MODELOS HÍBRIDOS BASADOS EN MACHINE LEARNING PARA SUPERAR LAS LIMITACIONES DE LOS MÉTODOS DFT EN LA PREDICCIÓN DE PROPIEDADES QUÍMICAS

[2] José A. Perez Mendoza,[1,2] Bruno A. Franco,[2] Ezequiel R. Luciano,[2] María M. Zanardi,[2] Ariel M. Sarotti[1]

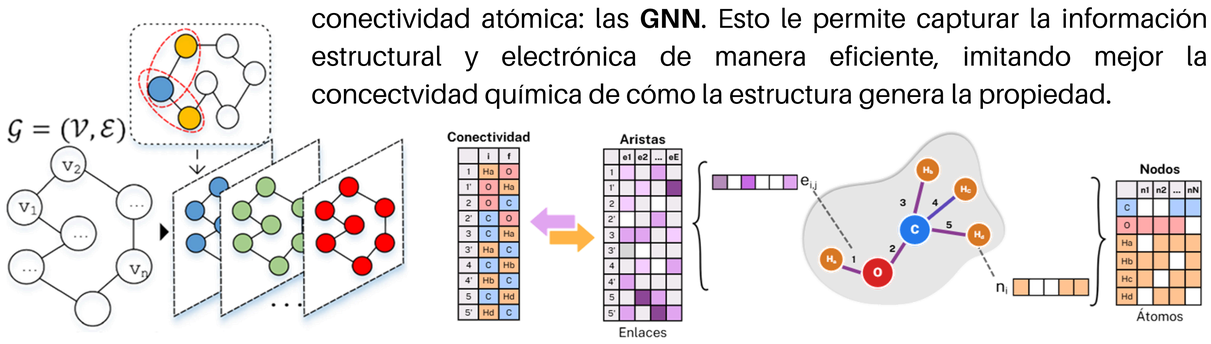
[1] Instituto de Química Rosario (IQUIR, CONICET-UNR) and Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Suipacha 531, Rosario, República Argentina.

[2] Instituto de Investigaciones en Ingeniería Ambiental, Química y Biotecnología Aplicada (INGEBIO), Facultad de Química e Ingeniería del Rosario, Pontificia Universidad Católica Argentina, S2002QEO Rosario, Argentina.

XXV SINAQO Mar del Plata 2025

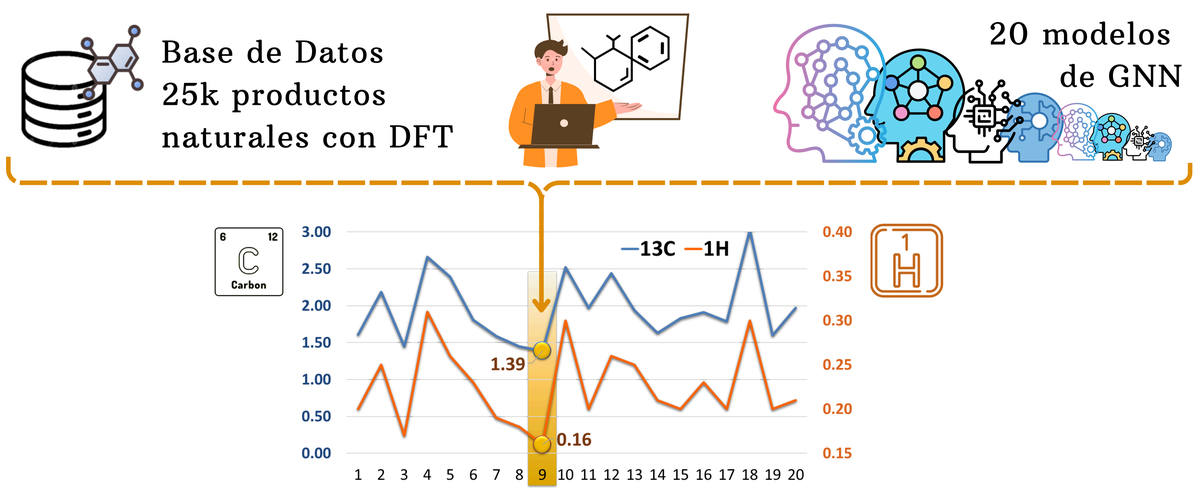
PREDICCIÓN DE RMN

Los **desplazamientos químicos** son altamente dependientes del entorno atómico (vecinos, enlaces, efectos a distancia). Existe una IA nativa de grafos capaz de procesar directamente la conectividad atómica: las **GNN**. Esto le permite capturar la información estructural y electrónica de manera eficiente, imitando mejor la concectvidad química de cómo la estructura genera la propiedad.

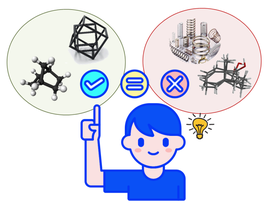


ENTRENAMIENTO CON MECÁNICA CUANTICA

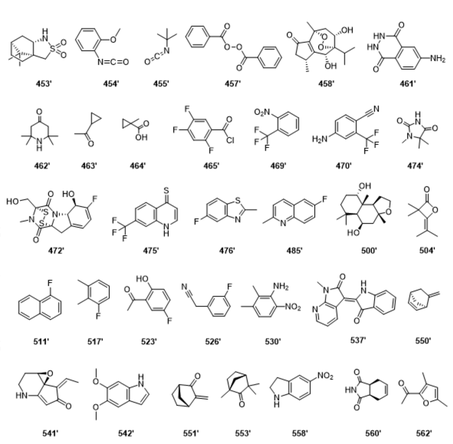
Se entrenó la GNN para predecir los desplazamientos químicos de 1H y13C a partir de datos DFT. Distintas arquitecturas de red fueron probadas para identificar el diseño con la mayor precisión, seleccionando el modelo con la mejor correlación con los valores de RMN.

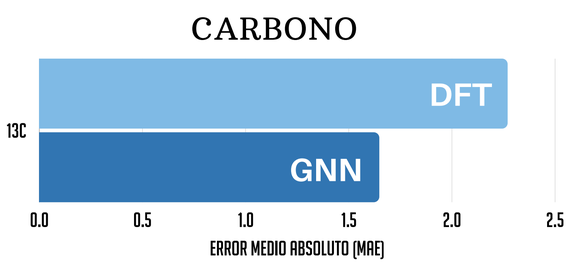


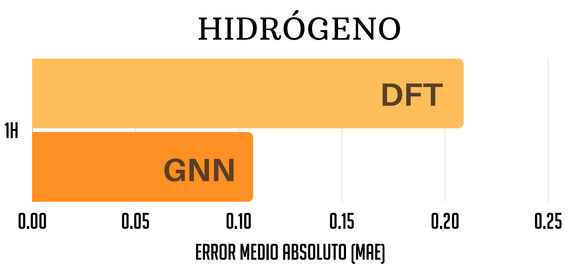
DESEMPEÑO A EXPERIMENTAL DE GNN



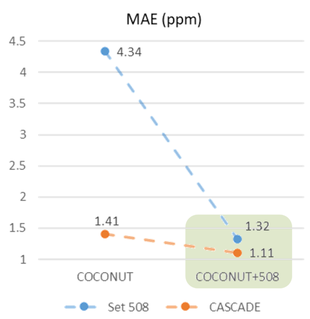
Para lograr una comparación directa y rigurosa, la validación del modelo se realizó en una base de datos externa de 140 moléculas orgánicas estructuralmente rígidas. Esta estrategia evita el complejo problema de la ponderación de Boltzmann de múltiples conformaciones. De esta manera, fue posible aislar el rendimiento predictivo del modelo y comparar directamente la precisión de la GNN y la DFT contra los datos experimentales de desplazamiento químico.



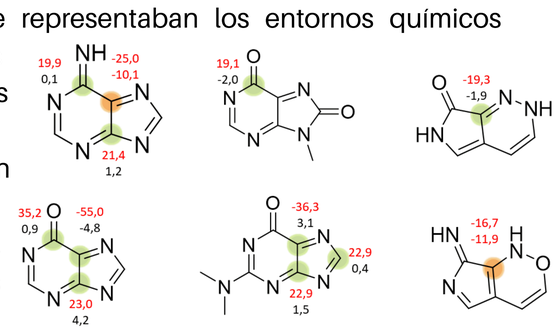


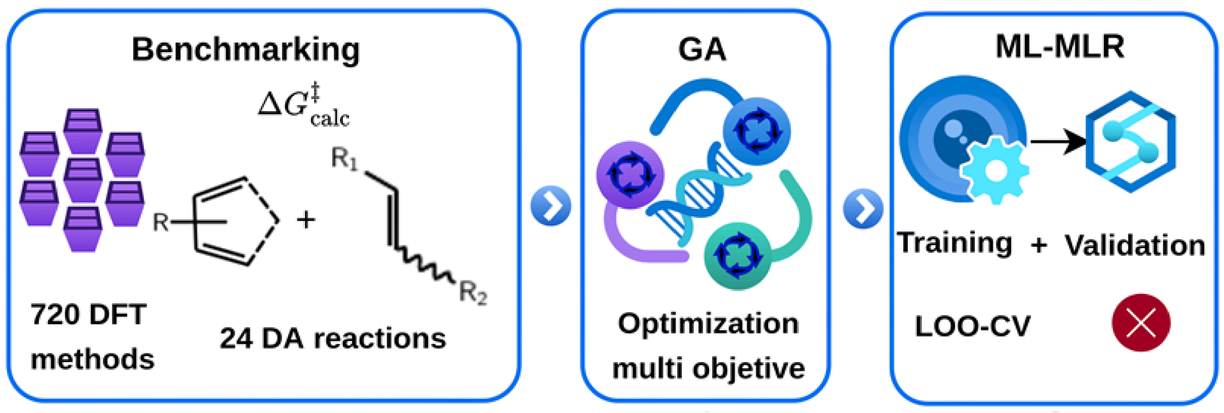


MEJORAS EN LA BASE DE DATOS



Se llevó a cabo una identificación recurrente de outliers en las predicciones de 13C. Para corregir estos errores, se anexaron dos conjuntos de estructuras que representaban los entornos químicos problemáticos (508 estructuras pequeñas y 220 purinas). Después de cada adición y re-entrenamiento, los modelos fueron capaces de corregir los errores de predicción previamente extremos, validando la importancia de datos de alta calidad.

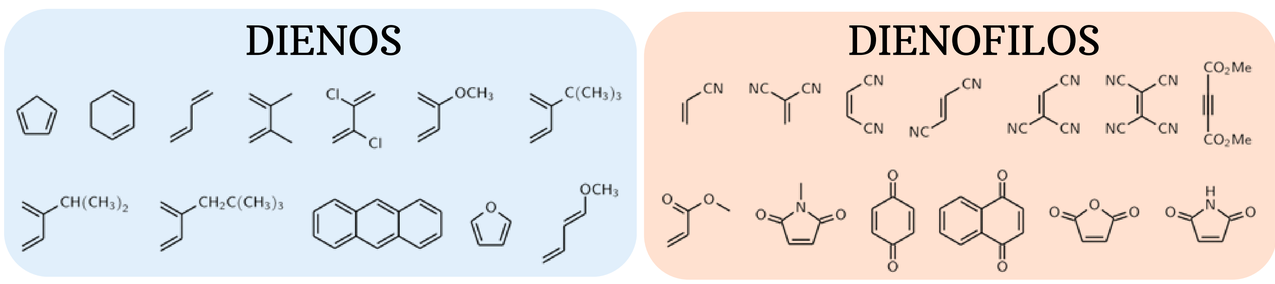




$\Delta G^{\ddagger}$ PARA DIELS-ALDER

La predicción precisa de la energía de Gibbs de activación ΔG‡ de reacciones Diels-Alder (DA) es crucial para la síntesis y el diseño de catalizadores. Pese a su uso, alcanzar la precisión química <1 kcal/mol consistentemente con los métodos de la Teoría del Funcional de la Densidad (DFT) sigue siendo un desafío. Recientemente, el **Machine Learning** (ML) y los **Algoritmos Genéticos** (GA) han surgido como potentes herramientas para superar las limitaciones de la DFT

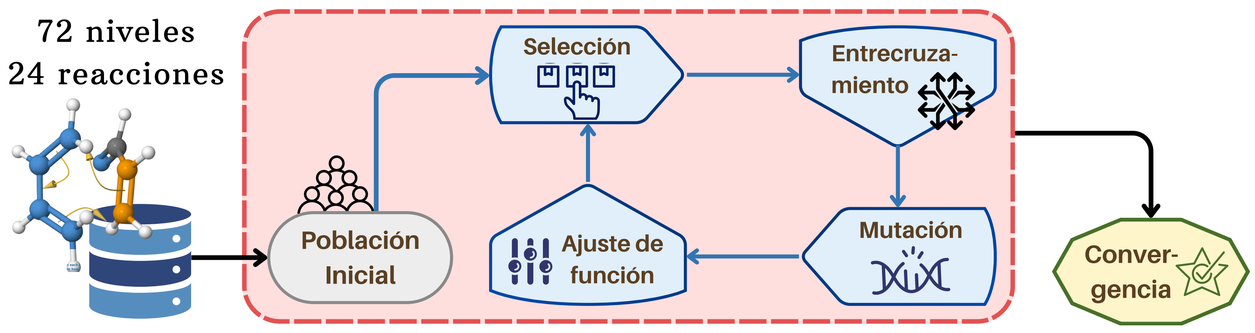
PREDICCIONES CON DENSIDAD FUNCIONAL (DFT)



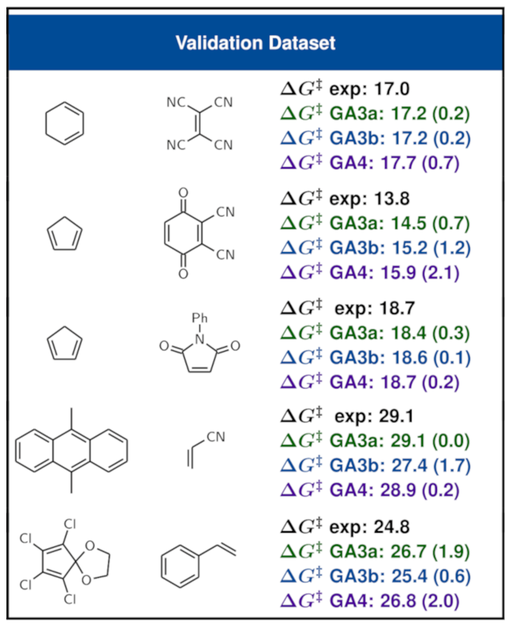
Se realizó un benchmarking exhaustivo de **720 niveles de teoría DFT** (combinando 15 funcionales, 16 bases y 3 modelos de solvatación) en 24 reacciones Diels-Alder. Este análisis demostró que **ningún nivel de DFT predice el ΔG‡** de manera universal y fiable. Aunque la referencia CBS-QB3 obtuvo un MAE de 0.9 kcal / mol (resultado bueno), la aparición de numerosos valores atípicos según el modelo de solvatación reafirma la limitación intrínseca del modelo DFT

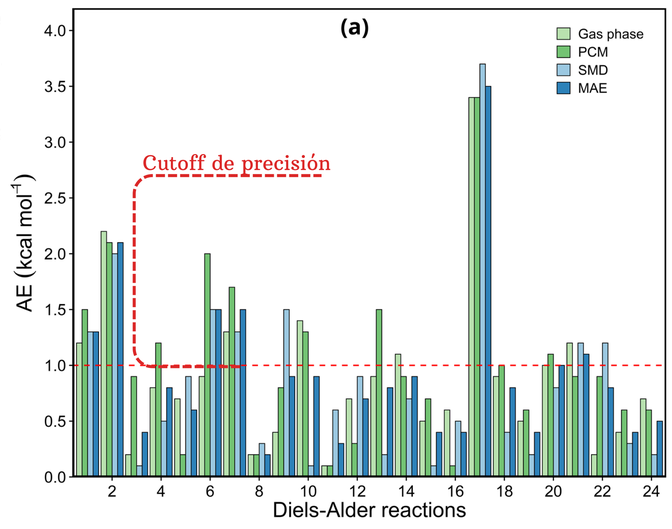
ASISTENCIA DE ALGORITMOS GENÉTIVOS Y ML

72 niveles
24 reacciones



Para superar estas limitaciones, se desarrollaron cinco modelos GA-ML inspirados en la "sabiduría de las masas", diseñados para identificar combinaciones óptimas de niveles de teoría de bajo costo.





Todos los modelos identificaron ensambles que, combinados linealmente, alcanzaron errores **inferiores a 1 kcal mol^{-1}** para las reacciones evaluadas. Entre ellos, el modelo GA4-ML se destacó al seleccionar un conjunto mínimo de cuatro niveles de teoría para lograr una precisión (MAE 0.4 kcal mol-1) comparable a la de métodos de alta precisión como CCSD(T).

REFERENCIAS

1. Houk, K. N.; Liu, F. Acc. Chem. Res. 2017, 50 (3), 539–543.
2. Lewis-Atwell, T.; Townsend, P. A.; Grayson. WIREs Comput. Mol. Sci. 2022, 12 (4), e1593.
3. Ai, W. J., Li, J., Cao, D., Liu, S., Yuan, Y. Y., Li, Y., ... & Wang, W. X. (2024). Journal of Natural Products, 87(4), 743-752.
4. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017, July). In International conference on machine learning (pp. 1263-1272). Pmlr.

FQO-038