic descriptors is necessary to capture the entirety of the physical phenomenon and surpass this performance bound.

Finally, it was confirmed that the preprocessing of the response variable is a highly advantageous strategy. The application of the Box-Cox transformation proved decisive for prediction precision, allowing RMSE to be reduced to minimum values close to 0.46 eV. This validates that normalizing the distribution of the target variable results in more reliable models with less bias against extreme values.

5 References

References

[Ahn et al., 2010] Ahn, S., Jung, S., Gwak, J., Cho, A., Shin, K., Yoon, K., Park, D., Cheong, H., and Yun, J. H. (2010). Determination of band gap energy (Eg) of Cu2ZnSnSe4 thin films: On the discrepancies of reported band gap values. *Applied Physics Letters*, 97(2):021905.

[Fasol, 1996] Fasol, G. (1996). Room-Temperature Blue Gallium Nitride Laser Diode. *Science*, 272(5269):1751–1752.

[Garza and Scuseria, 2016] Garza, A. J. and Scuseria, G. E. (2016). Predicting Band Gaps with Hybrid Density Functionals. *The Journal of Physical Chemistry Letters*, 7(20):4165–4170.

[Gerosa et al., 2015] Gerosa, M., Bottani, C. E., Caramella, L., Onida, G., Di Valentin, C., and Pacchioni, G. (2015). Electronic structure and phase stability of oxide semiconductors: Performance of dielectric-dependent hybrid functional DFT, benchmarked against G W band structure calculations and experiments. *Physical Review B*, 91(15):155201.

[Heyd and Scuseria, 2004] Heyd, J. and Scuseria, G. E. (2004). Efficient hybrid density functional calculations in solids: Assessment of the Heyd–Scuseria–Ernzerhof screened Coulomb hybrid functional. *The Journal of Chemical Physics*, 121(3):1187–1192.

[Kiselyova et al., 2016] Kiselyova, N. N., Dudarev, V. A., and Korzhuyev, M. A. (2016). Database on the bandgap of inorganic substances and materials. *Inorganic Materials: Applied Research*, 7(1):34–39.

[Perdew, 2009] Perdew, J. P. (2009). Density functional theory and the band gap problem. *International Journal of Quantum Chemistry*, 28(S19):497–523.

[Polman et al., 2016] Polman, A., Knight, M., Garnett, E. C., Ehrler, B., and Sinke, W. C. (2016). Photovoltaic materials: Present efficiencies and future challenges. *Science*, 352(6283):aad4424.

[Radisavljevic et al., 2011] Radisavljevic, B., Radenovic, A., Brivio, J., Giacometti, V., and Kis, A. (2011). Single-layer MoS2 transistors. *Nature Nanotechnology*, 6(3):147–150.

[Schubert and Kim, 2005] Schubert, E. F. and Kim, J. K. (2005). Solid-State Light Sources Getting Smart. *Science*, 308(5726):1274–1278.

[Seidl et al., 1996] Seidl, A., Görling, A., Vogl, P., Majewski, J. A., and Levy, M. (1996). Generalized Kohn-Sham schemes and the band-gap problem. *Physical Review B*, 53(7):3764–3774.

[Strehlow and Cook, 1973] Strehlow, W. H. and Cook, E. L. (1973). Compilation of Energy Band Gaps in Elemental and Binary Compound Semiconductors and Insulators. *Journal of Physical and Chemical Reference Data*, 2(1):163–200.

[Zhuo et al., 2018] Zhuo, Y., Mansouri Tehrani, A., and Brgoch, J. (2018). Predicting the band gaps of inorganic solids by machine learning. *The journal of physical chemistry letters*, 9(7):1668–1673.