Universidad de **Valladolid**

FACULTAD DE CIENCIAS

TRABAJO FIN DE GRADO

GRADO EN ESTADÍSTICA

Optimización de materiales para el almacenamiento de hidrógeno mediante modelos de aprendizaje automático

Autor: **Adrián Villalaín Moradillo**

Tutor: **Francisco Hernando Gallego**

2025

Agradecimientos

A mi tutor, por su ayuda y sus valiosos consejos durante estos meses, que han sido fundamentales para poder completar este trabajo.

A las universidades de Valladolid, Granada y Padova, y a todos los miembros de estas instituciones que me han formado a lo largo de estos años.

A mis amigos, por su apoyo, confianza y cariño. En especial, a aquellos que he conocido durante mi etapa universitaria: sois, sin duda, lo mejor de todo este camino.

Y, por supuesto, a mi familia, en especial a mis padres, por su ayuda incondicional, pase lo que pase.

Gracias a todos.

Resumen

El hidrógeno es un vector energético prometedor para una transición hacia fuentes sostenibles, pero su almacenamiento eficiente sigue siendo un desafío. Este trabajo aplica técnicas de aprendizaje automático para modelar y predecir la capacidad de almacenamiento de hidrógeno en materiales porosos tipo MOFs (Metal-Organic Frameworks), a partir de propiedades como densidad, porosidad y área superficial. Se emplean modelos de regresión como Ridge, LASSO y Random Forest, y se evalúa su desempeño mediante validación cruzada. Además, se exploran estrategias para la generación de nuevos candidatos mediante interpolación y simulación. Los resultados evidencian la capacidad de los modelos para identificar materiales con propiedades prometedoras y facilitar la búsqueda de compuestos óptimos para aplicaciones energéticas.

Abstract

Hydrogen is a promising energy vector in the transition to sustainable sources, although its efficient storage remains a challenge. This work applies machine learning techniques to model and predict the hydrogen storage capacity of porous materials known as MOFs (Metal-Organic Frameworks), using features such as density, porosity, and surface area. Regression models including Ridge, LASSO, and Random Forest are used and evaluated through cross-validation. Additionally, strategies for generating new candidate materials through interpolation and simulation are explored. The results highlight the models' ability to identify promising materials and support the discovery of optimal compounds for energy applications.

Índice

Índice de figuras

1. Introducción

El hidrógeno se ha posicionado como un vector energético clave en la transición hacia una economía descarbonizada, gracias a su alta densidad energética por unidad de masa (120 MJ/kg, casi tres veces mayor que la de la gasolina) que permitiría reducir emisiones de carbono. Su combustión solo produce agua H_2O, lo que lo convierte en una alternativa limpia y sostenible frente a los combustibles fósiles.

Sin embargo, su baja densidad volumétrica y alta permeabilidad plantean desafíos críticos para su almacenamiento eficiente. Para alcanzar densidades útiles, se requieren presiones muy altas (hasta 700 bar) o criogenización (-253 °C), lo que incrementa los costos y riesgos operativos. Además, su inflamabilidad y tendencia a fugarse complican el diseño de sistemas seguros. Los métodos de almacenamiento sólido ofrecen una alternativa prometedora, pero su desarrollo se ve limitado por la baja capacidad de almacenamiento y la falta de estabilidad en condiciones prácticas.

Este trabajo aborda estos desafíos mediante un enfoque centrado en técnicas de aprendizaje automático (Machine Learning), combinando técnicas de regresión avanzada con restricciones basadas en principios científicos para optimizar materiales de almacenamiento de hidrógeno. Nos centramos en Metal-Organic Frameworks (MOFs), materiales porosos formados por iones metálicos y ligandos orgánicos, cuya elevada superficie específica (hasta 7000 m^2/g) y porosidad ajustable los convierten en candidatos ideales para el almacenamiento de gases.

A partir de un conjunto de datos de 107 MOFs, caracterizados por propiedades clave como densidad, porosidad y área superficial, entrenamos modelos de regresión para predecir la capacidad gravimétrica (usablegc) y volumétrica (usablevc) de almacenamiento de hidrógeno. Para garantizar predicciones realistas, integramos restricciones termodinámicas y emplearemos técnicas de regularización (Ridge, LASSO) para evitar sobreajuste y seleccionar las características más relevantes.

1.1. Conjunto de Datos y Simulaciones

El conjunto de datos utilizado en este estudio se obtuvo a partir de simulaciones Grand Canonical Monte Carlo (GCMC) realizadas sobre una colección de 107 MOFs sintetizados experimentalmente, bajo condiciones estandarizadas: temperatura ambiente (298.15 K) y presión de almacenamiento de 25 MPa. Las simulaciones se ejecutaron utilizando un código interno validado, previamente calibrado con datos experimentales [1, 2].

Cada entrada en el conjunto de datos incluye parámetros estructurales y fisicoquímicos detallados, como:

- Densidad aparente (ρ)
- Porosidad
- Radio promedio de poro (Ri)
- Área superficial específica (SSA)
- Volumen de poro específico

Las variables de respuesta clave son:

- Capacidad gravimétrica usable (ugc, en wt.%)
- Capacidad volumétrica usable (uvc, en kg/L)

Estas métricas han sido calculadas mediante la diferencia en la cantidad de hidrógeno adsorbido entre 25 MPa y 0.5 MPa, reflejando así la capacidad entregable en condiciones realistas de operación para vehículos de pila de combustible [3, 4].

1.2. Metodología de Aprendizaje Automático

A partir de este conjunto de datos, entrenamos modelos de regresión para predecir ugc y uvc, integrando:

- Restricciones termodinámicas para garantizar la viabilidad física de las predicciones.
- Técnicas de regularización (Ridge, LASSO) para evitar sobreajuste y seleccionar características relevantes.

Dado el tamaño limitado del conjunto de datos (106 MOFs), adoptamos una estrategia de ”small data”, donde el conocimiento experto y la optimización cuidadosa del modelo permiten

extraer información valiosa incluso con muestras reducidas. Este enfoque no solo mejora la generalización del modelo, sino que también facilita la identificación de MOFs prometedores que podrían cumplir con los objetivos de la DOE (5.5 wt.% en usablegc y 0.040 kg/L en usablevc).

Con estos modelos exploraremos nuevas combinaciones de los MOFs existentes para obtener nuevos materiales que sean posibles candidatos que cumplan las restricciones mínimas que se buscamos.

2. Estado del Arte

2.1. Revisión de Trabajos Previos

El interés por el hidrógeno como combustible ha impulsado la aplicación de herramientas computacionales avanzadas con el objetivo de optimizar materiales de almacenamiento, particularmente los *Metal-Organic Frameworks* (MOFs).

2.1.1. Predicción de Propiedades y Almacenamiento de Gases

Diversos estudios han empleado *Machine Learning* (ML) para predecir la capacidad de almacenamiento de gases en MOFs. Técnicas como el *computational screening* han sido utilizadas para identificar materiales con alto rendimiento en el almacenamiento de metano [5]. Modelos de *transfer learning* han sido aplicados para predecir las propiedades de difusión de estos materiales [6], mientras que modelos auto-supervisados como MOFormer se han centrado en mejorar la predicción y síntesis de MOFs [7]. Además, investigaciones recientes han demostrado la utilidad de herramientas generativas de IA como GPT-4 para asistir en la generación de nuevas estructuras de MOFs, aunque sin centrarse directamente en capacidades de almacenamiento [16].

2.1.2. Uso de Redes Neuronales y Grandes Bases de Datos

Las *neural networks* (NN) han demostrado su eficacia en la predicción de capacidades de almacenamiento a partir de bases de datos experimentales y simuladas [8]. Bases de datos de isotermas de adsorción permiten estudiar la relación entre la estructura y las propiedades de los MOFs [9]. En estudios como los de Burner et al., se han manejado más de 300,000 estructuras con descriptores geométricos y químicos, lo cual es una cantidad gigante de datos, para predecir capacidades de adsorción de nuevos materiales utilizando las técnicas ya mencionadas [17].

2.1.3. Descomposición en Fragmentos y Uso de Descriptores Computacionales

Una estrategia consiste en descomponer MOFs en fragmentos, facilitando la predicción de propiedades mediante ML, desescalando los datos. Esquemas como MOFid permiten codificar estructuras y mejorar el *data mining* [10]. Modelos basados en descriptores han empleado propiedades como electronegatividad y radios covalentes para predecir la adsorción de gases [11]. Además, se han incorporado descriptores inspirados en el *computer-aided drug design* (CADD), como áreas polares y potenciales electrostáticos [18].

2.1.4. Modelos Avanzados e Iterativos

Se han desarrollado enfoques basados en *iterative learning* y redes neuronales convolucionales (CNN) para procesar estructuras cristalinas y predecir capacidades de almacenamiento [14, 12, 13]. Además, métodos de *iterative prescreening* combinados con simulaciones *Grand Canonical Monte Carlo* (GCMC) han optimizado la búsqueda de MOFs prometedores [15]. Majumdar et al. generaron una base de datos de 22,836 estructuras hipotéticas para entrenar redes neuronales multioutput que predicen coeficientes de Henry [19].

2.2. Comparación y Análisis Crítico

Los estudios previos mencionados han demostrado avances significativos en la aplicación de *Machine Learning* para la predicción de propiedades de MOFs, destacando enfoques como el *computational screening*, *transfer learning*, y el uso de descriptores complejos. Sin embargo, muchos de estos trabajos se centran exclusivamente en predecir propiedades de materiales ya conocidos sin abordar de manera integral la generación de nuevas estructuras optimizadas.

Entre las fortalezas de los trabajos existentes se encuentran el manejo de grandes bases de datos y la implementación de técnicas avanzadas como redes neuronales y CNNs. No obstante, presentan debilidades como la dependencia de descriptores específicos, la falta de integración de restricciones físicas fundamentales y la escasa validación experimental de los materiales predichos.

El presente trabajo se diferencia en el hecho de que propone un modelo que, además de predecir capacidades de almacenamiento, genera nuevas estructuras sin restricciones de fragmentación o descriptores, utilizando directamente las propiedades físicas esenciales de los MOFs. De esta manera, se busca ampliar el horizonte de los materiales conocidos.

2.3. Justificación del Enfoque del Trabajo

A pesar de los avances recientes en la aplicación de modelos de aprendizaje automático al diseño y caracterización de materiales porosos, persiste una limitación clave en el campo: la escasez de metodologías capaces de generar nuevas estructuras de MOFs optimizadas para el almacenamiento de hidrógeno, sin depender de técnicas basadas en la fragmentación de componentes o en el uso extensivo de descriptores computacionales específicos.

Este trabajo propone superar dicha limitación mediante un enfoque sencillo, interpretable y físicamente fundamentado: el uso de modelos de regresión lineal con técnicas de regularización,

concretamente Ridge y LASSO, sobre propiedades estructurales esenciales de los MOFs. Estas propiedades —densidad, porosidad, radio medio de poro, área superficial específica (SSA) y volumen de poro— han demostrado ser factores clave en la determinación de la capacidad de adsorción de hidrógeno, y presentan la ventaja de estar disponibles o simulables para la mayoría de estructuras, sin necesidad de representaciones intermedias.

La elección de modelos lineales regularizados responde a varios criterios estratégicos. Por un lado, estos modelos permiten una selección explícita de variables (en el caso de LASSO), lo que mejora la interpretabilidad de los resultados y ayuda a identificar qué propiedades influyen más en cada tipo de capacidad (volumétrica o gravimétrica). Por otro lado, su bajo número de hiperparámetros, junto con su robustez frente al sobreajuste, los hace especialmente adecuados para entornos con pocos datos, alta colinealidad y estructuras complejas, como ocurre en este trabajo.

Además, se ha optado por una estrategia de *small data* como enfoque deliberado, orientado a maximizar el valor informativo de conjuntos de datos limitados pero cuidadosamente seleccionados. En lugar de entrenar modelos complejos sobre millones de estructuras simuladas, se propone extraer conocimiento a partir de un número más reducido de muestras reales o simuladas con alta fidelidad, lo que permite controlar la calidad de los datos, reducir el ruido y evitar conclusiones artefactuales. Este enfoque también facilita el seguimiento posterior de las predicciones en el laboratorio, al priorizar estructuras viables y bien definidas.

En conjunto, este enfoque combina simplicidad computacional, interpretabilidad científica y relevancia físico-química, ofreciendo una alternativa eficiente y rigurosa frente a técnicas más opacas o costosas. Su objetivo no es competir con modelos complejos como redes neuronales profundas o algoritmos generativos, sino complementarlos con una herramienta ligera que pueda ser empleada como filtro inicial o guía en la búsqueda racional de nuevos materiales con propiedades óptimas para el almacenamiento de hidrógeno.

3. Problema Planteado

3.1. Definición del Problema

El presente trabajo aborda el problema de encontrar nuevos materiales porosos, concretamente estructuras metal-orgánicas (MOFs), que sean capaces de almacenar hidrógeno de forma eficiente a temperatura ambiente (RT) y presiones moderadas (25–35 MPa). La motivación de este problema radica en la necesidad de desarrollar tecnologías que permitan a los vehículos eléctricos de pila de combustible (FCV) alcanzar autonomías comparables a las de los vehículos convencionales de gasolina, en torno a 600 km, sin comprometer la seguridad, el volumen o el peso del sistema de almacenamiento.

Como ya hemos explicado en el apartado anterior, la propuesta del trabajo es utilizar herramientas de aprendizaje automático, específicamente modelos de regresión regularizados (Ridge y LASSO), para predecir la capacidad de almacenamiento de hidrógeno en MOFs a partir de sus propiedades estructurales. El objetivo final es generar nuevas estructuras candidatas que cumplan los umbrales definidos, contribuyendo así a la transición energética hacia una economía libre de emisiones de carbono, mediante la implementación del hidrógeno como vector energético.

Los objetivos se centran en encontrar MOFs que alcancen:

- **Double Tank Targets**: capacidad volumétrica útil (*usablevc*) $\geq$ 0,020 kg/L y capacidad gravimétrica útil (*usablegc*) = 5.5 wt. %.
- **2025 Targets**: capacidad volumétrica útil (*usablevc*) $\geq$ 0,040 kg/L y capacidad gravimétrica útil (*usablegc*) = 5.5 wt. %.

Buscamos estos objetivos dado que las capacidades volumétrica y gravimétrica útiles son parámetros fundamentales para evaluar el desempeño de los materiales en el almacenamiento de hidrógeno. La capacidad volumétrica útil (*usablevc*) mide la cantidad de hidrógeno que puede almacenarse por unidad de volumen del material, siendo clave para maximizar el aprovechamiento del espacio en depósitos de hidrógeno compactos. Por otro lado, la capacidad gravimétrica útil (*usablegc*) indica la cantidad de hidrógeno almacenada por unidad de masa del material, lo que es crucial para aplicaciones donde el peso del sistema de almacenamiento es determinante, como en vehículos de pila de combustible.

La selección de estos valores numéricos específicos se fundamenta en criterios técnicos y normativos clave. Los umbrales de 5.5 wt. % para *usablegc* y 0.020-0.040 kg/L para *usablevc* se alinean con los objetivos del Departamento de Energía de EE.UU. (DOE) para sistemas

de almacenamiento vehicular, donde: (1) el 5.5 wt.% garantiza autonomías competitivas sin penalizar el peso del sistema, acercándose a la densidad energética de la gasolina; y (2) los valores volumétricos equilibran la compactibilidad del depósito, evitando soluciones criogénicas o de ultra-alta presión. Estos valores representan un compromiso óptimo entre las limitaciones físico-químicas de los MOFs y las demandas reales de aplicaciones energéticas, centrándose específicamente en la capacidad útil (diferencia entre 25 MPa y 0.5 MPa) que refleja el hidrógeno realmente aprovechable en condiciones operativas.

Actualmente, no existen materiales sólidos conocidos que alcancen simultáneamente una capacidad gravimétrica igual o superior a 2.8 wt.% y una capacidad volumétrica igual o superior a 0.020 kg/L en condiciones de temperatura ambiente y presiones moderadas. A pesar de que estos valores representan aproximadamente la mitad de los objetivos establecidos por el Departamento de Energía de los Estados Unidos (DOE) para el año 2025, su cumplimiento se considera tecnológicamente ambicioso y relevante.

Optimizar ambas propiedades permitiría desarrollar sistemas de almacenamiento eficientes, ligeros y con alta densidad energética, requisitos esenciales para la viabilidad del hidrógeno como vector energético en transporte y otras aplicaciones industriales. Así, se pretende contribuir al desarrollo de depósitos de hidrógeno más eficientes, con impacto directo en la movilidad sostenible y la mitigación del cambio climático.

3.2. Hipótesis o Preguntas de Investigación

- ¿Es posible predecir la capacidad de almacenamiento de hidrógeno en MOFs utilizando modelos de regresión lineal con restricciones físicas?
- ¿Cómo influyen las técnicas de regularización (Ridge y LASSO) en la selección de características y la estabilidad del modelo?
- ¿Puede un enfoque de *small data* maximizar la capacidad predictiva del modelo en conjuntos de datos limitados?
- ¿Es viable proponer nuevas estructuras de MOFs que cumplan con los requisitos mínimos de almacenamiento (2.8 wt.% y 0.020 kg/L) basándose únicamente en predicciones computacionales?

3.3. Restricciones y Alcance

Este trabajo se centra exclusivamente en la optimización de materiales tipo MOF para el almacenamiento de hidrógeno, sin considerar otras familias de materiales como hidruros metáli-

cos, grafenos dopados o nanotubos de carbono. Las predicciones se realizan a partir de un conjunto de datos limitado en número pero detallado en propiedades estructurales como densidad, porosidad, área superficial específica (SSA) y volumen de poro.

El alcance del proyecto se limita a:

- La recopilación y simulación de datos de almacenamiento mediante métodos como GCMC para materiales ya conocidos.
- La construcción de modelos predictivos basados en aprendizaje automático.
- La generación computacional de nuevas estructuras candidatas.

No se contempla en este trabajo la síntesis experimental de los materiales propuestos ni su validación en laboratorio. Sin embargo, se discute la viabilidad técnica y energética de los materiales candidatos generados desde una perspectiva teórica.

4. Metodología y Desarollo

Este trabajo implementa modelos de aprendizaje automático para optimizar el almacenamiento de hidrógeno en materiales, aplicando regresión lineal con regularización y técnicas avanzadas de preprocesamiento de datos.

4.1. Selección y Procesamiento de Datos

El conjunto de datos utilizado incluye propiedades clave de materiales, como:

- **Densidad** ($0{,}1 \leq density \leq 4{,}1$ g/cm^3)
- **Porosidad** ($0{,}5 \leq porosity \leq 0{,}8$)
- **Radio de poro** ($3 \leq Ri \leq 20$ Å)
- **Área superficial específica** ($4000 \leq ssa \leq 10000$ m^2/g)
- **Volumen específico del poro** ($1{,}0 \leq specific_pore_volume \leq 3{,}0$ cm^3/g)

Estos datos se normalizan mediante *MinMaxScaler* y se transforman con *PolynomialFeatures* para capturar interacciones no lineales entre las variables. Además, se ha implementado un filtro de selección de datos que garantiza que los valores empleados en el modelo cumplen con restricciones físicas establecidas para asegurar la viabilidad experimental de los materiales propuestos.

4.2. Modelos y Optimización

Se han implementado y optimizado los siguientes modelos:

- **Ridge Regression**: Regularización L2 para evitar sobreajuste.
- **LASSO**: Regularización L1 para seleccionar características relevantes.
- **Random Forest**: Modelo de ensamble que captura relaciones no lineales.

La optimización de hiperparámetros se realiza mediante *GridSearchCV*, evaluando diferentes valores de `alpha` para Ridge/LASSO y parámetros como `n_estimators`, `max_depth` y `min_samples_split` en Random Forest. La inclusión de estos parámetros mejora la capacidad del modelo para capturar relaciones no triviales y refinar la precisión de las predicciones.

4.3. Evaluación de Modelos

Los modelos se evalúan mediante:

- **Error Absoluto Medio (MAE)**: Medida de la precisión de las predicciones.
- **Coeficiente de Determinación (R^2)**: Indica qué tan bien se ajusta el modelo a los datos.
- **Cumplimiento de objetivos**: Se verifica si las predicciones cumplen con los umbrales de *usablegc* ($\geq$ 5,5 wt.%) y *usablevc* ($\geq$ 0,020 kg/L).

Adicionalmente, se analiza el porcentaje de predicciones que cumplen con los valores esperados, lo que permite evaluar la aplicabilidad real de los materiales propuestos.

4.4. Resultados Esperados

Se espera que Random Forest proporcione el mejor rendimiento debido a su capacidad de modelar relaciones complejas. Ridge y LASSO ayudan a interpretar la importancia de las variables y mejorar la estabilidad del modelo.

Se ha observado que los modelos de regresión lineal pueden verse limitados en su capacidad de generalización cuando se trabaja con datos reducidos, lo que hace que la combinación con modelos de aprendizaje profundo o técnicas de ensamblado pueda representar una posible mejora futura. Además, el uso de validación cruzada y un mayor número de iteraciones en la optimización de hiperparámetros podría contribuir a mejorar la robustez del modelo y reducir la varianza de las predicciones.

5. Resultados y Discusión

Los resultados obtenidos se analizarán en función de los objetivos definidos y se contrastarán con estudios previos. Se presentarán métricas detalladas de rendimiento de los modelos y se evaluará la capacidad de predicción en función de los valores reales del conjunto de datos. Además, se discutirán las limitaciones del enfoque utilizado y posibles mejoras a futuro.

5.1. Análisis Exploratorio

Como acercamiento inicial a los datos, vamos a ver cómo son las características de los MOFs entre nuestros 107 candidatos. Para ello, estudiaremos cómo están distribuidas nuestras variables, mediante un histograma, que incluye una estimación suavizada de la densidad de las variables.

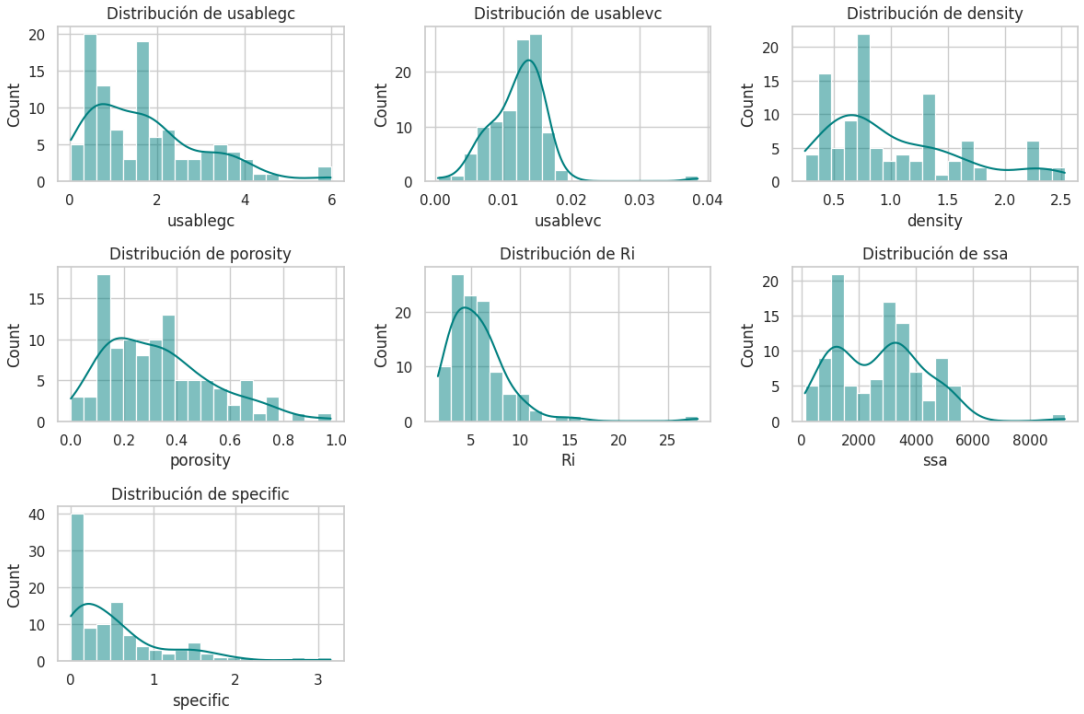


Figura 1: Distribución de variables clave. (a) *usablegc* muestra valores concentrados (0–4 wt. %); (b) *porosity* con distribución uniforme; (c) *usablevc* con valores predominantemente $< 0{,}02$ kg/L; (d) *density* y *ssa* presentan sesgo. Justifica la necesidad de normalización.

5.2. Importancia de Variables

Tras ver cómo están distribuidas nuestras variables, podemos entrar a valorar la importancia que tiene cada una a la hora de predecir las variables clave, usablegc y usablevc. Esto tendrá

gran relación con el nivel de correlación entre ellas. Para realizar este análisis, usaremos el modelo de random forest, ya que es el más equilibrado y une los resultados que obtendríamos comparando el Ridge y el Lasso.

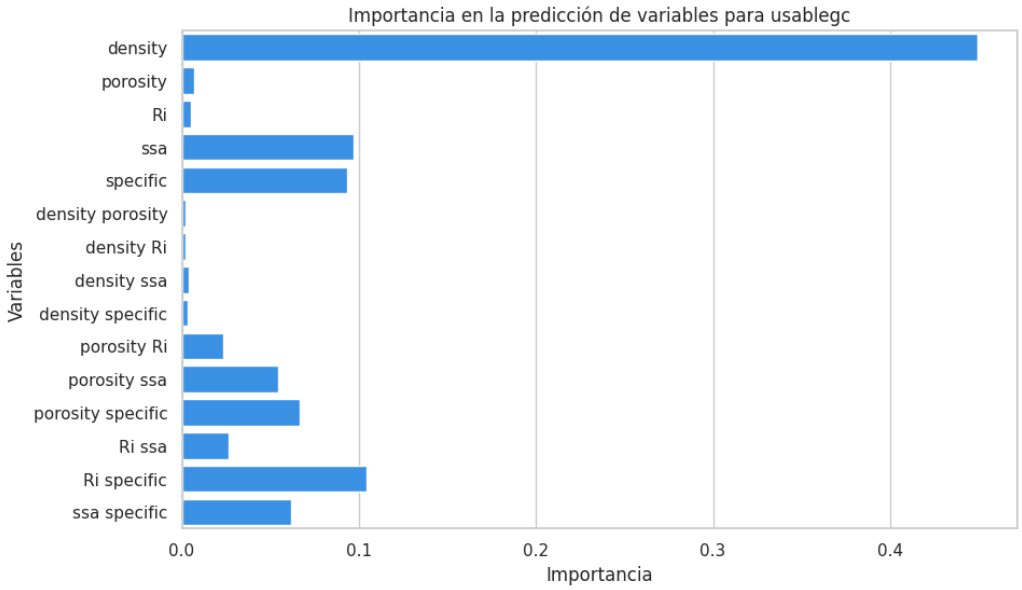


Figura 2: Importancia de variables en Random Forest para predecir usablegc.

Para usablegc podemos ver que la variable density es la más importante, con un peso ligeramente superior al 40 %. Después vemos que el área superficial y el 'specific' también tienen valores significativos, que no llegan al 10 %. Las otras 2 variables específicas no tienen casi peso. El resto del peso está repartido de manera uniforme entre las variables que relacionan las variables básicas entre sí, en especial las que están asociadas con 'specific'.

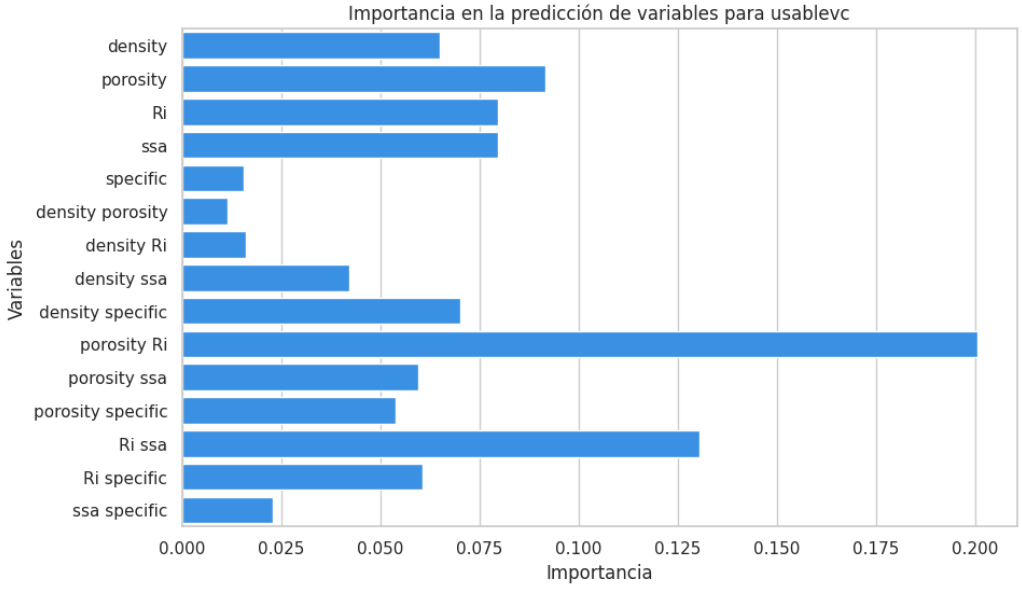


Figura 3: Importancia de variables en Random Forest para predecir usablevc.

Para usablevc destacan la variable compuesta por la relación entre porosity y Ri, junto con la compuesta por Ri y ssa. El resto de variables están muy igualadas con valores cercanos al 7 %, con la excepción de 'specific' que de nuevo tiene un valor muy bajo cercano al 0 %.

El modelo Random Forest no tiene la característica de 'sparsity' que hace que las variables que no tienen importancia lleguen a alcanzar el valor de 0, siendo eliminadas del modelo tras un número de iteraciones. Pero podemos suponer que si usáramos LASSO para evaluar la métrica de la importancia, variables como 'specific' desaparecerían del modelo.

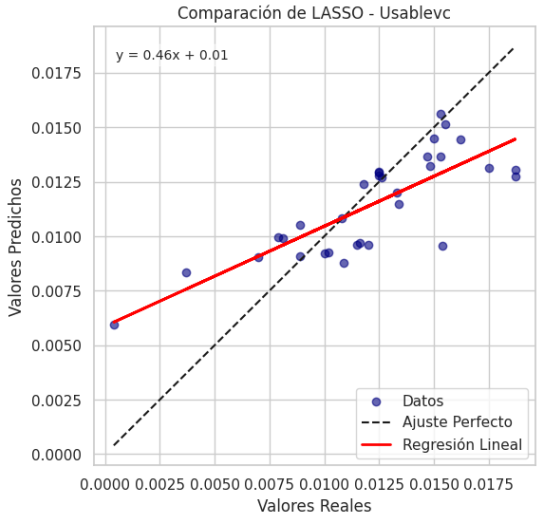
5.3. Comparación de Modelos

Vamos a ver cómo varía el rendimiento de ambos modelos. Esto nos dirá cómo se relacionan las variables. A la hora de evaluar el rendimiento de los modelos hay varias métricas que podemos usar. Hay muchas métricas que evalúan el rendimiento de los modelos, pero para este caso concreto he considerado que la más adecuada es MAE (Mean Absolute Error). El MAE mide el promedio de los errores absolutos entre las predicciones y los valores reales. Es decir, cuantifica cuánto se desvía, en promedio, una predicción del valor real.

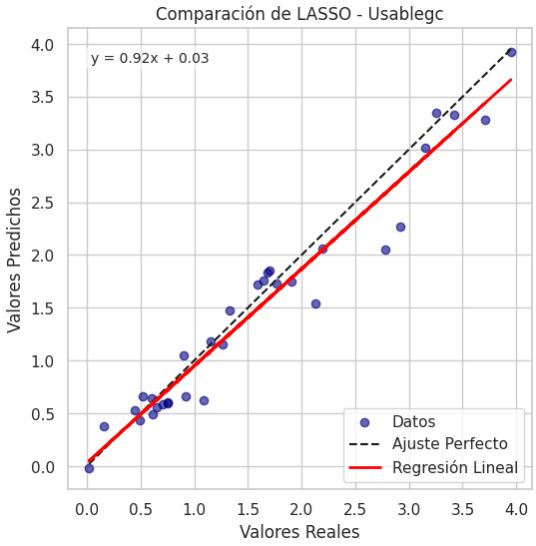
$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Como características del MAE a resaltar incluiremos que es muy intuitivo y robusto frente a valores extremos, aunque menos que otras métricas como el medAE (error absoluto mediano). También hay que destacar que se mide en las mismas unidades que la variable dependiente.

El rendimiento lo expresar gráficamente a través de un gráfico que contenga en el eje x los valores reales de las variables y en el eje y los valores predichos. Si las predicciones fuesen perfectas, todos los puntos estarían sobre la recta x=y. Vamos a ver qué diferencias hay y qué regresor funciona mejor.

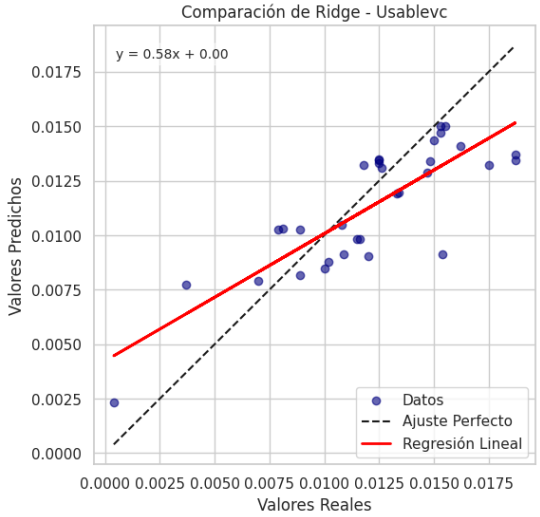


(a) LASSO para capacidad volumétrica: elimina 10 de 15 variables. MAE = 0.0038 kg/L

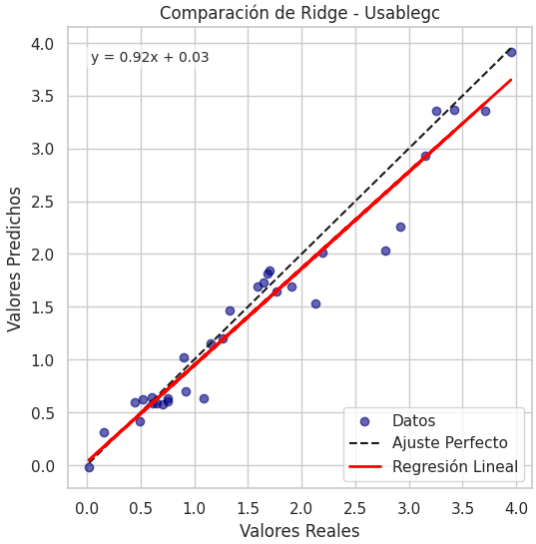


(b) LASSO para capacidad gravimétrica: elimina 10 de 15 variables. MAE = 0.0032 wt %

Figura 4: Rendimiento del modelo LASSO: prioriza interpretabilidad eliminando variables irrelevantes.



(a) Ridge para capacidad volumétrica: mantiene todas las variables. MAE = 0.0029 kg/L



(b) Ridge para capacidad gravimétrica: mantiene todas las variables. MAE = 0.0025 wt %

Figura 5: Rendimiento del modelo Ridge: conserva todas las variables para maximizar precisión.

La primera observación significativa es que el usabelgc es más fácil de predecir para ambos modelos, siendo la recta roja muy parecida a la línea rayada de ajuste perfecto. En cambio, para usablevc obtenemos una línea más horizontal de lo deseado, por lo que el valor del error será mayor. Aun así, podemos ver que el error es pequeño para ambos modelos, con lo que funcionan relativamente bien,

Observamos que el modelo Ridge alcanza un menor error absoluto medio (MAE) en ambas tareas de predicción, lo que indica una mayor capacidad de ajuste a los datos. Para la predicción de la capacidad volumétrica, Ridge logra un MAE de 0.0029 kg/L frente a los 0.0038 kg/L obtenidos por LASSO. En el caso de la capacidad gravimétrica, Ridge también supera a LASSO con un MAE de 0.0025 wt % frente a 0.0032 wt %.

Para un contexto en el que se busque identificar los factores más determinantes, LASSO resultará más útil, ya que eliminará variables que no son relevantes a la hora de predecir los valores de las variables de interés. En este caso, el modelo se queda con 2/3 de las variables iniciales, que no serán las mismas para usablevc y usablegc.

Pese a esta característica, la mayor robustez del modelo RIDGE y su mejor funcionamiento hacen que sea el modelo más adecuado a la hora de predecir los valores de las variables de los nuevos materiales que buscamos crear.

5.4. Generación de Materiales Idílicos Mediante Extrapolación Controlada

Aunque los modelos entrenados, tanto los lineales como el Random Forest, muestran buen rendimiento predictivo dentro del dominio observado, los objetivos más estrictos establecidos por los *2025 Targets* (usablegc $\geq$ 5.5 wt % y usablevc $\leq$ 0.04 kg/L) no pueden alcanzarse sin salir del espacio de diseño real. Esto se constató incluso generando 100,000 combinaciones aleatorias dentro de los rangos físicos de las variables.

Para abordar esta limitación, se implementó una estrategia de extrapolación controlada. Se ampliaron los límites de las variables físicas entre un 5 % y un 20 % más allá de los valores máximos y mínimos observados en los datos reales, y se generaron 100,000 combinaciones aleatorias para cada nivel de ampliación. Cada combinación fue evaluada con los modelos entrenados.

Los resultados muestran que:

- Con los **rangos originales**: **0 candidatos** cumplen simultáneamente los dos objetivos.
- Con un **5 %** de ampliación: alrededor de **10 candidatos** cumplen los objetivos.
- Con un **10 %** de ampliación: superamos la cantidad de **50 candidatos** que cumplen los objetivos.
- Con un **15 %** de ampliación: estaremos ya en más de **150 candidatos** válidos.

- Con un **20 %** de ampliación: conseguimos hasta **335 candidatos**.

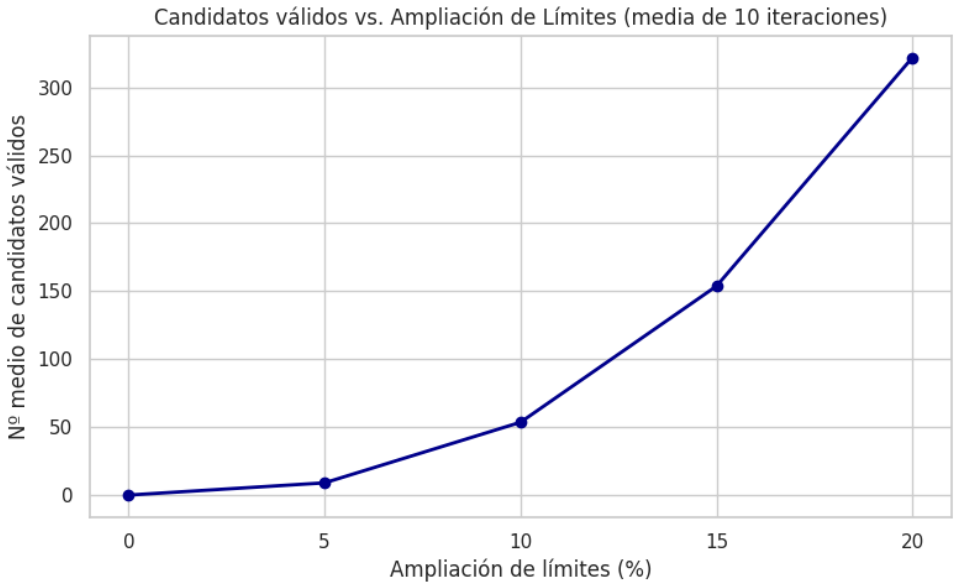


Figura 6: Evolución del número medio de candidatos válidos en función del porcentaje de ampliación de los límites de las variables.

Podemos observar que el aumento de candidatos posibles no tiene una relación lineal con el porcentaje de aumento de los límites, sino que parece seguir una distribución exponencial.

La relación no lineal observada entre el porcentaje de ampliación de los límites y el aumento de candidatos válidos encuentra su explicación en la naturaleza multidimensional del problema. Cuando expandimos simultáneamente los rangos de todas las variables estructurales relevantes (densidad, porosidad, área superficial específica, etc.), estamos efectuando una expansión no en una sola dimensión, sino en un espacio de alta dimensionalidad donde cada parámetro representa un eje independiente.

Este fenómeno puede entenderse mediante una analogía geométrica. Imaginemos un cubo en tres dimensiones: si aumentamos cada uno de sus lados en un 10 %, el volumen no crece linealmente, sino que lo hace de forma cúbica ($1{,}1 \times 1{,}1 \times 1{,}1 \approx 1{,}33$), resultando en un 33 % más de volumen. En nuestro caso, trabajamos con un "hipercubo" conceptual de cinco dimensiones (correspondientes a las cinco variables estructurales clave), donde un modesto incremento del 10 % en cada parámetro individual conduce a un aumento mucho más significativo ($1{,}1^5 \approx 1{,}61$) en el "volumen" total del espacio de búsqueda, es decir, en el número de combinaciones posibles.

Esta característica intrínseca de los espacios multidimensionales explica por qué incluso ampliaciones modestas de los límites individuales producen un crecimiento exponencial en el

número de candidatos potenciales. La estrategia de extrapolación controlada aprovecha precisamente esta propiedad matemática para explorar regiones del espacio de diseño que, aunque cercanas a los parámetros conocidos, permanecían inexploradas en los estudios convencionales. El resultado es la identificación de nuevas combinaciones estructurales prometedoras que cumplen simultáneamente con los exigentes objetivos de almacenamiento, demostrando cómo pequeños ajustes en múltiples dimensiones pueden generar mejoras sustanciales en el rendimiento global del material.

Esto demuestra que, aunque los materiales existentes no satisfacen los requisitos, es posible proponer materiales teóricos cuyos valores de propiedades predichas sí lo hagan. A modo de ejemplo, uno de los candidatos encontrados presenta las siguientes propiedades:

Propiedad	**Valor**
Density (g/cm^3)	2.522
Porosity	0.598
Ri (Å)	31.77
SSA (m^2/g)	10614
Specific (cm^3/g)	0.600
`usablegc` predicho	6.045 wt. %
`usablevc` predicho	0.0424 kg/L

Cuadro 1: Ejemplo de material idílico generado mediante extrapolación controlada.

Este resultado valida el uso de modelos predictivos para **diseñar materiales idílicos**, con potencial aplicación experimental.

5.5. Impacto Visual: Comparación de Candidatos por Nivel de Extrapolación

En esta sección se presenta un análisis gráfico del efecto de extrapolar ligeramente los límites de las variables estructurales sobre la aparición de nuevos candidatos. En todas las figuras, los materiales reales se representan en gris, mientras que los candidatos extrapolados que cumplen ambos objetivos se muestran en verde. Las líneas rojas y azules marcan los umbrales establecidos para las capacidades gravimétrica (5.5 wt. %) y volumétrica (0.04 kg/L), respectivamente.

Para una ampliación del 5 % de los valores límite, únicamente emergen 12 candidatos válidos dentro del conjunto de materiales extrapolados, lo que pone de manifiesto lo restringido del espacio de búsqueda original.

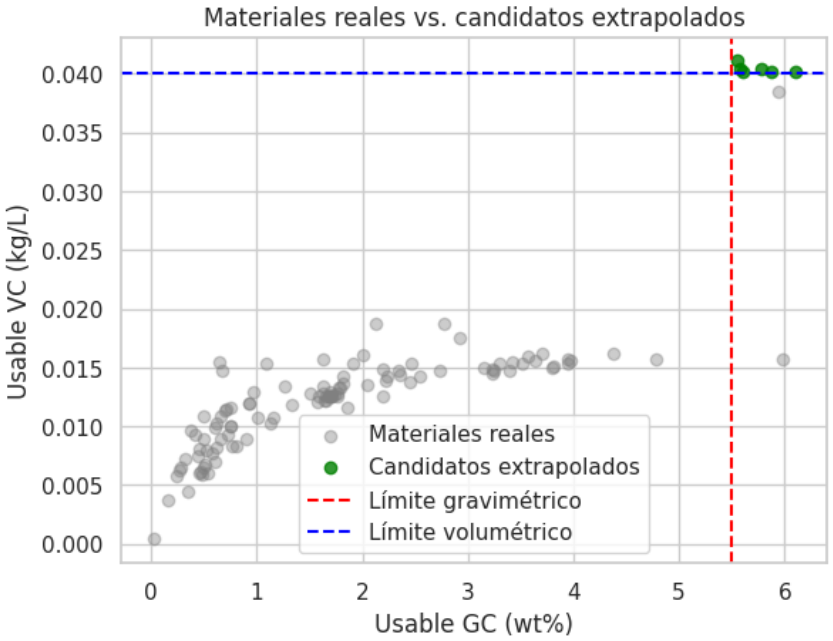


Figura 7: Distribución de materiales reales y candidatos extrapolados en el espacio `usablegc` vs. `usablevc` con un 5 % de ampliación en los límites.

Al incrementar la extrapolación al 10 %, se observa una expansión notable de la región factible, con nuevos candidatos que se alejan del punto de intersección de los umbrales. Aunque la mayoría de los candidatos aún se sitúan cerca de los límites, comienzan a aparecer materiales que destacan claramente en al menos una de las capacidades.

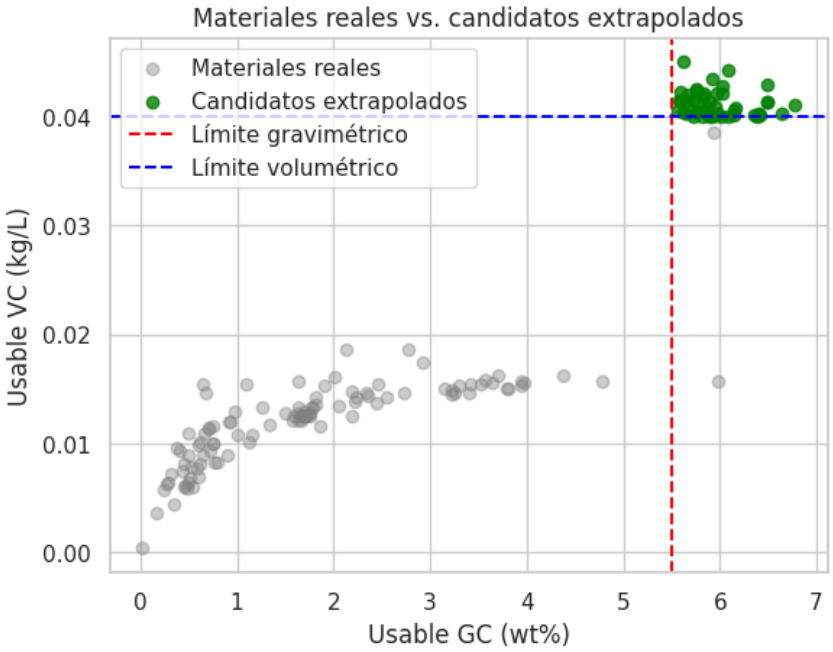


Figura 8: Materiales reales y extrapolados con un 10 % de ampliación en los límites. Se observa una mayor dispersión y aparición de candidatos viables.

El patrón se mantiene al ampliar los límites al 15 %, con un incremento tanto en la cantidad como en la calidad de los candidatos. Algunos materiales comienzan a sobresalir notablemente por encima de ambos umbrales.

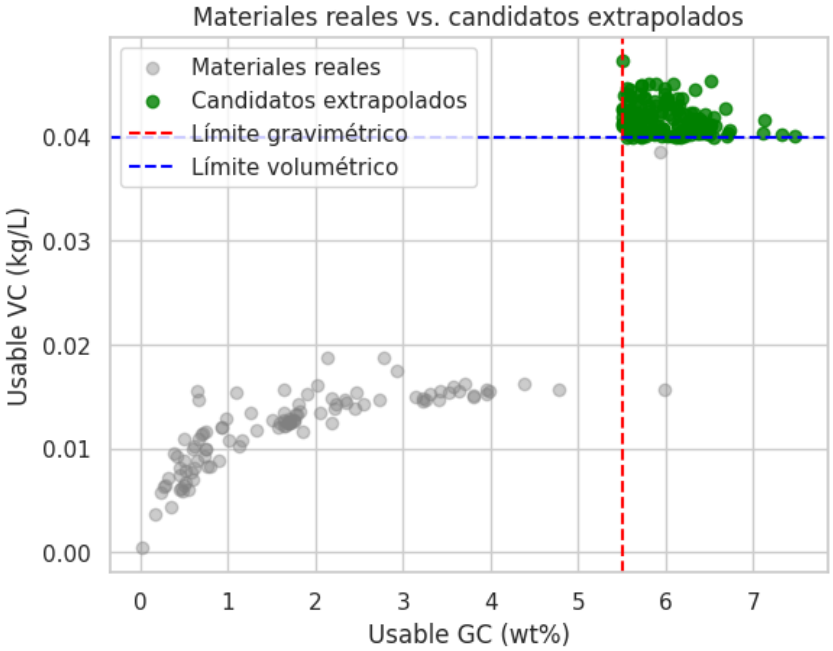


Figura 9: Candidatos extrapolados con un 15 % de ampliación. Aparecen materiales que superan ampliamente uno o ambos objetivos.

Finalmente, al permitir una ampliación del 20 %, la región verde se amplía significativamente. Se identifican materiales cuyas propiedades se alejan considerablemente de los valores típicos observados en los datos reales, lo que sugiere un potencial de descubrimiento elevado en esa región del espacio de diseño.

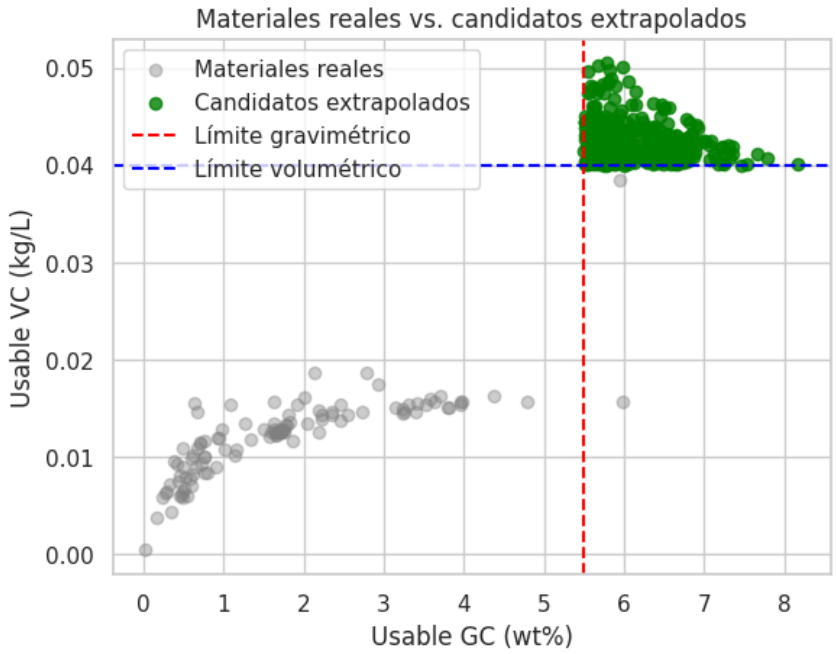


Figura 10: Visualización con una ampliación del 20 %. Los nuevos candidatos ocupan regiones previamente inexploradas.

En conjunto, estos gráficos muestran cómo los materiales reales se agrupan sistemáticamente por debajo de los umbrales objetivo, mientras que los extrapolados comienzan a ocupar zonas del espacio de diseño que antes estaban vacías. Este comportamiento sugiere que el modelo predictivo puede utilizarse para explorar regiones de alto rendimiento fuera del dominio observado, generando así hipótesis útiles para orientar futuras síntesis experimentales.

Como se evidencia en la Figura 6, el número de candidatos válidos crece rápidamente con la ampliación de los límites, incluso con incrementos tan modestos como el 5 %. Este fenómeno sugiere que las restricciones impuestas por el dominio de los datos reales limitaban la identificación de materiales óptimos. Por tanto, pequeñas extrapolaciones permiten desbloquear combinaciones estructurales prometedoras, hasta ahora invisibles para el modelo entrenado solo en datos observados.

Este análisis refuerza la utilidad de los modelos de aprendizaje automático no solo como herramientas de predicción, sino también como instrumentos para la exploración y generación

de hipótesis en el diseño de nuevos materiales.

5.6. Recomendaciones Experimentales

Con base en estos resultados, se recomienda explorar materiales con:

- Densidades superiores a 2.4 g/cm^3
- Porosidades alrededor de 0.6
- Radios de poro ampliados ($\geq$ 30 Å)
- Áreas superficiales mayores a 10,000 m^2/g

Estas propiedades no se dan en los materiales del dataset, ya que de ser así no tendría sentido esta búsqueda, pero los modelos sugieren que se permite alcanzar simultáneamente los altos valores de almacenamiento gravimétrico y volumétrico deseados.

En conjunto, esta sección demuestra que el aprendizaje automático no solo permite predecir comportamientos, sino también **proponer nuevas regiones del espacio de diseño** para orientar futuras investigaciones.

Además de las propiedades estructurales mencionadas, se sugiere priorizar combinaciones en las que las variables mantengan coherencia físico-química entre sí, evitando extrapolaciones extremas que comprometan la viabilidad sintética.

En particular:

- Se recomienda emplear simulaciones GCMC sobre los candidatos generados por extrapolación, para validar la predicción de capacidad de almacenamiento bajo condiciones realistas de presión y temperatura.
- Deberían ser considerados estudios de estabilidad térmica y química de los materiales propuestos, dado que algunas combinaciones estructurales podrían presentar limitaciones prácticas pese a sus buenas propiedades predictivas.
- En fases experimentales, comenzar con materiales cercanos a los valores extrapolados en un 5-10 %, que representan un balance adecuado entre innovación y realismo.
- La colaboración entre grupos de síntesis experimental y modelado computacional será clave para verificar la utilidad práctica de estos materiales idílicos y ajustar el modelo predictivo según nuevos datos reales.

6. Conclusiones y Trabajo Futuro

Este trabajo ha demostrado la aplicabilidad del aprendizaje automático, y en particular de modelos de regresión regularizada y ensambles no lineales, en la predicción y optimización de materiales porosos para el almacenamiento de hidrógeno. A partir de un conjunto de datos limitado y bien curado, se han implementado modelos robustos capaces de identificar relaciones estructurales relevantes y generar candidatos con propiedades idóneas para aplicaciones energéticas.

Entre los principales logros alcanzados se destacan:

- La identificación de las variables estructurales más influyentes en la capacidad de almacenamiento, donde la **densidad** y el **área superficial específica** dominan la predicción de la capacidad volumétrica, y la **interacción porosidad-radio de poro** en la gravimétrica.
- El modelo **Ridge** se posiciona como el más preciso en términos de MAE, mientras que **LASSO** proporciona interpretabilidad útil al eliminar variables redundantes.
- Se constató que, bajo las condiciones estructurales observadas en los datos reales, ningún material satisface simultáneamente los objetivos definidos. No obstante, mediante **extrapolación controlada** se han generado múltiples candidatos teóricos que sí lo hacen.
- El número de candidatos válidos aumenta exponencialmente con pequeñas ampliaciones en el dominio de las variables, revelando regiones inexploradas del espacio de diseño con alto potencial.

Como líneas futuras de trabajo, se proponen las siguientes direcciones:

1. Incorporar modelos más complejos como redes neuronales profundas, XGBoost o modelos basados en atenciones (Transformers), que podrían capturar relaciones más sutiles entre variables.
2. Ampliar el conjunto de datos con nuevas entradas provenientes de simulaciones o experimentos, mejorando la generalización del modelo y reduciendo su sesgo.
3. Aplicar esta metodología a otros materiales porosos de interés (como COFs, zeolitas o materiales híbridos) para extender su aplicabilidad.
4. Validar experimentalmente los materiales generados mediante extrapolación, comenzando por aquellos más cercanos al dominio original y con propiedades destacadas.

En conclusión, este estudio ha aportado una metodología reproducible y eficaz para el descubrimiento racional de nuevos materiales energéticos, integrando de manera sinérgica el conocimiento físico-químico y las capacidades predictivas del aprendizaje automático. Esta aproximación tiene el potencial de acelerar significativamente el desarrollo de soluciones tecnológicas sostenibles para la economía del hidrógeno.

Apéndice A

Código en Python

El código completo utilizado para realizar el análisis y la elaboración de los gráficos puede consultarse en el siguiente repositorio de GitHub:

Enlace al Repositorio GitHub

7. Bibliografía

Referencias

[1] A. Granja-DelRío and I. Cabria, *Grand canonical Monte Carlo simulations of the hydrogen and methane storage capacities of JLU-MOF120 and JLU-MOF121*. Int. J. Hydrogen Energy, vol. 61, pp. 57–72, 2024.

[2] I. Cabria, M. J. López, and J. A. Alonso, *Comparison of theoretical methods of the hydrogen storage capacities of nanoporous carbons*. Int. J. Hydrogen Energy, vol. 46, no. 22, pp. 12192–12205, 2021.

[3] P. Halder et al., *Hydrogen storage: Technical challenges and pathways*. Int. J. Hydrogen Energy, vol. 52D, p. 973, 2024.

[4] U.S. Department of Energy, *DOE Technical Targets for Onboard Hydrogen Storage for Light-Duty Vehicles*, 2018. [Online]. Available: `https://www.energy.gov/eere/fuelcells/doe-technical-targets-onboard-hydrogen-storage-light-duty-vehicles`

[5] J. Lee and A. Smith, *Machine Learning for Gas Storage in MOFs*. Berlin, Germany: Springer, 2021.

[6] K. Lim and B. Johnson, "Transfer Learning for Diffusion Properties in MOFs,"*Journal of Computational Chemistry*, vol. 43, no. 7, pp. 1234–1245, 2022.

[7] X. Cao and Y. Chen, "MOFormer: A Self-Supervised Model for MOF Design,"*Materials Science Letters*, vol. 12, pp. 56–67, 2023.

[8] M. Schmidt and T. Becker, "Neural Networks for Gas Adsorption Prediction,"*AI in Materials Science*, vol. 10, pp. 99–110, 2021.

[9] R. Oliveira and F. Santos, .ªdsorption Isotherm Databases for MOF Analysis,"*Chemical Engineering Journal*, vol. 330, p. 112233, 2023.

[10] A. Bucior and L. Jones, "MOFid: Encoding MOF Structures for Data Mining,"*Journal of Molecular Informatics*, vol. 7, no. 3, pp. 245–257, 2019.

[11] G. Fanourgakis and H. Gubbins, "Descriptor-Based Characterization of Nanoporous Materials,"*Langmuir*, vol. 36, no. 5, pp. 1298–1310, 2020.

[12] R. Anderson and J. Cooper, Ïterative Learning for Gas Capture in MOFs,"*Nature Materials*, vol. 19, pp. 478–485, 2020.

[13] S. Moosavi and K. Meredig, ”Machine Learning-Driven MOF Discovery,”*Advanced Functional Materials*, vol. 30, no. 11, p. 2001234, 2020.

[14] T. Lu and Z. Wang, “CNNs for Crystal Structure Analysis,” *Computational Materials Science*, vol. 195, p. 110489, 2022.

[15] D. Thornton and M. Patel, ”Prescreening Methods for Hydrogen Storage MOFs,”*Energy & Fuels*, vol. 33, no. 9, pp. 8721–8732, 2019.

[16] Z. Zheng and Z. Rong, ”GPT-4 Assisted Design of MOFs,”*Angewandte Chemie International Edition*, vol. 62, p. e20231198, 2023.

[17] J. Burner and K. Smith, ”Deep Learning for CO2 Adsorption in MOFs,”*Journal of Physical Chemistry C*, vol. 124, no. 51, pp. 27996–28005, 2020.

[18] D. Kim and J. Park, ”High-Level Descriptors for Gas Adsorption Modeling,”*Catalysis Today*, vol. 120, pp. 317–323, 2007.

[19] S. Majumdar and S. Moosavi, ”Diversifying MOF Databases with ML,”*ACS Applied Materials and Interfaces*, 2021.