VOLUMEN 37
XXX Verano De la Ciencia
ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

Predicción de propiedades de materiales semiconductores para aplicaciones en celdas solares mediante ML y cálculos *ab-initio*

Prediction of semiconductor material properties for solar cell applications using ML and ab-initio calculations

J. Hernández-Varela[1*], J. E. Castellanos-Águila[1**]
[1]División de Ingenierías, Campus Irapuato – Salamanca, Universidad de Guanajuato. Carr. Salamanca-Valle Km 3.5+1.8, Comunidad de Palo Blanco, 36700 Salamanca, Gto., México.
*j.hernandezvarela@ugto.mx, **je.castellanos@ugto.mx

Resumen

En este trabajo se propone una metodología híbrida que combina cálculos ab-initio y técnicas de aprendizaje automático para predecir el band-gap de compuestos binarios tipo AB. Se construyó una base de datos de 48 compuestos con estructura zinc blenda, utilizando propiedades atómicas pristinas como descriptores. Se entrenaron modelos simbólicos mediante la librería PySR, empleando datos de Materials Project y simulaciones propias con Quantum ESPRESSO. Los resultados muestran que el modelo basado en MP presenta una mayor capacidad explicativa (R^2 = 0.8947) y menor error (MAE = 0.2896) en comparación con el modelo basado en QE. Asimismo, nuestros resultados confirman que las propiedades periódicas de los elementos, como la electronegatividad y la afinidad electrónica, pueden ser utilizadas para construir modelos predictivos precisos e interpretables del *band-gap*, lo que abre nuevas posibilidades para el diseño racional de materiales semiconductores.

Palabras clave: Semiconductores, Band-gap, ML, Propiedades Periódicas.

Introducción

El ancho de banda prohibida o *band-gap* (E_g) es una propiedad fundamental en materiales semiconductores, ya que determina su comportamiento electrónico y óptico. Conocer su valor permite optimizar materiales para una amplia gama de aplicaciones tecnológicas, incluyendo celdas solares, dispositivos optoelectrónicos, sensores y materiales termoeléctricos. En particular, en el caso de las celdas solares, contar con un *band-gap* adecuado es crucial para maximizar la absorción de luz solar y, por ende, la eficiencia de conversión energética. Una predicción precisa del *band-gap* permite seleccionar o diseñar materiales con propiedades electrónicas óptimas, evitando costosos procesos de prueba y error.

Tradicionalmente, la determinación del *band-gap* se ha basado en métodos experimentales o en cálculos de primeros principios como la teoría del funcional de la densidad (DFT), los cuales, si bien son precisos, pueden resultar costosos en tiempo y recursos computacionales, especialmente cuando se pretende explorar grandes cantidades de materiales [1–3].

En años recientes, el uso de técnicas de inteligencia artificial (IA) ha emergido como una alternativa poderosa para acelerar el descubrimiento de materiales y la predicción de sus propiedades [4]. Sin embargo, una limitación importante de muchos modelos de IA es su falta de interpretabilidad física, lo que dificulta entender los mecanismos subyacentes que rigen las propiedades electrónicas de los materiales [5].

Una estrategia prometedora para mejorar la interpretabilidad consiste en utilizar propiedades atómicas intrínsecas, como la electronegatividad, la energía de ionización, la afinidad electrónica y el radio atómico, que están directamente relacionadas con la posición de los elementos en la tabla periódica. Estas propiedades, conocidas como propiedades pristinas, reflejan tendencias periódicas que influyen de manera significativa en el comportamiento electrónico de los compuestos. Estudios recientes han demostrado que existe una correlación clara entre estas propiedades y el *band-gap*, lo que permite construir modelos predictivos más comprensibles y físicamente fundamentados [6].

En este contexto, la regresión simbólica se presenta como una herramienta especialmente prometedora, ya que permite generar expresiones matemáticas explícitas que relacionan propiedades atómicas con propiedades macroscópicas como el *band-gap*, manteniendo al mismo tiempo un alto grado de interpretabilidad.

En este trabajo se construyó una base de datos de 48 compuestos binarios, tipo AB, generados combinando elementos de los grupos III–V y II–VI de la tabla periódica, restringidos a la fase cúbica tipo zinc blenda para garantizar consistencia estructural. A partir de esta base, se aplicó el algoritmo PySR [7] para generar descriptores simbólicos interpretables. Con base a lo anterior se evaluó la capacidad predictiva del modelo y se analizó cómo las propiedades atómicas de los elementos A y B influyen en el *band-gap* de estos compuestos.

Metodología

Para llevar a cabo este estudio, se construyó una base de datos de compuestos binarios tipo AB, con el objetivo de generar descriptores simbólicos que expliquen el comportamiento del band gap en materiales semiconductores. Se seleccionaron un total de 48 compuestos binarios tipo AB, generados a partir de combinaciones de elementos pertenecientes a los grupos III–V y II–VI de la tabla periódica. Una característica clave es que todos los materiales fueron modelados en la estructura cristalina cúbica tipo zinc blenda (grupo espacial F-43m). Esta decisión se tomó con el fin de mantener una coherencia estructural estricta entre todos los compuestos, facilitando así la comparación entre materiales y la interpretación de los descriptores generados.

Posteriormente, los compuestos fueron organizados en dos subconjuntos según el origen de sus datos electrónicos:

- En el primer conjunto, el valor del *band-gap* fue obtenido directamente desde la base de datos de Materials Project (https://next-gen.materialsproject.org/api) [8], seleccionando únicamente aquellas entradas cuya simetría fuera consistente con la fase zinc blenda o que, según la literatura, presentaran configuraciones estructurales compatibles con dicha geometría.
- Los cálculos basados en la teoría DFT para semiconductores del tipo AB se realizaron utilizando el software de Quantum Espresso [9], que emplea ondas planas para representar la función de onda electrónica. La representación de la interacción de los electrones de valencia y el núcleo atómico se realizó mediante los pseudopotenciales ultrasuaves del tipo PAW. Finalmente, para el termino de intercambio correlación se empleó la aproximación GGA [10]. Cada sistema semiconductor del tipo AB se modeló utilizando una celda unitaria, donde se relajaron completamente tanto las posiciones atómicas como los parámetros de red para garantizar la precisión de los cálculos. Asimismo, se utilizó un mallado de puntos k de 4×4×4. Esta metodología nos permitió obtener datos de banda prohibida de alta precisión, lo que proporciona una base sólida para un análisis más profundo de las propiedades de los materiales y facilita la comprensión y predicción del rendimiento de estos materiales en diversas aplicaciones.

Tabla 1. *Propiedades atómicas utilizadas para el entrenamiento del modelo de IA*

Nombre	Abreviación	Unidad
Número atómico	AN	-
Número de electrones de valencia	NVE	-
Energía de ionización	IE	kJ/mol
Pesó atómico	AW	a.m.u.
Radio atómico calculado	ARC	pm
Radio covalente	CR	pm
Afinidad electrónica	EA	kJ/mol
Electronegatividad de Pauling	EN	-

Ambos conjuntos comparten la misma representación de características atómicas (Tabla 1), lo cual permite realizar una comparación directa tanto en términos de distribución del *band-gap* como en el rendimiento de los modelos predictivos generados. Los datos fueron limpiados para eliminar entradas con valores faltantes en la variable objetivo o en las características predictoras. En la Tabla 2 se presenta la lista completa de los compuestos incluidos en la base de datos.

Tabla 2. *Listado de los 48 compuestos binarios tipo AB utilizado, clasificados por su tipo químico.*

Compuestos III-V				
BN	BP	Bas	BSb	BBi
AlN	AlP	AlAs	AlSb	AlBi
GaN	GaP	GaAs	GaSb	GaBi
InN	InP	InAs	InSb	InBi
Compuestos II-VI				
BeO	BeS	BeSe	BeTe	MgO
MgS	MgSe	MgTe	CaO	CaS
CaSe	CaTe	SrO	SrS	SrSe
SrTe	BaO	BaS	BaSe	BaTe
ZnO	ZnS	ZnSe	ZnTe	CdO
CdS	CdSe	CdTe		

Para modelar la relación entre las propiedades atómicas y el *band-gap*, se utilizó el algoritmo “PySR” que hace uso de la programación genética para realizar regresión simbólica. Esto genera expresiones matemáticas interpretables. El modelo fue configurado con los parámetros listados en la Tabla 3.

Tabla 3. *Configuración del algoritmo PySR*

Parámetro	Valor
Operadores binarios	+, -, *, /
Operadores unarios	sqrt, log, exp
Iteraciones	50
Tamaño de población	400
Número de poblaciones	20
Tamaño máximo de la expresión	30

El objetivo fue encontrar un descriptor simbólico que prediga el *band-gap* a partir de las propiedades atómicas, maximizando la precisión y la interpretabilidad.

Para evaluar la robustez del proceso de generación de descriptores, se implementó una validación cruzada de 5 pliegues. Por lo que en cada iteración:

1. Se dividió el conjunto de datos de manera aleatoria en 80% para entrenamiento y 20% para validación.
2. Se entrenó un nuevo modelo PySR usando los datos de entrenamiento.
3. Se generó un descriptor simbólico específico para ese pliegue.

4. Se evaluó el desempeño del descriptor sobre el conjunto de validación.

Las métricas utilizadas para evaluar el rendimiento fueron: RMSE (R*oot Mean Squared Error*), MAE (*Mean Absolute Error*) y R^2 (*Coeficiente de determinación*).

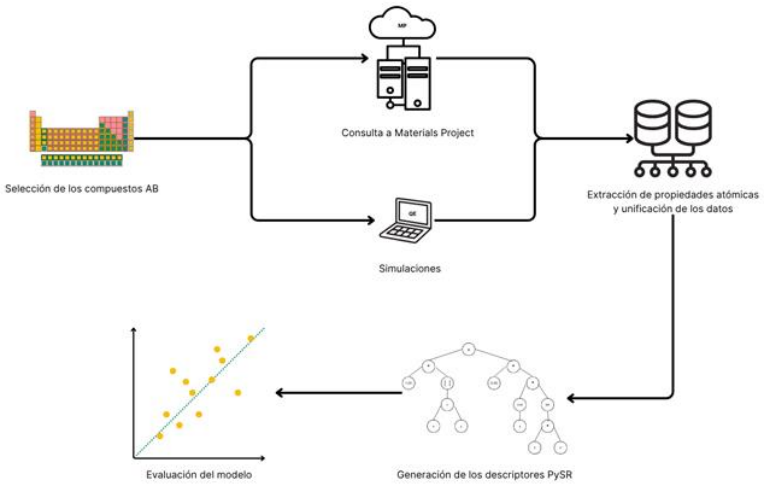

Figura 1. Flujo de trabajo para la predicción del band-gap de los compuestos AB. El proceso incluye la selección de compuestos a partir de la tabla periódica, obtención de datos a partir de la API de Materials Project y de simulaciones particulares con Quantum Espresso, extracción y unificación de propiedades atómicas, generación de descriptores con PySR, y evaluación del modelo.

Resultados

Para llevar a cabo el estudio se compararon los valores del band-gap para diversos compuestos semiconductores, derivados de datos experimentales, simulaciones con Quantum Espresso (QE) y predicciones de Materials Project (MP). En la Figura 2 se presentan los resultados de los 48 compuestos binarios considerados. Particularmente, los valores experimentales de *band-gap*, que oscilan aproximadamente entre 0.2 y 10 eV, sirvieron como estándar de referencia, para evaluar el modelo predictivo basado en las propiedades atómicas de la Tabla 1.

Partiendo de la hipótesis de que el band-gap debe reflejar las propiedades periódicas de los elementos que conforman los compuestos, tales como la electronegatividad, la afinidad electrónica y los radios atómicos, se evaluó la precisión de los dos métodos computacionales considerados (Fig. 2a). De la comparación con respecto a los valores experimentales, se determinó que MP presenta una correlación mayor (r = 0.95) con los datos experimentales, en comparación con QE (r = 0.63), aunque ambos métodos muestran errores absolutos similares (MAE ≈ 1.4 eV) (Fig. 2b). Estos resultados respaldan la idea de que las propiedades atómicas prístinas, derivadas de la posición de los elementos en la tabla periódica, pueden ser utilizadas como descriptores efectivos para modelar el band-gap, y que modelos basados en inteligencia artificial pueden beneficiarse de esta relación para mejorar su capacidad predictiva e interpretabilidad.

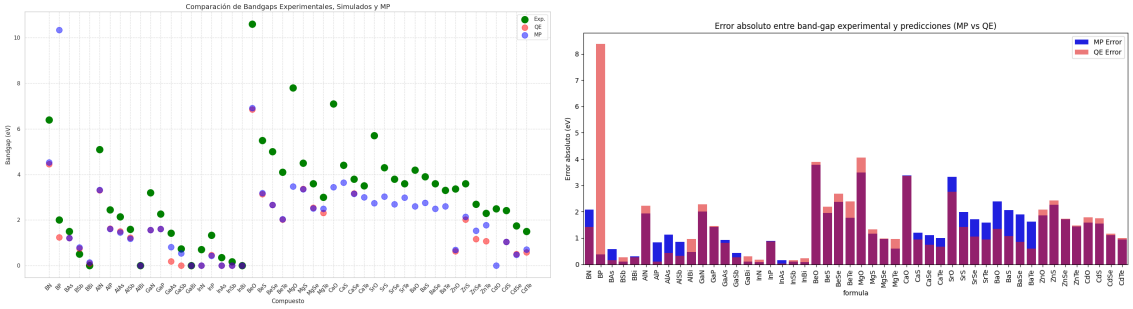

Figura 2. a) Distribución del band-gap y b) error absoluto de ambos conjuntos de datos con respecto a valores experimentales.

Puesto que se trata de validar las tendencias periódicas de los elementos en el band-gap de los semiconductores considerados, se buscó una correlación entre la electronegatividad y E_g. Las Figuras 3 y 4 muestra la distribución del band gap agrupadas por los elementos A y B, respectivamente. Ambas figuras conservan el mismo comportamiento de la electronegatividad de acuerdo con posición en la tabla periódica, incrementa de izquierda a derecha a lo largo de un período y disminuye de arriba hacia abajo dentro de un grupo.

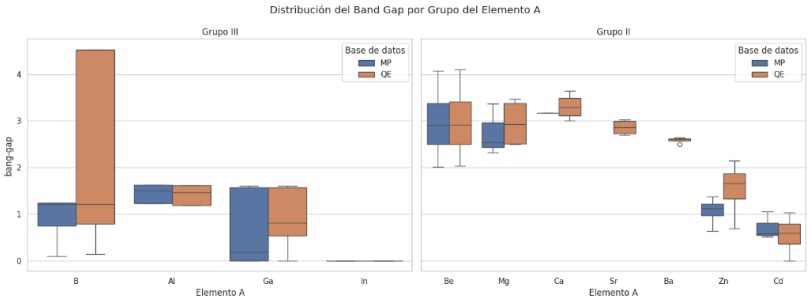

Figura 3. Distribución del band gap por elemento A en ambos conjuntos de datos.

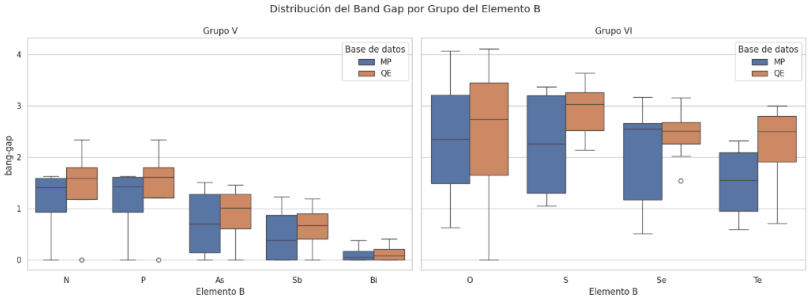

Figura 4. Distribución del band gap por elemento B en ambos conjuntos de datos.

Resultados del modelo PySR

Los resultados obtenidos mediante la librería PySR muestran diferencias significativas en el rendimiento predictivo entre los modelos entrenados con datos provenientes de Materials Project y Quantum ESPRESSO. Ambos modelos fueron ratificados mediante validación cruzada de 5 pliegues, y sus métricas promedio se resumen en la Tabla 4.

Tabla 4. *Resultados promedio de los 5 pliegues por métrica en cada conjunto de datos.*

Métrica	Materials Project	Quantum Espresso
RMSE	0.1363	0.4590
MAE	0.2896	0.5233
R^2	0.8947	0.6401

Se observa que el modelo basado en MP alcanzó un coeficiente de determinación R^2 de 0.8947, lo que indica una capacidad explicativa cercana al 90 % de la varianza del *band-gap*, mientras que el modelo entrenado con datos de QE obtuvo un R^2 de 0.6401. Esta diferencia se refleja también en las métricas de error, donde el modelo MP presenta valores inferiores tanto en RMSE (0.1363) como en MAE (0.2896), en comparación con el modelo QE (RMSE = 0.4590, MAE = 0.5233).

Estos resultados refuerzan la hipótesis de que el *band-gap* de los materiales semiconductores está fuertemente influenciado por las propiedades periódicas de los elementos que los conforman. Dado que las propiedades prístinas como la electronegatividad, la afinidad electrónica y los radios atómicos presentan variaciones sistemáticas a lo largo de la tabla periódica, es razonable esperar que el *band-gap* también exhiba tendencias periódicas cuando se forman compuestos binarios tipo AB. El modelo entrenado con datos de MP, que refleja mejor estas tendencias, logra capturar con mayor precisión las relaciones subyacentes entre las propiedades atómicas y el comportamiento electrónico del material. En contraste, las simulaciones propias con QE, aunque útiles, podrían estar sujetas a limitaciones metodológicas o de parametrización que afectan su capacidad para representar fielmente estas relaciones.

Descriptores simbólicos

Como resultado del proceso de regresión simbólica aplicado con PySR, se obtuvieron dos descriptores independientes, uno para cada conjunto de datos. Ambos descriptores expresan el *band-gap* como una función explícita de propiedades atómicas fundamentales, y muestran estructuras funcionales distintas, lo cual refleja la naturaleza de los datos con los que fueron entrenados.

El descriptor generado a partir del conjunto de datos obtenido a partir de **MP** es el siguiente:

$$-0.171EN_b + \mathrm{NV}E_b - \log\left(AW_a + 6.78\sqrt{AW_b}EN_a + 6.78EA_b - 10.1NVE_b\right) + \frac{2.20}{NVE_a}$$

Este descriptor combina propiedades de ambos sitios atómicos (A y B), incluyendo electronegatividad (EN), energía de afinidad electrónica (EA), peso atómico (AW) y número de electrones de valencia (NVE). La forma funcional, particularmente el uso del logaritmo y de una fracción inversa, sugiere una relación no lineal entre estas propiedades y el band gap, consistente con observaciones físicas previas.

Por otro lado, el descriptor obtenido a partir del conjunto de datos obtenidos de **QE** es:

$$-3.54\log(EN_a) + 1.80 + \frac{92.3}{AN_b + AW_a - \dfrac{14.2}{AN_a - AN_b + AW_a}}$$

Este modelo también incluye propiedades fundamentales como electronegatividad (EN), peso atómico (AW) y número atómico (AN), aunque centrado más en el sitio A y en relaciones más complejas entre AN y AW.

La Figura 5 compara las predicciones generadas por ambos descriptores contra los valores experimentales del *band-gap*. Se incluye la línea ideal $y = x$, que representa una predicción perfecta. Puede observarse que el descriptor entrenado con datos del conjunto MP (puntos azules) presenta una mayor concentración de puntos cercanos a la línea ideal, lo que concuerda con su mejor desempeño cuantitativo. El descriptor del conjunto QE (puntos naranjas), aunque presenta una mayor dispersión, logra capturar adecuadamente las tendencias generales, especialmente en el rango bajo del *band-gap*.

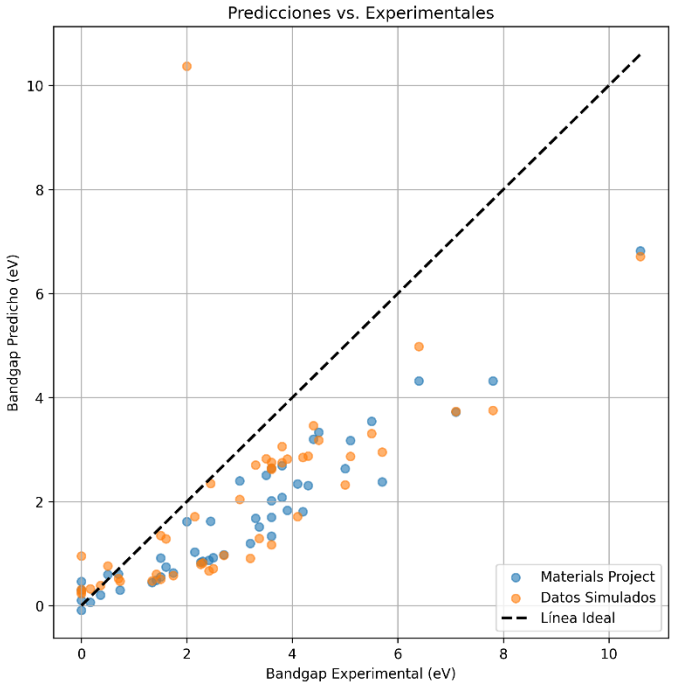

Figura 5. Comparación entre los valores de band-gap predichos por el modelo y los valores experimentales para las dos bases de datos evaluadas.

Conclusiones

Los resultados obtenidos en este estudio demuestran que el uso de propiedades atómicas prístinas como descriptores en modelos de regresión simbólica permite predecir con alta precisión el band-gap de compuestos semiconductores binarios. El modelo entrenado con datos de Materials Project superó en rendimiento al modelo basado en simulaciones con Quantum ESPRESSO, lo que sugiere que la calidad y consistencia de los datos de entrada son factores determinantes en la capacidad predictiva. La correlación observada entre el *band-gap* y las propiedades periódicas tales como la electronegatividad y la afinidad electrónica respalda la hipótesis de que estas características fundamentales pueden capturar tendencias estructurales y electrónicas relevantes. Por lo tanto, el enfoque propuesto ofrece una vía eficiente y físicamente fundamentada para acelerar el descubrimiento de materiales funcionales en aplicaciones fotovoltaicas y optoelectrónicas.

Bibliografía

[1] C. Fritz, J. Fernandez-Serra, M. Soler, J. Chem. Phys. 144, 224101 (2016).

[2] L. Fiedler, K. Shah, M. Bussman, A. Cangi, Phys. Rev. Materials 6, 040301 (2022).

[3] W. Kohn, L. J. Sham, Phys. Rev. 140, A1133-A1138 (1965).

[4] S.I. Simak, E.K. Delczeg-Czirjak, O. Eriksson, *Sci Rep* **15**, 17212 (2025).

[5] A. Mazheika, Y.-G. Wang, R. Valero, F. Viñes, F. Illas, L. M. Ghiringhelli, S. V. Levchenko and M. Scheffler, Nat. Commun., 13, 419 (2022).

[6] Fu, X., Wang, J., Hu, X., Xu, W., Levchenko, S. V., & Han, Z.-K. Chem. Commun., 61, 5122–5125, (2025).

[7] M. Cranmer, Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl. arXiv [Astro-Ph.IM]. Retrieved from http://arxiv.org/abs/2305.01582 (2023).

[8] Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., … Persson, K. A. (07 2013). Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. APL Materials, 1(1), 011002. doi:10.1063/1.4812323.

[9] P. Giannozzi, O. Baseggio, P. Bonfà, D. Brunato, R. Car, I. Carnimeo, C. Cavazzoni, S. de Gironcoli, P. Delugas, F. Ferrari Ruffino, A. Ferretti, N. Marzari, I. Timrov, A. Urru, S. Baroni; J. Chem. Phys. **152**, 154105 (2020).

[10] X. Hua, X. Chen, W. A. Goddard, Phys. Rev. B **55**, 16103 (1997).