Universidad Zaragoza
1542

Trabajo Fin de Grado

Modelos de difusión para generación de parámetros en redes bioquímicas

Difussion models for biochemical network parameter generation

Autora

Patricia Siwinska

Director

Alexandru Ioan Oarga Hategan

Ponente

Jorge Emilio Júlvez Bueno

ESCUELA DE INGENIERÍA Y ARQUITECTURA

2024

AGRADECIMIENTOS

Me gustaría agradecer a todas las personas que han cruzado mi camino y han confiado en mi potencial, tanto en el ámbito personal como académico. A mis profesores, por hacer crecer mi interés, conocimientos y pasión por la informática a lo largo de la carrera, aunque la empezara por curiosidad.

A mis padres, familiares y a mi compañero de cuatro patas por su apoyo constante e incondicional. Y, por último, pero no menos importante, a mi pareja y amigos en la universidad, por su confianza en mí y por ayudarme a reconocer capacidades y fortalezas que no siempre veía en mí misma.

RESUMEN

Las redes bioquímicas se representan mediante grafos, donde las moléculas biológicas, como compuestos y metabolitos, actúan como nodos conectados por interacciones físicas o funcionales. A través de estos procesos, se libera una energía conocida como energía libre de Gibbs, la cual ofrece información crucial sobre la termodinámica de las reacciones químicas, incluyendo su espontaneidad, la energía requerida y el equilibrio.

La estimación precisa de esta energía tiene múltiples aplicaciones en bioquímica, facilitando el estudio de la termodinámica de compuestos y la modelización computacional. En farmacología, es particularmente relevante para diseñar y validar rutas de síntesis de compuestos, ayudando a identificar y optimizar posibles reacciones químicas.

A pesar de que existen diversos algoritmos para predecir la energía libre de Gibbs, muchos de ellos no consideran de forma adecuada la estructura del grafo o requieren recursos computacionales elevados, además de que la cantidad de datos disponibles para este fin es limitada. Ante esta situación, los modelos de difusión surgen como una solución eficaz, ya que permiten la generación de múltiples muestras válidas que ayudan a estimar la incertidumbre en la predicción de esta energía. En lugar de buscar una predicción exacta, nuestro enfoque se centra en aproximarnos a la distribución de probabilidad de la energía libre de Gibbs. Dado que los grafos presentan simetrías, el uso de modelos deterministas puede conducir a resultados que, pese a partir de entradas distintas al tener la misma estructura ofrecen la misma salida.

Este trabajo explora la integración de distintos tipos de redes de grafos con modelos de difusión y analiza si esta combinación permite obtener una distribución de probabilidad fiable. A lo largo del estudio, se han realizado varias pruebas para evaluar la viabilidad de esta aproximación y comprobar su efectividad. Se examina también la influencia de distintas características en el entrenamiento y en las predicciones de la energía libre de Gibbs, así como el efecto de las perturbaciones en los modelos para, finalmente, exponer y discutir los resultados obtenidos.

En conclusión, este trabajo presenta un enfoque innovador en el campo de la bioquímica y la predicción de la energía libre de Gibbs, al combinar modelos de difusión con redes de grafos. Este enfoque puede suponer un avance significativo, ya que aprovecha la estructura inherente de las redes y la capacidad de manejar la incertidumbre, en lugar de depender únicamente de datos exactos. De esta manera, se abre la posibilidad de realizar predicciones más robustas y flexibles, lo que podría mejorar la comprensión de sistemas bioquímicos complejos y optimizar el diseño de reacciones químicas en contextos donde los datos son limitados o imprecisos.

Índice

Capítulo 1

Introducción

Las redes bioquímicas son sistemas complejos formados por interacciones moleculares, que incluyen proteínas, ácidos nucleicos y metabolitos. Estas redes, que se representan comúnmente como grafos, facilitan una variedad de procesos celulares, interconectando componentes que influyen mutuamente para mantener la función celular y la homeostasis. Un caso particularmente relevante de estas redes son las redes metabólicas, que describen cómo los metabolitos se interconectan a través de rutas bioquímicas, permitiendo la gestión eficiente de la energía y la adaptación celular. Este enfoque sistémico es clave para comprender fenómenos biológicos complejos, permitiendo a las células adaptarse a estímulos tanto internos como externos.

1.1. Contexto, motivación y estado del arte

La **energía libre de Gibbs** es un parámetro fundamental en termodinámica y en el estudio de las redes bioquímicas, ya que proporciona información sobre la espontaneidad de las reacciones químicas, sus requerimientos energéticos y los equilibrios asociados. Esta magnitud es clave para entender cómo los sistemas vivos gestionan la energía para mantener sus funciones metabólicas, predecir la viabilidad de reacciones y regular los procesos metabólicos celulares. Esta información es clave para aplicaciones como:

- **Bioquímica**: Estudio de la termodinámica de las reacciones bioquímicas, modelado computacional de redes bioquímicas y análisis de vías metabólicas [1][2].
- **Síntesis química y farmacología**: Predicción de la viabilidad de reacciones para la fabricación de medicamentos [3][4][5].

A pesar de su relevancia, la aplicación de la termodinámica en el metabolismo se ve limitada por la escasez de datos experimentales accesibles [6]. Las bases de datos públicas, que contienen la energía libre estándar de formación de reacciones, son muy limitadas, con solo unas pocas centenas de reacciones documentadas. Este déficit de datos resalta la necesidad de desarrollar métodos computacionales que estimen la energía libre de Gibbs de manera precisa y eficiente.

Existen varios enfoques computacionales para predecir la energía libre estándar de Gibbs basados en reglas de aditividad (redes bayesianas, modelos combinados, etc.) para la estimación de propiedades moleculares [7][8]. Sin embargo, estos métodos

suelen ser deterministas y no modelan adecuadamente la **incertidumbre** de las predicciones. Algunos intentan mitigar esta limitación entrenando múltiples modelos, lo que incrementa el costo computacional y la complejidad del proceso.

El enfoque propuesto en este trabajo se diferencia de los anteriores en que no depende de la estructura molecular de cada compuesto. En lugar de eso, utiliza la red de reacciones bioquímicas considerando la estructura global de la red, lo que permite realizar predicciones sin necesidad de datos moleculares explícitos. Este enfoque tiene la ventaja de que la información estructural de la red puede ser suficiente para estimar la energía libre de Gibbs de manera eficaz, ampliando así las aplicaciones sin depender de datos experimentales complejos.

Como consecuencia, proponemos un modelo basado en **modelos de difusión** para aproximar probabilísticamente la energía libre de Gibbs estándar en reacciones bioquímicas. Los modelos de difusión destacan por su capacidad para aprender distribuciones probabilísticas de manera efectiva, y su estabilidad y escalabilidad superior en comparación con otros enfoques generativos. Estos modelos han mostrado un rendimiento destacado en tareas complejas, como la generación de imágenes, lo que los convierte en una opción atractiva para el modelado de sistemas bioquímicos.

Una ventaja adicional de los modelos de difusión es su capacidad para realizar tareas de orientación [9] sin necesidad de reentrenar el modelo, lo que mejora su flexibilidad y reduce el costo computacional, permitiendo ajustar el modelo a nuevas tareas de manera eficiente.

Además, las **redes neuronales de grafos** (GNNs) han emergido como herramientas poderosas para el aprendizaje sobre grafos, y dado que las redes bioquímicas pueden representarse como grafos, las GNNs resultan ser una herramienta prometedora para identificar patrones y predecir el comportamiento de estos sistemas.

Trabajos previos han desarrollado diversas aproximaciones, como el uso de procesamiento de lenguaje natural para la síntesis de compuestos químicos [10], o el empleo de redes bayesianas y redes neuronales bayesianas para la estimación de la energía de reacciones [11] como se menciona anteriormente. Sin embargo, estas aproximaciones generalmente emplean conjuntos de datos y estructuras distintas.

Por lo tanto, nuestro enfoque reduce la complejidad computacional y el tiempo de entrenamiento, lo que representa una ventaja significativa en términos de eficiencia y precisión en la estimación de la energía libre de Gibbs.

1.2. Objetivos

Los objetivos principales de este trabajo son, en primer lugar, implementar un modelo basado en modelos de difusión [12][13] y redes de grafos (GNN) [14][15][16][17] para generar distribuciones probabilísticas de la energía libre de Gibbs en reacciones bioquímicas y conseguir modelar la incertidumbre.

En segundo lugar, se pretende verificar si es posible generar distribuciones de la energía libre de Gibbs partiendo de todos los parámetros del grafo, así como evaluar la capacidad de generar estas distribuciones utilizando únicamente la energía de formación de los compuestos químicos o la fórmula de los compuestos que intervienen y su carga.

Este trabajo también incluirá eventualmente el estudio de la viabilidad en escenarios con información limitada, para determinar cómo la falta de datos afecta la precisión de las predicciones. En los experimentos llevados a cabo durante la elaboración de

este trabajo se probarán modelos distintos para poder comparar su rendimiento e investigar si existe alguno que sea capaz de adaptarse mejor a los objetivos. Todos estos experimentos se harán sobre una red bioquímica de aproximadamente 24 mil nodos reconstruida a partir de varias bases de datos [18][19].

1.3. Estructura de la memoria

Esta memoria se organiza en cinco capítulos principales.

El **Capítulo 2** proporciona los conocimientos previos necesarios sobre los modelos de difusión, las redes neuronales de grafos (GNNs) y el conjunto de datos de reacciones bioquímicas, con el fin de contextualizar y facilitar la comprensión del proyecto.

El **Capítulo 3** se centra en la adaptación del modelo de difusión para predecir la energía libre de Gibbs y en la generación de la energía partiendo de todos los parámetros, lo que corresponde con el primer objetivo del proyecto. En esta sección, se integran las GNNs en el proceso, explicando las modificaciones realizadas al modelo y las pruebas iniciales con diferentes configuraciones de GNNs, incluyendo un muestreo inicial con tres muestras para visualizar la incertidumbre.

El **Capítulo 4** aborda el entrenamiento y la evaluación de un modelo obtenido al generar la energía libre de Gibbs utilizando como única característica la energía de formación de los compuestos. En esta sección se discuten las implicaciones de los resultados obtenidos.

El **Capítulo 5** plantea un desafío más complejo, en el que se intenta predecir la energía libre de Gibbs utilizando únicamente la fórmula química y la carga de los compuestos. Se evalúan los resultados y se reflexiona sobre la precisión del modelo utilizando esta información limitada.

El **Capítulo 6** presenta y discute los resultados finales y se analiza el comportamiento de los modelos en un escenario con datos que han sido perturbados.

Finalmente, en el **Capítulo 7** se redactan las conclusiones de este trabajo, además de una discusión sobre posibles líneas futuras de investigación y las mejoras que podrían implementarse en el enfoque propuesto.

Capítulo 2

Conocimientos Previos

2.1. Redes bioquímicas

Las reacciones químicas generalmente implican diferentes números de moléculas. Por ejemplo, la reacción 2 H2 + O2 →2 H2O implica 2 moléculas de hidrógeno (H) y 1 molécula de oxígeno (O2) para producir 2 moléculas de agua (H2O). Estos números se conocen como coeficientes estequiométricos de la reacción.

La estructura de la red se organiza como un **grafo** dirigido. Una arista que conecta un compuesto o metabolito con una reacción indica que el compuesto es consumido por la reacción (denotado como reactantes), mientras que una arista de una reacción hacia un compuesto indica que el compuesto es producido por la reacción (denotado como productos). Por ejemplo, en la siguiente figura ilustrativa, la reacción 2 consume los metabolitos R y C y produce el metabolito I, es decir, la reacción 2 se puede representar como R + C →I. De la misma forma, la reacción 3 consume I y produce los compuestos P y C y la reacción 1 consume R y produce P.

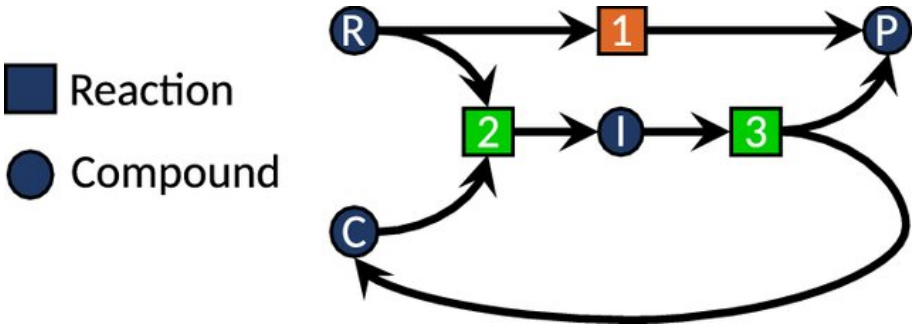

Figura 2.1: Representación del conjunto de datos mediante una red de Petri en la que se observa el consumo y producción de compuestos dada por reacciones.
[20]

El conjunto de datos utilizado en este trabajo está basado en una red que representa **reacciones bioquímicas**. Las reacciones se extrajeron de la base de datos MetanetX, mientras que la información de los compuestos proviene de MetaCyc.

Este conjunto de datos incluye 23,952 nodos y 68,496 aristas, donde cada arista está asociada a características específicas relacionadas con la reacción o metabolito correspondiente. En total, se utilizan 33 características para describir los nodos, lo que permite determinar si un nodo representa un compuesto o un reactante, su fórmula

química, la carga del compuesto y la energía libre de Gibbs estándar de formación. Estas características proporcionan una representación detallada de las propiedades bioquímicas de los nodos en la red.

2.2. Energía Libre de Gibbs

La **energía libre de Gibbs** (ΔG°) es una función termodinámica que proporciona información crucial sobre la espontaneidad de las reacciones químicas a temperatura y presión constantes. Esta magnitud termodinámica se define de la siguiente forma [21] :

$$\Delta G^\circ = \Delta H^\circ - T\Delta S^\circ \tag{2.1}$$

Donde:

- ΔG°: Variación de la energía libre de Gibbs de la reacción.
- ΔH°: Variación de la entalpía de la reacción.
- T: Temperatura.
- ΔS°: Variación de la entropía de la reacción.

Adicionalmente, se puede calcular la variación de la energía libre de Gibbs en una reacción química utilizando la siguiente relación, siempre que se disponga de los datos de todos los reactantes :

$$\Delta G^\circ = \sum_{c \in productos} \Delta G_f^\circ(c) - \sum_{c \in reactantes} \Delta G_f^\circ(c) \tag{2.2}$$

Donde:

- ΔG°: Variación de la energía libre de Gibbs de la reacción.
- $\Delta G_f^\circ(c)$: Energía libre de Gibbs de formación del compuesto c.
- $productos$: Conjunto de productos de la reacción.
- $reactantes$: Conjunto de reactantes de la reacción.

En nuestro caso, poseemos todos los datos de formación de los compuestos ($\Delta_f G^\circ$), lo que permite calcular de manera directa la energía libre de la reacción (ΔG°) mediante la ecuación correspondiente, ya que la red aprende a emularla. Sin embargo, nuestro objetivo no es únicamente replicar este cálculo, sino permitir que la red aprenda patrones subyacentes en los datos. Esto incluye generalizar a situaciones en las que ciertas propiedades no estén disponibles directamente, se enfrenten restricciones de datos o se opere con información incompleta o ruidosa. La eficacia del modelo se evaluará observando su capacidad para predecir ΔG° en el conjunto de evaluación, destacando su potencial en escenarios más complejos y prácticos.

2.3. Modelos de Difusión

En el contexto del aprendizaje automático, los **modelos de difusión** son una clase de modelos generativos utilizados para aprender a generar nuevas muestras que replican las características de un conjunto de datos original, como imágenes, sonidos o cualquier otro tipo de información estructurada.

El funcionamiento básico de este modelo tiene dos etapas que se detallan a continuación.

Difusión hacia adelante

Las distribuciones normales, o gaussianas, se describen por dos parámetros clave: la media, que representa el valor central de los datos, y la varianza, que indica la dispersión de estos valores. En el proceso de difusión, específicamente durante la fase inicial conocida como **difusión hacia adelante**, los valores se generan de manera gradual añadiendo ruido gaussiano a los valores previos en cada paso temporal. Este ruido se selecciona de tal forma que los valores generados se ajusten a una distribución normal. Así, comenzamos con un valor claro y, a medida que avanza el proceso, este se va distorsionando progresivamente hasta convertirse en un valor completamente aleatorio debido al ruido.

La forma en que se añade el ruido no es constante; varía en cada paso, lo que se denomina "programación de varianza"[22]. En nuestro caso, el ruido añadido sigue una programación lineal, lo que significa que la varianza del ruido aumenta de manera uniforme durante el proceso de difusión. Esto implica que la cantidad de ruido que se añade en cada paso de difusión es proporcional al paso en el que nos encontramos, con un incremento constante a lo largo de todas las iteraciones. Este tipo de programación es sencilla de implementar y ofrece un buen rendimiento. Sin embargo, también existen otras alternativas, como la programación sigmoide, cuadrática o de tipo coseno, cada una con diferentes formas de ajuste en el crecimiento del ruido durante el proceso [23].

Al final de la secuencia de pasos, los datos se aproximan a una distribución gaussiana isotrópica. La ecuación que describe el proceso de difusión hacia adelante,

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \tag{2.3}$$

Donde:

- $\mathbf{x}_0$: Representa el dato original, es decir, el valor limpio sin ningún tipo de ruido.
- $\mathbf{x}_t$: Es el dato en el paso de tiempo t, que contiene ruido gaussiano progresivamente añadido durante el proceso de difusión.
- β_t: Es el coeficiente que controla la cantidad de ruido añadido en cada paso temporal t. Este valor varía entre 0 y 1, y ajusta la cantidad de ruido que se añade en cada paso. Cuando β_t es cercano a 0, el ruido añadido es mínimo, y cuando β_t se aproxima a 1, el ruido se incrementa.
- $\mathbf{I}$: Es la matriz identidad. En este contexto, se usa para representar la varianza de la distribución normal, indicando que el ruido es aditivo y no tiene correlación entre las diferentes dimensiones de los datos.

- $\mathcal{N}$: Indica una distribución normal. El valor medio de esta distribución es $\sqrt{1-\beta_t}\mathbf{x}_{t-1}$, lo que significa que el valor de $\mathbf{x}_t$ se aproxima al valor anterior $\mathbf{x}_{t-1}$, escalado por $\sqrt{1-\beta_t}$. La varianza de la distribución es $\beta_t\mathbf{I}$, lo que implica que la dispersión de $\mathbf{x}_t$ depende del valor de β_t.

Como se ha mencionado, β_t varía entre 0 y 1, controlando así el nivel de ruido en cada paso de difusión. Cuando β_t es cercano a 0, el ruido añadido es mínimo, lo que implica que el modelo realiza pocas alteraciones en los datos; mientras que cuando β_t se acerca a 1, el ruido añadido es más significativo. En nuestro caso, β_t sigue una programación lineal entre 0 y 1, lo que permite un control gradual sobre el nivel de ruido durante el proceso de difusión. Este comportamiento lineal se debe a la forma en que hemos definido la programación de la varianza. La pendiente de este cambio depende de la definición precisa de β_t, y se puede ajustar para evaluar el rendimiento con diferentes programaciones. El proceso de difusión hacia adelante se puede ver ilustrado en la Figura 2.2.

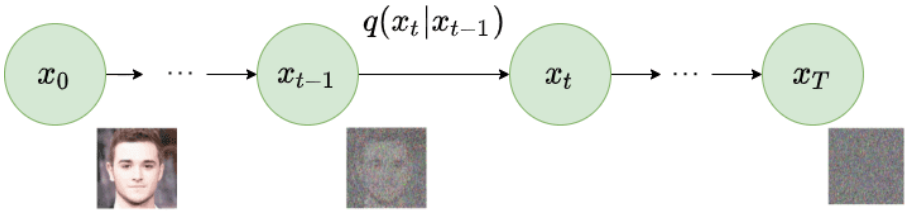

Figura 2.2: Proceso de difusión hacia adelante aplicado a una imagen [13]

Difusión hacia atrás o muestreo

La segunda etapa, llamada **difusión hacia atrás** se resume en que, gracias a los parámetros que ha aprendido esta red neuronal, se elimina el ruido y se reconstruye la imagen original a partir del estado de ruido puro. Es decir, partimos de unos datos que son muestras de ruido gaussiano, y de la forma inversa que en el primer paso, el modelo intenta deshacer el ruido con pasos pequeños. La fórmula que representa este paso es la siguiente, que también se encuentra en la Figura 2.3 :

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)), \tag{2.4}$$

Donde:

- $\mathbf{x}_t$: Es el dato en el paso de tiempo t, que contiene ruido gaussiano que fue añadido durante el proceso de difusión hacia adelante.

- $\mathbf{x}_{t-1}$: Es el dato en el paso temporal anterior $t-1$, que representa el valor estimado de los datos antes de que se añadiera ruido.

- $\mu_\theta(\mathbf{x}_t, t)$: Es la media de la distribución normal que define la predicción de $\mathbf{x}_{t-1}$ a partir de $\mathbf{x}_t$ y el tiempo t. Este término representa la aproximación del modelo a la información original de los datos, basándose en el estado ruidoso actual y el paso temporal.

- $\Sigma_\theta(\mathbf{x}_t, t)$: Es la varianza de la distribución normal, que describe la incertidumbre en la estimación de $\mathbf{x}_{t-1}$. Este valor depende del estado $\mathbf{x}_t$ y del tiempo t, lo

que implica que la cantidad de incertidumbre en la predicción varía durante el proceso de difusión hacia atrás.

- $\mathcal{N}$: Indica una distribución normal. La media y la varianza de esta distribución están determinadas por los parámetros $\mu_\theta(\mathbf{x}_t, t)$ y $\Sigma_\theta(\mathbf{x}_t, t)$, los cuales son predichos por el modelo de difusión hacia atrás.

El objetivo de este proceso es que, a partir de los valores ruidosos $\mathbf{x}_t$, el modelo de difusión hacia atrás recupere las representaciones anteriores de los datos. La media de la distribución $\mu_\theta(\mathbf{x}_t, t)$ proporciona la estimación más probable de $\mathbf{x}_{t-1}$, mientras que la varianza $\Sigma_\theta(\mathbf{x}_t, t)$ refleja la incertidumbre sobre esta estimación. De este modo, el modelo ajusta continuamente su predicción de los datos originales a medida que avanza en el proceso de difusión hacia atrás.

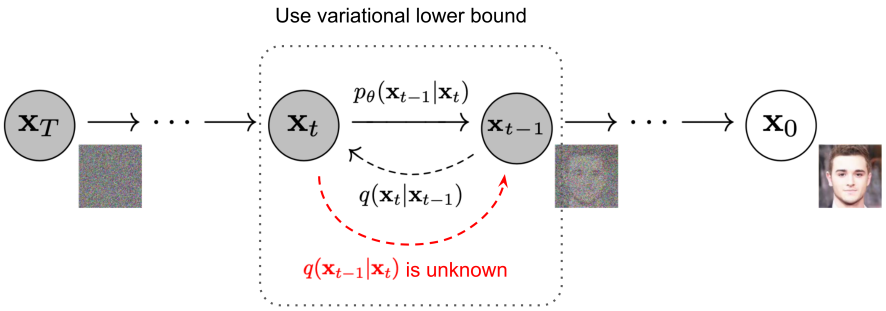

Figura 2.3: Proceso de difusión hacia atrás aplicado a ruido para reconstruir la imagen inicial [13]

Los dos algoritmos representados a continuación [13] describen el proceso en el que se basa este trabajo. El **Algoritmo 1** cubre la fase de difusión hacia adelante y el entrenamiento, mientras que el **Algoritmo 2** ilustra la fase de difusión hacia atrás, utilizada en el muestreo.

Algorithm 1 Entrenamiento

1: **repeat**
2: $x_0 \sim q(x_0)$
3: $t \sim \text{Uniform}(\{1, \ldots, T\})$
4: $\epsilon \sim \mathcal{N}(0, I)$
5: Realizar paso de descenso de gradiente en

$$\nabla_\theta \left\| \epsilon - \epsilon_\theta \left(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t \right) \right\|^2$$

6: **until** converge

Donde:

- x_0: Es una muestra de los datos originales, es decir, el dato limpio sin ruido.
- t: Es un paso temporal seleccionado aleatoriamente para simular la progresión del ruido en el proceso de difusión hacia adelante.

- ϵ: Representa el ruido gaussiano añadido a x_0 en el paso t. Este es el objetivo que el modelo debe aprender a predecir.
- ϵ_θ: Es la predicción del ruido realizada por el modelo con parámetros θ, dado un dato ruidoso y el paso temporal t.
- $\bar{\alpha}_t$: Es un parámetro que controla el grado de mezcla entre los datos originales y el ruido añadido en el paso t.

Algorithm 2 Muestreo

$x_T \sim \mathcal{N}(0, I)$
for $t = T, \ldots, 1$ **do**
    $\mathbf{z} \sim \mathcal{N}(0, I)$ si $t > 1$, de lo contrario $\mathbf{z} = 0$
    $x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) + \sigma_t \mathbf{z}$
end for
return x_0

Donde:

- x_T: Es el punto de partida del proceso de muestreo, que consiste en una muestra de ruido gaussiano puro.
- x_t: Es el dato en el paso de tiempo t, que contiene progresivamente menos ruido a medida que t disminuye.
- x_{t-1}: Es el dato en el paso temporal anterior, generado eliminando ruido de x_t.
- ϵ_θ: Es la predicción del ruido realizada por el modelo entrenado en el paso t.
- σ_t: Representa la escala del ruido residual añadido en cada paso de muestreo.
- $\mathbf{z}$: Es un ruido gaussiano adicional, que se añade en los pasos intermedios para mantener la variabilidad en la generación de datos.
- $\alpha_t, \bar{\alpha}_t$: Son parámetros que controlan la contribución relativa de los datos originales y el ruido en cada paso del proceso.

2.4. Redes neuronales de grafos

Aunque el concepto de redes neuronales de grafos se ha mencionado de forma superficial en la sección anterior, a continuación se presentan los detalles adicionales que complementan esta explicación [14][24].

Los grafos, en su forma más básica, son simplemente un grupo de nodos conectados por aristas, que pueden ser dirigidas o no, y tienen matrices de adyacencia. Por eso, es importante considerar su estructura para comprender su funcionamiento y las conexiones que modelan.

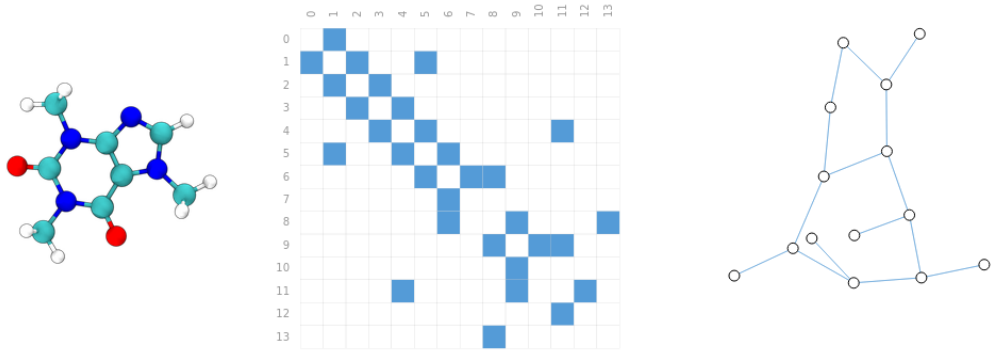

Figura 2.4: Ejemplo de representación de la molécula de cafeína, su matriz de adyacencia y su representación como un grafo [24]

Las **redes neuronales de grafos** son una gran opción porque, además de permitir que se representen las características de las aristas y nodos como vectores, tienen una función interesante: el paso de mensajes. Por lo general, los nodos con características similares suelen estar conectados. Durante el entrenamiento, estas redes aprenden por qué algunos nodos están interconectados y otros no, lo que las convierte en herramientas útiles para generar nuevos datos o clasificar información.

Para construir una red neuronal de grafos (GNN) básica, se utilizan vectores en lugar de escalares para representar los atributos del grafo, lo que permite procesar cada componente mediante un perceptrón multicapa (MLP) y crear una capa GNN. En esta capa, se aplica el MLP a cada vector de nodo y arista, generando nuevas representaciones, o *embeddings*, mientras que el grafo de salida mantiene la misma conectividad y el mismo número de vectores de características que el grafo de entrada. Para realizar predicciones, como en la clasificación binaria, se utiliza un clasificador lineal sobre las representaciones de los nodos [25].

Si solo hay información en las aristas y se necesitan hacer predicciones sobre los nodos, se emplea un proceso de agrupación o *pooling* que reúne y agrega las representaciones de las aristas relacionadas para obtener un nuevo *embedding* del nodo. Además, se incorpora el paso de mensajes, donde los nodos y aristas intercambian información en tres etapas: recopilación de representaciones de nodos vecinos, agregación de estas representaciones y actualización de las representaciones de los nodos con la información recopilada. Además, si añadimos el mecanismo de atención, podremos asignarles pesos a las conexiones y, por lo tanto, poder centrarnos más en unas que en otras.

Durante el desarrollo de este trabajo, utilizamos tres modelos distintos para evaluar el rendimiento de las redes neuronales de grafos: el modelo Higher-order Graph Neural Network (HOGNN) [26], el modelo Residual Gated Graph Convolutional Network (ResGNN) [27] y el modelo de Graph Attention Network (GAT) [16].

Las **HOGNN** se enfocan en la construcción de *embeddings* robustos tanto para los nodos como para las aristas a través de un aprendizaje no supervisado, maximizando la similitud entre nodos conectados y mejorando la representación de las características del grafo.

Las redes de tipo **ResGNN**, por su parte, emplean un mecanismo de atención local en cada nodo, ajustando dinámicamente los pesos de las conexiones durante el aprendizaje. Este mecanismo no solo incrementa la calidad de los *embeddings*, sino que también mejora el rendimiento en tareas como la clasificación de nodos y la predicción

de enlaces.

Por otro lado, el modelo **GAT** combina técnicas de atención y de convolución en grafos, permitiendo a la red asignar pesos distintos a los nodos vecinos en función de su relevancia. Esto optimiza la capacidad de la red para captar patrones complejos en los datos del grafo.

Capítulo 3

Adaptación del modelo y primera prueba

3.1. Preprocesamiento y adaptación

Los modelos de difusión han ganado gran popularidad en el ámbito de la generación de imágenes, lo que ha llevado a la necesidad de realizar varias modificaciones, especialmente en las etapas de generación, adición de ruido y muestreo. A diferencia de otros tipos de datos, las imágenes se representan como matrices, donde cada dimensión corresponde al número de píxeles; es decir, cada píxel tiene su propia celda en la matriz.

La principal dificultad para ajustar estos modelos para nuestro objetivo fue encontrar la forma de ajustar las dimensiones y aplicar correctamente el ruido a los datos, debido a que en vez de añadir ruido a cada elemento en una matriz, solo hay que añadirlo a un único valor. Antes de implementar cualquier modelo, se llevó a cabo un preprocesamiento y visualización de los datos. Se revisaron las dimensiones del conjunto, se analizó la distribución inicial, se realizó un escalado y, finalmente, se llevó a cabo una prueba preliminar de la función de adición de ruido para verificar la implementación de este proceso.

Distribución de los datos

La información inicial sobre el conjunto de datos es la siguiente:

Descripción	**Valor**
Número de grafos en el conjunto	1
Número de nodos en el grafo	23952
Número de aristas en el grafo	68496
Número de características de los nodos	33
Número de características de las aristas	1

Cuadro 3.1: Descripción inicial del conjunto de datos

En las siguientes figuras se presenta la visualización de la energía libre de Gibbs presente en este conjunto durante diferentes etapas de procesamiento. En la Figura 3.1 se muestra la distribución de los datos originales sin procesar, antes de normalizarlos en un rango entre 0 y 1. Esta normalización se realizó para estabilizar el entrenamiento

y mejorar la precisión de los resultados esperados. El resultado de esta normalización puede observarse en la Figura 3.2.

Finalmente, se evaluó la distribución resultante tras añadir ruido a los datos originales para emular la que tendría que aprender un modelo. Este resultado se presenta en la Figura 3.3.

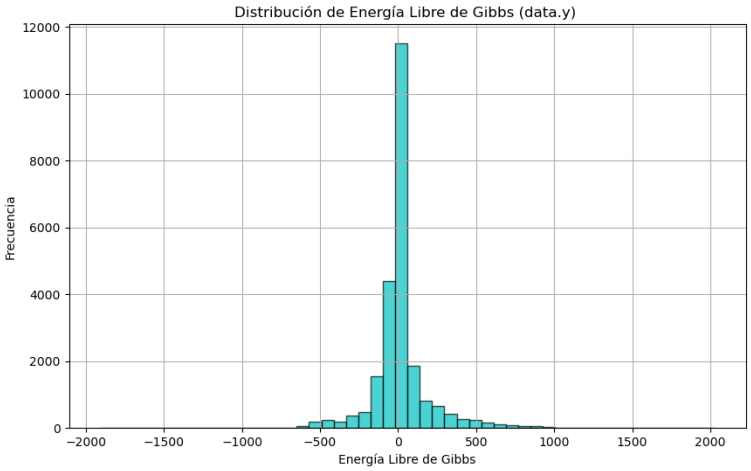

Figura 3.1: Distribución inicial de los datos de la energía libre de Gibbs

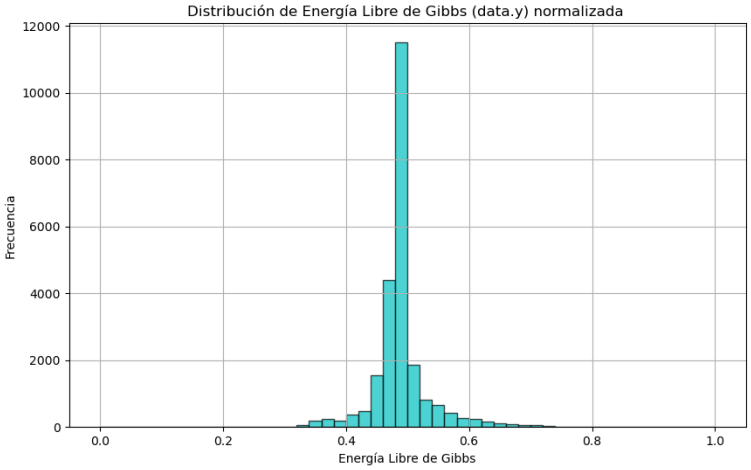

Figura 3.2: Distribución de los valores de la energía libre de Gibbs tras el escalado entre 0 y 1

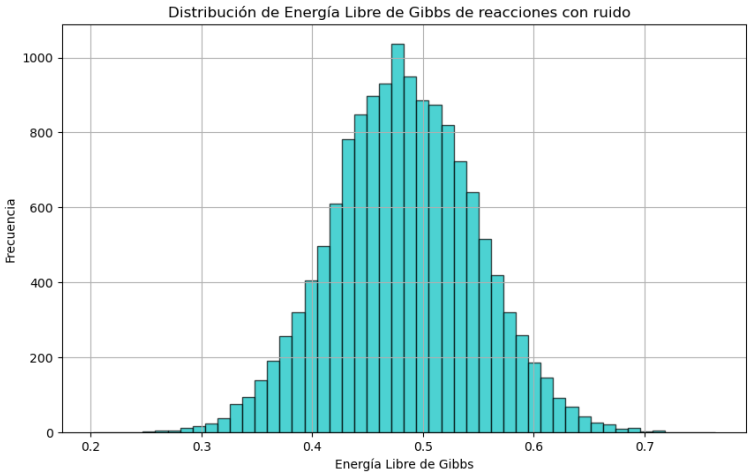

Figura 3.3: Energía libre de Gibbs con muestras de ruido gausiano añadidas

Para evaluar los resultados, utilizaremos dos métricas comunes: el **MAE (Mean Absolute Error)** y el **MSE (Mean Squared Error)**. El MAE mide la diferencia promedio entre los valores reales y los valores predichos, proporcionando una evaluación directa de la precisión. Por su parte, el MSE también mide esta diferencia, pero penaliza más severamente los valores atípicos, siendo útil en situaciones donde estos deben tener un impacto significativo.

3.2. Primera prueba

3.2.1. Objetivos

Después de realizar el primer ajuste en los datos, se llevó a cabo la primera aproximación propuesta en este trabajo: generar los valores de la energía libre utilizando todas las características disponibles. Para ello, se concatenó la energía libre con los datos originales de los nodos descritos en el conjunto de datos, y los modelos se entrenaron con esta información. Se probaron tres modelos diferentes con varios hiperparámetros como las iteraciones, la tasa de aprendizaje o decaimiento de los pesos (los cuales se encuentran explicados en detalle en el Anexo B), añadiendo distintos niveles de ruido en cada iteración para que el modelo pudiera aprender mejor la incertidumbre. Para cada modelo se ha seleccionado la mejor combinación de hiperparámetros según el error de test con la métrica MSE, los resultados se muestran a continuación:

Modelo	Test MSE	Test MAE	Iteraciones	LR	WD
GAT	0.0003	0.0125	100	0.01	0.0005
ResGNN	0.0001	0.0054	100	0.01	0.005
HOGNN	0.0461	0.1492	100	0.01	0.005

Cuadro 3.2: Resultados de entrenamiento en la primera prueba, utilizando como entrada del modelo todos los datos disponibles

3.2.2. Entrenamiento

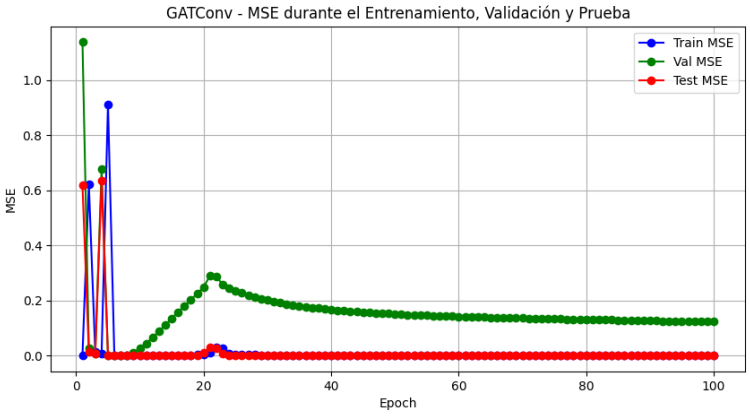

Figura 3.4: Resultados para el modelo GAT en la primera prueba

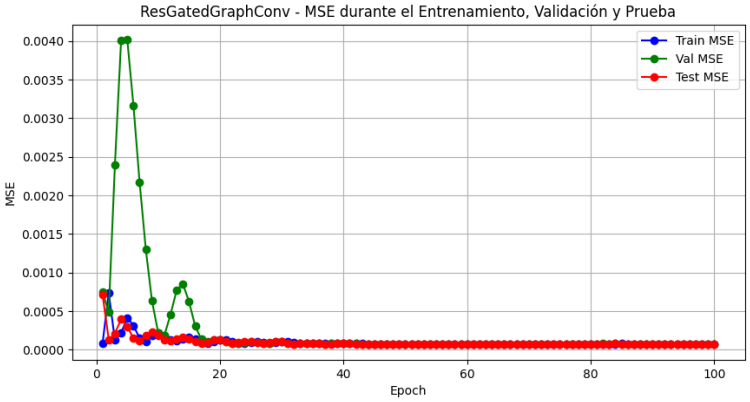

Figura 3.5: Resultados para el modelo ResGNN en la primera prueba

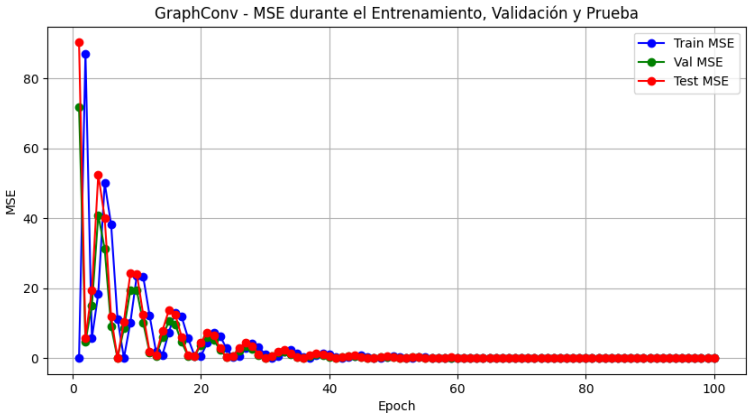

Figura 3.6: Resultados para el modelo HOGNN en la primera prueba

En los resultados cuantitativos presentados por el entrenamiento podemos observar que todos los errores son relativamente bajos. También observamos que los dos últimos modelos tienen los mejores resultados y que en las gráficas los errores mantienen la misma tendencia y descienden. El hecho de que los errores de prueba y validación sean similares a los de entrenamiento confirma que los modelos manejan bien valores nuevos sin incrementar significativamente el error, señal de que no existe un sobreajuste y nuestro modelo tiene la capacidad de generalizar bien o lo que es lo mismo, es capaz de captar bien la relación entre las características y resultado objetivo.

3.2.3. Predicción y resultados

En esta imagen se muestra un muestreo inicial usando solo tres nodos del grafo para visualizar la incertidumbre en los resultados. Aunque cada vez que se repite el muestreo, los valores predichos son bastante parecidos, nunca son exactamente iguales, lo que se debe al carácter no determinista del modelo de difusión utilizado. Esta variabilidad en los resultados refleja cómo el modelo genera estimaciones similares, pero no idénticas, lo cual es útil para entender la incertidumbre en las predicciones a partir de un conjunto de datos limitado.

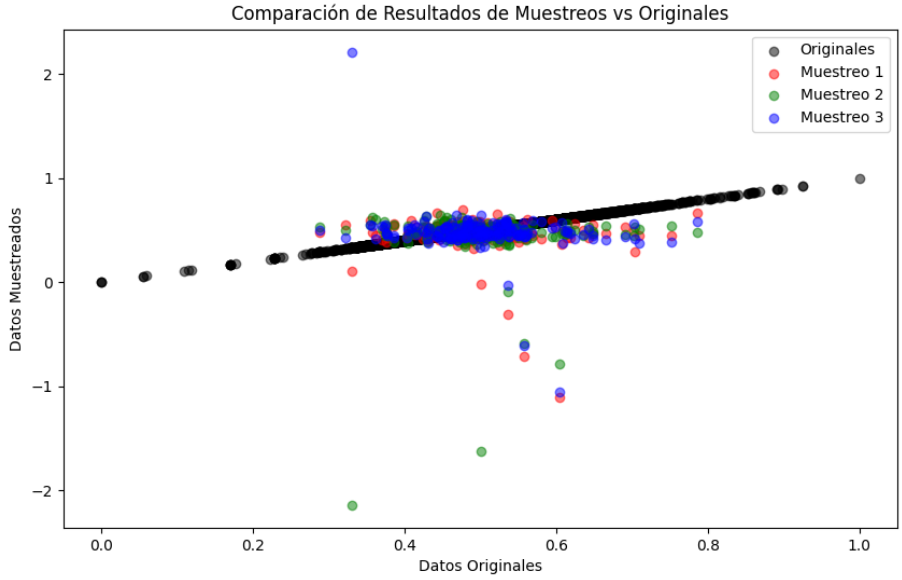

Figura 3.7: Valores de energía obtenidos respecto a los originales tras el muestreo en la primera prueba

La última parte de esta prueba consiste en usar este modelo para ser capaces de llegar desde ruido puro a los valores de energía libre de Gibbs reales. De acuerdo con el Algoritmo 2, utilizamos la fórmula del cuarto paso para predecir los datos iniciales.

Para utilizar el algoritmo, usamos el modelo que mejor resultado haya tenido en el entrenamiento y, posteriormente, calculamos el error:

El valor obtenido para el MAE es de **3.9402**, lo que representa el error promedio absoluto entre las predicciones y los valores reales. Este valor indica que, en promedio, la diferencia entre las predicciones del modelo y los valores reales es de aproximadamente 3.94 unidades. El error obtenido sugiere que el modelo tiene una precisión limitada al predecir los valores esperados y puede estar capturando solo parcialmente la variabilidad del conjunto de datos.

Por otro lado, el valor de MSE es de **60.5986**, que es considerablemente más alto que el MAE debido a que el MSE penaliza con mayor fuerza los errores de mayor magnitud. Este alto valor de MSE refleja la presencia de algunos valores atípicos o predicciones con errores significativos, que amplifican el error general del modelo.

La gran diferencia entre el MAE y el MSE subraya la sensibilidad del modelo ante valores extremos y evidencia que, aunque el modelo puede funcionar de forma razonable en la mayoría de los casos, existen algunas predicciones que están significativamente alejadas de los valores reales.

Este comportamiento se puede observar visualmente en la Figura 3.8, donde se muestran los valores de energía obtenidos por el modelo en comparación con los valores originales. En el gráfico se pueden distinguir algunos puntos que se distancian considerablemente de la zona de concentración de los datos originales, indicando que el modelo tiene dificultades para ajustar correctamente estos casos extremos.

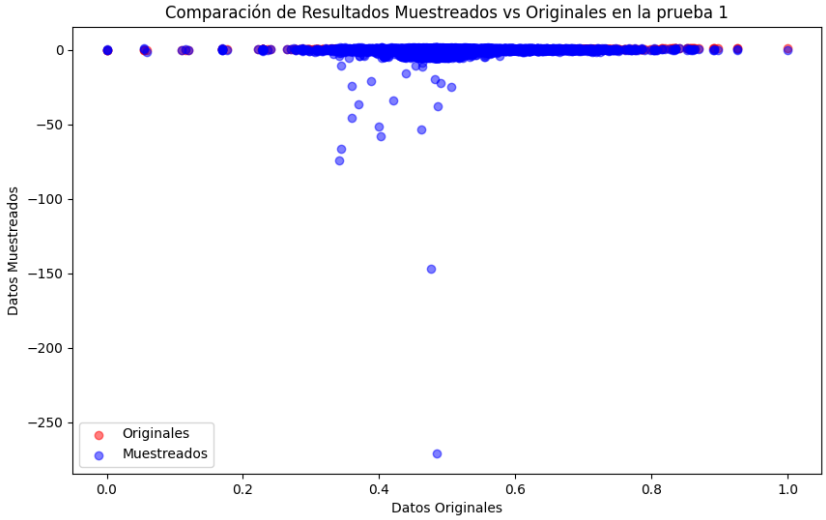

Figura 3.8: Valores de energía obtenidos respecto a los originales tras la predicción final en la primera prueba

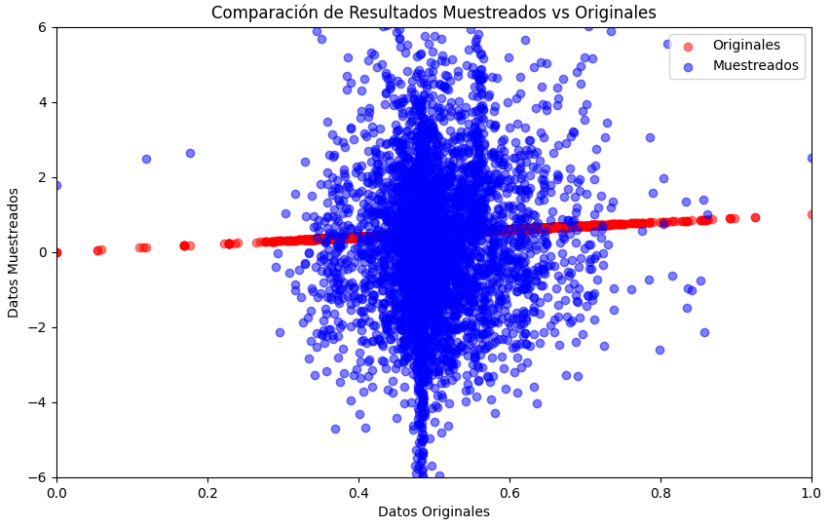

Figura 3.9: Valores de energía obtenidos aumentados entre -6 y 6

Aunque visualmente pueda parecer a priori que los datos muestreados se aproximan a los datos originales, al analizar la gráfica en detalle, se observa una notable desviación en la distribución de los puntos muestreados hacia los extremos. Esto indica que, aunque el modelo logra captar la tendencia central de los datos originales, presenta un grado considerable de dispersión, especialmente en los valores alejados del centro. Esta dispersión podría afectar la precisión del modelo en las predicciones para datos extremos e indica una posible limitación en la capacidad del modelo para ajustarse a la estructura de los datos originales.

3.2.4. Conclusión

En conclusión, los valores de MAE y MSE obtenidos en esta primera prueba indican que el modelo presenta poca precisión aceptable en general y es sensible a los valores atípicos, como lo sugiere el alto valor de MSE. Este análisis resalta la necesidad de

considerar técnicas adicionales, como regularización o ajustes en los parámetros del modelo, para mejorar su rendimiento en casos con errores elevados y minimizar su sensibilidad ante datos extremos.

Capítulo 4

Segunda prueba: predicción usando sólo la energía de formación

4.1. Objetivos

En esta segunda prueba, el objetivo es evaluar si es posible predecir la energía libre de Gibbs utilizando únicamente la energía de formación, la cual se presenta como una característica individual en el conjunto de datos.

Con esta estrategia, buscamos simplificar el modelo y reducir la dependencia de múltiples características, evaluando la efectividad de una predicción basada en una única variable. Para llevar a cabo este análisis, seguiremos el mismo procedimiento que en la prueba anterior, pero adaptando tanto los datos de entrada como los parámetros del modelo a esta nueva configuración.

A diferencia del modelo inicial, en lugar de entrenarlo con todas las características disponibles, hemos seleccionado exclusivamente la energía de formación. Posteriormente, se entrenó el modelo empleando el mismo procedimiento usado en el capítulo anterior. Los resultados de esta prueba se presentan a continuación:

4.2. Entrenamiento

Basándonos únicamente en los valores finales de Test MAE en la Tabla 4.1, los modelos muestran resultados similares en términos de error absoluto. Sin embargo, el número de iteraciones varía significativamente entre ellos, sugiriendo diferentes velocidades de convergencia.

Modelo	**Test MSE**	**Test MAE**	**Iteraciones**	**LR**	**WD**
GAT	0.0002	0.0074	70	0.01	0.0005
ResGNN	0.0001	0.0067	100	0.01	0.005
HOGNN	0.0001	0.0071	50	0.01	0.005

Cuadro 4.1: Resultados de entrenamiento en la segunda prueba

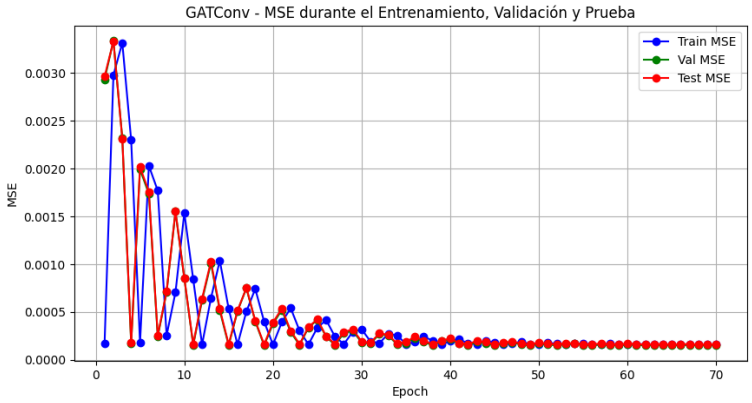

Figura 4.1: Resultados para el modelo GAT en la segunda prueba

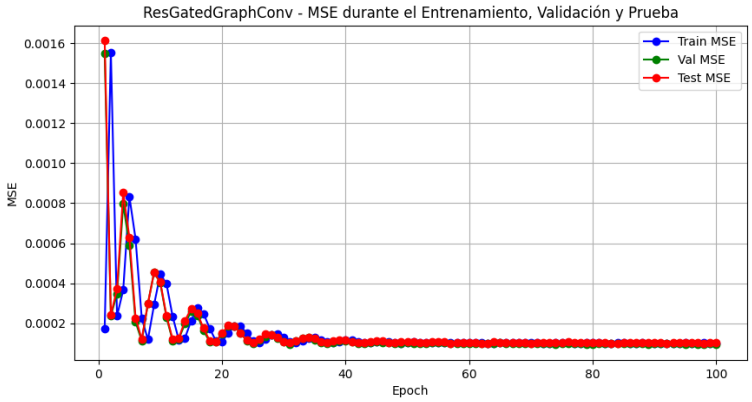

Figura 4.2: Resultados para el modelo ResGNN en la tercera prueba

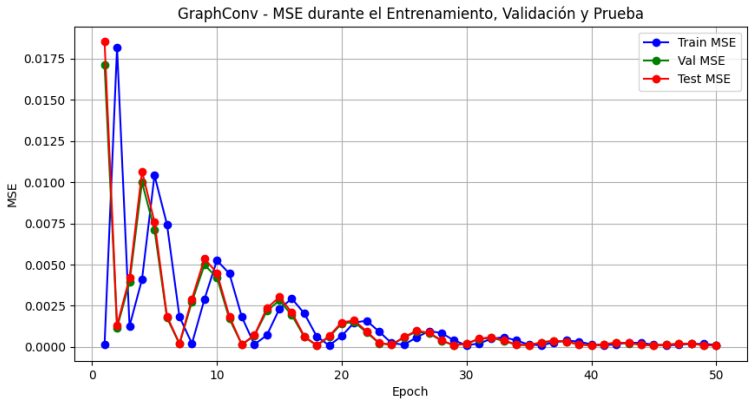

Figura 4.3: Resultados para el modelo HOGNN en la segunda prueba

Después de observar las gráficas podemos ver claramente que efectivamente aunque la forma en la que convergen es parecida, la velocidad con la que lo hacen es distinta, siendo HOGNN el más rápido, seguido de GAT y por último ResGNN. Comparando las gráficas de los dos procesos de entrenamiento que se han hecho hasta el momento, observamos que en esta prueba el error del conjunto de validación es más estable. Esto podría indicar que la forma en la que el modelo se ajusta a los datos es más lineal

y progresiva. Sin embargo, para poder hacer una comparación final se analizarán los resultados de las predicciones.

4.3. Predicción y resultados

Al analizar los errores, observamos que son bastante menores que los obtenidos en la prueba anterior, en la que se emplearon todas las características.

Esto sugiere que la energía de formación, por sí sola, aporta una cantidad significativa de información relevante para la predicción de la energía libre de Gibbs.

En particular, el valor del MAE es de **0.736970**, lo cual representa el error absoluto promedio entre las predicciones del modelo y los valores reales. Este valor indica que, en promedio, el modelo presenta una desviación moderada respecto a los valores esperados, lo cual puede considerarse aceptable dada la simplificación de la entrada.

Por otro lado, el valor del MSE obtenido es **0.826949**. Si bien el MSE penaliza los errores más grandes debido a su término cuadrático, su valor bajo sugiere que no existen grandes discrepancias en las predicciones, y que el modelo mantiene un rendimiento estable incluso con esta simplificación de las características. Esto indica que la energía de formación es una característica robusta y efectiva para la predicción en este contexto, aunque sigue siendo menos precisa que el modelo anterior con todas las variables.

Comparando las gráficas de los dos procesos de muestreo realizados hasta el momento, observamos que en esta prueba los datos muestreados se distribuyen de manera más estable a lo largo de todo el rango de valores originales. A diferencia del primer muestreo, en el cual se veía una tendencia a concentrarse en torno al centro de los datos originales, aquí los valores predichos se mantienen en un rango más amplio, sin acumularse excesivamente en ninguna región. Esto podría indicar que el modelo, en esta prueba, es capaz de capturar la variabilidad en los extremos de forma más uniforme.

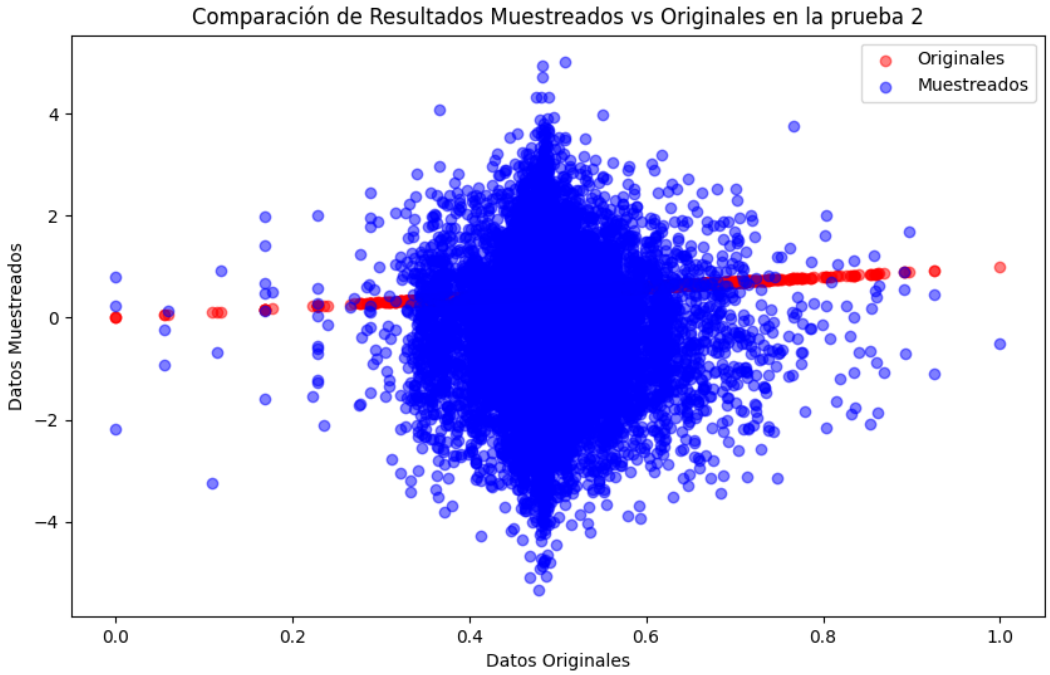

Figura 4.4: Valores de energía obtenidos respecto a los originales utilizando únicamente la energía de formación de Gibbs

4.4. Conclusión

En comparación con el modelo que utiliza todas las características, esta versión simplificada, que solo usa la energía de formación, muestra una precisión mucho mejor, como lo reflejan los valores de MAE y MSE.

En resumen, emplear la energía de formación como única entrada para predecir la energía libre de Gibbs ha resultado en un modelo con mejor rendimiento que el modelo más complejo. Esto sugiere que la energía de formación es una variable clave en la determinación de la energía libre de Gibbs, logrando predicciones con un margen de error muy bajo.

Sin embargo, al añadir otras características, como la fórmula de cada compuesto, el modelo no ignora estos datos (aunque no sean realmente necesarios), sino que sus predicciones dependen tanto de la fórmula como de la energía de formación. Esto indica que el modelo no está logrando la predicción más eficiente posible. Un enfoque que utilice solo las variables clave podría simplificar el modelo sin sacrificar precisión, lo que sería útil en aplicaciones que requieran eficiencia computacional o en escenarios con datos limitados.

Capítulo 5

Tercera prueba: predicción en base a las fórmulas químicas y sus cargas

5.1. Objetivos

El objetivo de esta tercera prueba es evaluar si las fórmulas químicas de los compuestos, junto con las cargas asociadas, pueden proporcionar información relevante para predecir la energía libre de Gibbs. En caso de que las fórmulas y las cargas tengan un impacto significativo en la predicción, esto podría indicar una correlación entre la energía libre de Gibbs y la estructura química básica de los compuestos, sugiriendo que el modelo ha capturado patrones inherentes en la relación entre estructura y energía.

Esta prueba es particularmente interesante porque, a diferencia de pruebas anteriores, en las que incluimos otras características que aportan contexto termodinámico o energético, aquí restringimos el modelo a información puramente estructural. En principio, esta información no debería ser suficiente para predecir con precisión la energía libre de Gibbs de una reacción; sin embargo, evaluar el rendimiento del modelo con este conjunto de datos limitado podría arrojar luz sobre el grado de sensibilidad que tiene la energía libre de Gibbs frente a características estructurales básicas.

5.2. Entrenamiento

Al igual que en la prueba anterior, procedemos a comparar los gráficos correspondientes.

Modelo	**Test MSE**	**Test MAE**	**Iteraciones**	**LR**	**WD**
GAT	0.0011	0.0254	50	0.01	0.0005
ResGNN	0.0001	0.0062	70	0.01	0.005
HOGNN	0.0002	0.0099	100	0.01	0.005

Cuadro 5.1: Resultados de entrenamiento en la tercera prueba

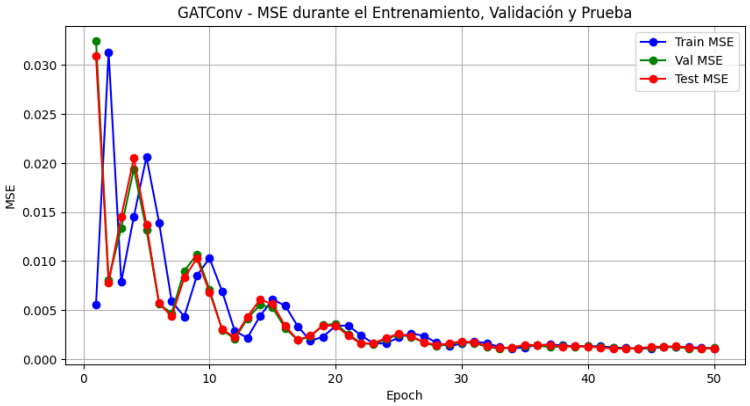

Figura 5.1: Resultados para el modelo GAT durante el entrenamiento de la tercera prueba

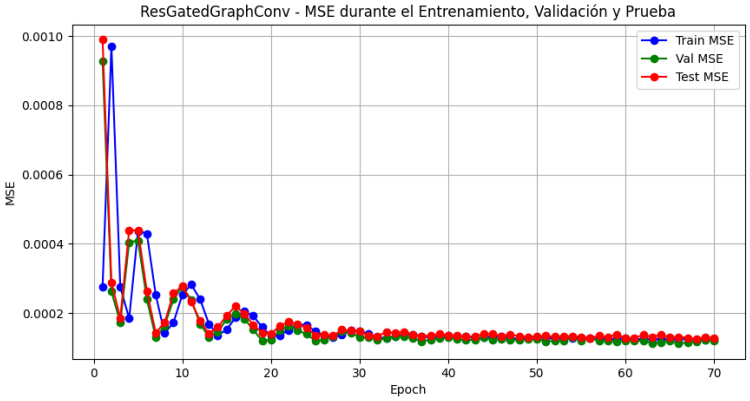

Figura 5.2: Resultados para el modelo ResGNN durante el entrenamiento de la tercera prueba

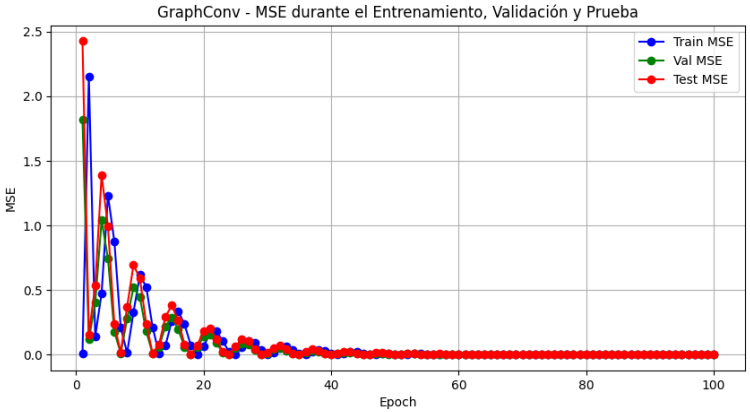

Figura 5.3: Resultados para el modelo HOGNN durante el entrenamiento de la tercera prueba

Inicialmente los valores de los errores para cada modelo tienen una diferencia ínfima con los obtenidos durante la segunda prueba. En general, todos los modelos entrenados tienen una buena capacidad de generalizar, lo cual se ve en cómo los errores

de validación y evaluación progresan con la misma dinámica y se mantienen en el mismo nivel que los errores de entrenamiento.

5.3. Predicción y resultados

El valor del MAE es de **0.8868**, lo cual indica que el modelo tiene una desviación promedio considerable al utilizar solo las fórmulas químicas y las cargas, lo que era esperable. Aun así, el valor de MAE relativamente bajo sugiere que el modelo ha logrado identificar algunos patrones generales, posiblemente relacionados con la estructura del grafo y su influencia en la energía.

Por otro lado, el valor de MSE es de **11.2593**, acercándose al obtenido en la prueba anterior. Este valor refleja que existen algunas predicciones con errores de mayor magnitud, lo cual puede interpretarse como un efecto de la limitación de los datos de entrada.

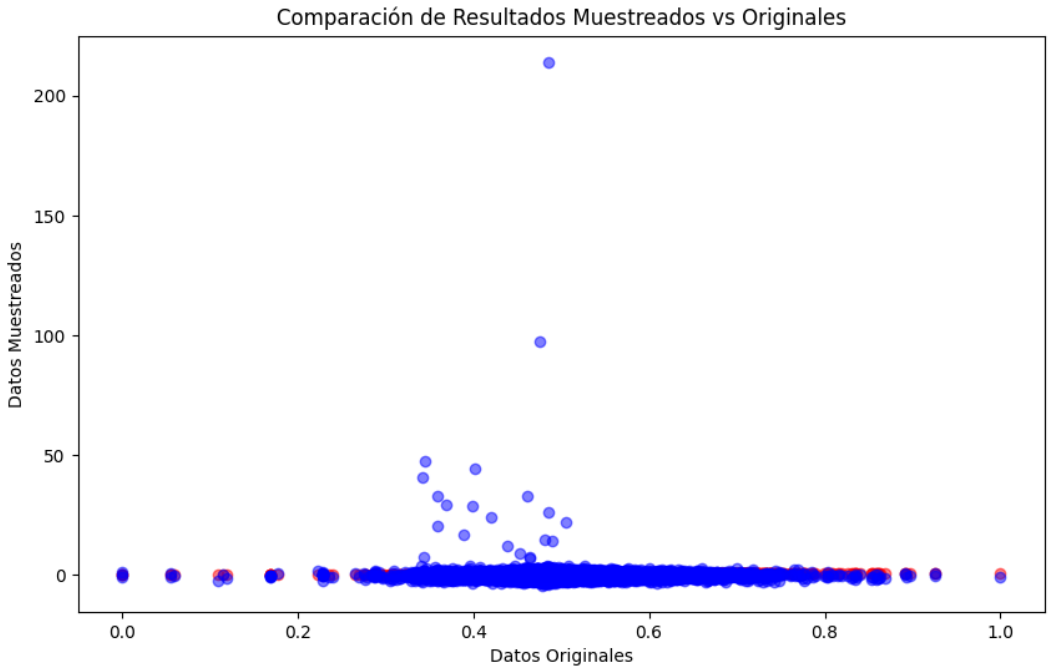

Figura 5.4: Valores de energía obtenidos respecto a los originales en la tercera prueba

Al aumentar la escala de la gráfica, podemos observar que el ajuste al patrón de los datos originales es visualmente el mejor hasta el momento. La distribución de los datos predichos es menos dispersa y muestra una mayor coherencia en comparación con pruebas anteriores, asemejándose bastante al resultado obtenido en la segunda prueba. Esto sugiere que el modelo está logrando captar la estructura de los datos originales con mayor precisión.

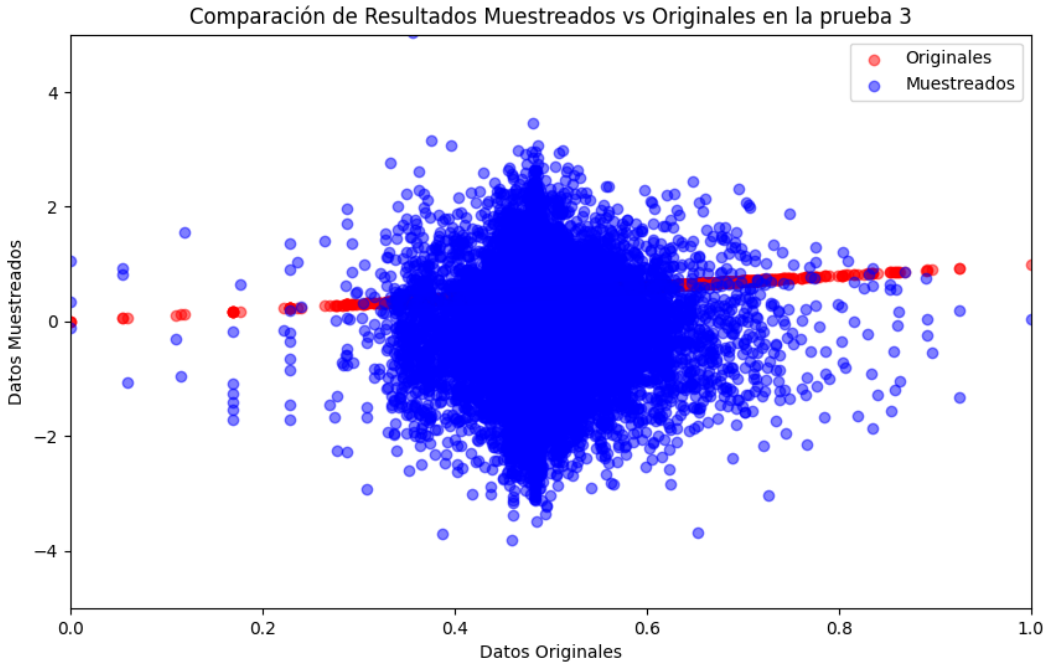

Figura 5.5: Valores de energía obtenidos respecto a los originales en la tercera prueba entre -4 y 4

Estos resultados son consistentes con la idea de que la energía libre de Gibbs está influenciada por la composición química de un compuesto, aunque esta información por sí sola no es suficiente para una predicción precisa. Sin embargo, el rendimiento del modelo con estos datos limitados sugiere que el modelo ha logrado captar y replicar en cierta medida esta relación.

Esta observación tiene implicaciones interesantes, ya que sugiere que, en ausencia de información energética detallada, la estructura química puede proporcionar una primera aproximación de la energía libre de Gibbs. Esta capacidad predictiva limitada podría ser útil en contextos donde no se dispone de todos los datos necesarios, permitiendo hacer una estimación preliminar basada únicamente en la fórmula química.

5.4. Conclusión

En resumen, esta tercera prueba ha demostrado que, aunque las fórmulas químicas y las cargas proporcionan una base débil para la predicción de la energía libre de Gibbs, el modelo es capaz de extraer cierta información relevante de estos datos. Los valores de MAE y MSE evidencian que el modelo ha captado algunos patrones generales, lo que subraya el potencial de la estructura química como predictor secundario de la energía libre en ausencia de información energética específica.

Capítulo 6

Resultados

En este capítulo se presentarán y analizarán los resultados obtenidos a través de las tres pruebas realizadas, destacando las diferencias y similitudes entre ellas. Para facilitar la comparación, los resultados se resumen en la siguiente tabla:

Prueba	**MAE**	**MSE**	**Observaciones**
Primera prueba	3.9402	60.5986	Tiene tendencia a concentrase en el centro de los datos originales y no cubre adecuadamente los datos de los extremos.
Segunda prueba	0.7369	0.8269	Mejor precisión y distribución de los datos muestreados respecto a la primera, con unos valores moderados y aceptables.
Tercera prueba	0.8868	11.2593	Precisión similar a la segunda prueba y una distribución que se asemeja a la segunda prueba con varios valores atípicos.

Cuadro 6.1: Síntesis de resultados obtenidos en las tres pruebas realizadas con valores de errores y observaciones.

A partir de estos datos, se puede concluir que la energía de formación de Gibbs es la característica que más influye en los valores muestreados. Esto se debe en parte a que no es necesario escalar la gráfica, consecuencia de que no presenta valores atípicos extremos, sino que se encuentra dentro de un rango aceptable. Este comportamiento sugiere que las mediciones están fuertemente relacionadas, y en un contexto de datos limitados, esta propiedad parece ser la de mayor relevancia.

Por otro lado, es importante destacar que las fórmulas químicas y las cargas juegan un papel clave en el ajuste de la distribución. Aunque, en ocasiones, se observa un valor algo más desviado de la media, la tendencia general se mantiene ajustada. En este sentido, se puede deducir que la estructura de la red tiene influencia en el comportamiento de los datos a la hora de orientar la distribución.

Una conclusión importante de este estudio es que los modelos que se han entrenado y evaluado no seleccionan las características que pueden ser más importantes o estar más relacionadas, sino que tratan de tenerlas todas en cuenta de manera simultánea. Esto puede llevar a que el modelo se vea sobrecargado con información irrelevante o ruidosa, lo que puede afectar negativamente a su capacidad de generalización.

Adicionalmente, se realizó un experimento en el que se comprometió un porcentaje de los datos en la segunda y tercera prueba, con el fin de observar cómo se comportan los modelos y analizar la influencia de dicha alteración en sus predicciones.

La alteración consistió en reemplazar un número determinado de valores, seleccionado de manera aleatoria según el porcentaje especificado, por ceros, dado que este valor es frecuente tanto en las fórmulas químicas como en los datos de energía de formación. Por lo tanto, esta estrategia no generaría la predicción de datos atípicos, sino que permitiría evaluar cómo el modelo maneja situaciones realistas con datos incompletos o con alteraciones. Se probó a corromper el 10 % y el 20 % de los valores, primero en el caso de la energía de formación de Gibbs y luego en las fórmulas químicas y las cargas, ya que son los modelos que mejor se han comportado.

Los resultados muestran un moderado aumento en los errores de predicción, y se observa que las nubes de puntos conservan una forma muy similar a la obtenida con los datos originales. Esto sugiere que los modelos elaborados consiguen mantener una robustez considerable frente a la perturbación de datos, preservando la coherencia en la distribución de los valores predichos. Esta capacidad de resiliencia es indicativa de que los modelos podrían ser aplicables en contextos con datos ruidosos o incompletos, manteniendo un rendimiento aceptable y precisión adecuada en sus predicciones.Estos resultados se presentan en la Tabla 6.2 y las figuras 6.1, 6.2, 6.3 y 6.4.

Característica a corromper	**Porcentaje**	**MAE**	**MSE**
Energía de formación de Gibbs	10	0.9823	2.4671
Energía de formación de Gibbs	20	1.0874	2.5466
Fórmula química y cargas	10	1.1985	12.7786
Fórmula química y cargas	20	1.5025	13.8832

Cuadro 6.2: Errores observados al corromper un determinado porcentaje de los valores de las características correspondientes a la segunda y tercera prueba.

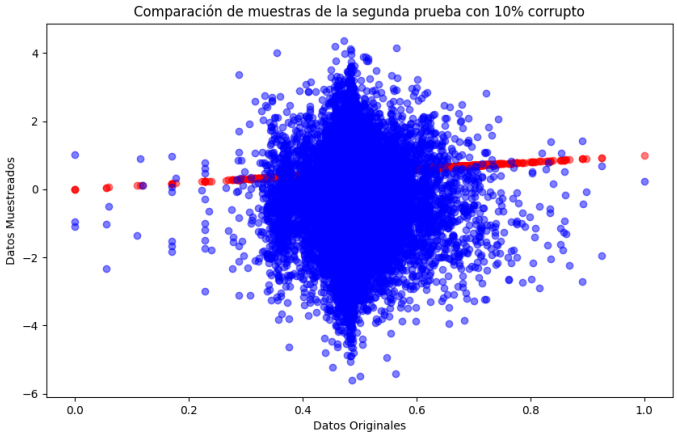

Figura 6.1: Resultados de predicción corrompiendo el 10 % de los datos de energía de formación de Gibbs

Como podemos observar, las nubes de puntos para las predicciones hechas con valores corruptos son prácticamente idénticas entre sí y se asemejan mucho a la nube de puntos sin corrupción. Esto indica que el modelo mantiene una consistencia notable

en la distribución de sus predicciones, incluso cuando se introducen perturbaciones en los datos.

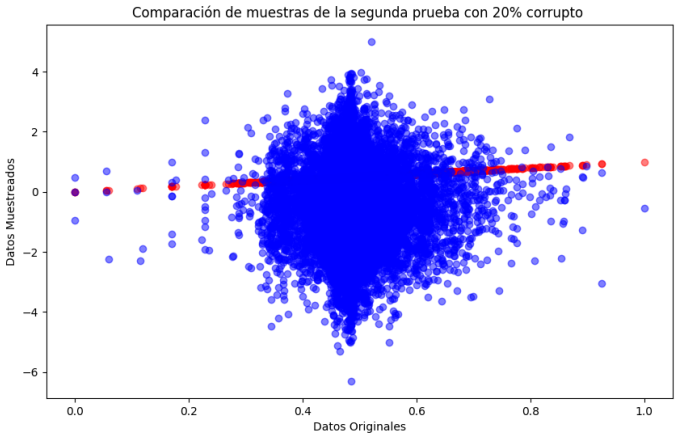

Figura 6.2: Resultados de predicción corrompiendo el 20 % de los datos de energía de formación de Gibbs

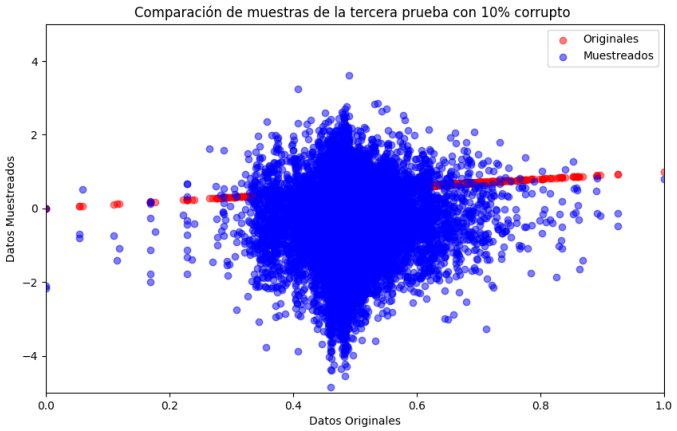

Figura 6.3: Resultados de predicción corrompiendo el 10 % de los datos de fórmulas y cargas entre -4 y 4.

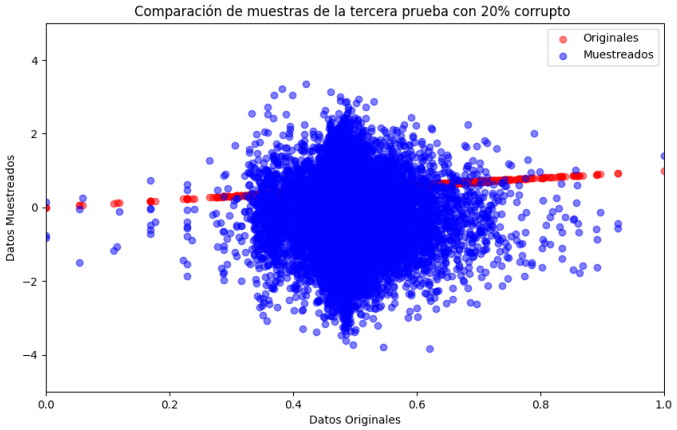

Figura 6.4: Resultados de predicción corrompiendo el 20 % de los datos de fórmulas y cargas entre -4 y 4.

Capítulo 7

Conclusiones

Este trabajo tuvo como objetivo desarrollar un modelo computacional basado en modelos de difusión y redes neuronales de grafos (GNNs) para predecir la energía libre de Gibbs en redes bioquímicas, superando las limitaciones actuales en la estimación de esta magnitud clave para el metabolismo celular.

Mediante un enfoque nuevo que no depende de datos moleculares específicos sino que se centra en su estructura, logramos generar distribuciones probabilísticas de la energía libre de Gibbs, permitiendo obtener un cierto grado de incertidumbre. Este avance es relevante para la investigación bioquímica, donde los datos experimentales son frecuentemente limitados.

Un reto importante fue adaptar técnicas de aprendizaje al campo bioquímico, más concretamente comprender el flujo de trabajo y adaptarlo a nuestras entradas. Después de evaluar diferentes modelos con diferentes hiperparámetros, conseguimos modelar las predicciones usando distintas características y profundizar en su relación con la energía de Gibbs.

Los resultados demuestran que los modelos son capaces de hacer estimaciones precisas de la energía libre de Gibbs, incluso con pocos datos, y que el enfoque propuesto es prometedor para la predicción de esta magnitud en redes bioquímicas complejas. Se observó que, al contrario de lo que se esperaba, únicamente con las fórmulas químicas y las cargas ya es posible obtener una precisión moderada, lo que subraya lo útil que resulta tener los datos en forma de grafo y poder utilizar la estructura para aprender relaciones. Este avance tiene aplicaciones potenciales en bioquímica, farmacología y síntesis química, mejorando la capacidad de identificar patrones y predecir comportamientos en sistemas dinámicos complejos.

7.1. Trabajos futuros

A partir de este TFG, se podrían explorar arquitecturas de GNN más complejas o diferentes variantes para evaluar su rendimiento en el modelo.En este sentido, se puede plantear la profundización de los modelos aplicados a las fórmulas y las energías, por ejemplo, mediante la adición de más capas en el modelo de redes neuronales. Esto podría ayudar a mejorar el ajuste de las características menos consistentes y reducir el impacto de los valores atípicos, logrando una mayor precisión en la predicción general. Además, la incorporación de capas adicionales podría permitir que el modelo capture de manera más eficiente la compleja interacción entre las variables, especialmente cuando se enfrenta a datos limitados o altamente variables. También se podría investigar el

impacto de trabajar con cantidades limitadas de datos, lo que permitiría evaluar la robustez y capacidad de generalización del modelo en escenarios de datos escasos, un desafío común en la investigación bioquímica. Los métodos desarrollados en este trabajo facilitan además la investigación en diversas áreas de interés como la predicción de otras propiedades bioquímicas, como la fórmula química a partir de valores de energía, lo que proporcionaría una visión más completa de los procesos bioquímicos y herramientas útiles para el diseño de compuestos.

Capítulo 8

Bibliografía

[1] Yi Fang. Protein folding: The gibbs free energy, 2012.

[2] Juan S. Jiménez and María J. Benítez. Gibbs free energy and enthalpy–entropy compensation in protein–ligand interactions. *Biophysica*, 4(2):298–309, 2024.

[3] Molecular recognition and binding free energy calculations in drug development. *Current Pharmaceutical Biotechnology*, 9(2), 2008.

[4] Deliang Zhou, Geoff G.Z. Zhang, Devalina Law, David J.W. Grant, and Eric A. Schmitt. Physical stability of amorphous pharmaceuticals: Importance of configurational thermodynamic quantities and molecular mobility. *Journal of Pharmaceutical Sciences*, 91(8):1863–1872, 2002.

[5] Sharad B. Murdande, Michael J. Pikal, Ravi M. Shanker, and Robin H. Bogner. Solubility advantage of amorphous pharmaceuticals: I. a thermodynamic analysis. *Journal of Pharmaceutical Sciences*, 99(3):1254–1264, 2010.

[6] Wenchao Fan, Chuyun Ding, Dan Huang, Weiyan Zheng, and Ziwei Dai. Unraveling principles of thermodynamics for genome-scale metabolic networks using graph neural networks, 01 2024.

[7] Laurent Valentin Jospin, Hamid Laga, Farid Boussaid, Wray Buntine, and Mohammed Bennamoun. Hands-on bayesian neural networks—a tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2):29–48, May 2022.

[8] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective, 2020.

[9] Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion, 2024.

[10] Igor V. Tetko, Pavel Karpov, Robert Van Deursen, and Gaston Godin. State-of-the-art augmented nlp transformer models for direct and single-step retrosynthesis. *Nature Communications*, 11:5575, November 2020.

[11] Wenchao Fan, Chuyun Ding, Dan Huang, Weiyan Zheng, and Ziwei Dai. Unraveling principles of thermodynamics for genome-scale metabolic networks using graph neural networks. *bioRxiv*, 2024.

[12] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015.

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[14] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

[15] Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

[16] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks, 2018.

[17] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2017.

[18] MetaCyc. Metacyc: A database of metabolic pathways and enzymes, 2023.

[19] MetanetX. Metanetx: A database of metabolic reactions and pathways, 2023.

[20] A minimal reaction network including a reactant r, catalyst c, intermediate i, and product p. *ResearchGate*.

[21] LibreTexts. Gibbs (free) energy, n.d.

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[23] Tiankai Hang and Shuyang Gu. Improved noise schedule for diffusion training, 2024.

[24] Benjamin Sanchez-Lengeling, Emily Reif, Adam Pearce, and Alex Wiltschko. A gentle introduction to graph neural networks. *Distill*, 6(9):e33, September 2021.

[25] Jiaxuan You, Rex Ying, and Jure Leskovec. Design space for graph neural networks, 2021.

[26] Christopher Morris, Martin Ritzert, Matthias Fey, William L. Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and leman go neural: Higher-order graph neural networks, 2021.

[27] Xavier Bresson and Thomas Laurent. Residual gated graph convnets, 2018.

[28] Matthias Fey and Jan Eric Lenssen. *Fast Graph Representation Learning with PyTorch Geometric*, May 2019. MIT License.

Lista de Figuras

Lista de Tablas

Anexo A

Planificación

El desarrollo de este trabajo ha requerido aproximadamente 330 horas, distribuidas a lo largo de casi cuatro meses. Las distintas fases del proyecto, junto con el tiempo invertido en cada una, se muestran en el diagrama de Gantt de la Figura A.1. Estas fases corresponden a los objetivos planteados en el Capítulo 1, además de la documentación y redacción de este informe. La organización del trabajo se ha estructurado en dos bloques principales: una fase inicial enfocada en el estudio y comprensión de modelos de difusión y redes neuronales gráficas (GNNs), así como en la integración y pruebas iniciales del modelo; y dos fases posteriores centradas en la evaluación y aplicación del modelo mediante diferentes métodos, finalizando con la documentación y redacción de la memoria.

	Agosto	Septiembre	Octubre	Noviembre	Horas totales
Estudio y comprensión de los modelos de difusión, GNN's y conjunto de datos inicial					53
Adaptación e integración del modelo y desarollo de la primera prueba					72
Evaluación y aplicación de la estructura con la energía de formación					59
Evaluación y aplicación de la estructura utilizando la fórmula química y las cargas					64
Documentación del código y redacción de la memoria					95

Figura A.1: Diagrama de Gantt

Anexo B

Detalles técnicos

B.1. Hiperparámetros del Modelo

Los hiperparámetros son parámetros que controlan el comportamiento y la capacidad de aprendizaje del modelo. Su ajuste es importante para optimizar el rendimiento del modelo sin caer en problemas de sobreajuste o subajuste. A continuación, se listan y explican los principales hiperparámetros utilizados en este proyecto:

- **Tamaño del *batch* (batch size)**: Se usó todo el conjunto de datos en vez de usar *batches*, esto se justifica debido al tamaño manejable del dataset, que permite cargarlo completamente en memoria, proporcionando estabilidad en la convergencia y reduciendo la variabilidad en las gradientes. De este modo, cada actualización de los parámetros refleja de manera más precisa la dirección óptima para minimizar la función de pérdida, lo que favorece una convergencia más estable y consistente.

- **Tasa de aprendizaje (learning rate)**: Se utilizaron valores como 0.5, 0.1 y 0.01. La tasa de aprendizaje determina el tamaño de los pasos que da el modelo al ajustar sus parámetros. Un valor demasiado alto puede hacer que el modelo no converja, mientras que un valor muy bajo puede ralentizar el entrenamiento.

- **Número de capas en la GNN**: Se empleó una sola capa para capturar las relaciones esenciales sin introducir una complejidad innecesaria. Dado el tamaño y la naturaleza del conjunto de datos, una capa es suficiente para lograr un buen rendimiento general, evitando el sobreajuste y reduciendo el costo computacional.

- **Decaimiento de Pesos (Weight Decay)**: Para mejorar la generalización del modelo y reducir el sobreajuste, se aplicó el decaimiento de pesos como método de regularización en el proceso de optimización. Este parámetro introduce una penalización en los pesos del modelo, empujándolos a mantenerse en valores pequeños durante el entrenamiento. Al añadir un término de regularización L2 en la función de pérdida, ayuda a evitar que los pesos crezcan demasiado, lo que podría llevar a un modelo que memorice los datos de entrenamiento en lugar de aprender patrones generales. En este proyecto, esta técnica resultó útil para mejorar la robustez del modelo sin añadir complejidad significativa.

- **Número de iteraciones (epochs)**: Se probaron 20, 50, 70 y 100 iteraciones. Las iteraciones representan cuántas veces el modelo ha pasado por el conjunto de datos completo. Un número adecuado permite que el modelo aprenda sin llegar a sobreajustarse.

B.2. Implementación en PyTorch Geometric

Para la implementación de la red, se utilizó **PyTorch Geometric** [28], una extensión de PyTorch diseñada específicamente para trabajar con datos en forma de gráfos. Esta biblioteca es especialmente útil para el tratamiento de datos donde las relaciones entre elementos (nodos y aristas) son relevantes.

- **Definición del grafo**: PyTorch Geometric facilita la definición de grafos a partir de datos, lo cual es esencial en este proyecto para representar las relaciones entre las muestras de datos.
- **Capas de la GNN**: Se emplearon capas de tipo `GraphConv`, `ResGatedGraphConv` y `GATConv`, que son algunos de las mcuhas variantes que están optimizadas para captar relaciones en grafos. Estas capas permiten al modelo aprender representaciones a partir de la estructura del grafo, lo cual es fundamental en modelos de difusión.
- **Función de pérdida y optimización**: Para ajustar los parámetros del modelo, se utilizaron dos funciones de pérdida estándar: la pérdida cuadrática media (*Mean Squared Error*, `MSELoss`) y el error absoluto medio (*Mean Absolute Error*, `L1Loss`), junto con el optimizador `Adam`, todos ellos disponibles en PyTorch. Ambas funciones de pérdida fueron elegidas para proporcionar una evaluación equilibrada del rendimiento del modelo. Mientras que `MSELoss` pone más énfasis en las desviaciones grandes, ayudando a ajustar el modelo a valores extremos, `L1Loss` es menos sensible a esos errores grandes, favoreciendo una mayor consistencia en las predicciones generales. Por otro lado, el optimizador `Adam` se encargó de ajustar los pesos del modelo en función de la pérdida. Se eligió `Adam` porque, tras probar otras variantes, no se observaron mejoras significativas, por lo que se concluyó que no aportaban ventajas adicionales.

B.3. Configuración del Entorno de Ejecución

Para asegurar la reproducibilidad de los resultados, se utilizó el siguiente entorno de desarrollo:

- **Librerías principales**:
  - `torch`: Framework principal para la creación y entrenamiento de modelos de aprendizaje profundo.
  - `torch_geometric` : Extensión de PyTorch especializada en el tratamiento de grafos.

- **Hardware**: El modelo fue entrenado utilizando la CPU en el entorno de Google Colab, lo cual permitió acelerar el proceso de entrenamiento y reducir significativamente el tiempo de cómputo.