FACULTAD DE INGENIERÍA
CARRERA INGENIERÍA CIVIL INFORMÁTICA

# Evaluación de Modelos de Aprendizaje Automático Para el Diseño de Catalizadores Para la Reacción de Evolución de Hidrógeno Verde

Tesis para optar al Título de Ingeniera Civil Informática

Director: Dr. Álvaro Rafael Muñoz Castro
Codirector: Dr. Alfredo Jesús Pereira Toloza
**Estudiante: Constanza Isidora Concha Casas**

Santiago - Chile
2025

# HOJA DE CALIFICACIÓN

En _____, el_____ de. _____ de _____ los abajo firmantes dejan constancia que la estudiante *Constanza Concha Casas* de la carrera o programa de *Ingeniería Civil Informática* ha aprobado la tesis para optar al título o grado académico de *Ingeniera Civil Informática* con una nota de _____.

_____
Profesor Evaluador

_____
Profesor Evaluador

_____
Profesor Evaluador

## DEDICATORIA

Dedico este trabajo primeramente a Dios y a mi madre, padre y hermana, por su amor y apoyo incondicional.

También dedico este trabajo a mi novio, quien desde que conocí me ha animado en mis momentos más complicados.

Finalmente, a mi profesor tutor, por su inigualable ánimo y apoyo, además de todo el conocimiento que me otorgó.

## AGRADECIMIENTOS

Agradezco primeramente a Dios por llegar hasta acá.

A mi profesor tutor, por el apoyo y el ánimo que, él sin saberlo, yo tanto necesitaba.

A mi profesor co-tutor Alfredo, que compartió sus conocimientos conmigo y me guió mucho, sobre todo en términos informáticos.

A todos mis profesores que tuve a lo largo de esta carrera en la universidad. Cada uno me ha otorgado distintos conocimientos que puedo aprovechar en el presente.

A mi directora de carrera, que me ha ayudado y apoyado en todo este proceso.

A mi familia y novio, que se alegran con cada logro que tengo.

# TABLA DE CONTENIDO

# ÍNDICE DE FIGURAS

## RESUMEN

El hidrógeno verde (H2V) se presenta como una solución energética sustentable para reducir emisiones de gases de efecto invernadero, donde Chile destaca por su alto potencial en energías renovables y costos de producción. En este contexto, la optimización de catalizadores para la reacción de evolución de hidrógeno (HER) es fundamental, especialmente ante el alto costo y escasez de materiales nobles como el platino. Como alternativa, los boruros bidimensionales de metales de transición (MBenes) surgen como candidatos prometedores. En este trabajo se utilizan modelos de aprendizaje automático entrenados con descriptores químicos derivados de simulaciones computacionales sobre 180 sistemas MBene, incluyendo estructuras dopadas, para predecir la energía libre de adsorción de hidrógeno ($\Delta G_H$) y así identificar catalizadores potencialmente eficientes y económicos. La investigación reproduce la metodología de Sun et al. (2020) e incorpora cuatro algoritmos (SVR, Random Forest, XGBoost y MLP). Los resultados indican que XGBoost y Random Forest son los modelos con mejor desempeño, alcanzando altos valores de $R^2$ y bajos RMSE en el conjunto de prueba, mientras que SVR y, especialmente, MLP muestran un rendimiento inferior. La aplicación de esquemas de evaluación reforzada evidenció además la sensibilidad de XGBoost a la partición de datos y permitió obtener una estimación más realista de su capacidad de generalización. En conjunto, el estudio demuestra que es posible predecir con buena precisión $\Delta G_H$ en MBenes mediante aprendizaje automático y resalta la importancia de utilizar evaluaciones robustas basadas en múltiples particiones. Como proyección, se propone ampliar el conjunto de datos con nuevas simulaciones y/o resultados experimentales, extender la metodología a otras familias de electrocatalizadores y explorar modelos más avanzados e interpretables que apoyen el diseño racional de catalizadores para hidrógeno verde.



## Palabras Clave:

Hidrógeno verde, Reacción de evolución de hidrógeno (HER), Electrocatalizadores, MBenes, Materiales bidimensionales (2D), Aprendizaje automático (Machine Learning), XGBoost, Random Forest, Support Vector Regression (SVR), Energía libre de adsorción de hidrógeno ($\Delta G_H$*).

## ABSTRACT

Green hydrogen (H2V) is presented as a sustainable energy solution to reduce greenhouse gas emissions, with Chile standing out due to its high potential in renewable energies and competitive production costs. In this context, the optimization of catalysts for the hydrogen evolution reaction (HER) is crucial, especially given the high cost and scarcity of noble materials such as platinum. As an alternative, two-dimensional transition-metal borides (MBenes) emerge as promising candidates. In this work, machine learning models are trained with chemical descriptors derived from computational simulations on 180 MBene systems, including doped structures, to predict the free energy of hydrogen adsorption ($\Delta G_H$) and thus identify potentially efficient and cost-effective catalysts. The study reproduces the methodology of Sun et al. (2020) and incorporates four algorithms (SVR, Random Forest, XGBoost, and MLP). The results indicate that XGBoost and Random Forest are the best-performing models, achieving high $R^2$ values and low RMSE on the test set, whereas SVR and, in particular, MLP show lower performance. The application of reinforced evaluation schemes also reveals the sensitivity of XGBoost to data partitioning and enables a more realistic estimate of its generalization capability. Overall, the study demonstrates that $\Delta G_H$ in MBenes can be predicted with good accuracy using machine learning and highlights the importance of employing robust evaluation strategies based on multiple partitions. As a projection, it is proposed to expand the dataset with new simulations and/or experimental results, extend the methodology to other families of electrocatalysts, and explore more advanced and interpretable models to support the rational design of catalysts for green hydrogen production.



### Keywords:

Green hydrogen, Hydrogen evolution reaction (HER), Electrocatalysts, MBenes, Two-dimensional (2D) materials, Machine learning, XGBoost, Random Forest, Support Vector Regression (SVR), Hydrogen adsorption free energy ($\Delta G_H$*).

# INTRODUCCIÓN

El hidrógeno verde (H2V) se posiciona como una fuente energética renovable fundamental para combatir la crisis climática global, ofreciendo una alternativa limpia que contribuye a la reducción significativa de las emisiones de gases de efecto invernadero. A diferencia de la producción tradicional de hidrógeno, que depende en gran medida de fuentes fósiles no renovables, el H2V se genera exclusivamente a partir de energías renovables, lo que garantiza un suministro sostenible. Chile destaca en este ámbito gracias a su abundante potencial en energías solar, eólica e hidráulica, con una capacidad energética superior a 1.800 gigavatios (GW), muy por encima de su demanda interna (Arenas et al., 2024). Esta ventaja competitiva se traduce en costos de producción estimados alrededor de 1 USD/kg para 2030, situando al país en una posición favorable frente a competidores internacionales (Ministerio de Energía & Barhorst, 2016), (Acosta et al., 2022).

Actualmente, Chile cuenta con cinco proyectos de hidrógeno verde en evaluación ambiental, con inversiones aproximadas de 15.000 millones de dólares y una capacidad productiva que superaría las 300 kilotoneladas anuales (Arenas et al., 2024). Estos proyectos están estratégicamente ubicados en regiones como Magallanes, Antofagasta y Valparaíso, con inicio previsto entre 2025 y 2026. La demanda nacional e internacional de hidrógeno verde sigue en aumento, impulsada por la necesidad de descarbonizar sectores clave como transporte, minería e industria, en línea con la meta de carbono-neutralidad para 2050 (Benavides et al., 2021). Aunque el desarrollo del H2V presenta oportunidades económicas significativas, también enfrenta desafíos relacionados con la incertidumbre y la necesidad de políticas públicas sólidas que apoyen su consolidación en un contexto global competitivo. (Acosta et al., 2022).

En este sentido, realizar una búsqueda optima y acabada sobre posibles futuros materiales con el objetivo de abaratar costos de producción de hidrogeno verde, enfocándose en la obtención de catalizadores que permitan promover la reacción de evolución de hidrogeno (Hydrogen Evolution Reaction, HER, por sus siglas en ingles); reacción clave en el desarrollo de tecnologías en base a la producción de hidrogeno por medio de fuentes de energía renovables (Ferriday et al., 2021), (Qadeer et al., 2024). La producción de hidrógeno mediante electrolizadores de agua a partir de fuentes de energía renovables, representa una estrategia clave para la generación de electricidad sustentable, hacia una aplicación cotidiana de H2V. Este objetivo plantea desafíos fundamentales sobre los factores que controlan el desarrollo de la HER en condiciones reales. En este contexto, el diseño de materiales tomando en cuenta los principios básicos de la HER, permiten acceder a las propuestas concretas para lograr este objetivo. Para ello, se requiere involucrar descriptores químicos para explicar los factores relevantes que afecten al desarrollo de la HER, los cuales involucran el uso de técnicas avanzadas como la espectroscopía, simulaciones moleculares y enfoques químicos, que pueden dirigir el diseño de catalizadores prácticos y económicos. Estas estrategias no solo son relevantes para la HER, sino también para el desarrollo de dispositivos electroquímicos de almacenamiento energético que emplean electrolitos acuosos, ampliando así el impacto de esta investigación en el campo de la energía renovable.

Generalmente, se hace uso de materiales basados en platino (Pt), entre otros, los cuales son reconocidos como los catalizadores más efectivos para la reacción de evolución de hidrógeno (HER) debido a su bajo requerimiento en energía y a una alta velocidad catalítica, pero su alto costo prohibitivo y escasez limitan su aplicación en la vida cotidiana (Shi et al., 2025). Por ello, la búsqueda de materiales similares no metálicos nobles (Pt, Pd, Au) y altamente activos es un desafío crucial. En este contexto, el uso de materiales bidimensionales (2D), como los dicalcogenuros de metales de transición ($MS_2$), derivados de grafeno, entre otros, han

mostrado gran potencial debido a su alta área superficial y estructuras ajustables. Particularmente, el $MoS_2$ destaca como una alternativa prometedora, aunque su actividad está limitada por su baja conductividad eléctrica y escasez de sitios activos (Niu et al., 2021). En este sentido, materiales 2D, como los carburos (MCenes) y nitruros (MNenes) de metales de transición y los boruros (MBenes), han captado atención por sus únicas estructuras y propiedades prometedoras, especialmente los MBenes debido a la influencia del boro en su comportamiento electrónico. Se ha demostrado que MBenes como $Pd_2B$, $Mo_2B_2$ y $Fe_2B_2$ poseen gran potencial para HER gracias a su conductividad metálica y variedad estructural (Bai et al., 2021). Además, el dopaje con átomos metálicos individuales mejora la actividad catalítica al generar más sitios activos y nuevos estados electrónicos, aunque la selección óptima de dopantes mediante métodos tradicionales es costosa. La combinación de aprendizaje automático (Machine Learning, ML) surge como una herramienta eficaz para acelerar el diseño y selección de catalizadores HER basados en MBenes dopados, permitiendo predecir rápidamente la variación de la energía libre de Gibbs requerida para que ocurra el proceso de la reacción HER, como un indicador clave de actividad catalítica.

En este trabajo, se presenta una investigación cuantitativa, de carácter predictivo con el fin de ampliar un modelo de aprendizaje automático diseñado para predecir la energía libre de adsorción de hidrógeno ($\Delta G_H$) en materiales tipo MBene empleados como catalizadores en la reacción de evolución de hidrógeno (HER) (Li et al., 2023). A través de simulaciones computacionales, análisis de datos y técnicas de aprendizaje supervisado, se busca desarrollar modelos de regresión capaces de representar con precisión la relación entre descriptores fisicoquímicos y los valores de $\Delta G_H$ obtenidos mediante cálculos computacionales para estos sistemas. La metodología se estructura en tres etapas principales: recolección y preparación de datos, caracterización y estandarización de descriptores, y entrenamiento y evaluación de modelos. El conjunto de datos comprende 180 sistemas MBene (Sun et al., 2020), tanto puros como dopados con átomos metálicos individuales, con diversas propiedades estructurales y electrónicas. Para garantizar un desempeño robusto, los datos se dividen en subconjuntos de entrenamiento y prueba con particiones aleatorias controladas, y se aplican esquemas avanzados de validación, especialmente para modelos con tendencia al sobreajuste como XGBoost. Este enfoque no solo facilita una predicción fiable de la actividad catalítica, sino que también acelera el descubrimiento y optimización de catalizadores HER económicos y eficientes. La integración del aprendizaje automático con conocimientos teóricos contribuye al avance de tecnologías sostenibles para la producción de hidrógeno mediante un diseño informado de catalizadores.

Por lo cual, se establecen los siguientes objetivos:

## Objetivo general

Crear modelos de Machine Learning (aprendizaje automático) para predecir el $\Delta G_H$ dado un conjunto de moléculas generadas previamente.

## Objetivos específicos

**Objetivo 1:** Estandarizar y preparar el conjunto de datos de MBenes para su uso en ML.

**Objetivo 2**: Entrenar modelos de aprendizaje automático para predecir $\Delta G_H$.

**Objetivo 3:** Evaluar y comparar el desempeño de los modelos entrenados.

# REVISIÓN BIBLIOGRÁFICA

## Fundamentos conceptuales

### Hidrógeno verde y reacción de evolución de hidrógeno

El hidrógeno verde (H2V) se ha consolidado como un vector energético importante en estrategias de descarbonización, pues permite almacenar energía proveniente de fuentes renovables y utilizarla posteriormente en procesos industriales, transporte y generación eléctrica sin emisiones directas de $CO_2$. La producción de H2V se basa principalmente en la electrólisis del agua alimentada con electricidad renovable, lo que la diferencia de rutas convencionales dependientes de combustibles fósiles (Lee et al., 2025), (Khalafallah et al., 2025)

En un electrolizador, la reacción global se descompone en la reacción de evolución de oxígeno (OER) en el ánodo y la reacción de evolución de hidrógeno (HER) en el cátodo. La HER determina en gran medida la eficiencia global del proceso, ya que requiere catalizadores capaces de disminuir el sobrepotencial y mantener corrientes elevadas de forma estable. Tradicionalmente, el platino ha sido el material de referencia por su actividad casi óptima, pero su escasez y alto costo limitan su uso a gran escala, lo que ha impulsado la búsqueda de catalizadores alternativos basados en metales abundantes y materiales no nobles (C. Zhang et al., 2025).

Desde el punto de vista teórico, un descriptor central de la actividad hacia HER es la energía libre de adsorción de hidrógeno ($\Delta G_H$). De acuerdo con el principio de Sabatier, valores de $\Delta G_H$ cercanos a 0.0 eV indican un equilibrio adecuado entre adsorción y desorción: si la adsorción es demasiado débil, el hidrógeno no se fija al catalizador; si es demasiado fuerte, no se libera con facilidad. Por ello, gran parte de la literatura utiliza $\Delta G_H$ como variable objetivo a predecir y optimizar (Sun et al., 2020), (Yin et al., 2025).

### Materiales de baja dimensionalidad como electrocatalizadores

El desarrollo de catalizadores alternativos al platino ha puesto especial atención en materiales de baja dimensionalidad (0D, 1D y 2D), como nanopartículas, nanotubos y láminas atómicamente delgadas. Estos materiales presentan alta relación superficie/volumen, gran densidad de sitios activos y estructuras electrónicas ajustables mediante dopaje, defectos o formación de heteroestructuras (Li et al., 2023).

Dentro de este grupo destacan los MXenes y sus análogos ricos en boro conocidos como MBenes, compuestos por capas metálicas y capas de carbono o boro, con alta conductividad eléctrica y una química superficial versátil, considerados como materiales de 2D. La posibilidad de dopar con átomos metálicos aislados, modificar terminaciones superficiales o introducir vacancias genera un espacio de diseño muy amplio para optimizar la actividad catalítica hacia HER (J. Zhang et al., 2023).

Otros trabajos han estudiado heteroestructuras 2D más complejas, por ejemplo sistemas basados en g-$C_3N_4$ acoplado a dicalcogenuros de metales de transición ($MX_2$), donde el intercalado de átomos metálicos modula la estructura electrónica y la energía de adsorción de hidrógeno. Estudios recientes han mostrado que, combinando cálculos de estructura electrónica y modelos de aprendizaje automático, es posible identificar configuraciones intercaladas con $\Delta G_H$ cercanas a cero y actividad comparable a catalizadores nobles (Jyothirmai et al., 2024).

### Cálculos DFT y descriptores para HER

La Teoría del Funcional de la Densidad (DFT) es la herramienta estándar para estudiar la actividad electrocatalítica a nivel atómico. A partir de DFT se obtienen energías de adsorción,

estructuras optimizadas, densidades de estados y otros parámetros que permiten estimar la estabilidad y la reactividad de los sitios activos (Li et al., 2023).

En el contexto de HER, además de $\Delta G_H$, se han propuesto numerosos descriptores fisicoquímicos que capturan la relación entre estructura y actividad, tales como:

- carga de Bader en átomos metálicos y dopantes,
- energía cohesiva y energía de formación,
- centro de banda d y parámetros de red,
- propiedades elementales (radio atómico, electronegatividad, electrones de valencia, afinidad electrónica, energías de ionización, etc.).

Estos descriptores se discuten de forma sistemática en diversas revisiones recientes y estudios de caso sobre catalizadores para HER (Yin et al., 2025), (Li et al., 2023), (J. Zhang et al., 2023).

Estos descriptores pueden obtenerse directamente de cálculos DFT o de bases de datos de propiedades atómicas, y constituyen la base de entrada para los modelos de aprendizaje automático cuyo objetivo es predecir $\Delta G_H$ u otras magnitudes de interés.

**Aprendizaje automático aplicado a electrocatalizadores**

El aprendizaje automático (Machine Learning, ML, por sus siglas en inglés) se ha convertido en un componente central de los flujos de trabajo para el diseño de materiales. En HER, su objetivo principal es aprender la relación entre un vector de descriptores y propiedades como $\Delta G_H$, sobrepotencial o estabilidad, a partir de un conjunto de datos generado por DFT o por experimentos (C. Zhang et al., 2025), (Ram et al., 2025).

Modelos clásicos como Regresión Lineal Múltiple, Máquinas de Soporte Vectorial (SVM/SVR), Árboles de Decisión, Random Forest, Gradient Boosting y XGBoost se han utilizado de forma extensiva en la predicción de propiedades de catalizadores, aprovechando su capacidad para manejar relaciones altamente no lineales y espacios de características de alta dimensión. Paralelamente, se han explorado redes neuronales profundas y enfoques basados en graph neural networks para representar de manera más directa la estructura atómica (Khalafallah et al., 2025), (C. Zhang et al., 2025), (Jyothirmai et al., 2024).

Las revisiones recientes resaltan dos beneficios principales del ML en este contexto: (i) reduce drásticamente el número de cálculos DFT necesarios para explorar un espacio de materiales, y (ii) permite identificar descriptores clave mediante análisis de importancia de variables, SHAP u otras técnicas interpretables, aportando comprensión física adicional además de capacidad predictiva (Khalafallah et al., 2025).

## Estado del arte en el uso de aprendizaje automático para HER

**Revisiones recientes y tendencias generales**

En los últimos años han aparecido varias revisiones que sistematizan el uso de ML en electrocatalizadores para HER (Lee et al., 2025). Zhang y colaboradores describen un flujo de trabajo general que incluye: recopilación y curado de datos, selección de descriptores, entrenamiento de modelos supervisados y exploración del espacio de materiales mediante predicciones sobre estructuras no calculadas aún por DFT (J. Zhang et al., 2023).

Por su parte, otros autores se enfocan específicamente en materiales de baja dimensionalidad, destacando que la combinación de ML con cálculos de alto rendimiento permite evaluar miles de candidatos (MXenes, heteroestructuras 2D, catalizadores de un solo átomo) y filtrar aquellos con mejores $\Delta G_H$ y estabilidad (Yin et al., 2025), (Khalafallah et al., 2025), (Jyothirmai et al., 2024) (J. Zhang et al., 2023). Estas revisiones coinciden en que el ML no sólo acelera el cribado de nuevos catalizadores, sino que también facilita la construcción de

descriptores universales que explican tendencias de actividad a lo largo de familias completas de materiales.

Además, se ha observado un uso creciente de estrategias de ensemble learning (como Random Forest y XGBoost) y de técnicas de selección de características para mejorar la precisión y, especialmente, la robustez de los modelos cuando el número de muestras es limitado, como suele ocurrir en bases de datos derivadas de DFT (Khalafallah et al., 2025), (C. Zhang et al., 2025), (Jyothirmai et al., 2024).

**Cribado de alto rendimiento asistido por ML**

La integración de ML con esquemas de cribado de alto rendimiento (high-throughput screening, en inglés) ha demostrado ser una herramienta muy eficaz para explorar espacios de diseño grandes. En estos trabajos se realiza primero un conjunto de cálculos DFT sobre un subconjunto de materiales, se entrena un modelo de regresión y luego se utilizan sus predicciones para priorizar nuevos candidatos para cálculos adicionales (Yin et al., 2025), (Ram et al., 2025).

Por ejemplo, estudios recientes sobre heteroestructuras g-$C_3N_4$/$MX_2$ intercaladas con metales de transición (Jyothirmai et al., 2024) han construido conjuntos de cientos de configuraciones con diferentes sitios de adsorción de hidrógeno. Una fracción de estas configuraciones se evalúa mediante DFT y se emplea para entrenar un modelo de Random Forest, que luego predice $\Delta G_H$ en el resto del espacio, identificando rápidamente combinaciones de metal y sustrato con actividad teórica sobresaliente.

De manera similar, revisiones sobre MXenes muestran cómo la combinación de DFT y algoritmos de boosting ha permitido seleccionar, a partir de miles de candidatos, decenas de estructuras 2D con $|\Delta G_H|$ por debajo de 0,2 eV, algunas de las cuales superan incluso la actividad prevista para Pt en términos de estabilidad y energía de adsorción (J. Zhang et al., 2023), (Yin et al., 2025).

Estos enfoques reflejan cómo se articula el estado del arte mediante un flujo integrado que combina DFT, descriptores, modelos de ML y estrategias para explorar el espacio de materiales.

**Aplicaciones específicas a MXenes y MBenes: artículo de referencia**

Dentro de este panorama general, los materiales MXene y MBene han recibido atención particular como plataformas 2D para HER. Se han reportado numerosos estudios que utilizan ML para analizar el efecto del dopaje con metales de transición, la presencia de terminaciones superficiales y la sustitución de elementos en la capa metálica sobre $\Delta G_H$ y la estabilidad (Yin et al., 2025), (Khalafallah et al., 2025), (Ding et al., 2024).

Entre estos trabajos destaca el estudio de Sun et al. (2020), que constituye el artículo de referencia de esta tesis. En dicho trabajo, los autores combinan cálculos DFT y aprendizaje automático para realizar un cribado acelerado de catalizadores HER en materiales MBene basados en boro. A partir de un conjunto de estructuras puras y dopadas con un solo átomo metálico, calculan energías de adsorción y un conjunto amplio de descriptores estructurales y elementales, incluyendo cargas de Bader, energías cohesivas, centros de banda d y propiedades atómicas de los metales implicados.

Sobre esta base, entrenan varios modelos supervisados (entre ellos Support Vector Regression y Random Forest) para predecir $\Delta G_H$ a partir de los descriptores. Los resultados muestran que los modelos basados en árboles, especialmente Random Forest, logran errores de prueba del orden de 0,27 eV y permiten identificar materiales como $Co_2B_2$ y Mn/$Co_2B_2$ con valores de $\Delta G_H$ cercanos a cero en un amplio rango de coberturas de hidrógeno, lo que los posiciona como candidatos particularmente prometedores.

Más allá de la predicción, Sun et al. (2020) analizan la importancia relativa de los descriptores y concluyen que la variación de la carga de Bader y del centro de banda d en los átomos metálicos juega un papel determinante en la actividad catalítica. Esta combinación de cribado acelerado, interpretabilidad y foco en MBenes dopados sitúa al trabajo como una contribución central dentro del estado del arte, y proporciona el conjunto de datos y la selección de descriptores que se adoptan en el presente estudio.

## Síntesis del estado del arte y aporte del presente trabajo

La literatura revisada muestra que la integración de DFT y aprendizaje automático ha permitido avanzar rápidamente en el diseño de electrocatalizadores para HER, en particular en sistemas de baja dimensionalidad como MXenes, MBenes y heteroestructuras 2D. Existen flujos de trabajo bien establecidos para: construir conjuntos de datos a partir de simulaciones de alto costo, definir descriptores fisicoquímicos relevantes, entrenar modelos de regresión (SVR, RFR, XGBoost, redes neuronales, entre otros) y explorar de manera sistemática grandes espacios de materiales (C. Zhang et al., 2025), (Ram et al., 2025).

Sin embargo, también se identifican brechas. Muchos trabajos, incluido el propio estudio de Sun et al. (2020), se centran en demostrar el rendimiento de uno o dos algoritmos sobre un conjunto de datos dado, pero profundizan menos en la comparación sistemática entre modelos bajo un mismo esquema de validación. Asimismo, en conjuntos de tamaño moderado (del orden de cientos de muestras), la evaluación suele basarse en una o pocas particiones del conjunto de datos, lo que puede subestimar la variabilidad estadística y conducir a estimaciones optimistas del desempeño real en generalización (Yin et al., 2025), (Ding et al., 2024).

En este contexto, el presente trabajo de título se posiciona como una contribución principalmente metodológica dentro del estado del arte. Por un lado, replica la metodología del artículo de referencia de Sun et al. (2020) para los modelos de Random Forest y SVR, utilizando el mismo conjunto de datos de 180 sistemas MBene y los mismos descriptores estandarizados derivados de DFT, lo que asegura una comparación directa con los resultados originales. Por otro lado, extiende el análisis incorporando dos modelos adicionales de uso extendido en ciencia de datos, XGBoost y un perceptrón multicapa (MLP), ambos ampliamente utilizados en regresión sobre conjuntos de datos tabulares y en predicción de propiedades de materiales (Chen & Guestrin, 2016), (Fu et al., 2022), (Lu et al., 2024). Asimismo, se aplica a XGBoost un esquema de evaluación reforzado basado en múltiples particiones aleatorias y estadísticos resumen (promedios y desviaciones estándar de las métricas), en línea con las recomendaciones metodológicas sobre validación cruzada repetida y estrategias de evaluación robusta para evitar conclusiones demasiado optimistas basadas en una sola partición de entrenamiento/prueba (Kohavi, 2000), (Sweet et al., 2023).

# METODOLOGÍA

La metodología de este trabajo se diseña como un estudio cuantitativo, de carácter no experimental y alcance explicativo, basado en la replicación y extensión del modelo de aprendizaje automático propuesto en el artículo de referencia (Sun et al., 2020), cuyo objetivo es predecir la energía libre de adsorción del hidrógeno ($\Delta G_H$) en materiales tipo MBene utilizados como catalizadores de la reacción de evolución de hidrógeno (HER). La investigación se desarrolla íntegramente mediante simulación computacional, análisis de datos y aplicación de técnicas de Machine Learning.

En términos de diseño, el estudio utiliza aprendizaje supervisado, ya que la variable objetivo $\Delta G_H$ se encuentra disponible para cada muestra en el conjunto de datos. La meta principal es entrenar modelos de regresión capaces de aproximar la relación entre los descriptores fisicoquímicos (características de entrada) y el valor objetivo (etiqueta), para así guiar en un futuro el diseño de materiales para su uso en H2V. La metodología se estructura en tres etapas principales: (1) recolección y preparación de los datos, (2) caracterización y estandarización de los descriptores, y (3) entrenamiento y evaluación de modelos de aprendizaje automático.

## Área de estudio y diseño de la investigación

El área de estudio corresponde al diseño, entrenamiento y evaluación de modelos de regresión supervisada para predecir $\Delta G_H$ a partir de descriptores de materiales MBenes, tanto puros como dopados con átomos metálicos individuales. El conjunto de datos y las definiciones de las variables provienen directamente del artículo de referencia, donde $\Delta G_H$ se calcula mediante Teoría del Funcional de la Densidad (DFT).

El estudio se clasifica como:

- Cuantitativo: se trabaja con variables numéricas continuas y métricas estadísticas para evaluar el rendimiento de los modelos.
- No experimental: no se manipulan materiales en laboratorio; se emplean datos ya generados mediante simulaciones DFT.
- Correlacional–explicativo: se busca determinar qué modelos capturan mejor la relación entre los descriptores y $\Delta G_H$.

Cada sistema MBene constituye una unidad de análisis individual, representada por un vector de descriptores y un valor escalar $\Delta G_H$.

## Recolección de datos

### Origen del conjunto de datos

El conjunto de datos empleado se basa en el set de datos publicado en el artículo de referencia, compuesto por 180 sistemas MBene (puros y dopados con un solo átomo). En el estudio original, estos sistemas se obtienen a partir de:

- Modelación DFT de estructuras $M_2B_1$ y $M_2B_2$ para metales de las series 3d, 4d y 5d.
- Cálculo de $\Delta G_H$ combinando energía de adsorción, energía de punto cero y contribución entrópica.
- Obtención de descriptores estructurales y elementales como energías cohesivas, cargas de Bader, centro de banda *d*, radios atómicos y propiedades electrónicas.

El repositorio oficial de los autores incluye el archivo (ML_Figure5.xlsx) que contiene los descriptores originales. Este archivo presenta características con escalas heterogéneas, lo cual perjudica el rendimiento de los modelos. Por ello, se realizó un proceso de estandarización de todas las variables numéricas, generando un nuevo archivo que constituye la base utilizada en las etapas posteriores.

La lectura y manipulación de este archivo se realizó mediante la librería pandas, separando el conjunto en una matriz de características, y una variable objetivo, correspondiente a $\Delta G_H$.

**División en conjuntos de entrenamiento y prueba**

Para los modelos SVR, RFR y MLP, se replica la metodología del artículo de referencia, dividiendo el conjunto de datos en 75 % para entrenamiento y 25 % para prueba. Esto se realiza mediante partición aleatoria controlada por una semilla fija (42), lo que asegura la reproducibilidad del análisis en diferentes ejecuciones.

Sin embargo, el modelo XGBoost requirió un esquema de validación más detallado. Debido a su mayor capacidad de modelar relaciones complejas y a su propensión al sobreajuste en conjuntos moderados, se estableció una evaluación basada en:

- 50 particiones distintas del conjunto de datos
- con divisiones 70% / 30% para entrenamiento y prueba
- cada una definida por una semilla diferente
- posterior generación de métricas promedio y desviación estándar.

Este procedimiento permitió obtener métricas más confiables y representativas del rendimiento real del modelo, lo cual era fundamental dada la expectativa teórica de que XGBoost fuese el método con mejor capacidad predictiva entre todos los modelos implementados.

## Caracterización y estandarización de los datos

**Descriptores utilizados**

Los descriptores disponibles provienen del artículo de referencia (Sun et al., 2020), donde se combinan propiedades obtenidas mediante DFT con propiedades elementales. Si bien estos descriptores tienen un trasfondo químico importante, en el contexto de esta tesis su uso se limita al ámbito computacional: funcionan como variables predictoras numéricas para los modelos de Machine Learning.

Dentro de este conjunto se distinguen dos categorías:

a) Descriptores calculados mediante DFT

- Energía cohesiva del material
- Carga de Bader del metal superficial y del átomo dopante
- Centro de banda d
- Parámetros de red
- Longitudes de enlace metal–metal y metal–boro

b) Descriptores elementales

- Masa atómica, radio atómico y electrones de valencia

- Electronegatividad
- Afinidad electrónica
- Primera energía de ionización

Estos descriptores constituyen una representación suficiente para capturar relaciones químicas relevantes sin necesidad de profundizar en análisis desde la química de materiales, ya que el propósito de este trabajo se sitúa en identificar patrones mediante técnicas informáticas.

**Estandarización de características**

El artículo de Sun et al. (2020) establece que los descriptores deben regularizarse antes del entrenamiento para mejorar la estabilidad numérica. En este trabajo se replicó este criterio aplicando estandarización mediante z-score. Con este método, cada variable adquiere media cero y desviación estándar uno, lo que homogeniza las escalas y favorece el desempeño de modelos sensibles (como SVR, MLP y XGBoost). Este procedimiento se realizó mediante la librería scikit-learn.

El proceso consistió en:

1. Identificar todas las columnas numéricas.
2. Ajustar un escalador para calcular media y desviación estándar del conjunto de datos original.
3. Transformar cada descriptor a su versión estandarizada.
4. Conservar la variable objetivo sin modificar.
5. Guardar el resultado en un archivo nuevo para asegurar trazabilidad.

## Entrenamiento y evaluación de modelos de aprendizaje automático

La etapa final consiste en entrenar, optimizar y comparar cuatro modelos supervisados: Random Forest (RFR), Support Vector Regression (SVR), Multilayer Perceptron (MLP) y XGBoost (XGB). La implementación se basa en el uso de scikit-learn y xgboost, en conjunto con técnicas de validación cruzada y evaluación estadística.

**Entorno de implementación**

El desarrollo se llevó a cabo en Python, utilizando pandas y numpy para manipulación de datos; scikit-learn para modelos, preprocesamiento y validación; xgboost para el modelo basado en boosting de gradiente y matplotlib para generación de gráficos.

**Esquema general de entrenamiento**

Para los modelos SVR, RFR y MLP se empleó el siguiente flujo:

1. Optimización mediante GridSearchCV (CV de 10 pliegues): Se definieron hiperparámetros para cada modelo y se seleccionaron los mejores según la métrica $R^2$.
2. Entrenamiento del mejor modelo: Cada modelo resultante se ajusta solamente sobre el conjunto de entrenamiento.
3. Evaluación: Se calculan métricas de rendimiento (RMSE y $R^2$) en entrenamiento y prueba para identificar sobreajuste o subajuste.
4. Visualización mediante gráficos de paridad: Cada modelo genera un gráfico predicción–valor real, siguiendo el estilo del artículo base.

## Modelos replicados del artículo: SVR y RFR

Los modelos SVR y RFR se implementan siguiendo la metodología del estudio original:

- SVR, utilizando Kernel RBF, regularización C y parámetros gamma para capturar relaciones no lineales.
- RFR, empleando múltiples árboles y aleatorización para controlar la varianza del modelo.

Ambos modelos se entrenaron con validación cruzada y posterior evaluación en el conjunto de prueba.

### Extensión metodológica: MLP y XGBoost

*MLP Regressor:*

El modelo MLP se incorpora como una aproximación adicional basada en redes neuronales densas. Se optimizaron:

- arquitectura (capas y neuronas)
- función de activación
- parámetros de regularización L2
- tasa de aprendizaje

Se utilizó entrenamiento con detención temprana para evitar sobreajuste.

*XGBoost Regressor:*

Debido a su alta capacidad predictiva, XGBoost se sometió a un proceso más riguroso:

1. Búsqueda inicial de hiperparámetros (CV 10-fold)
   Se optimizaron parámetros relacionados con profundidad, regularización, submuestreo y tasa de aprendizaje.
2. 50 repeticiones con semillas distintas
   Cada repetición utiliza una nueva partición aleatoria 70/30.
   Esto permite evaluar la estabilidad estadística del modelo.
3. Cálculo de métricas medias y desviaciones estándar
   Se genera un resumen del desempeño medio del modelo.
4. Selección de la mejor repetición para visualización
   El gráfico que se presenta corresponde a la ejecución con $R^2$ más alto en el conjunto de prueba.

Este enfoque proporciona resultados más confiables para un modelo potencialmente superior pero más propenso al sobreajuste.

## Esquema metodológico

El proceso seguido en esta tesis puede resumirse en las siguientes etapas:

1. Obtención del conjunto de datos original.
2. Estandarización de los descriptores.
3. Separación de características y variable objetivo.

4. División del conjunto de datos en entrenamiento y prueba.
5. Entrenamiento de los modelos SVR, RFR y MLP mediante validación cruzada.
6. Implementación de un pipeline único para XGBoost basado en múltiples repeticiones.
7. Evaluación comparativa mediante métricas y gráficos de paridad.

Este diseño metodológico es replicable, sistemático y coherente con las buenas prácticas de aprendizaje automático, además de aportar una contribución original mediante la evaluación reforzada del modelo XGBoost.

# RESULTADOS

## Análisis descriptivo del conjunto de datos

### Distribución de los descriptores estandarizados

La Figura 1 muestra los diagramas de violín correspondientes a las variables empleadas como descriptores en los modelos de aprendizaje automático, separando en la parte superior las características calculadas mediante DFT y en la parte inferior las características elementales. En ambos casos, los valores se representan en términos de puntaje z.

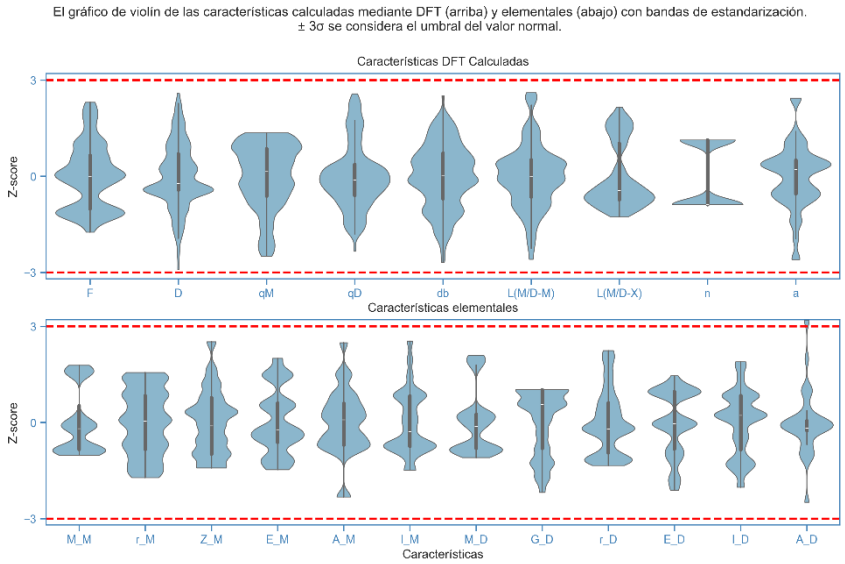


*Figura 1 – Violines.*

Se observa que, para la mayoría de los descriptores, la masa principal de los datos se concentra en el entorno de 0 y dentro del intervalo aproximado $[-3\sigma, +3\sigma]$. Algunas variables presentan colas algo más extendidas, pero en general la estandarización produce distribuciones centradas y comparables en escala entre sí. Esta representación confirma que el conjunto de entrada utilizado en los modelos está previamente normalizado y que no se aprecian valores extremos aislados a gran distancia del resto de las observaciones.

**Correlación entre descriptores**

La Figura 2 presenta el mapa de correlaciones de Pearson entre los descriptores estandarizados, representado como triángulo inferior con círculos coloreados. El tamaño de cada marcador es proporcional al valor absoluto del coeficiente de correlación, mientras que el color indica el signo (correlaciones positivas en tonos rojizos y negativas en tonos azulados).

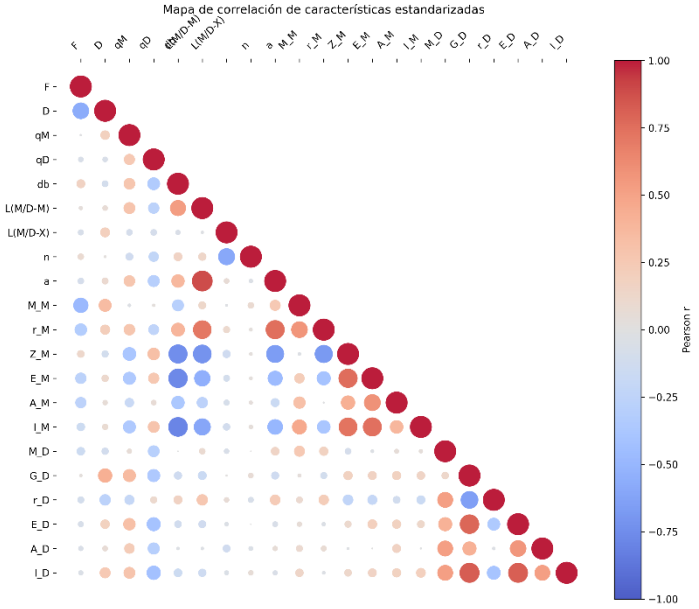


*Figura 2 - Mapa de correlación de Pearson.*

El mapa evidencia la existencia de varios pares de descriptores con correlaciones moderadas o altas, especialmente dentro de grupos de variables que comparten origen físico similar (por ejemplo, parámetros estructurales o propiedades electrónicas de los mismos elementos). En contraste, otros pares muestran marcadores de pequeño tamaño, lo que indica correlaciones próximas a cero y, por tanto, baja redundancia lineal entre esas características. Estos patrones de correlación constituyen el contexto estadístico en el que se entrenan posteriormente los modelos de regresión.

## Resultados de los modelos de regresión

En esta sección se presentan los resultados obtenidos para los distintos modelos supervisados utilizados para predecir la energía libre de adsorción de hidrógeno $\Delta G_{H^*}$ a partir de los descriptores definidos, con el fin de evaluar la espontaneidad del proceso. En todos los casos se emplea una partición del conjunto de datos en entrenamiento y prueba, y se reportan el error cuadrático medio (RMSE) y el coeficiente de determinación $R^2$para ambas particiones. Los gráficos de paridad muestran en el eje horizontal los valores predichos por el modelo y en el eje vertical los valores calculados por DFT, junto con la recta identidad como referencia.

**Modelo MLP**

La Figura 3 corresponde al gráfico de paridad del perceptrón multicapa (MLP). En la esquina superior izquierda se indican las métricas obtenidas: el modelo presenta un RMSE del orden de 0,26 eV en entrenamiento y 0,20 eV en prueba, con valores de $R^2$aproximados de 0,39 y 0,59, respectivamente. Los puntos verdes representan las muestras de entrenamiento y los puntos azules las muestras de prueba.

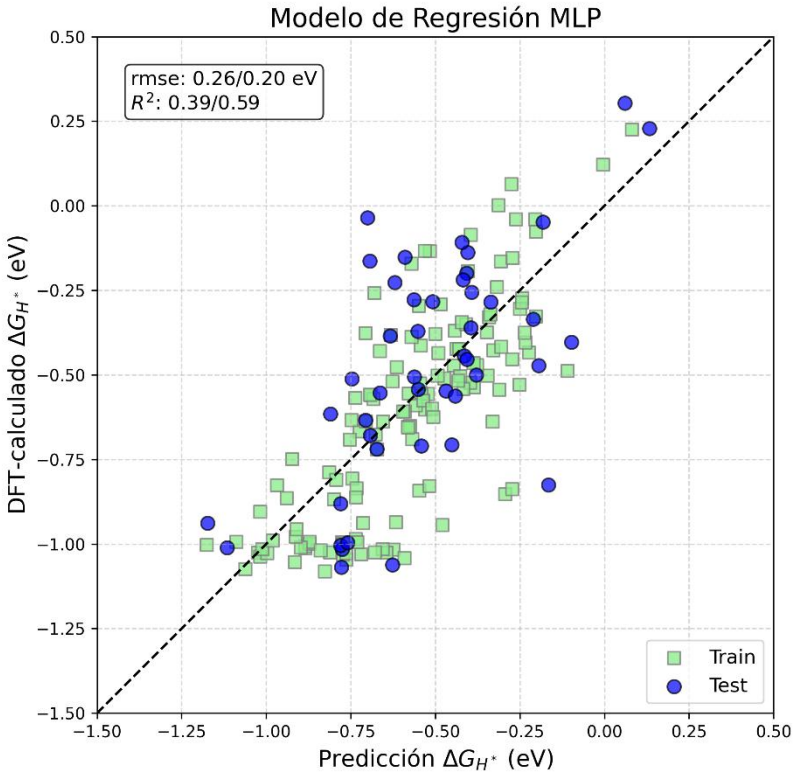


*Figura 3 - Modelo de Regresión MLP.*

En el diagrama (Figura 3) se aprecia una dispersión apreciable alrededor de la recta identidad, tanto en entrenamiento como en prueba. Aun así, la mayoría de las observaciones se sitúan en torno al rango de $\Delta G_{\mathrm{H}^*}$ cubierto por el conjunto de datos original, sin concentraciones anómalas en regiones específicas del eje de predicción.

**Modelo Random Forest Regression**

La Figura 4 muestra el gráfico de paridad correspondiente al modelo de Random Forest Regression (RFR). En este caso, las métricas indicadas en la figura señalan valores de RMSE cercanos a 0,19 eV para el conjunto de entrenamiento y 0,08 eV para el conjunto de prueba, mientras que los coeficientes $R^2$se sitúan aproximadamente en 0,67 (entrenamiento) y 0,94 (prueba).

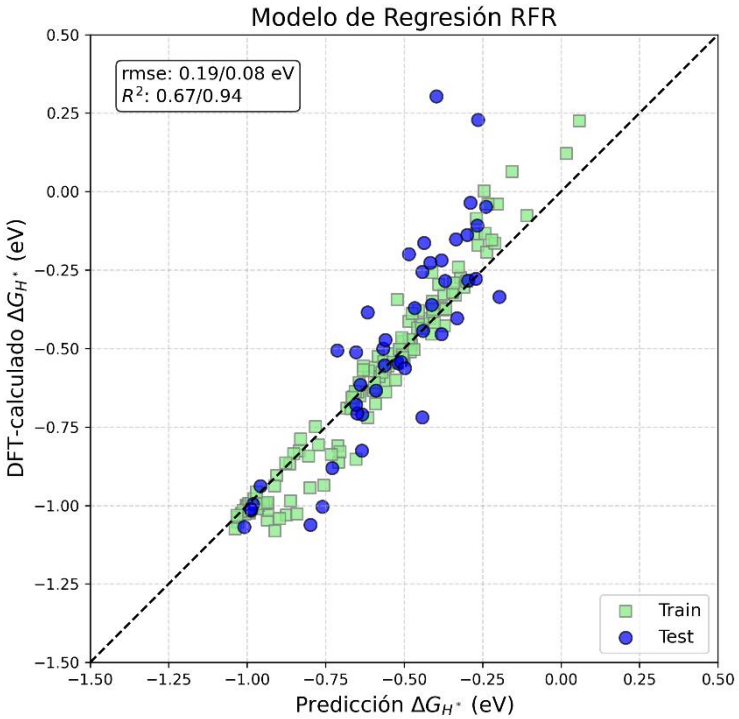


*Figura 4 - Modelo RFR.*

En el diagrama de dispersión (Figura 4) se observa que los puntos de entrenamiento se agrupan en torno a la recta identidad, y las muestras de prueba también se alinean de forma más estrecha con respecto a la referencia en comparación con el caso del MLP. El rango de valores de $\Delta G_{H^*}$ representado es similar al de los otros modelos, y no se aprecian vacíos evidentes en regiones particulares del espacio de predicción.

**Modelo Support Vector Regression**

La Figura 5 presenta el gráfico de paridad para el modelo de Support Vector Regression (SVR). Las métricas mostradas en el recuadro indican un RMSE aproximado de 0,17 eV en entrenamiento y 0,13 eV en prueba, con valores de $R^2$alrededor de 0,75 y 0,83, respectivamente.

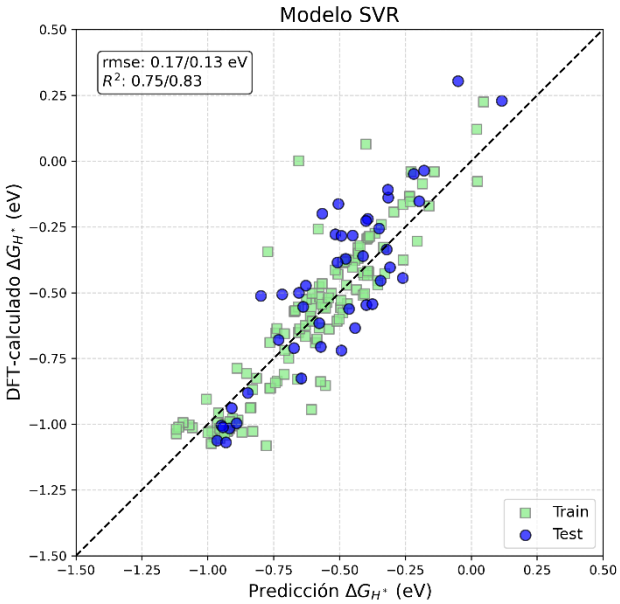


*Figura 5 - Modelo SVR.*

En este caso, tanto las muestras de entrenamiento como las de prueba se distribuyen cercanas a la recta identidad, con una dispersión intermedia si se compara visualmente con los casos de MLP y Random Forest. El modelo SVR mantiene una alineación razonablemente uniforme a lo largo de todo el rango de valores de $\Delta G_{\mathrm{H}^*}$ considerados.

**Modelos XGBoost**

Las Figuras 6–8 recopilan los resultados obtenidos con las distintas configuraciones evaluadas del modelo XGBoost. Cada gráfico corresponde a un ajuste concreto del modelo, manteniendo en todos los casos la misma estructura general: puntos verdes para el conjunto de entrenamiento, puntos azules para el conjunto de prueba y la recta identidad como referencia.

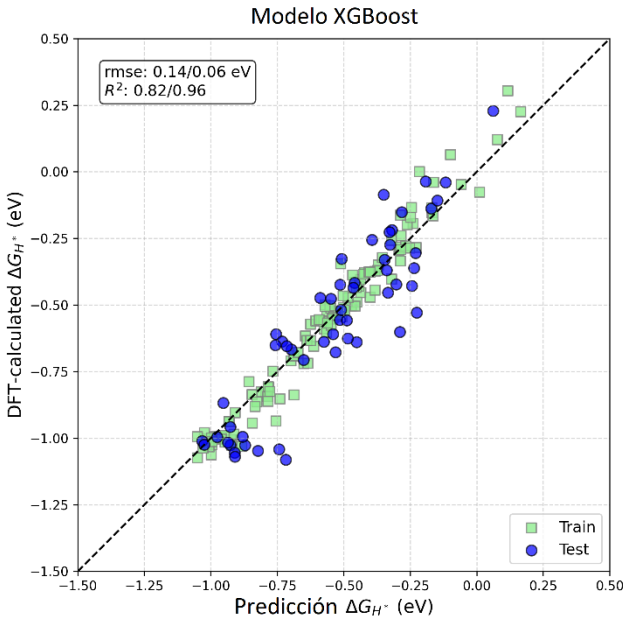


*Figura 6 - Primer Modelo XGBoost.*

En la primera configuración (Figura 6), asociada al modelo obtenido a partir de los hiperparámetros seleccionados mediante validación cruzada de 10 pliegues, las métricas indicadas muestran un RMSE cercano a 0,14 eV para el conjunto de entrenamiento y alrededor de 0,06 eV para el conjunto de prueba, con valores de $R^2$en torno a 0,82 (entrenamiento) y 0,96 (prueba). La nube de puntos aparece concentrada alrededor de la recta identidad, especialmente para las muestras de prueba, y las predicciones cubren todo el rango de valores de $\Delta G_{H^*}$presente en el conjunto de datos.

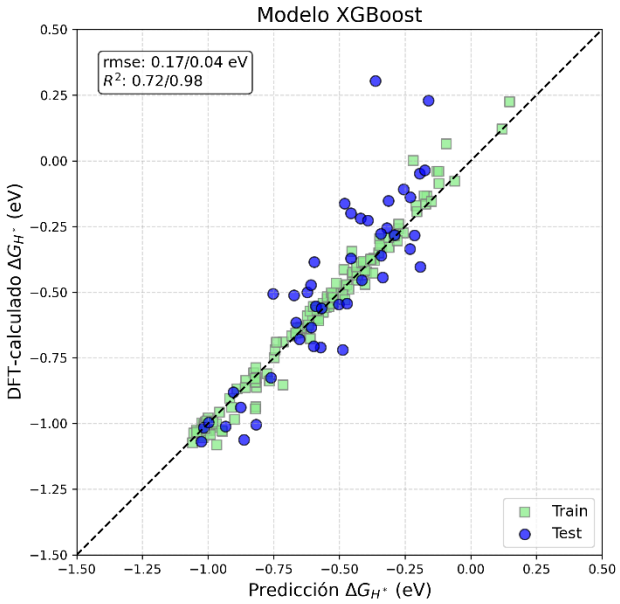


*Figura 7 - Segundo modelo XGBoost.*

En una segunda configuración (Figura 7), asociada a un ajuste alternativo del modelo con una partición distinta del conjunto de datos, las métricas se sitúan en un RMSE aproximado de 0,17 eV para el conjunto de entrenamiento y de 0,04 eV para el conjunto de prueba, con coeficientes $R^2$cercanos a 0,72 y 0,98, respectivamente. En este caso, la dispersión de las muestras de entrenamiento respecto de la recta identidad es baja y la nube de puntos de prueba se mantiene alineada con dicha recta, aunque con una distribución algo diferente a la observada en la primera configuración.

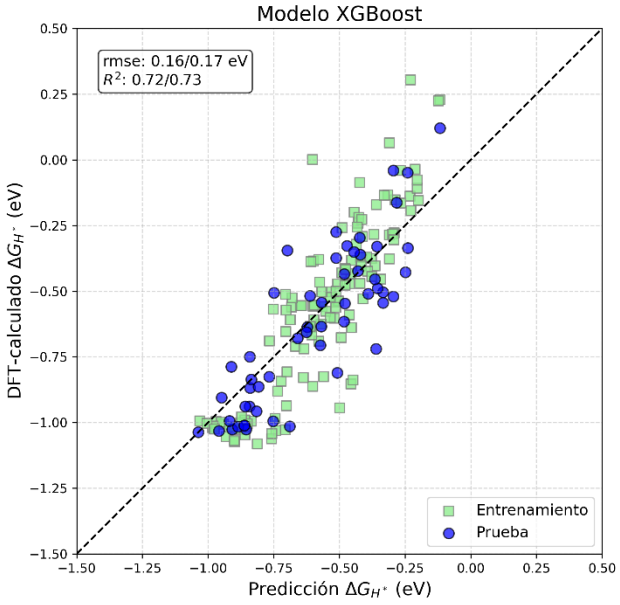


*Figura 8 - Tercer modelo XGBoost.*

Finalmente, la Figura 8 muestra el gráfico de paridad correspondiente a una tercera configuración de XGBoost en la que se emplea un esquema de evaluación reforzado. En esta configuración se generan múltiples particiones aleatorias del conjunto de datos y se selecciona una de las ejecuciones para representar gráficamente los resultados. El recuadro de métricas

indica valores de RMSE del orden de 0,16 eV en entrenamiento y 0,17 eV en prueba, con coeficientes $R^2$próximos a 0,72 y 0,73, respectivamente. De manera complementaria, el resumen estadístico de las 100 repeticiones realizadas muestra valores promedio de $R^2$cercanos a 0,97 en entrenamiento y 0,63 en prueba, lo que cuantifica el comportamiento medio del modelo bajo distintos cortes aleatorios del conjunto de datos.

## Importancia de las características

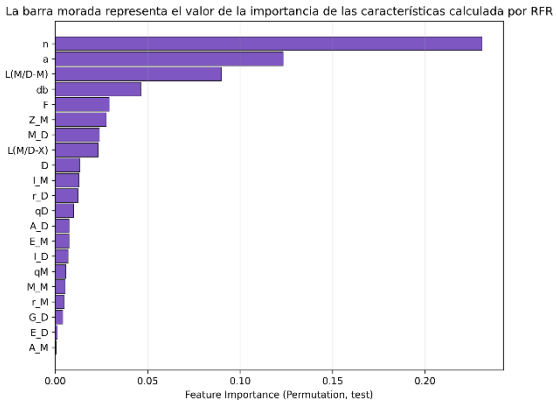


*Figura 9 - Importancia características (extendido).*

La Figura 9 presenta la importancia de las características calculada mediante el modelo de Random Forest a partir de la técnica de "permutation importance" aplicada sobre el conjunto de prueba. Las barras horizontales muestran el descenso promedio en la métrica de desempeño cuando se permuta aleatoriamente cada descriptor, ordenadas de mayor a menor impacto.

En este gráfico se observa que los descriptores etiquetados como n, a y L(M/D–M) se sitúan en las primeras posiciones, con valores de importancia superiores al resto de las variables. A continuación aparecen descriptores como db, F, Z_M y M_D, con valores de importancia intermedios, mientras que el grupo restante de variables presenta contribuciones progresivamente menores.

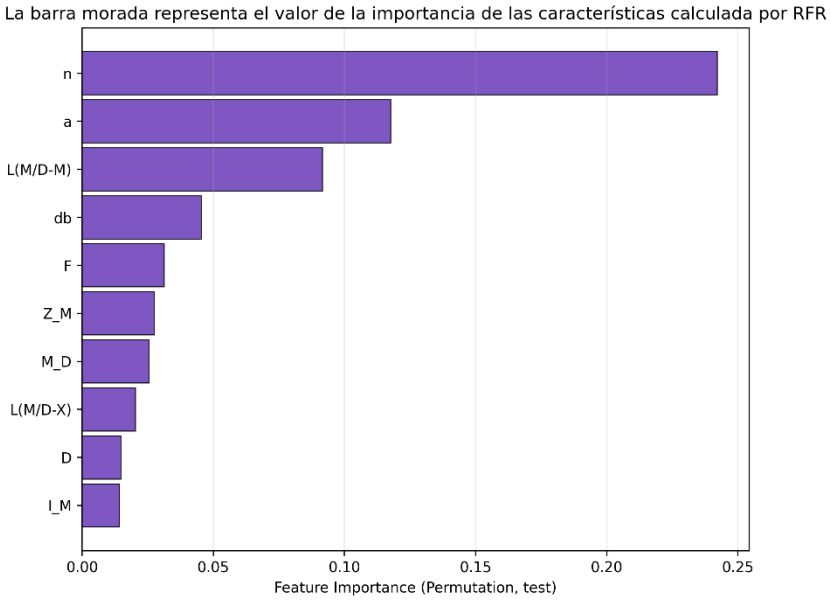


*Figura 10 - Importancia características (top 10).*

La Figura 10 muestra una versión reducida del mismo análisis, en la que se representan únicamente los diez descriptores con mayor importancia según la métrica de permutación. Esta representación resume visualmente qué variables ejercen el mayor efecto sobre las predicciones de $\Delta G_{H^*}$ dentro del modelo de Random Forest, manteniendo el mismo orden relativo observado en la figura completa.

## DISCUSIONES (con la literatura y análisis)

Los resultados obtenidos confirman, en primer lugar, que los algoritmos basados en árboles y márgenes máximos se adaptan mejor al tamaño y naturaleza del conjunto de datos que el perceptrón multicapa. Mientras MLP alcanza un desempeño moderado, Random Forest, SVR y especialmente XGBoost logran errores claramente menores y coeficientes de determinación más altos. Esto es consistente con el estado del arte, donde se destaca que, para conjuntos de datos de algunas centenas de muestras, los modelos de ensamble y los métodos kernel suelen superar a las redes neuronales densas, que requieren volúmenes de datos mayores para explotar su capacidad representacional (Khalafallah et al., 2025), (Li et al., 2023), (C. Zhang et al., 2025).

Dentro de este grupo, los resultados sitúan a XGBoost como el modelo con mejor capacidad predictiva puntual, ya que en la configuración seleccionada mediante validación cruzada alcanza los valores más altos de $R^2$ en el conjunto de prueba. Sin embargo, el análisis con múltiples semillas revela que este rendimiento es sensible a la partición entrenamiento/prueba: al promediar sobre distintas divisiones del conjunto de datos, el desempeño de XGBoost se acerca al observado en Random Forest y SVR y aparece un desfase claro entre el ajuste en entrenamiento y en prueba. Esto indica que la evaluación basada en una única partición puede conducir a una visión demasiado optimista de la capacidad de generalización del modelo (Khalafallah et al., 2025), (Ram et al., 2025).

Desde una perspectiva metodológica, el esquema de repeticiones aplicado a XGBoost muestra ser una herramienta útil para cuantificar esta variabilidad y aporta una visión más honesta del comportamiento del algoritmo. En consecuencia, aunque la configuración "óptima" de XGBoost entrega las mejores métricas en un caso concreto, la configuración evaluada mediante repeticiones se interpreta como el mejor modelo de XGBoost en sentido robusto, y es la que resulta más adecuada para comparar con Random Forest y SVR (Sun et al., 2020).

# CONCLUSIONES

En relación con el Objetivo 1 (crear/obtener el set de datos), la tesis logra reconstruir de forma consistente el conjunto de 180 sistemas MBene descrito en el estudio de referencia, incorporando los mismos descriptores estructurales y elementales en formato estandarizado. Respecto del Objetivo 2 (entrenar modelos), se implementan y ajustan cuatro algoritmos de aprendizaje supervisado —SVR, Random Forest, MLP y XGBoost— siguiendo prácticas actuales de selección de hiperparámetros y validación. Finalmente, en cuanto al Objetivo 3 (evaluar modelos), se obtienen y comparan métricas de error y coeficientes de determinación en entrenamiento y prueba, lo que permite identificar a XGBoost, Random Forest y SVR como los modelos más competitivos, dejando al MLP como alternativa menos adecuada para el tamaño de muestra disponible.

En conjunto, estos resultados cumplen el objetivo general de la tesis: se construyen modelos de Machine Learning capaces de predecir $\Delta G_{H^*}$en la familia de MBenes estudiada, y se demuestra que su desempeño es comparable e incluso superior al de los modelos originalmente reportados para este conjunto de datos. La investigación responde así a la pregunta planteada, mostrando que es posible reproducir la metodología previa y, al mismo tiempo, extenderla mediante la incorporación de XGBoost y de un esquema de evaluación más exigente basado en repeticiones.

La principal limitación del trabajo es el tamaño y alcance del conjunto de datos: se dispone de solo 180 estructuras pertenecientes a una misma familia de materiales y todas las etiquetas provienen de cálculos DFT. Esto restringe la generalización de los modelos a otros tipos de electrocatalizadores y hace que cualquier sesgo asociado a las simulaciones se traslade directamente a las predicciones. Además, el análisis con repeticiones se aplica en detalle solo a XGBoost; extender el mismo procedimiento a SVR y Random Forest permitiría una comparación aún más equilibrada entre algoritmos.

A partir de estos resultados se abren varias oportunidades de trabajo futuro. Entre ellas, ampliar el conjunto de datos combinando nuevas simulaciones DFT y datos experimentales, aplicar esquemas de validación cruzada anidada a todos los modelos, incorporar técnicas de interpretabilidad avanzadas y explorar arquitecturas específicas para materiales basadas en grafos. Estas extensiones permitirían mejorar la capacidad de generalización de los modelos, trasladar el enfoque a otras familias de catalizadores para HER y consolidar el uso de aprendizaje automático como herramienta de apoyo en el diseño computacional de electrocatalizadores para hidrógeno verde (H2V). Con esto, las herramientas computacionales basadas en aprendizaje automático favorecen un desarrollo en vías de fomentar la implementación del hidrógeno verde (H2V) como una fuente energética renovable fundamental para combatir la crisis climática global, el que es un eje estratégico de Chile.

## REFERENCIAS

Acosta, K., Salazar, I., Saldaña, M., Ramos, J., Navarra, A., & Toro, N. (2022). Chile and its Potential Role Among the Most Affordable Green Hydrogen Producers in the World. *Frontiers in Environmental Science*, *10*. https://doi.org/10.3389/fenvs.2022.890104

Arenas, J., Bauducco, S., Contreras, G., Griffith-Jones, S., & Guerra-Salas, J. (2024). *Hidrógeno verde en Chile: perspectivas de demanda e inversión*. https://www.bcentral.cl/en/w/hidrogeno-verde-chile

Bai, S., Yang, M., Jiang, J., He, X., Zou, J., Xiong, Z., Liao, G., & Liu, S. (2021). Recent advances of MXenes as electrocatalysts for hydrogen evolution reaction. *Npj 2D Materials and Applications*, *5*(1). https://doi.org/10.1038/s41699-021-00259-4

Benavides, Cifuentes Luis, Diaz Manuel, & Gonzalez Luis. (2021). *Opciones para lograr la neutralidad en carbono para 2050 en Chile en condiciones de incertidumbre Modelación y Análisis*. 124.

Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Ding, R., Chen, J., Chen, Y., Liu, J., Bando, Y., & Wang, X. (2024). Unlocking the potential: machine learning applications in electrocatalyst design for electrochemical hydrogen energy transformation. *Chemical Society Reviews*, *53*(23), 11390–11461. https://doi.org/10.1039/d4cs00844h

Ferriday, T. B., Middleton, P. H., & Kolhe, M. L. (2021). Review of the Hydrogen Evolution Reaction—A Basic Approach. *Energies*, *14*(24), 8535. https://doi.org/10.3390/en14248535

Fu, Z., Liu, W., Huang, C., & Mei, T. (2022). A Review of Performance Prediction Based on Machine Learning in Materials Science. *Nanomaterials*, *12*(17), 2957. https://doi.org/10.3390/nano12172957

Jyothirmai, M. V., Dantuluri, R., Sinha, P., Abraham, B. M., & Singh, J. K. (2024). Machine-Learning-Driven High-Throughput Screening of Transition-Metal Atom Intercalated g-C3N4/MX2 (M = Mo, W; X = S, Se, Te) Heterostructures for the Hydrogen Evolution Reaction. *ACS Applied Materials and Interfaces*, *16*(10), 12437–12445. https://doi.org/10.1021/acsami.3c17389

Khalafallah, D., Lai, F., Huang, H., Wang, J., Wang, X., Tong, S., & Zhang, Q. (2025). Machine learning-driven breakthroughs in water electrolysis and supercapacitors. *Materials Chemistry Frontiers*, *9*(15), 2322–2353. https://doi.org/10.1039/d5qm00326a

Kohavi, R. (2000). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *American Journal of Orthodontics and Dentofacial Orthopedics*, *118*(4), 456–461. https://doi.org/10.1067/mod.2000.109032

Lee, J., Lee, S. A., Lee, T. H., & Jang, H. W. (2025). Unlocking the potential of chemical-assisted water electrolysis for green hydrogen production. *Industrial Chemistry & Materials*, *3*(3), 277–310. https://doi.org/10.1039/d4im00163j

Li, J., Wu, N., Zhang, J., Wu, H. H., Pan, K., Wang, Y., Liu, G., Liu, X., Yao, Z., & Zhang, Q. (2023). Machine Learning-Assisted Low-Dimensional Electrocatalysts Design for Hydrogen Evolution Reaction. *Nano-Micro Letters*, *15*(1), 1–27.

https://doi.org/10.1007/s40820-023-01192-5

Lu, B., Xia, Y., Ren, Y., Xie, M., Zhou, L., Vinai, G., Morton, S. A., Wee, A. T. S., van der Wiel, W. G., Zhang, W., & Wong, P. K. J. (2024). When Machine Learning Meets 2D Materials: A Review. *Advanced Science*, *11*(13). https://doi.org/10.1002/advs.202305277

Ministerio de Energía, C., & Barhorst, N. (2016). Green hydrogen. *39th World Energy Engineering Conference, WEEC 2016*, *2*, 886–897. https://doi.org/10.4018/979-8-3693-8980-5.ch010

Niu, S., Cai, J., & Wang, G. (2021). Two-dimensional MOS2 for hydrogen evolution reaction catalysis: The electronic structure regulation. *Nano Research*, *14*(6), 1985–2002. https://doi.org/10.1007/s12274-020-3249-z

Qadeer, M. A., Zhang, X., Farid, M. A., Tanveer, M., Yan, Y., Du, S., Huang, Z.-F., Tahir, M., & Zou, J.-J. (2024). A review on fundamentals for designing hydrogen evolution electrocatalyst. *Journal of Power Sources*, *613*, 234856. https://doi.org/10.1016/j.jpowsour.2024.234856

Ram, S., Lee, A. S., Lee, S. C., & Bhattacharjee, S. (2025). Advanced Multifunctional Electrocatalysts: Integrating DFT and Machine Learning for OER, HER, and ORR Reactions. *Chemistry of Materials*, *37*(10), 3608–3621. https://doi.org/10.1021/acs.chemmater.4c03213

Shi, J., Bao, Y., Ye, R., Zhong, J., Zhou, L., Zhao, Z., Kang, W., & Aidarova, S. B. (2025). Recent progress and perspective of electrocatalysts for the hydrogen evolution reaction. *Catalysis Science & Technology*, *15*(7), 2104–2131. https://doi.org/10.1039/D4CY01449A

Sun, X., Zheng, J., Gao, Y., Qiu, C., Yan, Y., Yao, Z., Deng, S., & Wang, J. (2020). Machine-learning-accelerated screening of hydrogen evolution catalysts in MBenes materials. *Applied Surface Science*, *526*(May), 146522. https://doi.org/10.1016/j.apsusc.2020.146522

Sweet, L., Müller, C., Anand, M., & Zscheischler, J. (2023). Cross-Validation Strategy Impacts the Performance and Interpretation of Machine Learning Models. *Artificial Intelligence for the Earth Systems*, *2*(4). https://doi.org/10.1175/AIES-D-23-0026.1

Yin, G., Zhu, H., Chen, S., Li, T., Wu, C., Jia, S., Shang, J., Ren, Z., Ding, T., & Li, Y. (2025). Machine Learning-Assisted High-Throughput Screening for Electrocatalytic Hydrogen Evolution Reaction. *Molecules*, *30*(4), 1–25. https://doi.org/10.3390/molecules30040759

Zhang, C., Bao, L., Liu, L., Alomar, M., Qin, N., Guo, W., Li, F., Zhang, H., Liu, H., Zhang, C., Bao, L., & Liu, L. (2025). Recent advances in machine learning-driven discovery of alloy electrocatalysts for hydrogen evolution reaction. *Chinese Chemical Letters*, 112021. https://doi.org/10.1016/j.cclet.2025.112021

Zhang, J., Wang, Y., Zhou, X., Zhong, C., Zhang, K., Liu, J., Hu, K., & Lin, X. (2023). Accurate and efficient machine learning models for predicting hydrogen evolution reaction catalysts based on structural and electronic feature engineering in alloys. *Nanoscale*, *15*(26), 11072–11082. https://doi.org/10.1039/d3nr01442h

## NOTAS AL PIE DE PÁGINA:

Se proporcionan términos para facilitar la comprensión del lector:

-Machine Learning (ML): Machine learning (aprendizaje automático, en español) es un conjunto de métodos computacionales que permiten a un modelo "aprender" patrones a partir de datos, sin ser programado explícitamente para cada tarea específica. En este trabajo se usa para entrenar modelos que predicen propiedades (por ejemplo, $\Delta G_H^*$) a partir de descriptores fisicoquímicos.

-MXenes: familia de materiales bidimensionales (2D) basados en carburos, nitruros o carbonitruros de metales de transición, con fórmula general $M_{n+1}X_n$ (donde M es un metal de transición y X es C y/o N). Se caracterizan por su alta conductividad eléctrica, estructura en capas y superficies funcionalizables, lo que los hace atractivos para aplicaciones electroquímicas y catalíticas. (Los MBenes son análogos ricos en boro de esta familia, reemplazando C/N por B).

-Coeficiente de determinación $R^2$: métrica que mide qué fracción de la variabilidad de los datos observados es explicada por el modelo. Toma valores entre $-\infty$ y 1, donde 1 indica una predicción perfecta y valores cercanos a 0 indican que el modelo explica poca varianza (similar a usar la media como predicción).

-Desviación estándar ($\sigma$): medida de dispersión que indica cuánto se alejan, en promedio, los datos respecto de su valor medio. Cuanto mayor es la desviación estándar, más "extendidos" están los datos alrededor de la media.

-Error cuadrático medio de la raíz (RMSE): métrica de error que se define como la raíz cuadrada del promedio de los errores al cuadrado entre valores predichos y valores reales. Se expresa en las mismas unidades que la variable objetivo (por ejemplo, eV para $\Delta G_H^*$) y valores menores indican un mejor ajuste del modelo.

-pipeline: es una secuencia de etapas de procesamiento donde la salida de una etapa se convierte en la entrada de la siguiente, permitiendo procesar datos o instrucciones de forma continua y paralela, aumentando la eficiencia.

## ANEXOS

Repositorio del código del artículo de referencia: https://github.com/xilingyi/sx_MBene