UNIVERSIDAD AUTÓNOMA DEL ESTADO DE HIDALGO

INSTITUTO DE CIENCIAS AGROPECUARIAS

Instituto de Biotecnología
UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

LICENCIATURA EN INGENIERÍA EN BIOTECNOLOGÍA

TESIS DE LICENCIATURA

"PREDICCIÓN DE ACTIVIDAD HIDROLASA O TRANSFERASA EN AMILASAS MEDIANTE MODELOS DE LENGUAJE DE PROTEÍNAS"

Para obtener el grado de

Licenciado(a) en Ingeniería en Biotecnología

PRESENTA

Odeth Giovanna Valencia Garcia

Director (a)

Dr. Paul Misael Garza López

Codirector (a)

Dr. Alejandro Garciarrubio Granados

Asesores

Dra. Silvia Armenta Jaime

Dra. Josefa Espitia López

Santiago Tulantepec de Lugo Guerrero, Hgo., México., 04 de noviembre de 2025

Universidad Autónoma del Estado de Hidalgo
Instituto de Ciencias Agropecuarias
Institute of Agricultural Sciences
Área Académica de Ciencias Agrícolas y Forestales
Academic Area of Agricultural and Forestry Sciences

Santiago Tulantepec de Lugo Guerrero, Hgo., a 04 de noviembre de 2025
Asunto: Autorización de impresión

MTRA. OJUKY DEL ROCÍO ISLAS MALDONADO
DIRECTORA DE ADMINISTRACIÓN ESCOLAR DE LA UAEH

Por este conducto y con fundamento en el Título Cuarto, Capítulo I, Artículo 40 del Reglamento de Titulación, le comunico que el jurado que le fue asignado al pasante de Licenciatura en Ingeniería en Biotecnología, **Odeth Giovanna Valencia García**, quien presenta el trabajo de Tesis denominado **"Predicción de actividad hidrolasa o transferasa en amilasas mediante modelos de lenguaje de proteínas"**, que después de revisarlo en reunión de comité de tesis, ha decidido autorizar la impresión de este, hechas las correcciones que fueron acordadas.

A continuación, se anotan las firmas de conformidad de los miembros del comité de tesis:

DIRECTOR	Dr. Paul Misael Garza López
CODIRECTOR	Dr. Alejandro Garciarrubio Granados
ASESOR	Dra. Josefa Espitia López
ASESOR	Dra. Silvia Armenta Jaime

Sin otro particular por el momento, me despido de usted.

Atentamente
"Amor, Orden y Progreso"

Dr. Paul Misael Garza López
Coordinador del P.E. de Ingeniería en Biotecnología

Avenida Universidad #133, Col. San Miguel Huatengo,
Santiago Tulantepec de Lugo Guerrero, Hidalgo,
México. C.P. 43775.
Teléfono: 7717172001 Ext. 42173
profe_5566@uaeh.edu.mx

PREDICCIÓN DE ACTIVIDAD HIDROLASA O TRANSFERASA EN AMILASAS MEDIANTE MODELOS DE LENGUAJE DE PROTEÍNAS

"Imaginar es la facultad del descubrimiento. Es lo que penetra en lo invisible, lo que ve más allá de la superficie."

- Ada Lovelace

AGRADECIMIENTOS

En primer lugar, agradezco profundamente a Dios, por ser mi guía constante y darme la fortaleza necesaria para superar los desafíos que surgieron a lo largo de este camino.

A mi madre, por su amor incondicional y por ser mi mayor ejemplo de perseverancia. A mi familia, por su cariño, comprensión y por acompañarme con paciencia y orgullo en este proceso.

A mi director de tesis, el Dr. Alejandro, le expreso mi sincero agradecimiento por su orientación, compromiso y confianza en este trabajo. Agradezco también a los miembros de mi comité tutoral, por sus valiosas observaciones, sugerencias y por el tiempo dedicado a mejorar cada aspecto de esta investigación.

Al Instituto de Biotecnología por el espacio y el ambiente académico que hicieron posible la realización de esta tesis.

Y, finalmente, a mi amado David, gracias por tu amor y tu comprensión. Gracias por ser mi refugio en los momentos de cansancio y mi impulso para seguir adelante, no imagino un futuro sin ti.

A todos ustedes, gracias por formar parte de esta etapa de mi vida.

DEDICATORIA

A mi madre:

Quien me enseñó a pensar con el corazón y a luchar con la mente. Gracias por ser mi mayor inspiración.

A David:

Mi compañero de vida y de mis sueños.

ÍNDICE GENERAL

ABREVIACIONES

En la siguiente tabla se presenta un listado exhaustivo de las abreviaturas empleadas en el presente documento, con el fin de asegurar claridad y uniformidad en su uso.

CAZy	Enzimas Carbohidrato-activas
CBM	Módulo de Unión a Carbohidrato
SBDs	Dominios de Unión al Almidón
ESM	Modelo de Escalamiento Evolutivo
ESM-C	Modelo de Escalamiento Evolutivo-Cambrian
GH13	Familia de Hidrolasas de Glúcidos 13
T/H	Transglicosilación/Hidrólisis
EC	Clasificación Enzimática
AF2	AlphaFold 2
ProteinMPNN	Red Neuronal de Propagación de Mensajes para Proteínas
LigandMPNN	Red Neuronal de Propagación de Mensajes para Ligandos
PDB	Banco de Datos de Proteínas
Barril TIM	Barril de Triosafosfato Isomerasa

MUSCLE	Comparación de Secuencias Múltiples mediante Log-Expectativa
ML	Aprendizaje Automático
PCA	Análisis de Componente Principales
ARI	Índice Rand Ajustado
NMI	Información Mutua Normalizada
t-SNE	Inserción Estocástica Vecina en el Espacio de Dimensiones Reducidas
UMAP	Proyección de Aproximación de Manifold Uniforme
PLM	Modelos de Lenguajes de Proteínas

RESUMEN

Las enzimas de la familia GH13, tradicionalmente clasificadas como hidrolasas, también incluyen miembros con actividad transferasa. Esta dualidad funcional plantea un reto en la predicción de su especificidad catalítica a partir de la secuencia primaria. En este trabajo se emplean modelos de lenguaje de proteínas (PLMs), particularmente ESM Cambrian (ESMC), junto con herramientas estructurales como AlphaFold2 y FoldMason, para predecir si una amilasa actuará como hidrolasa o transferasa. Se construyó un conjunto de datos no redundante de proteínas caracterizadas, se alinearon sus secuencias y estructuras, y se identificaron posiciones diferenciadoras mediante análisis estadísticos robustos. Los embeddings generados por ESMC fueron analizados mediante técnicas de reducción de dimensionalidad y clustering, revelando capas específicas del modelo capaces de separar las funciones enzimáticas. Asimismo, se identificaron residuos clave, tanto a nivel de secuencia como estructural, que podrían determinar la especificidad funcional. Este enfoque demuestra el potencial de los modelos de lenguaje y el análisis bioinformático integrado para predecir funciones enzimáticas con aplicaciones biotecnológicas y de diseño racional de proteínas.

Palabras clave: amilasas, modelos de lenguaje de proteínas, aprendizaje profundo.

ABSTRACT

Enzymes from the GH13 family, traditionally classified as hydrolases, also include members with transferase activity. This functional duality presents a challenge in predicting catalytic specificity from primary sequence alone. This study applies protein language models (PLMs), particularly ESM Cambrian (ESMC), alongside structural tools like AlphaFold2 and FoldMason, to predict whether a given amylase will function as a hydrolase or transferase. A non-redundant dataset of biochemically characterized proteins was curated, aligned, and analyzed to identify discriminating positions through robust statistical tests. Embeddings generated by ESMC were analyzed using dimensionality reduction and clustering techniques, revealing specific model layers capable of separating enzymatic functions. Key residues were identified at both sequence and structural levels, potentially determining functional specificity. This integrated bioinformatic approach highlights the power of language models in enzyme function prediction, with implications for biotechnology and rational protein design.

Keywords: amylases, protein language models, deep learning.

CAPÍTULO 1. INTRODUCCIÓN

Las enzimas de la familia GH13 (glicosil hidrolasas tipo 13), son catalizadores clave en la degradación de polisacáridos y oligosacáridos, desempañando un papel esencial en la conversión de carbohidratos completos en unidades más simples. Estas enzimas son ampliamente distribuidas en diversos organismos, incluidos bacterias, hongos y plantas y se encuentran involucradas en procesos biológicos complejos vitales como la digestión, la biosíntesis de glicanos y la remodelación de estructuras celulares. Aunque tradicionalmente se clasifican como hidrolasas debido a su capacidad de romper enlaces glicosídicos a través de la adición de agua, algunas de las enzimas de esta familia también presentan actividad transferasa.

La actividad transferasa consiste en transferir fragmentos de carbohidratos a sustratos distintos al agua, un proceso que amplía la diversidad de sus funciones catalíticas. Esta dualidad funcional ha sido un tema creciente de interés en la investigación, ya que comprender los factores moleculares responsables de una actividad u otra podría tener implicaciones significativas para aplicaciones industriales y biotecnológicas como la producción de biocombustibles, la ingeniería de proteínas y la síntesis de nuevos compuestos bioactivos (Cantarel et al., 2009). Esta dualidad ha interesado a la comunidad científica pues estas enzimas son de gran importancia en procesos como la digestión del almidón y la modificación de carbohidratos en la industria alimentaria y farmacéutica. Sin embargo, la naturaleza exacta de los determinantes moleculares que definen si una amilasa adoptará una función hidrolasa o transferasa sigue siendo un tema poco comprendido. Si bien se han propuesto varios mecanismos de regulación enzimática, la identificación precisa de las características estructurales y secuenciales que conducen a la especificidad funcional sigue siendo un reto.

El trabajo previo de Arreola-Barroso et al. (2021), antecedente directo a este proyecto, sugiere que ciertos residuos clave en las secuencias de aminoácidos y configuraciones espaciales en las estructuras tridimensionales de las enzimas pueden estar involucrados en la modulación de la actividad. A pesar de décadas de investigación, la predicción precisa de la especificidad funcional (hidrolítica vs transferasa) a partir de únicamente la secuencia de una enzima GH13

sigue siendo un desafío computacional y experimental no resuelto. Si bien se han propuesto ciertos residuos o motivos como determinantes, no existen reglas universales, y las bases moleculares que gobiernan esta dualidad catalítica permanecen poco comprendidas. Este "vacío de conocimiento" limita nuestra capacidad para explotar el vasto universo de secuencias enzimáticas disponibles y para diseñar biocatalizadores a la medida.

En este contexto, la aplicación de modelos de aprendizaje automático ofrece un enfoque prometedor para desentrañar la compleja relación de las secuencias de aminoácidos, las estructuras tridimensionales y la actividad enzimática. Los avances recientes en el campo de la inteligencia artificial y el modelado computacional han permitido la creación de herramientas más sofisticadas para predecir la actividad de las enzimas basándose en sus características estructurales y de secuencia. Se postula que los patrones sutiles de secuencia y estructura que gobiernan la dualidad catalítica en la familia GH13, aunque difíciles de discernir con métodos tradicionales, pueden ser descifrados mediante la aplicación integrada de análisis bioinformáticos comparativos y modelos de lenguaje de proteínas (PLMs) de última generación.

Esta investigación pretende aplicar modelos de aprendizaje profundo para predecir si una amilasa específica exhibirá actividad hidrolasa o transferasa. Al examinar patrones secuenciales y estructurales de las amilasas, este proyecto busca identificar las características clave que influyen en la especificidad de la reacción. Basándose en investigaciones previas sobre la modulación de la actividad de glicosil hidrolasas, particularmente el trabajo reportado por Arreola et al. (Arreola-Barroso et al., 2021), ya mencionado, este proyecto aprovechará herramientas computacionales como *Evolutionary Scale Modeling* (ESMC), AlphaFold2 (AF2), FoldMason (FM), y otros, para evaluar y comparar el rendimiento y resultados de cada uno de los modelos (Hou et al., 2022).

La capacidad de predecir con fiabilidad la función de una amilasa a partir de su secuencia tendría un impacto significativo. Este enfoque integrado de predicción de actividad que combina aprendizaje automático con información estructural y evolutiva, no sólo permitirá mejorar nuestra comprensión de los determinantes moleculares de la actividad enzimática en amilasas, sino que también abrirá nuevas posibilidades para la ingeniería de enzimas con actividades específicas.

CAPÍTULO 2. MARCO TEÓRICO

Introducción a la familia GH13.

La familia GH13 de la base de datos de *Carbohydrate-Active Enzymes* (CAZy) es la mayor familia de α-amilasas, pues abarca más de 184,000 secuencias en conjunto con ~800 miembros caracterizados bioquímicamente (Drula et al., 2022). Esta familia incluye enzimas que actúan sobre enlaces α-1,4 glicosídicos en polímeros y oligómeros de glucosa. Entre sus miembros hay proteínas con distinta especificidad de reacción. Sus estructuras tienen obligatoriamente tres dominios conservados de los cuales el dominio A, llamado "barril de triosafosfato isomerasa" (TIM), es el más importante. La estructura de estos dominios se discute más abajo. También pueden contener dominios adicionales como los módulos de unión de carbohidratos (CBM). Los CBM juegan un papel importante en la interacción de la enzima y los polisacáridos, mejorando la unión del sustrato y aumentando la eficiencia catalítica. Estos módulos permiten que los catalizadores actúen más eficientemente sobre sustratos insolubles como el almidón y la celulosa, lo que es particularmente importante para las amilasas y otras enzimas transferasas en sistemas biológicos e industriales (Z. Zhang et al., 2024). Existen también los llamados dominios de unión al almidón (SBDs), que en la clasificación de CAZy se agrupan en diversas familias de módulos de unión a carbohidratos. En términos generales, los SBDs ayudan al dominio catalítico mediante uno o dos sitios de unión, facilitando la degradación de sustratos derivados del almidón (Mareček et al., 2024).

Clasificación de GH13 por tipo de reacción.

Las glicosil hidrolasas pueden clasificarse según el resultado estereoquímico de la reacción de hidrólisis: así, pueden clasificarse como enzimas que retienen o invierten la configuración. También pueden clasificarse como exo o endo actuantes, dependiendo de si actúan en el extremo (usualmente no reductor) o en el centro de una cadena oligo/polisacarídica, respectivamente. También pueden clasificarse por secuencia o estructura. Sin embargo, la principal clasificación de las glicosiladas GH13 es por su tipo de reacción enzimática, representada por el número EC (Enzyme Commission number).

α-amilasas (EC 3.2.1.2)

Son enzimas con capacidad hidrolítica que rompen los enlaces α-1,4 glucosídicos en almidones y polisacáridos de manera interna (endohidrolasa). Esto las convierte en enzimas clave para la degradación de carbohidratos, obteniendo en consecuencia fragmentos de maltosa, glucosa y dextrinas. Tiene acción aleatoria en las cadenas de glucosa y su pH óptimo es entre 6.7 y 7.0 (Hernández-Heredia, 2018).

β-amilasas (EC 3.2.1.2)

Rompen enlaces α-1,4 glucosídicos desde los extremos no reductores de los polisacáridos. A diferencia de las α-amilasas, son consideradas exohidrolasas. Liberan maltosa (dos moléculas de glucosa) de manera consecutiva. Su acción es lenta y ordenada, siendo primordial en la maduración de semillas y el proceso industrial de los carbohidratos (Hernández-Heredia, 2018).

Amiloglucosidasas (EC 3.2.1.3)

Catalizan la hidrólisis de enlaces α-1,4 y α-1,6 glucosídicos en polisacáridos como el almidón, liberando glucosa libre. Estas enzimas son exohidrolasas, ya que actúan en los extremos no reductores de las cadenas de carbohidratos. Su acción es crucial para procesos de fermentación y en la producción de jarabes de glucosa, debido a su capacidad de convertir almidón en glucosa de manera eficiente (Hernández-Heredia, 2018).

Pululanasas (EC 3.2.1.41)

También conocidas como pululan-glucanohidrolasa, hidrolizan enlaces glucosídicos α-1,6 en la molécula de pululano (polisacárido formado por unidades de maltotriosa) del almidón y tiene como productos principales la maltotriosa y la maltosa. Se usan para producir maltosa a partir de almidón (Hernández-Heredia, 2018).

Ciclodextrinas glucanotransferasas (EC 2.4.1.19)

Catalizan la ciclización de fragmentos de amilosa formando ciclooligosacaridos. Son caracterizadas por ejecutar diversas reacciones catalíticas; como ciclación, transglicosilación,

hidrólisis, acoplamiento y desacoplamiento. Transforman el almidón en diversos polisacáridos pequeños, principalmente ciclodextrinas (CDs) (Hernández-Heredia, 2018).

Transglicosilasas (EC 2.4.1.25)

También conocidas como 4-α-glucano transferasas, catalizan la transferencia de residuos de glucosa desde un polisacárido donador a un aceptor que puede ser otro azúcar, un oligosacárido o incluso un alcohol. En vez de hidrolizar llevan a cabo una reacción de transglicosilación, pues forman un nuevo enlace glicosídico. Sus principales sustratos son el almidón, glucógeno y otros glucanos que contienen enlaces α-1,4 (Hernández-Heredia, 2018).

Sacarosa fosforilasa (EC 2.4.1.7)

Estas transferasas transfieren un grupo glicosilo de la sacarosa a un fosfato inorgánico, lo que resulta de la glucosa-1-fosfato. Este producto es intermediario clave en varias rutas metabólicas como la gluconeogénesis, y la glucólisis, lo que subraya la importancia de esta enzima en el metabolismo de carbohidratos (Hernández-Heredia, 2018).

Enzimas ramificadoras (EC 2.4.1.18)

Las enzimas ramificadoras del glucógeno o glicosil transferasas ramificadoras son enzimas clave en la síntesis de glucógeno y otros polisacáridos ramificados. Catalizan la formación de enlaces glicosídicos α-1.6 a partir de una cadena lineal de glucosa unida por enlaces α-1,4, introduciendo ramificaciones en la molécula (Hernández-Heredia, 2018).

Las amilasas son proteínas caracterizadas por ser metaloenzimas, es decir, dependen de iones metálicos, típicamente calcio (Ca^{2+}), necesarios para mantener la estabilización de la arquitectura de la hendidura catalítica y la termoestabilidad. Principalmente son caracterizadas por favorecer la hidrólisis de los enlaces glucosídicos en los carbohidratos, descomponiendo moléculas complejas como el almidón y el glucógeno en azúcares más simples como la maltosa y la glucosa (Xiong et al., 2024).

Schormann et al., (2023) reportaron que en dos glucanosucrasas de la familia GH70 (homólogas a GH13), se introdujo una mutación puntual en uno de los residuos de aspartato (Asp411 y

Asp437) los cuáles están involucrados en unión del Ca2+. Esta modificación afectó su actividad, lo que sugiere que la estabilidad estructural de estas enzimas depende del ion de calcio.

Por ende, el calcio juega un papel importante, ya que proporciona estabilidad a la proteína y ayuda a mantener la correcta configuración del sitio activo. Además, algunas α-amilasas pueden requerir otros iones metálicos como sodio (Na^+) o cloruro (Cl^-) para su máxima actividad (Ugwuoji et al., 2024).

Estructura tridimensional de las glicosil hidrolasas GH13.

Zhang, Han y Xiao (2017) mencionaron que cada familia GH tiene su propio mecanismo molecular conservado y la estructura cristalina de la proteína está mejor conservada que las secuencias. Por lo que todas las enzimas de la familia GH13 deben cumplir con cuatro criterios.

a. Ruptura de enlaces glucosídicos con retención en la configuración α-anomérica.
b. Presencia de cuatro a siete regiones altamente conservadas.
c. Adopción de dominio catalítico de barril TIM (β/α)8.
d. Catálisis por tríada catalítica conservada
    - Ácido aspártico β4 (nucleófilo/base)
    - Ácido glutámico β5 (donador de protones)
    - Ácido aspártico β7 (estabilizador del estado de transición)

Hay tres dominios comunes a todas las proteínas de esta familia: **A, B** y **C.** El **dominio A** que corresponde al TIM barrel es el más importante pues es el que contiene el sitio catalítico. Los TIM barrels son comunes en muchas enzimas de todo tipo de especificidades, más allá de las glicosil hidrolasas. Se cree que el 10% de todas las enzimas utilizan un TIM barrel como dominio catalítico. Los TIM barrel también son llamados “barriles (β/α)-8”, ya que su unidad básica es una hebra β seguida de una hélice α, y esta unidad se repite 8 veces. Las repeticiones se acomodan alrededor de un eje central cerrando un barril de tal forma que la β8 queda contigua a la β1. Las betas son paralelas y ascienden por el centro del barril, las hélices alfa descienden por

la parte exterior del barril. La función de este dominio se explica más ampliamente en mecanismo de catálisis. El **dominio B** se inserta entre la hebra β-3 y la hélice α-3 del dominio **A**. El dominio B contribuye a la especificidad del sustrato y a la estabilidad estructural de la enzima. En algunas α-amilasas, participa en la unión de iones calcio (Ca^{2+}), que estabilizan la estructura terciaria y son esenciales para la actividad enzimática. Finalmente el **dominio C** se encuentra en el extremo C-terminal de la enzima. Es un dominio globular compacto, típicamente compuesto por una estructura de sándwich β-antiparalela de ocho cadenas. Aunque no participa directamente en la catálisis, el dominio C es importante para la estabilidad estructural de la enzima y puede desempeñar un papel en la unión al sustrato o en la interacción con otros componentes celulares (Q. Zhang et al., 2017).

Las comparaciones de las características de la familia de las amilasas, podrían indicar que existen de cuatro a siete regiones conservadas. Actualmente se han estudiado cuatro residuos de aminoácidos conservados entre todas las enzimas. De los cuáles; tres (β4-aspartato, β5-glutamato y β7-aspartato) son esenciales para actividades de la maquinaria catalítica común en la familia de las α-amilasas (Q. Zhang et al., 2017).

Regiones conservadas y distribución de residuos catalíticos en amilasas.

Según Samanta (2022), la familia de las α-amilasas tiene cuatro (I-IV) regiones de secuencia altamente conservadas en su estructura de barril TIM. Están presentes en los extremos terminales de las hebras β3, β4 y β5, y en el bucle de la hebra β7 con la hélice α7. Estas regiones forman el sitio de unión del sustrato y el centro catalítico. Los aminoácidos de la triada catalítica están siempre presentes en las regiones de secuencia conservadas II, III y IV, pero su posición se diferencia entre las amilasas. Estás regiones muestran contribución en la actividad catalítica y además ayudan a la capacidad de unión del calcio.

Además existen otras tres regiones presentes en muchas amilasas de GH13, incluyendo glicosil-hidrolasas, transferasas e isomerasas. (Samanta, 2022). Dos de ellas (VI y VII) se encuentran en la hoja β2 y la hoja β8 del *scaffold* (barril TIM β/α_8). La región V está presente cerca del C-terminal del dominio B conectando con la región de la hoja β3 y la hélice α3 cercano del aspartato que une el calcio (Samanta, 2022). La región V Estabiliza el sitio de unión del

calcio, esencial para la estructura global y la actividad catalítica. Contribuye a la flexibilidad del dominio B, que en muchas amilasas regula el acceso al sitio activo.Puede influir en la especificidad de sustrato y en la estabilidad térmica. Las regiones VI y VII participan en la formación del sitio activo. Contribuyen al posicionamiento preciso del sustrato durante la catálisis. Pueden afectar la orientación de los residuos catalíticos o del sustrato en el barril TIM. En algunos casos, actúan como regiones adaptativas que permiten la evolución de nuevas funciones (como la transglicosilación o la isomerización).

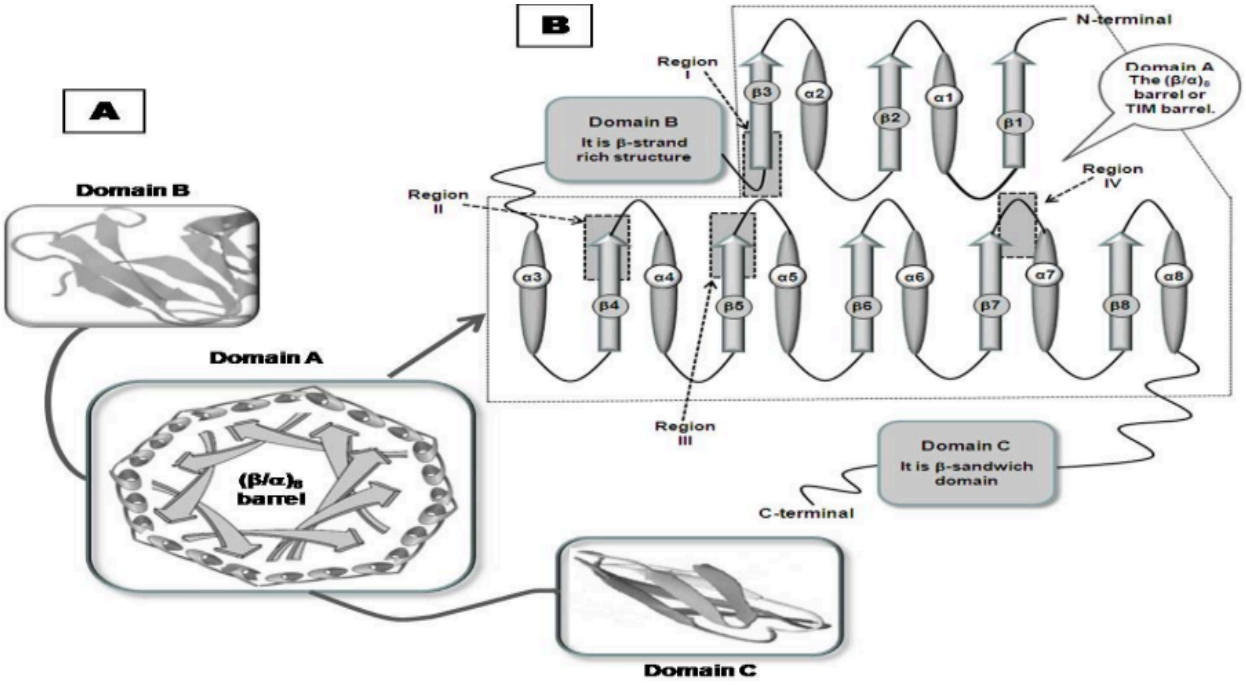

Figura 1. *Representación esquemática de la estructura molecular de una amilasa*

Fuente: (Samanta, 2022)

Nota. Se muestran tres dominios distintos con una configuración diferente. La estructura de barril (β/α)8 está presente en el dominio A. B: Topología del barril (β/α)8. Las posiciones de las cuatro secuencias conservadas (I–IV) se indican con cuadros sombreados.

Mecanismo de actividad catalítica de las amilasas.

La actividad hidrolítica de las amilasas se lleva a cabo mediante un mecanismo de doble desplazamiento que conserva la configuración α del enlace glucosídico. El sustrato, que requiere

de cuatro a diez unidades de glucosa del almidón, se une a una hendidura en el sitio activo, donde los subsitios de unión se designan con una nomenclatura específica: los subsitios negativos (-n) acomodan el extremo no reductor del sustrato, mientras que los subsitios positivos (+n) sostienen el extremo reductor. La catálisis ocurre entre los subsitios -1 y +1, donde la tríada catalítica, formada por el β4-aspartato, el β5-glutamato y el β7-aspartato, desempeña roles esenciales.

El primer paso del mecanismo, paso de glicosilación, el β5-glutamato actúa como catalizador ácido-base, donando un protón al oxígeno del enlace glucosídico en el subsitio -1, lo que facilita la salida del grupo saliente. Simultáneamente, el β4-aspartato, funcionando como nucleófilo, ataca el carbono C1 de la unidad de glucosa (carbono anomérico) en el subsitio -1, formando un intermediario covalente glucosilo-enzima. En la segunda fase, o desglicosilación, el mismo residuo de glutamato actúa ahora como catalizador base, activando a la molécula atacante (generalmente agua, aunque en las transferasas puede ser otro azúcar o un alcohol) al desprotonarla. Esta molécula activada hidroliza el intermediario, liberando el producto y regenerando la enzima (Samanta, 2022).

Aplicaciones de las amilasas.

Las amilasas están entre los catalizadores comerciales más importantes, representando entre el 25 y 30% del mercado mundial de enzimas industriales. Su aplicación en la biotecnología moderna es muy amplia, ya que incluye procesos comerciales e industriales, tales como la fabricación de papel, panificación, licuefacción y sacarificación del almidón, así como en el área química médica y clínica. En la industria papelera la amilasa se usa para eliminar el almidón, que interfiere con los procesos de texturificación y blanqueado. Esta enzima ofrece una alternativa más económica que los procesos basados en tratamientos químicos (Z. Zhang et al., 2024).

En farmaceútica, las amilasas suelen formar parte de complejos enzimáticos que favorecen la digestión. También se usan en la formulación de diversos fármacos de liberación prolongada. Por otro lado, en la producción de bioetanol por fermentación, se usan sustratos lignocelulósicos y almidonados, como cereales, bagazo y patatas. Para optimizar la producción, estos sustratos deben descomponerse en azúcares fermentables, un proceso que comienza con la acción de la α-amilasa, que los convierte en moléculas pequeñas fácilmente fermentables (Paul et al., 2021).

El Problema Bioquímico: La Relación Transglicosilación/Hidrólisis (T/H).

Como ya se mencionó, la reacción catalítica de una enzima GH13 procede a través de un intermediario covalente glico-enzima. El destino de este intermediario define la función de la enzima. Sí el intermediario es atacado por una molécula de agua, el resultado es la hidrólisis. Sí es atacado por el grupo hidroxilo de otra molécula de azúcar (el aceptor), el resultado es la transglicosilación. La preferencia por uno u otro aceptor se cuantifica mediante la relación T/H. Una T/H alta indica una transferesa eficiente, mientras que una T/H baja caracteriza a una hidrolasa. Se cree que la arquitectura del sitio activo, incluyendo su accesibilidad al solvente y su afinidad por el aceptor de azúcar, es el factor determinante de esta relación, pero los determinantes estructurales específicos siguen siendo objeto de intenso estudio.

Aprendizaje profundo y herramientas computacionales para el análisis de proteínas.

Los modelos de aprendizaje profundo están revolucionando la biología computacional al permitir el análisis de patrones complejos en secuencias y estructuras de proteínas. Estos modelos aprenden relaciones intrincadas entre las secuencias de aminoácidos y las funciones, estructuras e interacciones de las proteínas facilitando predicciones que antes eran inalcanzables.

Se han desarrollado varias herramientas computacionales avanzadas para mejorar el análisis de proteínas, entre las que destacan:

AlphaFold2

AlphaFold2 representa un avance significativo en la predicción de estructuras de proteínas, logrando una precisión notable, similar a la de los métodos experimentales. Al predecir estructuras 3D de alta calidad a partir de secuencias de aminoácidos, AlphaFold2 permite a los científicos conocer la base estructural de la función proteica, incluso para proteínas sin estructuras experimentales disponibles (Jumper et al., 2021).

AlphaFold3

Una versión más reciente, amplía estas capacidades al incorporar información de ARN, interacciones proteínas-proteína y ligandos pequeños, como metales, cofactores y otros, lo que permite predicciones más integrales y biológicamente relevantes (Abramson et al., 2024).

RFdiffusion

Es un modelo generativo basado en redes de difusión entrenado sobre la arquitectura de RoseTTAFold, que permite diseñar nuevas proteínas a partir de restricciones espaciales, como sitios activos o motivos estructurales (Watson et al., 2022).

ProteinMPNN

(Red Neuronal de Propagación de Mensajes para Proteínas) es una herramienta que diseña secuencias proteicas condicionadas en una estructura tridimensional. Genera secuencias que adoptarán preferentemente esa conformación, optimizando la estabilidad, solubilidad y expresabilidad en organismos como *E. coli* (Dauparas et al., 2022).

FoldMason y el sistema de símbolos 3Di para análisis estructural

FoldMason es un programa para generar alineamientos múltiples de estructuras proteicas. Es extremadamente rápido. Esto se debe a que convirtió el problema tridimensional a uno unidimensional. Para ello codifica las interacciones 3D de los aminoácidos en una secuencia de símbolos que llama 3Di de forma tal que la representación tiene exactamente la misma longitud que la secuencia. Una vez así codificado utiliza MMseqs2 para generar el alineamiento. Mucha gente ha descubierto que los 3Di son muy útiles para otros fines más allá que hacer alineamientos pues contienen importante información de las estructuras. Los tipos de interacciones existentes de los símbolos 3Di se encuentran en suplementos (Figura S1)

Modelos de Lenguaje de Proteínas (PLM)

Los PLM son algoritmos de aprendizaje profundo entrenados sobre grandes cantidades de secuencias de aminoácidos que tratan a las proteínas como un lenguaje por analogía con los lenguajes humanos. Estos modelos se basan típicamente en arquitectura tipo *transformer,*

introducidas por Vaswani et al. (2017), que permiten capturar relaciones complejas de dependencias a largo plazo dentro de las secuencias. Modelos como los de la familia de ESM, aprenden representaciones de las secuencias en espacios de alta dimensión, capturando tanto información evolutiva como propiedades estructurales y funciones relevantes (ESM Team, 2024). Al igual que los LLMs (Large Language Models), como ChatGPT, los PLM son autoregresivos, es decir, generan tokens (en este caso aminoácidos) uno a la vez y el último token se convierte en parte de la entrada para la siguiente predicción.

Modelo de Escala Evolutiva 2 (ESM2)

Es un modelo de lenguaje de proteínas entrenado para predecir secuencias de proteínas que se plegarían como las proteínas naturales. Fue entrenado con millones de secuencias de proteínas por lo que adquirió un gran conocimiento de cómo son estas. Aunque no es su función, frecuentemente se usan sus embeddings como representaciones enriquecidas de proteínas, siendo útiles para tareas cómo predicción de estructuras, anotación funcional y clasificación de familias enzimáticas (Lin et al., 2023).

Modelo de Escala Evolutiva 3 (ESM3)

La versión más reciente, es un modelo multicanal que permite simulaciones evolutivas, predicción de estructura, secuencia y función de manera conjunta. Este modelo se ha usado para simular cientos de millones de años de evolución, generando variantes plausibles a partir de secuencias reales, lo cual lo convierte en una herramienta especialmente poderosa para el diseño racional de proteínas y las predicciones de cambios funcionales (Hayes et al., 2025). ESM3 se entrenó para predecir a voluntad del investigador ya sea la secuencia, la estructura, o la función de una proteína.

ESM Cambrian (ESMC): especializado en representaciones de proteínas.

Este modelo se entrenó en paralelo con ESM3. Su función específica es generar representaciones de proteínas, como los embeddings arriba descritos. De ESM3 solo conserva el canal de secuencia omitiendo los de estructura y función. ESMC está diseñado para producir representaciones computacionales de la biología subyacente de las proteínas. Al escalar tanto los

datos de entrenamiento como los recursos computacionales, ESMC logra mejoras significativas en rendimiento frente a su predecesor, ESM2. En cierta forma los embeddings permiten análisis que anteriormente requerían aproximaciones experimentales. (ESM Team, 2024).

Cuando se entrena un modelo se suele generar variantes de distintos tamaños. Teóricamente, el mayor es el más capaz pero las variantes menores mantienen buena parte de esa capacidad. Cabe aclarar que los modelos muy grandes requieren hardware costoso que solo existen en las grandes compañías. En el caso de ESMC este existe en tres distintas escalas:

ESMC 300M: modelo abierto con 300 millones de parámetros

ESMC 600M: modelo abierto con 600 millones de parámetros, que iguala el rendimiento de modelos mucho más grandes de generaciones anteriores.

ESMC 6B: Modelo con 6 mil millones de parámetros, disponible en plataformas académicas y comerciales, superando ampliamente el desempeño de modelos anteriores.

Mapas de contacto: Las distancias interatómicas capturan la estructura de una proteína.

Los mapas de contacto, y más generalmente, los mapas de distancia atómica son una matriz que representa la distancia euclidiana, en la estructura, entre los distintos residuos de una proteína. En sus encabezados de columnas y filas se encuentra la numeración de los residuos. Los valores de la tabla son distancias en Amstrongs. Dado el gran número de restricciones geométricas en esta tabla, esta suele ser compatible con una y solo una estructura. De hecho, se podría afirmar que el transformer de AlphaFold2 y 3 tiene como principal función predecir una matriz de distancia para la proteína. Los mapas de distancias (y de contacto) son muy útiles para el análisis de estructuras proteicas. Es más fácil trabajar con ellos que con coordenadas atómicas.

En el trabajo de Arreola-Barroso et al. (2021) se usaron estas matrices para determinar los contactos entre residuos y sobreponer a esta información si los residuos variaban de forma correlacionada. De aquí se predijo algunos residuos que podrían ser determinantes de la relación H/T. El análisis experimental posterior dio cierto soporte a esas predicciones. Este es sólo un ejemplo de cómo los mapas de distancia sirven para descubrir patrones que ayudan a los

investigadores a predecir la función enzimática y diseñar mutaciones dirigidas para alterar la especificidad.

CAPÍTULO 3. ANTECEDENTES

Un desafío significativo en enzimología es la capacidad de modificar y diseñar enzimas con especificidad hacia reacciones particulares. La evolución dirigida y el diseño de novo, han demostrado avances, sin embargo, comprender las dinámicas que dictan la especificidad enzimática sigue siendo un desafío. Las investigaciones han demostrado que existen residuos fuera del sitio catalítico que pueden tener un impacto crítico en la estructura y función de las proteínas, ampliando la investigación sobre especificidad enzimática más allá del núcleo catalítico (Q. Zhang et al., 2017).

Modulación de la actividad GH13.

El estudio de Arreola et al. (2021) se centró en la modulación de la actividad de las hidrolasas entre la hidrólisis y las reacciones de transferencia realizando un enfoque evolutivo. Se utilizó α-amilasa (TmAmyA) y la glucanotransferasa (TmGTase) de *Thermotoga maritima* como enzimas modelos, los investigadores desarrollaron un enfoque computacional basado en el análisis de contactos residuo-residuo para identificar posiciones clave de aminoácidos responsables de la especificidad de reacción. Este método permitió identificar residuos que coevolucionaron y fueron considerados esenciales para su especificidad, aunque se localizaran fuera del sitio activo, prediciendo así que mutaciones podrían alterar la proporción de transglicosilación/hidrólisis (T/H) (Arreola-Barroso et al., 2021).

Se identificaron y se mutaron las posiciones **Lys98** y **Asp99** en TmAmyA, adicionales a la ya conocida posición **His222**. Se generó una triple mutante con estas mutaciones: K98P (esto se lee como Lisina [K] en la posición 98 mutada a Prolina [P]), D99A, y H22Q. Se demostró experimentalmente que esta triple mutante duplicó la proporción T/H en comparación con la enzima nativa. Cabe aclarar que este cambio en T/H se debió más a una disminución en la actividad de transferasa, que a un aumento en la actividad de hidrolasa.

Por otra parte, para transformar TmGTase, que es predominantemente una transferasa, en una enzima con mayor actividad hidrolítica, se probaron mutaciones en residuos **Phe72**, **Val86**, **Thr274** y **Met279**. De ellas, la variante **M279N** aumentó la actividad hidrolítica en un 25% y disminuyó la actividad transglicosídica, resultando en un incremento de cinco veces la relación H/T (Arreola-Barroso et al., 2021).

Importancia de residuos conservados en la actividad de transglicosilación de amilasas

En el trabajo de Casa-Villegas y colaboradores se analizó la amilasa maltogénica OPMA-N de *Bacillus sp.,* también conocida como amilasa multifuncional-N, que tiene la capacidad de utilizar almidón como sustrato para producir isomaltotriosa e isomaltotetraosa. En su estudio, identificaron que el residuo **Trp358**, altamente conservado, juega un papel crucial en la actividad de transglicosilación. Este residuo se encuentra cercano a la triada catalítica conformada por **Asp327** (nucleófilo), **Glu356** (ácido/base) y **Asp423** (estabilizador del intermedio catalítico), los cuales están ubicados en el barril TIM, específicamente en el subsitio 2+. Esta región es clave para la interacción con los sustratos, y la posición estratégica de **Trp358** resalta su importancia en la funcionalidad de esta enzima (Casa-Villegas et al., 2018).

Estas investigaciones demostraron que es posible predecir y modificar la relación H/T identificando sitios importantes, incluso cuando se encuentran alejados del sitio activo, subrayando el papel de los contactos residuo-residuo en la especificidad enzimática, además de identificar residuos que no son necesariamente parte de la triada catalítica pero tienen un papel importante para la especificidad de actividad.

CAPÍTULO 4. JUSTIFICACIÓN E HIPÓTESIS

Justificación

Distinguir entre la actividad transferasa e hidrolasa en las amilasas es crucial para maximizar su potencial biotecnológico. Aunque tradicionalmente se han considerado hidrolasas capaces de degradar carbohidratos complejos, se ha demostrado que algunas amilasas también exhiben actividad transferasa, lo que permite la síntesis oligosacáridos funcionales. La predicción del tipo de actividad enzimática resulta fundamental para optimizar su uso en procesos industriales

-incluyendo la producción de jarabes, la síntesis de compuestos bioactivos y el desarrollo de alimentos funcionales- , reducir la necesidad de pruebas experimentales, ahorrando así tiempo y recursos en el desarrollo de biocatalizadores, y obtener una comprensión más profunda de la especificidad enzimática mediante el análisis de las diferencias estructurales entre ambas actividades. En este contexto, el presente estudio se enfoca en la familia glicosil hidrolasa GH13, que comprende ocho tipo de enzimas y en la que predominan las amilasas, con el fin de establecer una base amplia para su análisis y caracterización.

Hipótesis

Las nuevas herramientas para diseño de proteínas basadas en modelos de aprendizaje profundo mejoran la predicción de la especificidad hidrolasa/transferasas en enzimas de la familia GH13.

CAPÍTULO 5. OBJETIVOS

Objetivo general

Identificar y caracterizar patrones estructurales y de secuencia específicos que determinan la actividad hidrolasa o transferasa en amilasas utilizando modelos de lenguaje de proteínas y técnicas de aprendizaje profundo.

Objetivos particulares

Construir un conjunto de datos integral curado y no redundante de hidrolasas y transferasas de la familia GH13.

Desarrollar e implementar un pipeline computacional para identificar posiciones de aminoácidos y estados estructurales (3Di) que se correlacionan estadísticamente con cada función.

Integrar el modelo de lenguaje de proteínas ESM Cambrian para generar embeddings de secuencias y evaluar su capacidad para separar las clases funcionales en un espacio latente.

Comparar las regiones identificadas por el análisis comparativo con las representaciones aprendidas por el modelo de lenguaje para obtener una visión unificada de los determinantes funcionales.

Evaluar y comparar el rendimiento de diferentes aproximaciones computacionales para la clasificación funcional.

CAPÍTULO 6. METODOLOGÍA Y DATOS

PARTE 1. PROCESAMIENTO DE DATOS

Selección del conjunto de estudio.

En la base de datos de enzimas glicosídicas CAZy (Cantarel et al., 2009), la familia *"Glycoside Hydrolase* 13" (GH13) incluye 211,882 proteínas, de las cuales se eligieron las 852 más confiables, con la etiqueta "Characterized". Las secuencias correspondientes se obtuvieron de UniProt (The Uniprot Consortium, 2019) basados en el identificar UniProt declarado por CAZY (UniProtID). A las secuencias se les dio un nombre corto, donde la primera letra indica su actividad (**H** para hidrolasas y **T** para transferasas, la segunda, en minúsculas indica su origen taxonómico (**a** para Archaea, **b** para eubacteria, y **e** para eucariotes), y un número final acompleta un identificador único. Las secuencias renombradas se filtraron por tamaño, conservando aquellas con longitud de 400 a 1000 aminoácidos. La redundancia se redujo usando el protocolo "easy-cluster" de MMseqs2 (Steinegger & Söding, 2017), agrupando con una identidad máxima de 80% (-i 0.8). La colección final incluyó las secuencias de los líderes de cada cluster, más amilasa de *Thermotoga maritima* (**Hb216**) que no quedó como líder. Así, la colección final incluyó 273 hidrolasas (13, 187 y 73 de Archaea, Bacteria y Eucariotes, respectivamente) y 89 transferasas (4, 66 y 19 de Archaea, Bacteria y Eucariotes, respectivamente).

Alineamiento de secuencias.

Para comparar las secuencias y descubrir en ellas posiciones destacadas por su conservación, se alinearon con MUSCLE (Edgar, 2004). Las columnas con exceso de gaps se eliminaron usando TrimAl (Capella-Gutiérrez et al., 2009) con parámetros de default. La inspección de este alineamiento (y posteriores) se hizo con Jalview (Waterhouse et al., 2009).

Construcción de filogenias moleculares.

A partir del alineamiento, se construyó una filogenia en el servidor de NGPhylogeny.fr (https://ngphylogeny.fr/; Lemoine et al., 2019) usando la modalidad "*A la carte*" y creando el pipeline [TrimAl -> PhyML+SMS(bootstrap 100X) -> Newick Display]. El árbol final se descargó en formato Newick ('.nwx'). La visualización se hizo en iTOL (https://itol.embl.de/).

Modelos de pre-computados de AlphaFold.

Para obtener las estructuras 3D de las proteínas, se descargaron los modelos en formato **PDB** de "AlphaFold Protein Structure Database" (https://alphafold.ebi.ac.uk/; Varadi et al., 2022), usando el **API** programático de ese sitio, basados en el UniProtID de cada proteína, según manifestado por CAZy. Todas las visualizaciones de estructura proteicas se hicieron con UCSF ChimeraX (Pettersen et al., 2021).

Alineamiento estructural con FoldMason.

Las estructuras se alinearon con el protocolo "easy-msa" de FoldMason (Gilchrist et al., 2024) el cual generó dos alineamientos en formato fasta. Uno con las secuencias de aminoácidos alineadas mediante gaps ('-'), otro con las secuencias de '3Di', con gaps en idénticas posiciones que el primero. Cabe recordar que los símbolos '3Di', representan, en 20 símbolos, el contexto estructural de cada residuo es la estructura tridimensional de la proteína (una síntesis de ángulos y distancias).

Se desarrolló un programa en Python (HT_differentiating_positions.py), para identificar posiciones de las proteínas que distingan hidrolasas de transferasas. El programa lee el alineamiento en formato FASTA y lo convierte, internamente, en una tabla (panda.DataFrame), que permite trabajar con columnas fácilmente. En seguida se eligen las mejores columnas (good_cols) por el hecho de tener pocos gaps (max_gaps=5). A partir de ese momento, el programa ignora las columnas que no estén en good_cols. La idea es que las regiones importantes para la función deben estar presentes en la mayoría de las proteínas (no tener gaps). Para cada columna en good_cols, el programa compara la frecuencia de símbolos entre hidrolasas y transferasas haciendo una prueba de chi-cuadrada. Esto le asigna a cada columna

una probabilidad de que las frecuencias sean iguales (provengan de la misma distribución) para ambos tipos de enzimas. Cuando la probabilidad es muy baja, esto indica que en esa columna las hidrolasas son muy distintas que las transferasas. En vez de tomar esa probabilidad como tal, el programa usa la "Corrección de Bonferroni" para compensar por el hecho de que hay muchas comparaciones (una por cada columna en good_cols) generando probabilidades "ajustadas", mucho más confiables.

Las pruebas de chi-cuadrada se realizan sobre frecuencias de símbolos. En un primer análisis los símbolos son los 20 aminoácidos, tal cual. En análisis posteriores, los aminoácidos se representan por su clase dentro de distintas clasificaciones fisicoquímicas. En las clasificaciones binarias los aminoácidos se traducen a dos símbolos (0|1); en la clasificación "positivo:negativo:sin-carga" se traducen a tres (0|1|2). Se realizan análisis para estas 11 clasificaciones:

= hidrofóbico:hidrofílico

= polar:no-polar

= cargado:neutral

= formador-de-hélice:no-formador

= formador-de-hebra:no-formador

= formador-de-coil:no-formador

= enterrado:expuesto

= aromático:no-aromático

= con-azufre:sin-el

= positivo:negativo:sin-carga

= pequeño:grande

En un análisis final los símbolos son los 3Di asignados por FoldMason (20 clases). Excepto por los símbolos usados, todos los análisis son iguales: chi-cuadrada, seguida de Bonferroni. El

programa reportó las columnas con probabilidad menor a 1E-10 para cada análisis y las frecuencias de símbolos para hidrolasas y transferasas en esas columnas.

Visualización y anotación de estructuras 3D de proteínas.

Todas las estructuras se inspeccionaron y anotaron con ChimeraX versión 1.10.

PARTE 2. EXTRACCIÓN DE LOS EMBEDDINGS DE ESMC Y VISUALIZACIÓN EN BAJA DIMENSIÓN.

Implementación del modelo.

Se empleó el modelo de lenguaje pre-entrenado ESM Cambrian (ESMC-300M, v1.0.0), el cual cuenta con 300 millones de parámetros y 30 capas de atención (transformers layers). El código original fue adaptado para procesar el conjunto de datos.

Extracción de embeddings por capa.

Para obtener los embeddings de cada capa se usó un script de python, modificado de ejemplo provisto por los autores del modelo en su sitio de GitHub (https://github.com/evolutionaryscale/esm). El programa descargó automáticamente el modelo "esmc_300m" y los parámetros pre-entrenados . El programa recibe los datos de entrada en un archivo CSV, con las siguientes columnas ['name','sequence','activity'], el cual es leído como un DataFrame de pandas: **inputDF**. A continuación el script procesa con ESMC cada secuencia de las proteínas que se le dieron. Para cada una, el resultado es una estructura que incluye los embedding crudos para las 30 capas de ESMC. Otra función en ese script permitió extraer el embedding específico de la capa que se le haya pedido.

En cada capa, el embedding por proteína es una matriz con **L** filas (la **longitud** de la proteína) y **960** columnas (960 características que el modelo calcula por aminoácido). Para comparar proteínas de diferentes longitudes, se hizo **average pooling**: se promedió sobre la dimensión de la secuencia (L). De este modo, una matriz de tamaño **L×960** se convierte en un **vector de 960** que representa a la proteína.

Enfocándonos en cierta capa, para comparar todas las proteínas, se juntaron sus embedding en una sola estructura cuyo orden es paralelo al de las secuencias en **inputDF.** En combinación, estas dos estructuras paralelas nos permiten realizar los análisis que se describen a continuación. Esos análisis se realizaron para las distintas capas, especialmente las internas (de la 7 a la 27) para ver cual capa provee más información para separar hidrolasas de transferasas.

Análisis de clustering

Para evaluar la información contenida en los embeddings, se realizó un análisis de clustering no supervisado. Se aplicó el algoritmo de k-means a los vectores de representación de cada capa. Se exploró sistemáticamente la capacidad de agrupación para k, con valores de 2,3,4 y 5. Las agrupaciones resultantes se compararon con las etiquetas hidrolasa-transferasa de las proteínas, para ver si estas agrupaciones son puras; es decir, se forman de proteínas de la misma actividad.

Reducción de dimensionalidad

Para visualizar la distribución de las proteínas en el espacio latente de alta dimensionalidad, se utilizaron tres técnicas de reducción de dimensionalidad, que permiten proyectar las 960 dimensiones a dos (o tres): Análisis de Componentes Principales (PCA) y t-SNE se realizaron con scikit-learn 1.7.2 y UMAP se realizó con UMAP 0.5.8. Cabe aclarar que PCA hace una transformación lineal y escoge los ejes que generan la mejor proyección, aquellas que logran la mejor separación de los embeddings. En cambio t-SNE y UMAP usan transformaciones no lineales con la intención de que los embeddings cercanos en el espacio de 960 dimensiones queden cercanos en el espacio de baja dimensionalidad y aquellos que estaban lejanos, queden lejanos en este nuevo espacio.

CAPÍTULO 7. RESULTADOS

Análisis a nivel de secuencia.

Para comparar posiciones equivalentes entre las estructuras de hidrolasas y transferasas, se alinearon sus modelos estructurales generados por AlphaFold usando FoldMason, el cual además regresa el alineamiento de secuencias, y el alineamiento de símbolos 3Di. Sobre las columnas del

alineamiento de secuencias se buscaron las posiciones que más difieren entre hidrolasas y transferasas. La evaluación se hizo usando la prueba de chi-cuadrada y corrección de Bonferroni.

Se identificaron alrededor de 1370 posiciones diferenciadoras, con una p ajustada menos a 1e-10.

Tabla 1.

Comparación de posiciones secuenciales entre hidrolasas y transferasas

PA	PSOH	PSOT	P-value	P-value ajustado	TAA-H	AMA-H	MCA-H	TAA-T	AMA-T	MCA-T
1997	**257**	**260**	2.5891e-44	3.529e-41	255	F	173	97	I	57
2258	348	354	5.9209e-34	4.0351e-31	228	A	67	96	Q	36
2117	-	292	3.9118e-30	1.7772e-27	181	E	94	47	W	34
2123	291	294	2.2216e-29	7.57e-27	194	I	41	84	L	30
1648	132	172	4.3382e-29	1.1826e-26	256	F	57	93	P	62
2375	387	389	6.1454e-27	1.396e-24	247	T	77	97	V	42
1620	109	149	4.7783e-26	9.3041e-24	253	K	50	97	P	23
1464	58	97	2.2669e-25	3.8623e-23	156	L	60	42	D	29
2167	318	320	8.2696e-25	1.2524e-22	245	R	49	97	M	36
2612	430	446	9.8309e-25	1.3399e-22	217	L	74	43	R	33

Nota. Análisis de posiciones específicas en el alineamiento secuencial entre hidrolasas (H) y transferasas (T), con sus correspondientes valores p y p ajustados. **PA:** Posición en el Alineamiento, **PSOH:** Posición de la Secuencia Original en Hidrolasas, **PSOT:** Posición de la Secuencia Original en Transferasas**, TAA-H:** Total de Aminoácidos de Hidrolasas, **AMA-H:** Aminoácido Más Abundante de Hidrolasas, **MCA-H:** Máxima Cantidad de Aminoácidos de Hidrolasas, **TAA-T:** Total de Aminoácidos de Transferasas, **AMA-T:** Aminoácido Más Abundante de Transferasas, **MCA-T:** Máxima Cantidad de Aminoácidos de Transferasas

La tabla 1 muestra las 10 posiciones más significativas. Como proteínas de referencia se usó la Hb61 como hidrolasa y la Tb332 como transferasa. Las posiciones se indican respecto a la secuencia de estas proteínas. Por ejemplo, en la posición 257 de Hb61 predominó una fenilalanina (F), mientras que en la posición equivalente (260) de Tb332 predominó una isoleucina (I).

Cabe hacer notar que en algunas de estas posiciones muy diferenciadas el residuo dominante puede tener propiedades estructurales muy distintas. Llama la atención la presencia de prolina en las transferasas en las posiciones 1648 y 1620 del alineamiento pues este residuo causa un quiebre en el esqueleto de la proteína. También llama la atención que en algunas posiciones dos aminoácidos dominantes son muy distintos: 2117 E/W, 1464 L/D y 2167 R/M. En cambio, la posición más diferenciada (1997) tiene residuos similares, fenilalanina en las hidrolasas e isoleucinas en transferasas. Esto no le resta importancia a la diferenciación, solo marca que a veces las proteínas requieren cambios más sutiles.

Una de las hipótesis sobre la relación H/T es que las transferasas deben tener una cavidad mayor para alojar al donador (azúcar o alcohol) que las hidrolasas que solo deben permitir la entrada del agua. Cabría suponer que las transferasas utilizan residuos más pequeños para generar esa mayor cavidad. Sin embargo, al menos en estas 10 posiciones no se aprecia esa tendencia, quizás porque no caen directamente en la superficie de la cavidad.

Análisis a nivel estructural

Los símbolos 3Di no tienen una interpretación exacta porque los asigna una red neuronal en función del ambiente de un residuo según las distancias y ángulos a residuos cercanos. Sin embargo contienen mucha de la información estructural de la proteína.

Para ver si la información estructural local es un importante diferenciador entre hidrolasas y transferasas, se analizó el alineamiento 3Di provisto por FoldMason con una estrategia similar a la que se usó en el alineamiento de secuencias.

Tabla 2.

Comparación de posiciones estructurales entre hidrolasas y transferasas

PA	PSOH	PSOT	P-value	P-value ajustado	TS-H	SMA-H	MCS-H	TS-T	SMA-T	MCS-T
1479	72	111	2.374e-31	3.2357e-28	252	Q	145	93	E	54
2085	**277**	**280**	8.6007e-27	5.8614e-24	172	N	65	90	C	70
1565	-	130	3.7348e-26	1.6969e-23	184	S	54	68	C	40
2156	309	311	6.9475e-26	2.3674e-23	235	G	68	81	L	33
1827	192	212	1.9912e-25	5.4281e-23	127	A	70	86	Q	39
2128	294	-	1.1245e-23	2.3645e-21	200	D	99	75	A	18
1954	235	239	1.2144e-23	2.3645e-21	247	P	151	67	Q	21
2496	-	407	3.4701e-23	5.9123e-21	96	D	36	40	R	37
2117	-	292	4.807e-23	7.1606e-21	181	Q	60	47	D	40
1642	127	167	5.2535e-23	7.1606e-21	256	**A**	164	97	**E**	94

Nota. Análisis de posiciones específicas en el alineamiento estructural entre hidrolasas (H) y transferasas (T), con sus correspondientes valores p y p ajustados. **PA:** Posición en el Alineamiento, **PSOH:** Posición de la Secuencia Original en Hidrolasas, **PSOT:** Posición de la Secuencia Original en Transferasas, **TS-H:** Total de Símbolos para Hidrolasas, **SMA-H:** Símbolo Más Abundante de Hidrolasas, **MCS-H:** Máxima Cantidad de Símbolos de Hidrolasas, **TS-T:** Total de Símbolos para Transferasas**, SMA-T:** Símbolo Más Abundante de Transferasas, **MCS-T:** Máxima Cantidad de Símbolos de Transferasas.

Los resultados se muestran en la tabla 2. Se identificaron varias posiciones diferenciadoras. La más destacada fue la posición 127 en Hb61, que corresponde a la 167 en Tb332. En esta posición las transferasas presentaron mayoritariamente el símbolo "E" mientras que las hidrolasas presentaron el símbolo "A". Aunque no se sabe qué significa este símbolo la diferencia indica que en esa posición el ambiente del residuo es distinto entre hidrolasas y transferasas. Cabe hacer notar que esta posición no está entre las 10 más diferenciadas a nivel de secuencia. De hecho, las 10 más diferenciadas por secuencia no coinciden ni en un caso con las

10 posiciones más diferenciadas por estructura. Otro hecho notable es que las posiciones estructurales se centran en el TIM barrel mientras las de secuencia están más distribuidas por toda la proteína.

Hb61 y Tb332 como proteínas de referencia

Para ver donde mapean las posiciones diferenciadas sobre las hidrolasas y las transferasas se desarrolló un script de Python, que las coloca sobre la hidrolasa de referencia (Hb61=1BL1) y la transferasa de referencia (Tb332=1CGT). Usando ChimeraX se identificaron las posiciones y se colorearon para su más fácil visualización.

Las posiciones más destacadas tanto por secuencia como por 3Di se localizaron en el barril TIM (dominio A), donde se encuentra el sitio activo. Las posiciones 1997 (**Phe257** en Hb61, **Ile260** en Tb332) y 2085 (**Val277** en Hb61, **Ile280** en Tb332) fueron espacialmente cercanas y potencialmente interactivas. Cabe hacer notar que mientras la posición detectada por secuencia se sobrepone perfectamente entre la hidrolasa y la transferasa de referencia la superposición no es perfecta para la posición detectada por 3Di lo cual podría explicar porque se les asigna símbolos 3Di distintos.

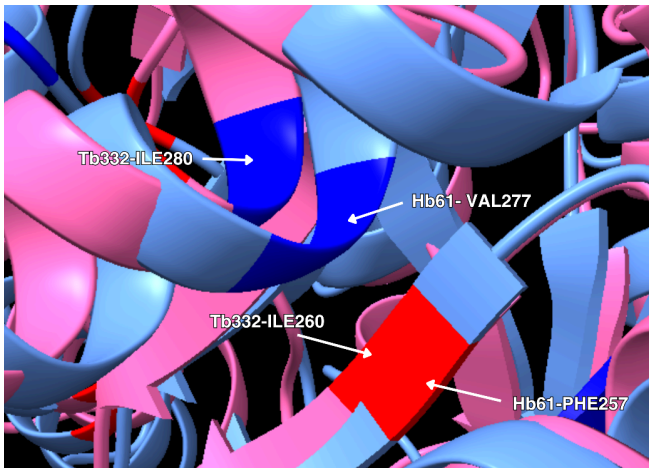

Figura 2. *Proteína Tb332 (rosa) y Hb61 (azul) y con posiciones de interacción importantes (1997 y 2085)*

Nota. El color rojo indica las posiciones de aminoácidos y el color azul las posiciones de símbolos 3Di.

En la tabla 3, se puede observar la interacción entre los aminoácidos que se mencionaron anteriormente. Estos puntos son relevantes para la investigación, pues se plantea que, debido a su proximidad espacial, podrían interactuar entre sí. Esta interacción potencial podría desempeñar un papel importante en la predicción de la actividad enzimática, destacando su importancia para la predicción funcional de las proteínas analizadas.

Tabla 3.

Aminoácidos que interaccionan cerca del sitio activo de Hb61 y Tb332.

	Hb61	Tb332
Análisis a nivel de secuencia	Phe257	Ile260
Análisis a nivel estructural	Val277	Ile280

Nota. En Hb61 a nivel de secuencia, la Phe257 interacciona con la Val277 a nivel estructural, asimismo, en Tb332, a nivel de secuencia: Ile260 interacciona con Ile280 a nivel estructural.

Para Hb61, Phe257 y Val277 se encuentran en el barril TIM (dominio A catalítico). Aunque su interacción no es muy fuerte y directa, ya que ambas tienen cadenas no polares y carecen de grupos funcionales que participen en interacciones específicas como puentes de hidrógeno o interacciones iónicas, sí puede interaccionar a través de fuerzas hidrofóbicas o por efectos de apilamiento. Ya que al encontrarse el dominio A catalítico (específicamente en el núcleo) la proximidad de estos aminoácidos podría jugar un papel clave.

La fenilalanina podría contribuir a la orientación del sustrato mediante interacciones hidrofóbicas con otras regiones del dominio. La valina al ser pequeña podría mantener la flexibilidad local o ajustar el empaquetamiento hidrofóbico alrededor del sustrato. Por ende, aunque no interactúen directamente con el sustrato, su proximidad podría influir en cómo otros residuos catalíticos se orientan para actuar como hidrolasa o transferasa.

MODELO ESM CAMBRIAN PARA CLASIFICACIÓN FUNCIONAL

Análisis de representaciones

Los grandes modelos de lenguaje de proteínas (PLM), han conocido millones de proteínas de las cuales han aprendido relaciones y propiedades que no somos capaces de describir. Cuando estos modelos reciben una secuencia nueva la "visten" con su conocimiento previo generando una versión muy enriquecida que en el modelo se representa por un vector llamado *embedding*.

Para ver sí esta información agregada por el PLM ayuda a la diferenciación y agrupamiento de hidrolasas y transferasas, se usó el modelo de ESM Cambrian (versión 300M) para analizar los embeddings que genera para estas proteínas. Los PLMs están organizados por capas de neuronas y cada capa tiene su "opinión particular" sobre la proteína; es decir, cada capa produce un embedding distinto.

Se extrajeron para cada proteína los 30 embeddings de las 30 capas de ESM Cambrian y se analizaron para saber cuál de las capas es más informativa. Para ello, a partir de cada capa, se agruparon los embeddings de todas las proteínas usando distintos valores de k-means (k=2,3,4,5). Asimismo, para cada capa se usó algún método de reducción de la dimensionalidad para proyectar los embeddings (960 dimensiones) a dos o tres dimensiones. Esto se hizo tanto con PCA, como con t-SNE y UMAP.

Se identificaron capas específicas del modelo donde los embeddings separaban claramente las funciones. En particular, las capas 12, 13, 17 y 18 mostraron una buena separación entre hidrolasas y transferasas, validada visualmente en PCA, t-SNE y UMAP.

Como se esperaba las representaciones internas del modelo ESM Cambrian capturan diferencias funcionales entre enzimas y permiten ver las dos actividades sin supervisión. Sin embargo, el PCA de la mejor capa desde el punto de vista visual, la 12, las hidrolasas forman un grupo conectado mientras las transferasas forman tres nubes separadas. Se intentó ver si el tercer componente del PCA lograba una mejor separación entre hidrolasas y transferasas, pero el resultado no fue satisfactorio.

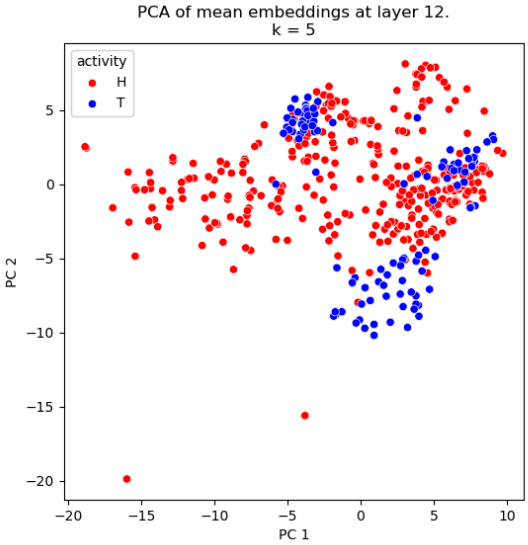

Figura 3. *PCA del promedio de embeddings en la capa 12 (k = 5) con proyecciones PC1-PC2*

Existen alternativas al PCA que prometen mantener la estructura de los datos de manera más favorable; intentan que lo que está próximo en el espacio de muchas dimensiones se mantenga próximo en el de baja dimensión a la vez que lo que está lejos en el espacio de alta dimensión aparezca lejano en el de baja dimensión. Se logró probar t-SNE y UMAP para ver si se lograba una mejor agrupación/separación de hidrolasas y transferasas. La figura 4 muestra el t-SNE de 3 grupos de transferasas formados por k-means con k=5, ocultando otros dos grupos específicos para hidrolasas. Esto comprueba que los tres grupos de transferasas que se observó en el PCA son verdaderos subconjuntos de transferasas y no un artefacto del método de proyección.

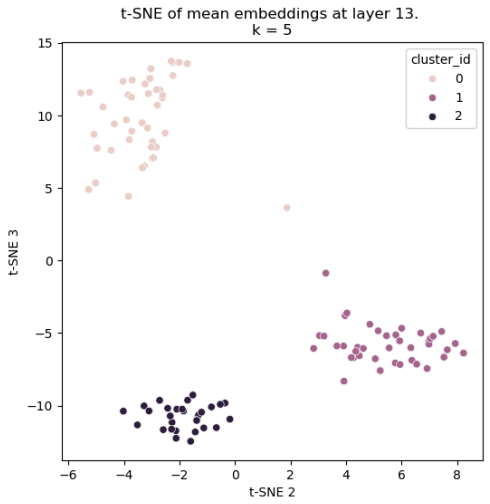

Figura 4. *t-SNE del promedio de embeddings (transferasas) en la capa 13 (k = 5)*

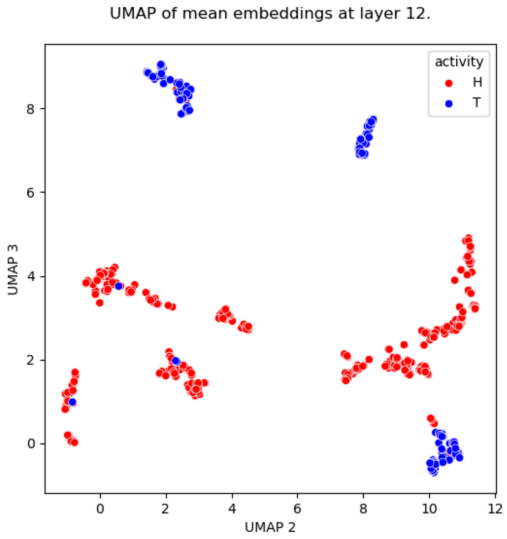

Figura 5. *UMAP del promedio de embeddings en la capa 12 (proyecciones UMAP2-UMAP3)*

Basados en la misma capa 12, se usó UMAP con un parámetro extremo para enfatizar la separación entre grupos. Como se puede ver en la figura 5, las transferasas siguen formando 3 grupos distintos. Sin embargo, bajo estas condiciones, se aprecia que las hidrolasas también están formadas por múltiples subconjuntos. Es posible que formen 5 grupos o más. Sin embargo, se ve que forman caminos de un grupo hacia otro lo cual sugiere que se han ido separando paulatinamente. También hay que notar que un pequeño grupo de hidrolasas sobrelapa con las transferasas que aparecen abajo a la derecha. Esto sugiere que son difíciles de separar. No se podía descartar que esto sea un problema de anotación que viene directamente de CAZy, y que en realidad sean transferasas. Una explicación más interesante es que hayan cambiado de transferasas a hidrolasas recientemente en su evolución.

Resumen cuantitativo del rendimiento del modelo.

Dado que la naturaleza de la tesis fue explorativa, los resultados se reportan tal y como se obtuvieron. Posteriormente, se aplicó un enfoque más cuantitativo con el fin de determinar qué capa produce los mejores embeddings, aún cuando a simple vista no parecieran los más

adecuados. Para todas las capas se calcularon las métricas estadísticas de ARI (*adjusted rand index*) y NMI (*normalized mutual information*), las cuales informan sobre el agrupamiento en el espacio de 960 dimensiones, es decir, sin necesidad de reducir la dimensionalidad.

De acuerdo con la teoría, las capas iniciales captaron características de bajo nivel mientras que las capas finales se especializaron en el objetivo de entrenamiento del modelo (para ESM3, la predicción de aminoácidos enmascarados). Los resultados indican que, para los fines del presente trabajo las mejores capas son las 17 y 18, ya que se alinean con la función biológica. El valor máximo de ARI (0.27 para k=4 en la capa 17) y NMI (0.13 para k=4 en la capa 18) se registra en dichas capas. Aunque modestos en términos absolutos, estos valores resultan altamente significativos para un método no supervisado y demuestran la presencia de una fuerte señal funcional en los embeddings.

CAPÍTULO 8. DISCUSIÓN

El presente estudio aborda un desafío fundamental en enzimología computacional: predecir si una amilasa de la familia GH13 actuará predominantemente como hidrolasa o como transferasa basándose en su secuencia primaria y estructura tridimensional. A pesar de décadas de caracterización bioquímica, este problema ha permanecido sin resolver debido a la extraordinaria conservación estructural del barril TIM y la tríada catalítica que define a esta familia. Los resultados demuestran que la integración de análisis comparativo de secuencias y estructuras con modelos de lenguaje de proteínas puede capturar señales funcionales sutiles que escapan a los métodos tradicionales basados en homología.

Posiciones diferenciadas distribuidas.

El análisis estadístico riguroso identificó aproximadamente 1,370 posiciones del alineamiento con diferencias significativas (p ajustado <1x10-10) entre hidrolasas y transferasas. Este número sustancial sugiere que la especificidad funcional no está codificada por unos pocos residuos "clave" sino por una firma distribuida a lo claro de la secuencia. Esto es consistente con el trabajo pionero de Arreola-Barroso et al. (2021), antecedente directo de esta investigación, que

demostró experimentalmente que residuos alejados del sitio activo (**Lys98, Asp99 en TmAmyA**) pueden modular la relación T/H cuando se mutan en combinación. Este enfoque computacional generaliza esta observación mediante un muestreo sistemático de la familia completa, identificando no solo las posiciones previamente reportadas sino ampliando significativamente el repertorio de sitios potencialmente moduladores.

Un hallazgo particularmente intrigante es la presencia recurrente de transferasas en posiciones donde las hidrolasas presentan residuos más flexibles (posiciones 1648 y 1620). La prolina, que induce restricciones conformacionales, podría estar "bloqueando" conformaciones que favorecen la entrada de aceptores de azúcar sobre moléculas de agua. Esta observación es relevante a la luz del trabajo de Leemhuis et al. (2003), quienes evidenciaron mediante evolución dirigida en CGTasas que mutaciones que afectan la flexibilidad de bucles específicos pueden alterar dramáticamente la relación T/H, sugiriendo que este mecanismo de restricción conformacional podría ser un principio general de GH13.

Complementariedad entre información de secuencia y estructura.

El análisis de símbolos 3Di proporcionó una perspectiva complementaria crucial. La falta de sobreposición entre las 10 posiciones más discriminativas por secuencia y las 10 más discriminativas por estructura indica que ambos niveles de análisis aportan información independientes y complementarias, coherente con estudios recientes sobre utilidad de representaciones estructurales y simbólicas (van Kempen et al., 2024).

La concentración de posiciones estructuralmente diferenciadoras en el barril TIM tiene sentido mecanístico, ya que alberga el sitio activo donde ocurre la catálisis. Esta observación es consistente con estudios estructurales y de evolución dirigida en enzimas de la familia GH13, donde Leemhuis et al. (2003) demostraron que el subsitio +1 es más espacioso en transferasas, permitiendo acomodar aceptores de azúcar, mientras que en hidrolasas, o en variantes hidrolíticas de CGTasa, esta región es más restringida por residuos voluminosos que dificultan la unión del aceptor. Estudios previos han demostrado que mutaciones en subsitios distales pueden modular la especificidad del producto en CGTasas (van der Veen et al., 2000). El análisis computacional de 362 estructuras predichas por AlphaFold2 generaliza estas observaciones cristalográficas

limitadas, identificando patrones estructurales conservados que correlacionan con la función a lo largo de toda la familia de GH13.

Captura de patrones evolutivos latentes.

ESM Cambrian reveló que los embeddings de capas internas específicas (17-18) contienen información suficiente para separar parcialmente hidrolasas y transferasas. incluso sin entrenamiento supervisado. Los valores de ARI (0.27) y NMI (0.13), aunque modestos, son estadísticamente significativos y demuestran que el modelo capturó señales funcionales relevantes. Esto es coherente con estudios que demuestran que los PLM's capturan implícitamente información evolutiva y estructural (Rives et al., 2021).

Se usó ESMC-330M por consideraciones prácticas de recursos computacionales y tiempo de análisis, siguiendo precedentes en la literatura donde modelos de escala similar han demostrado ser suficientes para tareas de clasificación funcional exploratoria. Capela et al., (2025) evaluaron sistemáticamente varios PLM's para predicción de números EC, encontrando que ESM2 superó consistentemente a otros modelos, particularmente cuando la identidad de secuencia con enzimas conocidas era menor al 25%. Reportaron además que los PLMs son complementarios a BLASTp, cada uno destacando en subconjuntos distintos de enzimas. Esto sugiere que un enfoque híbrido combinando predicciones de ESMC con alineamientos de secuencias podría superar a cualquiera de los dos individualmente.

La identificación de las capas intermedias (17-18) como las más informativas es consistente con estudios sobre interpretabilidad de modelos de lenguaje, donde las capas iniciales capturan características de bajo nivel, las intermedias codifican propiedades estructurales y funcionales, y las finales se especializan en la tarea de entrenamiento (Vig et al., 2020). Para clasificación funcional, puede ser más efectivo extraer embeddings de capas intermedias que de la capa final.

Heterogeneidad funcional y evolutiva.

Un patrón recurrente fue la mayor homogeneidad de las transferasas comparada con las hidrolasas en análisis de clustering y proyecciones dimensionales. Esta observación tiene interpretación evolutiva: la transglicosilación es una actividad más "especializada" que la

hidrólisis. Todas las GH14 deben hidrolizar enlaces α-1,4 glicosídicos, su función ancestral, pero solo un subconjunto evolucionó la capacidad de transferir eficientemente a aceptores distintos del agua. Esta especialización pudo requerir convergencia evolutiva hacia soluciones estructurales específicas, resultando en mayor similitud entre transferasas. Los análisis filogenéticos apoyan esta hipótesis, mostrando que las actividades transferasa han emergido independientemente múltiples veces dentro de GH13 (Stam et al., 2006).

Alternativamente, algunas hidrolasas podrían estar mal anotadas o ser enzimas “promiscuas” con actividad mixta. Casa-Villegas et al. (2018) caracterizaron experimentalmente la amilasa maltogénica OPMA-N, demostrando que presenta tanto actividades hidrolíticas como transglicosilasas significativas, desafiando su clasificación simple. La heterogeneidad en hidrolasas también podría reflejar genuinamente múltiples “formas de ser hidrolasa” dentro de GH13, donde diferentes subfamilias difieren sustancialmente en especificidad del sustrato y patrón de acción.

CAPÍTULO 9. CONCLUSIONES

Las herramientas basadas en modelos de aprendizaje profundo demostraron mejorar la predicción de la especificidad hidrolasa o transferasa en enzimas de la familia GH13. El uso de modelos de lenguaje de proteínas permitió identificar patrones de secuencia y estructura que distinguen las funciones hidrolasa y transferasa. La construcción del conjunto de datos curado y no redundante facilitó el análisis comparativo y la detección de posiciones clave asociadas a cada actividad.

El pipeline computacional implementado integró representaciones de secuencia (embeddings) y descriptores estructurales (3Di), mostrando correlaciones significativas con la función enzimática. Asimismo, el modelo ESM Cambrian logró separar de forma efectiva las clases funcionales en el espacio latente, evidenciando su potencial para la clasificación funcional enzimática.

La comparación entre distintas aproximaciones computacionales indicó que los métodos basados en aprendizaje profundo ofrecen mayor precisión y capacidad de interpretación frente a estrategias tradicionales, consolidando su relevancia en el diseño racional y la ingeniería de proteínas de la familia GH13.

CAPÍTULO 10. PERSPECTIVAS

Es crucial reconocer las limitaciones inherentes. Primero, dependemos de anotaciones funcionales de CAZy heredando cualquier error o inconsistencia. Un aspecto fundamental que complica tanto la anotación como la predicción es que la dicotomía hidrolasa/transferasa representa una simplificación de continuo funcional. Numerosas enzimas GH13 exhiben actividad dual significativa, con relaciones T/H que varían desde valores cercanos a cero hasta superiores a 10 (Arreola-Barroso et al., 2021). La necesidad de clasificar estas enzimas como hidrolasa o transferasa basándose en su actividad predominante introduce ruido inevitable y establece un límite superior teórico para cualquier clasificador binario. De hecho, el sobrelapamiento en subgrupos de hidrolasas con transferasas en las proyecciones de UMAP podría reflejar la realidad biológica de enzimas con actividad mixta genuina. Esto sugiere que nuestros valores de ARI (0.27) y NMI (0.13) deben interpretarse no solo como limitaciones del

enfoque no supervisado, sino también como reflejo de la inherente ambigüedad funcional de la familia GH13.

El enfoque fue exploratorio y no supervisado. Como demuestra Capela et al. (2025), el entrenamiento supervisado sobre embeddings de PLMs puede mejorar significativamente el rendimiento. Entrenar un clasificador supervisado sobre los embeddings de la capa 17-18 de ESMC es una extensión natural y prometedora.

Aunque AlphaFold2 ha revolucionado la predicción estructural, sus modelos no son perfectos. Arreola-Barroso et al. (2021) demostraron mediante dinámica molecular que mutaciones alejadas del sitio activo pueden alterar la flexibilidad de bucles cercanos a la región catalítica. Nuestro análisis estático no captura esta dimensión dinámica.

Finalmente todas las conclusiones son computacionales y requieren validación experimental. La mutagénesis dirigida de los residuos identificados particularmente (Phe257/Ile260, Val277/Ile280, y las posiciones con prolina), seguida de la caracterización cinética rigurosa, será esencial para confirmar su rol causal en la especificidad funcional.

REFERENCIAS BIBLIOGRÁFICAS

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., … Jumper, J. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, *630*, 493–500. https://doi.org/10.1038/s41586-024-07487-w

Arreola-Barroso, R. A., Llopiz, A., Olvera, L., & Saab-Rincón, G. (2021). Modulating glycoside hydrolase activity between hydrolysis and transfer reactions using an evolutionary approach. *Molecules (Basel, Switzerland)*, *26*(21), 6586. https://doi.org/10.3390/molecules26216586

Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V., & Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Research*, *37*(Database issue), D233–D238. https://doi.org/10.1093/nar/gkn663

Capela, J., Zimmermann-Kogadeeva, M., van Dijk, A. D. J., de Ridder, D., Dias, O., & Rocha, M. (2025). Comparative assessment of protein large language models for enzyme commission number prediction. *BMC Bioinformatics*, *26*(1), 68. https://doi.org/10.1186/s12859-025-06081-9

Casa-Villegas, M., Marín-Navarro, J., & Polaina, J. (2018). *Amylases and related glycoside hydrolases with transglycosylation activity used for the production of isomaltooligosaccharides*. https://doi.org/10.1515/amylase-2018-0003

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Courbet, A., de Haas, R. J., Bethel, N., Leung, P. J. Y., Huddy, T. F., Pellock, S., Tischer, D., Chan, F., Koepnick, B., Nguyen, H., Kang, A., Sankaran, B., … Baker, D. (2022). Robust deep learning-based protein sequence design using ProteinMPNN. *Science (New York, N.Y.)*, *378*(6615), 49–56. https://doi.org/10.1126/science.add2187

Drula, E., Garron, M.-L., Dogan, S., Lombard, V., Henrissat, B., & Terrapon, N. (2022). The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Research*, *50*(D1), D571–D577. https://doi.org/10.1093/nar/gkab1045

Hernández-Heredia, S. (2018). *Estudio del dominio C y de la región C-terminal de la α-amilasa de Bacillus amyloliquefaciens JJC33M*. Universidad de Papaloapan.

Hou, Q., Waury, K., Gogishvili, D., & Feenstra, K. A. (2022). Ten quick tips for sequence-based prediction of protein properties using machine learning. *PLoS Computational Biology*,

18(12), e1010669. https://doi.org/10.1371/journal.pcbi.1010669

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Leemhuis, H., Rozeboom, H. J., Wilbrink, M., Euverink, G.-J. W., Dijkstra, B. W., & Dijkhuizen, L. (2003). Conversion of cyclodextrin glycosyltransferase into a starch hydrolase by directed evolution: the role of alanine 230 in acceptor subsite +1. *Biochemistry*, *42*(24), 7518–7526. https://doi.org/10.1021/bi034439q

Mareček, F., Terrapon, N., & Janeček, Š. (2024). Two newly established and mutually related subfamilies GH13_48 and GH13_49 of the α-amylase family GH13. *Applied Microbiology and Biotechnology*, *108*(1), 415. https://doi.org/10.1007/s00253-024-13251-x

Paul, J. S., Gupta, N., Beliya, E., Tiwari, S., & Jadhav, S. K. (2021). Aspects and recent trends in microbial α-amylase: A review. *Applied Biochemistry and Biotechnology*, *193*(8), 2649–2698. https://doi.org/10.1007/s12010-021-03546-4

Pettersen, E. F., Goddard, T. D., Huang, C. C., Meng, E. C., Couch, G. S., Croll, T. I., Morris, J. H., & Ferrin, T. E. (2021). UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Science: A Publication of the Protein Society*, *30*(1), 70–82. https://doi.org/10.1002/pro.3943

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., & Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*

of the United States of America, *118*(15), e2016239118. https://doi.org/10.1073/pnas.2016239118

Samanta, S. (2022). Structural and catalytical features of different amylases and their potential applications. *Jordan Journal of Biological Sciences*, *15*(02), 311–337. https://doi.org/10.54319/jjbs/150220

Stam, M. R., Danchin, E. G. J., Rancurel, C., Coutinho, P. M., & Henrissat, B. (2006). Dividing the large glycoside hydrolase family 13 into subfamilies: towards improved functional annotations of alpha-amylase-related proteins. *Protein Engineering, Design & Selection: PEDS*, *19*(12), 555–562. https://doi.org/10.1093/protein/gzl044

Ugwuoji, E. T., Eze, I. S., Nwagu, T. N. T., & Ezeogu, L. I. (2024). Enhancement of stability and activity of RSD amylase from Paenibacillus lactis OPSA3 for biotechnological applications by covalent immobilization on green silver nanoparticles. *International Journal of Biological Macromolecules*, *279*(Pt 1), 135132. https://doi.org/10.1016/j.ijbiomac.2024.135132

van der Veen, B. A., Uitdehaag, J. C., Penninga, D., van Alebeek, G. J., Smith, L. M., Dijkstra, B. W., & Dijkhuizen, L. (2000). Rational design of cyclodextrin glycosyltransferase from Bacillus circulans strain 251 to increase alpha-cyclodextrin production. *Journal of Molecular Biology*, *296*(4), 1027–1038. https://doi.org/10.1006/jmbi.2000.3528

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., Söding, J., & Steinegger, M. (2024). Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*, *42*(2), 243–246. https://doi.org/10.1038/s41587-023-01773-0

Vig, J., Madani, A., Varshney, L. R., Xiong, C., Socher, R., & Rajani, N. F. (2020). BERTology meets biology: Interpreting attention in protein language models. In *arXiv [cs.CL]*.

https://doi.org/10.48550/ARXIV.2006.15222

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., & Barton, G. J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics (Oxford, England)*, *25*(9), 1189–1191. https://doi.org/10.1093/bioinformatics/btp033

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Hanikel, N., Pellock, S. J., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I., Torres, S. V., … Baker, D. (2022). Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. In *bioRxiv*. https://doi.org/10.1101/2022.12.09.519842

Xiong, Y., Tian, C., Zhu, J., Zhang, S., Wang, X., Chen, W., Han, Y., Du, Y., Wu, Z., & Zhang, K. (2024). Dynamic changes of starch properties, sweetness, and β-amylases during the development of sweet potato storage roots. *Food Bioscience*, *61*(104964), 104964. https://doi.org/10.1016/j.fbio.2024.104964

Zhang, Q., Han, Y., & Xiao, H. (2017). Microbial α-amylase: A biomolecular overview. *Process Biochemistry (Barking, London, England)*, *53*, 88–101. https://doi.org/10.1016/j.procbio.2016.11.012

Zhang, Z., Fan, H., Yu, Z., Luo, X., Zhao, J., Wang, N., & Li, Z. (2024). Metagenomics-based gene exploration and biochemical characterization of novel glucoamylases and α-amylases in Daqu and Pu-erh tea microorganisms. *International Journal of Biological Macromolecules*, *278*(Pt 1), 134182. https://doi.org/10.1016/j.ijbiomac.2024.134182

MATERIAL SUPLEMENTARIO

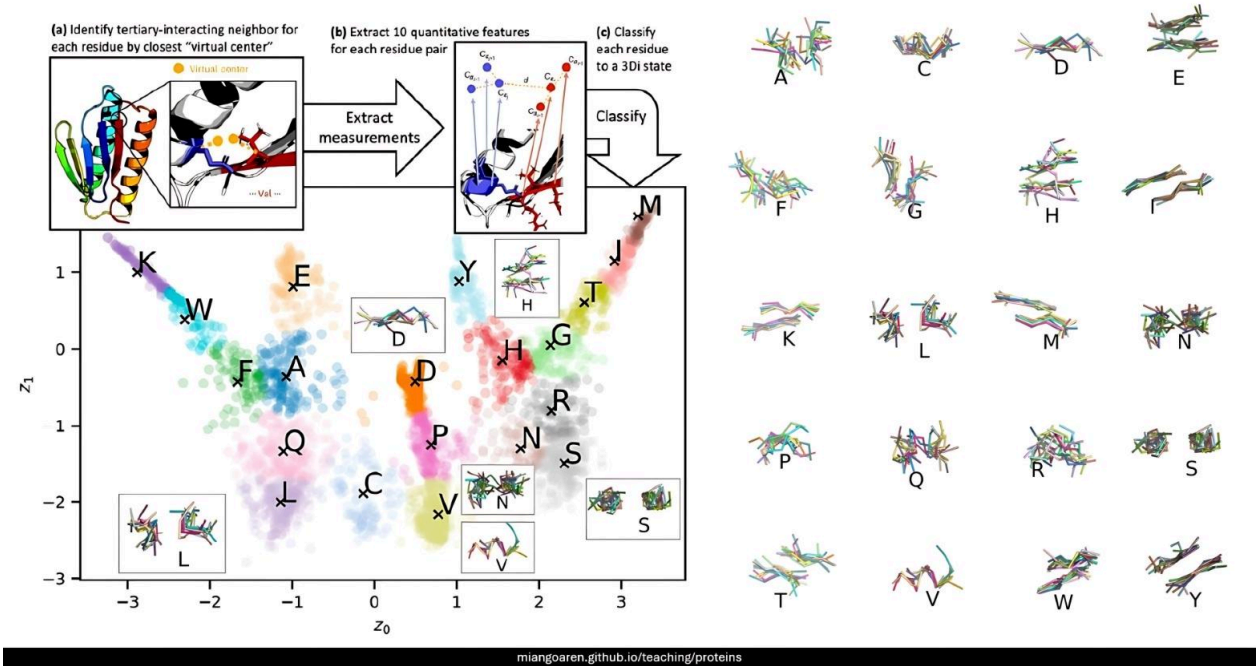

Figura S1. *Representación de los 20 estados de símbolos 3Di*

Fuente: (Puente-Lelievre et al., 2023; Van Kempen et al., 2023)

Nota: Sistema de Foldseek para clasificar interacciones terciarias en el alfabeto 3Di de 20 estados. (a) Para cada residuo de una proteínas, se calcula un centro virtual y se asigna como vecino el residuo cuyo centro virtual esté más próximo. (b) Se extraen 10 características numéricas que describen la interacción entre cada residuo y su vecino. © Estas características se clasifican en uno de los 20 estados 2Di utilizando un modelo pre entrenado desarrollado por Van Kempen et al., 2023. El panel principal ilustra los 20 estados en una proyección bidimensional, junto con ejemplos que muestran 10 fragmentos estructurales representativos para cada estado.

Tabla suplementaria 1.

Software empleado para el procesamiento, análisis y visualización de los datos.

Nombre	Descripción de uso	versión
ChimeraX	Visualización de proteínas	v1.10

MMseqs2	Agrupación y filtrado de secuencias por identidad	v15.6f466
Foldseek	Búsqueda de similitud estructural entre proteínas comparando sus estructuras 3D.	v9.33b118b
MUSCLE	Alineamientos múltiples de secuencias	v5.1
TrimmAl	Eliminación de columnas con *gaps* en alineamientos	v1.4.rev22
Jalview	Visualización y edición de alineamientos	v2.11.3.0
Python	Manipulación, procesamiento y visualización de datos: Torch, Numpy, SciKit-Learn, Pandas, MatplotLib, Seaborn, Os, Plotly, Mamba, Jupyter.	v3.11
ESMC	Modelo de generación de embeddings	ESMC-300M, v1.0.0