Predicción de la constante de acoplamiento en Resonancia Magnética Nuclear (RMN) para pares de átomos Hidrógeno-Nitrógeno, mediante Machine Learning (ML)

Alejandro Vargas Ocampo, *Ingeniería de Sistemas*
Universidad de Antioquia

***Abstract*— Nuclear Magnetic Resonance (NMR) is a fundamental technique for atomic-level structural characterization in chemistry, biology, and materials science. Among the parameters obtained from NMR, the scalar coupling constant is crucial for describing indirect interactions between atomic nuclei through chemical bonds. However, its calculation using quantum chemistry methods such as Density Functional Theory (DFT) is computationally expensive, limiting its application in large-scale studies. This work proposes a supervised learning approach to predict the scalar coupling constant using a molecular dataset with categorical and continuous variables representing atomic structure. Feature engineering and nonlinear models are employed to capture the complex relationships between molecular properties, accelerating the prediction process without sacrificing accuracy, and enabling large-scale analysis for chemical and biomolecular applications.**

***Index Terms*—Nuclear Magnetic Resonance (NMR), Scalar Coupling Constant, Density Funtional Theory (DFT), Machine Learning (ML)**

I. INTRODUCTION

La Resonancia Magnética Nuclear (RMN) es una técnica analítica ampliamente utilizada para el estudio detallado de estructuras moleculares, permitiendo obtener información sobre la dinámica, la geometría y el entorno químico a nivel atómico. Esta herramienta resulta esencial en campos como el diseño de fármacos, la química orgánica y la ciencia de materiales. Una de las propiedades más importantes derivadas de la RMN es la constante de acoplamiento escalar, que refleja la interacción indirecta entre núcleos atómicos a través de enlaces químicos, proporcionando indicios sobre la estructura tridimensional de las moléculas.

Sin embargo, la determinación precisa de esta constante mediante métodos de química cuántica, como la Teoría del Funcional de la Densidad (DFT), implica un alto costo computacional, lo que dificulta su aplicación en estudios a gran escala o en bases de datos moleculares extensas. Para superar esta limitación, se han desarrollado enfoques basados en aprendizaje automático (ML), que permiten predecir rápidamente las constantes de acoplamiento escalar a partir de características moleculares extraídas del análisis estructural.

Este proyecto propone un modelo supervisado de predicción de la constante de acoplamiento escalar utilizando un dataset con variables categóricas que representan tipos atómicos y variables continuas que describen distancias euclidianas entre átomos.

II. DESCRIPCIÓN DEL PROBLEMA

La Resonancia Magnética Nuclear (RMN) es una técnica analítica ampliamente utilizada en química, biología y ciencia de materiales para determinar la estructura, dinámica y entorno químico de las moléculas. Su capacidad para ofrecer información detallada a nivel atómico la convierte en una herramienta fundamental en el diseño de fármacos, análisis estructural de compuestos orgánicos, estudios biomoleculares y caracterización de materiales.

Dentro de los parámetros que se obtienen mediante RMN, la constante de acoplamiento escalar es esencial para describir la interacción indirecta entre núcleos atómicos a través de enlaces químicos, lo que permite inferir la estructura tridimensional de una molécula. Sin embargo, su cálculo preciso mediante métodos de química cuántica, como la Teoría del Funcional de la Densidad (DFT), requiere un alto costo computacional, dificultando su uso en estudios a gran escala.

Como alternativa al uso de la DFT, diferentes investigadores han propuesto soluciones basadas en aprendizaje automático (ML) para predecir las constantes de acoplamiento escalar de forma rápida y eficiente, reduciendo significativamente los tiempos de procesamiento y permitiendo el análisis masivo de estructuras moleculares. Esta alternativa ofrece un nuevo flujo de trabajo para elucidación estructural, acelerando la investigación científica, sin comprometer la precisión de los resultados.

A. Composición de la base de datos

El proyecto emplea tres datasets distintos: 1JHN, 2JHN y 3JHN, con 43,680, 119,059 y 166,613 muestras respectivamente. Cada uno contiene 73 variables, de las cuales 12 son categóricas y 61 numéricas continuas, incluyendo la variable objetivo scalar_coupling_constant.

Las variables categóricas corresponden a los tipos de átomos involucrados y a los índices de posición atómica en las moléculas. Estas se encuentran codificadas como enteros, por ejemplo: atom_2 toma valores en {1, 6, 7} en 1JHN, y en {6, 7, 8} en los otros datasets. Variables como atom_3, atom_4, ..., atom_9 pueden incluir el valor 0, que indica la ausencia de un átomo en cierta posición. Otras variables categóricas incluyen identificadores como atom_index_0, atom_index_1, molecule_index e id.

Las variables numéricas continuas representan principalmente distancias euclidianas entre pares de átomos, así como relaciones geométricas derivadas de la estructura molecular. Estas características cuantifican la disposición espacial de los átomos y son fundamentales para modelar sus interacciones.

La variable objetivo scalar_coupling_constant es una magnitud continua que mide la constante de acoplamiento escalar entre dos átomos dentro de una molécula. Esta propiedad refleja interacciones cuánticas que son clave para la predicción de propiedades químicas mediante modelos de aprendizaje automático.

- Datos faltantes e imputación:

La estrategia que se siguió para tratar potenciales datos faltantes fue la imputación con ceros (0), especialmente para variables de tipo atómico, que usan el valor 0 para indicar la ausencia o inexistencia de un átomo en esa posición. Esto fue posible gracias al procesamiento previo en la función take_n_atoms() y el uso de fillna(0).

Finalmente, no existen valores faltantes en el dataset final empleado.

B. Descripción inicial del conjunto de datos

En las figura 1, 2 y 3 se puede apreciar la distribución de la variable constante de acoplamiento (scalar_coupling_constant), para cada uno de los datasets.:

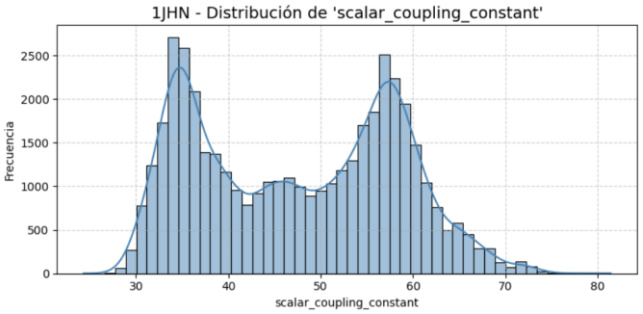

Figura 1. Distribución de la variable constante de acoplamiento para 1JHN

En el caso del acoplamiento directo entre átomos de hidrógeno y nitrógeno (1JHN), la distribución de la constante de acoplamiento escalar presenta un comportamiento bimodal, con dos modos prominentes centrados aproximadamente en 33–35 y 60 Hz, y un valle intermedio en el rango de 45–50 Hz. Esta distribución refleja la influencia de diferentes entornos electrónicos y conformaciones geométricas sobre el acoplamiento H–N. El primer pico (33–35 Hz) suele asociarse a enlaces H–N más débiles o geometrías con menor densidad electrónica compartida, mientras que el segundo pico (alrededor de 60 Hz) corresponde típicamente a enlaces más directos y efectivos, con mayor solapamiento orbital. Esta variabilidad es consistente con lo observado experimentalmente en compuestos orgánicos y estructuras tipo amida o amina, donde la geometría molecular y el grado de hibridación influyen directamente sobre los valores de acoplamiento escalar.

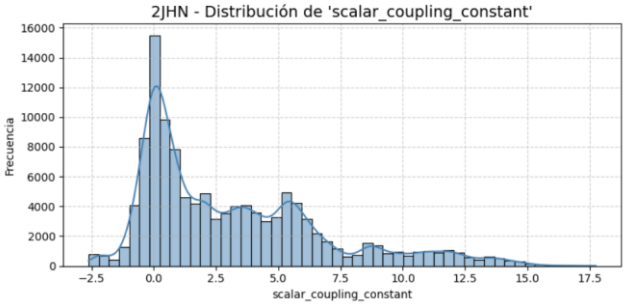

Figura 2. Distribución de la variable constante de acoplamiento para 2JHN

En el caso del acoplamiento escalar entre átomos de hidrógeno y nitrógeno separados por dos enlaces (2JHN), la distribución de la constante de acoplamiento presenta una forma asimétrica y centrada en valores bajos, con un modo principal entre aproximadamente -1 y 6 Hz. A diferencia del caso directo (1JHN), este tipo de acoplamiento se ve fuertemente influenciado por el ángulo de enlace, la conformación torsional y la presencia de efectos electrónicos de grupos vecinos. La menor magnitud del acoplamiento es esperable, dado que la interacción se transmite a través de dos enlaces, lo que reduce significativamente el solapamiento orbital efectivo.

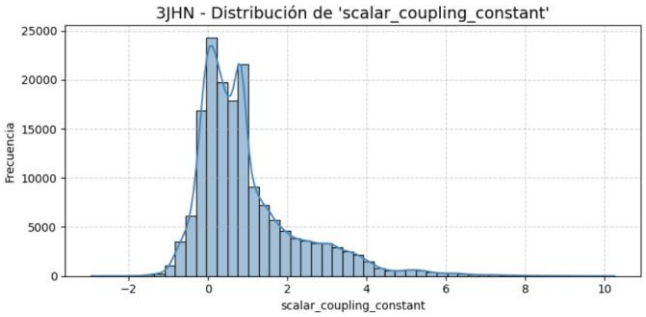

Figura 3. Distribución de la variable constante de acoplamiento para 3JHN

La distribución de la constante de acoplamiento escalar para 3JHN presenta una forma claramente asimétrica, centrada en valores bajos. El máximo de frecuencia se encuentra entre 0 y 1 Hz, con una leve bimodalidad dentro de ese mismo rango, evidenciando picos cercanos a 0 Hz y alrededor de 0.8 Hz. Se observa además una cola hacia valores más altos, que se extiende gradualmente hasta aproximadamente 8 Hz, aunque con frecuencia decreciente a partir de los 2 Hz. Esta distribución es coherente con la naturaleza del acoplamiento escalar a tres enlaces entre Hidrógeno y Nitrógeno (3JHN), cuya magnitud promedio es baja debido a la distancia electrónica involucrada en la transmisión del acoplamiento.

La matriz de correlación (Figura 4) evidencia la relación lineal entre las variables numéricas del dataset, que corresponden a distancias euclidianas entre átomos dentro de las moléculas, junto con la variable objetivo, la constante de acoplamiento escalar (scalar_coupling_constant).

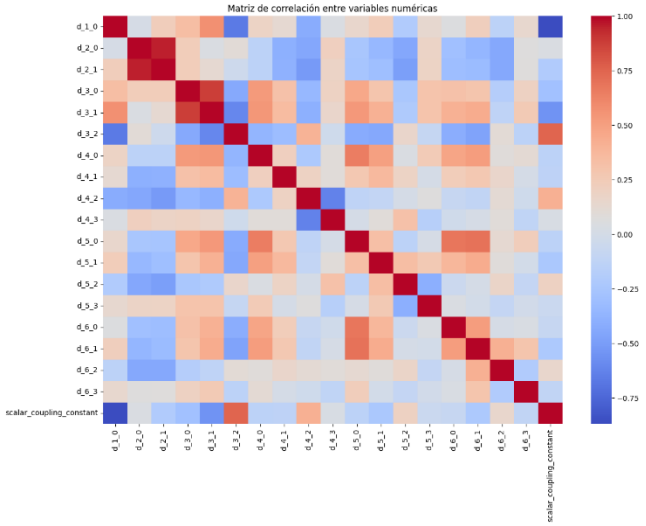

Figura 4. Matriz de correlación entre las variables

En la figura 4 se observa que varias distancias atómicas (d_i_j) están altamente correlacionadas entre sí, reflejando la estructura espacial coherente de las moléculas. Asimismo, algunas de estas distancias muestran una relación notable con la constante de acoplamiento escalar, señalando variables potencialmente relevantes para su predicción.

Adicionalmente, se observa que todas las variables que representan distancias entre átomos (d_i_j) presentan algún grado de correlación con la constante de acoplamiento escalar. por ejemplo, la variable d_6_3 muestra una correlación negativa significativa con scalar_coupling_constant, lo que sugiere que a medida que aumenta la distancia entre los átomos correspondientes, el valor del acoplamiento tiende a disminuir. en contraste, otras variables como d_4_2 presentan una correlación positiva moderada, indicando que en ciertos casos, mayores distancias pueden estar asociadas con valores más altos del acoplamiento, posiblemente debido a la geometría molecular o al tipo específico de interacción entre átomos.

C. Paradigma

Para llevar a cabo el proyecto, se seguirá un enfoque de aprendizaje supervisado para predecir la constante de acoplamiento escalar entre átomos, ya que cuenta con datos etiquetados que relacionan características moleculares con valores conocidos. Este paradigma permitirá que el modelo aprenda a partir de ejemplos para realizar predicciones precisas sobre datos nuevos.

III. ESTADO DEL ARTE

Fang *et al.* [1] abordan el problema de la predicción de la constante de acoplamiento escalar (SCC) utilizando técnicas de aprendizaje automático, con énfasis en la interpretación química. Los autores proponen un modelo denominado Graph Angle-Attention Neural Network (GAANN), el cual pertenece al paradigma de aprendizaje supervisado y está basado en redes neuronales de grafos (GNN) con un mecanismo de atención angular. Esta arquitectura permite representar estructuras moleculares, incorporando información sobre los ángulos entre átomos, lo cual enriquece la capacidad del modelo para predecir propiedades moleculares. Para la validación del modelo, los autores emplean un esquema de evaluación experimental con división entrenamiento/prueba sobre un conjunto de moléculas pequeñas. La métrica utilizada para evaluar el rendimiento es el logaritmo del error absoluto medio (log(MAE)), alcanzando un valor de log(MAE) = −2.52, comparable al obtenido mediante métodos de química cuántica como la teoría del funcional de la densidad (DFT).

Por su parte, Yiu *et al.* [2] abordaron el problema de la predicción de parámetros espectroscópicos de resonancia magnética nuclear (NMR), tales como los desplazamientos químicos y las constantes de acoplamiento escalar (SCC), mediante técnicas de aprendizaje automático, con el objetivo de reemplazar métodos de química cuántica como la teoría del funcional de la densidad (DFT). Los autores proponen un modelo denominado IMPRESSION generación 2

(IMPRESSION-G2), el cual se enmarca en el paradigma de aprendizaje supervisado. Esta red neuronal incorpora un mecanismo de atención que le permite modelar de forma simultánea y eficiente tanto los desplazamientos químicos (^{1}H, ^{13}C, ^{15}N y ^{19}F) como las constantes de acoplamiento J hasta cuatro enlaces de distancia, a partir de estructuras moleculares tridimensionales.

Para el entrenamiento del modelo, los autores utilizaron un conjunto de datos compuesto por más de 18.000 moléculas obtenidas de bases de datos como CSD, ChEMBL y OTAVA, con propiedades NMR simuladas por DFT, lo cual permitió generar más de 740.000 entornos de desplazamiento químico y 5.7 millones de valores de J. El modelo fue validado mediante una comparación directa con resultados de DFT y con datos experimentales de alrededor de 5.000 compuestos, mostrando errores medios absolutos (MAE) de 0.07 ppm para ^{1}H, 0.8 ppm para ^{13}C, y menores a 0.15 Hz para constantes de acoplamiento como ^{3}JHH, logrando así una precisión comparable a la de los métodos cuánticos.

En términos de eficiencia, IMPRESSION-G2 es capaz de realizar predicciones en menos de 50 milisegundos por molécula, lo que representa una mejora de hasta seis órdenes de magnitud en velocidad frente a DFT.

Zhang *et al.* [3] abordan el problema de la predicción de constantes de acoplamiento escalar entre pares atómicos en moléculas, proponiendo un modelo de red neuronal profunda basado en bloques Dense, apoyado en ingeniería de características para extraer atributos relevantes de las moléculas. Esta arquitectura permite predecir propiedades moleculares con alta precisión, superando a métodos populares de aprendizaje automático como XGBoost y LightGBM. Para la validación, utilizan un conjunto de datos molecular proveniente de una competencia de Kaggle, evaluando el desempeño mediante métricas de error. Los resultados muestran que su modelo reduce el error en 0.4 respecto a XGBoost y en 0.3 respecto a LightGBM, demostrando la eficacia de la red neuronal profunda para esta tarea.

IV. ENTRENAMIENTO Y EVALUACIÓN DE MODELOS.

En este proyecto los modelos seleccionados incluyen enfoques lineales, no paramétricos, y de aprendizaje profundo: regresión Ridge (que incorpora regularización para mitigar el sobreajuste), K-Nearest Neighbors (KNN, un modelo no paramétrico que predice en función de la similitud local), redes neuronales artificiales, y LightGBM (un modelo de ensamble basado en árboles de decisión).

A. Regresión lineal regularizada (Ridge Regression)

Se evaluó el desempeño del modelo de regresión Ridge para predecir constantes de acoplamiento escalar en tres tipos diferentes: 3JHN, 2JHN y 1JHN. Para cada tipo, se determinó el parámetro de regularización óptimo (λ o alpha) mediante validación cruzada y se compararon métricas clave como el error cuadrático medio (MSE) y el coeficiente de determinación (R^2) entre los conjuntos de entrenamiento, validación y test. Los resultados se presentan en la tabla 1.

Tipo de acoplamiento	λ óptimo (alpha)	MSE validación (CV)	MSE entrenamiento	R^2 entrenamiento	MSE test	R^2 test
3JHN	1.00000	1.18612	1.18460	0.31758	1.21800	0.31490
2JHN	0.01000	3.26459	3.25815	0.75906	3.28609	0.75627
1JHN	0.01000	4.52822	4.49314	0.96214	4.54437	0.96185

Tabla 1. Resultados para Ridge Regression

Respecto al parámetro λ, para el acoplamiento 3JHN se seleccionó un valor moderado de 1.0, indicando una mayor regularización, mientras que para 2JHN y 1JHN el valor óptimo fue mucho menor (0.01), lo que sugiere que en estos casos el modelo requiere poca penalización para ajustarse adecuadamente a los datos.

Al analizar el MSE, se observa que para los tres tipos, el error en validación y en entrenamiento es muy similar. En particular, para 3JHN el MSE es prácticamente idéntico en ambos conjuntos (alrededor de 1.18), lo que indica un buen balance entre ajuste y generalización sin indicios de sobreajuste. En 2JHN y 1JHN también se observa una concordancia estrecha entre MSE de entrenamiento y validación, confirmando que el modelo generaliza bien y no está sobreajustando.

En cuanto al coeficiente de determinación, los valores para 3JHN son modestos (R^2 cerca de 0.31 en entrenamiento y test), señalando que el modelo explica una fracción limitada de la varianza, posiblemente debido a la mayor complejidad o ruido de este tipo de acoplamiento. Por el contrario, para 2JHN y 1JHN, los R^2 alcanzan valores altos (aproximadamente 0.76 y 0.96 respectivamente), lo que indica un poder predictivo mayor para este acoplamiento y estabilidad del modelo en datos no vistos.

Finalmente, al evaluar la magnitud de los coeficientes estimados, se encuentra que el promedio y el valor máximo de los coeficientes absolutos aumentan considerablemente desde 3JHN hasta 1JHN. Esto refleja que para 1JHN el modelo permite coeficientes más grandes, probablemente porque la regularización es menor, mientras que para 3JHN la penalización más fuerte restringe la amplitud de los coeficientes. Esta diferencia también puede reflejar variaciones en la escala o relevancia de las características para cada tipo de acoplamiento. Esta situación se ve reflejada en las figuras 5, 6 y 7.

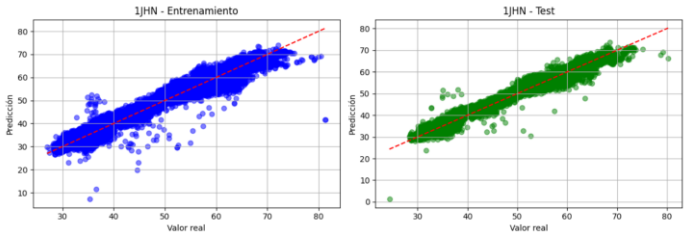

Figura 5. Predicción vs valor real para acoplamiento 1JHN

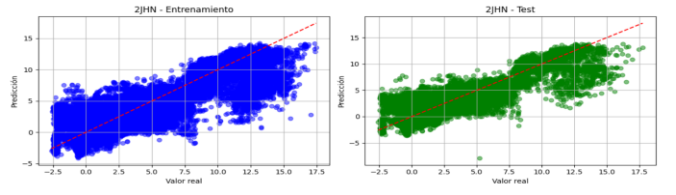

Figura 5. Predicción vs valor real para acoplamiento 2JHN

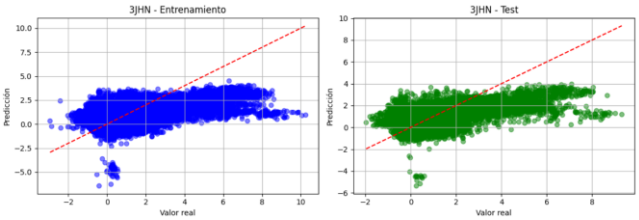

Figura 6. Predicción vs valor real para acoplamiento 3JHN

B. Modelo basado en K-vecinos más cercanos (KNN)

El modelo utilizado es un regresor basado en K-vecinos más cercanos (KNN), que predice la constante de acoplamiento escalar a partir de las características moleculares disponibles. Este método no paramétrico estima el valor objetivo promediando los valores de los puntos más cercanos en el espacio de características, ponderados por la distancia. En este caso, se estandarizaron las variables y se fijó un único conjunto de hiperparámetros óptimos (5 vecinos, métrica Euclidiana y pesos según la distancia) encontrados mediante validación cruzada con ShuffleSplit. Los resultados se presentan en la tabla 2:

Tipo de acoplamiento	MSE train	MSE test	R^2 train	R^2 test	Diff R^2 (Train-Test)	Overfitting Score
3JHN	0.00000	0.35706	1.00000	0.79916	0.20084	-0.37810
2JHN	0.00000	0.62632	1.00000	0.95355	0.04645	-0.69042
1JHN	0.00000	5.85583	1.00000	0.95084	0.04916	-6.23215

Tabla 2. Resultados para KNN

Para los tres tipos de acoplamiento (3JHN, 2JHN y 1JHN), el modelo KNN mostró un desempeño sobresaliente en el conjunto de entrenamiento, alcanzando un error cuadrático medio (MSE) prácticamente nulo y un R^2 perfecto (1.0).

Sin embargo, al evaluar el desempeño en test, se observa un incremento en el error (MSE) y una reducción en R^2, aunque los valores permanecen altos para 2JHN y 1JHN (R^2 test ≈ 0.95), mostrando excelente capacidad predictiva y generalización. En contraste, para 3JHN el R^2 test es menor (~0.80), indicando una predicción algo menos precisa para este tipo de acoplamiento, pero aún aceptable.

La diferencia entre R^2 de entrenamiento y test (Diff R^2) revela un descenso en la capacidad explicativa fuera de muestra, más notable para 3JHN (≈0.20) que para los otros casos (≈0.05). Esta situación refleja sobreajuste en el modelo.

El puntaje de sobreajuste calculado como la diferencia entre MSE de entrenamiento y validación cruzada es negativo en todos los casos, indicando que el error de entrenamiento es menor que el de validación.

C. Modelo LightGBM (Light Gradient Boosting Machine)

El modelo utilizado en este caso es LightGBM (Light Gradient Boosting Machine), un algoritmo de ensamble basado en árboles de decisión, diseñado para ofrecer alta precisión y velocidad en tareas de regresión y clasificación. Utiliza una estrategia de "boosting" que construye árboles secuencialmente, corrigiendo errores cometidos por modelos anteriores. En este análisis, se entrenó una versión optimizada del modelo utilizando una rejilla reducida de hiperparámetros (100 árboles, tasa de aprendizaje 0.1, sin límite de profundidad y 31 hojas por árbol) y se aplicó validación cruzada con 3 particiones para garantizar una evaluación robusta. Los resultados se reportan en la tabla 3:

	Tipo de acoplamiento	MSE train	MSE test	MSE CV (validación)	R2 train	R2 test
0	3JHN	0.16157	0.17469	0.17317	0.90692	0.90174
1	2JHN	0.22159	0.24304	0.25593	0.98361	0.98197
2	1JHN	0.62580	0.78966	0.84944	0.99473	0.99337

Tabla 3. Resultados para LightGBM

Los resultados muestran que LightGBM ofrece un alto desempeño predictivo en los tres tipos de acoplamiento, tanto en entrenamiento como en test. En particular, se observa una alta capacidad de generalización con R^2 superiores al 0.90 en todos los casos. Para el tipo 1JHN, el modelo alcanzó un R^2 de 0.9947 en entrenamiento y 0.9934 en test, con un MSE test bajo (0.7897), evidenciando una predicción bastante precisa. Algo similar ocurre con 2JHN, donde el R^2 en test fue de 0.9820 y el MSE de 0.2430, reflejando una buena concordancia entre los valores reales y predichos.

En el caso de 3JHN, el rendimiento fue ligeramente menor, pero aún sobresaliente: el modelo logró un R^2 de 0.9017 en test y un MSE de 0.1747, lo cual indica que más del 90% de la varianza de los datos fue explicada por el modelo. Además, los valores de MSE de validación obtenidos mediante cross-validation fueron consistentes con los resultados en test, lo que refuerza la estabilidad del modelo.

Un aspecto clave es que no se detectó sobreajuste en ninguno de los casos. La diferencia entre R^2 de entrenamiento y test fue pequeña, y el criterio definido para marcar sobreajuste (una diferencia de más de 0.15 en R^2 combinada con un RMSE sustancialmente menor en entrenamiento) no se cumplió en ningún escenario. Esto sugiere que la regularización implícita

de LightGBM, junto con la selección adecuada de hiperparámetros, permitió un ajuste sólido sin comprometer la capacidad de generalización.

D. Modelo de Red Neuronal Artificial (ANN)

Se implementó un modelo de Red Neuronal Artificial (ANN) utilizando Keras/TensorFlow con el objetivo de predecir las constantes de acoplamiento escalar a partir de características moleculares. La arquitectura adoptada fue intencionalmente sencilla para evitar sobreajuste y mantener la eficiencia: una red de tipo feedforward con una capa oculta de 64 neuronas y activación ReLU, seguida de una capa de salida lineal. Se utilizó el optimizador Adam y la función de pérdida fue el error cuadrático medio (MSE). Los datos fueron escalados previamente con StandardScaler, y se dividieron en entrenamiento (64%), validación (16%) y test (20%). Los resultados se presentan en la tabla 4.

Tipo de acoplamiento	MSE entrenamiento	MSE validación	MSE test	R^2 entrenamiento	R^2 validación	R^2 test
3JHN	0.22655	0.25188	0.25068	0.86951	0.85480	0.85900
2JHN	0.21768	0.25763	0.27521	0.98381	0.98135	0.97959
1JHN	0.91741	1.11532	1.13754	0.99225	0.99067	0.99045

Tabla 4. Resultados para ANN

Los resultados obtenidos indican que el modelo ANN logra un desempeño sólido y generaliza bien para los tres tipos de acoplamiento. Para el tipo 1JHN, el modelo alcanzó un R^2 de 0.9905 en el conjunto de test, acompañado de un MSE test de 1.1375, lo que refleja su capacidad predictiva. Las métricas de entrenamiento y validación fueron muy cercanas, indicando un ajuste balanceado sin indicios de sobreajuste.

Un comportamiento similar se observó en el tipo 2JHN, donde el modelo alcanzó un R^2 de 0.9796 en test y un MSE test de 0.2752. Nuevamente, las métricas entre los tres subconjuntos son consistentes, lo que sugiere una generalización adecuada del modelo y una buena estabilidad durante el entrenamiento.

Para el tipo 3JHN, el desempeño fue ligeramente inferior en comparación con los otros dos tipos, con un R^2 test de 0.8590 y un MSE test de 0.2507. Aunque menor, estos valores siguen siendo aceptables, especialmente considerando que este tipo de acoplamiento tiende a presentar mayor variabilidad y complejidad en los datos. En este caso también se mantuvo la coherencia entre las métricas de entrenamiento, validación y test.

E. Modelo de vectores de soporte (SVM)

Se aplicó un modelo de regresión con máquinas de vectores de soporte (SVR, por sus siglas en inglés), utilizando el kernel radial básico (RBF) para capturar relaciones no lineales entre las variables moleculares y la constante de acoplamiento escalar. Se escalaron las características con StandardScaler y se realizó una división en subconjuntos de entrenamiento (64%), validación (16%) y prueba (20%). Se utilizó una búsqueda en rejilla con validación cruzada para seleccionar los mejores hiperparámetros (C, gamma y epsilon), lo que permitió afinar la regularización y la flexibilidad del modelo. Los resultados se presentan en la tabla 5:

Tipo de acoplamiento	MSE entrenamiento	MSE validación	MSE test	R^2 entrenamiento	R^2 validación	R^2 test
3JHN	0.04365	0.88701	0.74726	0.97459	0.54272	0.60513
2JHN	0.07066	1.52056	1.64640	0.99493	0.89049	0.88257
1JHN	0.65176	3.34653	4.68517	0.99451	0.97213	0.96032

Tabla 5. Resultados para SVM

El desempeño del modelo SVR fue heterogéneo entre los distintos tipos de acoplamiento, mostrando muy buenos resultados en algunos casos y ciertas limitaciones en otros.

En el caso de 1JHN, el modelo alcanzó un R^2 de 0.9603 en el conjunto de test, lo que indica un buen ajuste a los datos, aunque ligeramente inferior al alcanzado por otros modelos como ANN o LightGBM. El MSE de test fue de 4.6852, y el rendimiento se mantuvo estable entre entrenamiento, validación y test, sin evidencia clara de sobreajuste.

Para 2JHN, el desempeño también fue favorable: el modelo obtuvo un R^2 de 0.8826 en test y un MSE de 1.6464, con resultados consistentes en los tres subconjuntos, aunque el rendimiento general fue más bajo que el de modelos como LightGBM y ANN. La menor capacidad explicativa podría deberse a la naturaleza no lineal y local del SVR, sumada a la reducción de tamaño muestral.

En contraste, para 3JHN, el modelo SVR tuvo un desempeño más limitado. A pesar de que el R^2 en entrenamiento fue alto (0.9746), cayó drásticamente a 0.6051 en test, con un MSE test de 0.7473. Esta caída en desempeño sugiere un caso claro de sobreajuste, donde el modelo aprende patrones del conjunto de entrenamiento que no generalizan bien al conjunto de prueba. La baja R^2 en validación (0.5427) refuerza esta hipótesis.

En todos los casos, los hiperparámetros óptimos seleccionaron valores altos de C (10), lo que implica menor regularización, junto con un epsilon muy pequeño (0.01), favoreciendo un ajuste muy cercano a los datos. Esto pudo haber contribuido al sobreajuste observado, particularmente en 3JHN.

V. Reducción de Dimensión

A. Evaluación con LightGBM

Con el objetivo de evaluar el impacto de la reducción de dimensión sobre el rendimiento predictivo del modelo, se aplicaron tres enfoques distintos utilizando LightGBM sobre submuestras de 5000 observaciones por tipo de acoplamiento: sin reducción (modelo completo), selección de las 20 variables más importantes según la importancia de atributos del propio

modelo, y reducción mediante Análisis de Componentes Principales (PCA), conservando el 95% de la varianza explicada. Cada uno de estos enfoques fue evaluado en términos del error cuadrático medio (MSE) tanto en entrenamiento como en prueba. Los resultados se pueden observar en la tabla 6:

Tipo	Reducción	Nº variables/componentes	MSE Train	MSE Test
1JHN	Sin reducción	71	0.28177	1.21896
1JHN	Top 20 Features (Importancia)	20	0.36747	1.07562
1JHN	PCA (95%)	45	2.13223	11.32829
2JHN	Sin reducción	71	0.10575	0.69863
2JHN	Top 20 Features (Importancia)	20	0.13308	0.63926
2JHN	PCA (95%)	46	0.45097	2.29776
3JHN	Sin reducción	71	0.06936	0.34873
3JHN	Top 20 Features (Importancia)	20	0.08922	0.33153
3JHN	PCA (95%)	47	0.19962	0.95166

Tabla 6. Resultados para modelo LightGBM tras realizar PCA

Para el acoplamiento 1JHN, el modelo sin reducción obtuvo un MSE de prueba de 1.219. Al aplicar selección de características basada en importancia, utilizando solo las 20 variables más relevantes, se logró una ligera mejora, alcanzando un MSE test de 1.076. Esto sugiere que gran parte de la información predictiva está concentrada en un subconjunto de atributos. En contraste, el uso de PCA (45 componentes) provocó un deterioro notable en el rendimiento del modelo, con un MSE test de 11.328 y un aumento considerable del error en entrenamiento, indicando una pérdida sustancial de relación con el target durante la proyección.

En el caso del acoplamiento 2JHN, se evidenció un patrón similar. El modelo completo alcanzó un MSE de prueba de 0.699, mientras que el uso de las 20 variables más importantes redujo ligeramente el error a 0.639. Nuevamente, esto indica que una selección informada de características puede mantener (o incluso mejorar) el desempeño con menor complejidad. No obstante, la reducción con PCA (46 componentes) elevó el error a 2.298, sugiriendo que las transformaciones lineales aplicadas por PCA no logran preservar adecuadamente las relaciones predictivas en este caso.

Para el tipo de acoplamiento 3JHN, los tres enfoques ofrecieron resultados más cercanos entre sí. El modelo completo obtuvo un MSE de prueba de 0.349, y la reducción por importancia a 20 variables no afectó negativamente el desempeño, registrando un error similar de 0.332. En cambio, PCA (47 componentes) incrementó el MSE test a 0.952, aunque en una magnitud menor que en los casos anteriores. Esto puede indicar que el patrón de variabilidad en 3JHN es más uniforme o menos sensible a proyecciones lineales.

En resumen, la selección de características mediante la importancia derivada del modelo se comportó de forma más robusta en los tres casos, permitiendo reducir la dimensionalidad sin sacrificar capacidad predictiva, e incluso mejorándola ligeramente en algunos casos. Por el contrario, la reducción por PCA mostró un impacto negativo constante, lo cual sugiere que esta técnica no es adecuada para este tipo de problema, posiblemente debido a la pérdida de interpretabilidad y relaciones no lineales esenciales para la predicción. La selección basada en importancia se posiciona, por tanto, como la estrategia de reducción de dimensión más recomendable en este contexto.

VI. Discusión

El presente estudio abordó la predicción de constantes de acoplamiento escalar en enlaces tipo H-N (1JHN, 2JHN, 3JHN) mediante diferentes modelos de aprendizaje supervisado, con énfasis en la comparación de técnicas basadas en regresión, métodos no paramétricos, ensambles y redes neuronales, así como la evaluación del impacto de estrategias de reducción de dimensión.

Entre los modelos evaluados, LightGBM se destacó consistentemente por su equilibrio entre precisión y eficiencia computacional. Para todos los tipos de acoplamiento, este algoritmo logró errores cuadráticos medios (MSE) bajos y coeficientes de determinación (R^2) superiores al 0.90 en conjunto de prueba, sin evidencias de sobreajuste. Estos resultados son consistentes con los reportes de Zhang et al. [3], quienes también posicionan a LightGBM como una técnica competitiva para tareas de predicción molecular, aunque su modelo de red neuronal profunda mostró mejoras adicionales al incorporar bloques densos y mayor capacidad expresiva. En este estudio, sin embargo, LightGBM resultó más interpretable y eficiente que las redes neuronales en su configuración reducida, especialmente cuando se evaluaron variantes con selección de atributos.

El modelo KNN, si bien conceptualmente simple, logró desempeños notables en los casos 1JHN y 2JHN, con R^2 superiores al 0.95. No obstante, presentó señales de sobreajuste en el tipo 3JHN, con una diferencia marcada entre el error de entrenamiento (cercano a cero) y el de prueba, lo cual sugiere una alta sensibilidad a la complejidad estructural de los datos. Esta limitación refuerza la necesidad de métodos más sofisticados para capturar la heterogeneidad espacial y topológica de los entornos moleculares, como los modelos GNN utilizados por Fang et al. [1], quienes emplean mecanismos de atención angular para representar explícitamente relaciones tridimensionales entre átomos.

Las redes neuronales artificiales (ANN) mostraron un rendimiento robusto y generalizable, especialmente para los acoplamientos más cortos (1JHN y 2JHN), alcanzando R^2 de prueba superiores a 0.98 y errores medios comparables a los de LightGBM. Aunque se utilizó una arquitectura sencilla (una capa oculta), la ANN logró capturar no linealidades relevantes del problema. Sin embargo, a diferencia del modelo IMPRESSION-G2 de Yiu et al. [2], que integra múltiples

propiedades espectroscópicas con estructuras moleculares 3D y atención contextual, el modelo aquí propuesto no explota información estructural directa, lo cual limita su aplicabilidad frente a arquitecturas de frontera. Pese a ello, su facilidad de implementación y precisión justifican su inclusión en pipelines de predicción de SCC.

Por otro lado, el modelo SVM, alcanzó desempeños aceptables en los tres tipos de acoplamiento. El mejor comportamiento se observó en 1JHN (R^2 de 0.96), aunque con un deterioro progresivo hacia 3JHN, donde el poder predictivo fue más limitado ($R^2 \approx 0.61$). Estos resultados reflejan la rigidez del kernel RBF para capturar patrones complejos en datos con alta dimensionalidad y correlación, situación en la que métodos como LightGBM o redes profundas muestran mayor flexibilidad.

Finalmente, se exploró la influencia de la reducción de dimensión, utilizando tres enfoques: el modelo completo, selección por importancia de atributos y extracción de componentes con PCA. En general, se observó que la selección de las 20 variables más importantes permitió mantener o incluso mejorar el rendimiento predictivo en comparación con el modelo completo, reduciendo la complejidad del modelo sin pérdida significativa de información. Por el contrario, la reducción mediante PCA condujo a una degradación sustancial del desempeño en todos los tipos de acoplamiento, particularmente en 1JHN, donde el MSE se incrementó casi diez veces. Este comportamiento se debe a que PCA proyecta las variables en combinaciones lineales que pueden diluir relaciones no lineales importantes para la predicción, lo cual es especialmente problemático en datos moleculares con interacción compleja entre atributos.

REFERENCES

[1] J. Fang, L. Hu, J. Dong, H. Li, H. Wang, H. Zhao, Y. Zhang y M. Liu, “Predicting scalar coupling constants by graph angle-attention neural network,” *Scientific Reports*, vol. 11, art. no. 18686, 2021, doi: 10.1038/s41598-021-97146-1.

[2] C. Yiu, B. Honoré, W. Gerrard, J. Napolitano-Farina, D. Russell, I. M. L. Trist, R. Dooley y C. P. Butts, “IMPRESSION generation 2 – accurate, fast and generalised neural network model for predicting NMR parameters in place of DFT,” *Chemical Science*, vol. 16, pp. 8377–8382, 2025, doi: 10.1039/d4sc07858f.

[3] T. Zhang, Y. Cui, M. Wang y W. Cheng, “Molecular Magnetic interactions Properties Predicting using Neural Network,” en *Proc. 2021 IEEE 3rd International Conference on Communications, Information System and Computer Engineering (CISCE)*, Beijing, China, 14–16 May 2021, pp. 875–878, doi: 10.1109/CISCE52179.2021.9445909.