TiES
Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

número
15

Congreso Internacional de SUPERCÓMPUTO

El Supercómputo en los Tiempos de la Inteligencia Artificial 1

Mayo 2026 • DOI: 10.22201/dgtic.26832968e.2026.15 • ISSN: 2683-2968.

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

Índice

TIES, núm 15, mayo 2026, https://doi.org/10.22201/dgtic.26832968e.2026.15

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

Congreso Internacional de supercómputo 2025, volumen 1.

En un contexto donde la inteligencia artificial y el análisis de grandes volúmenes de datos adquieren cada vez más relevancia, el supercómputo se ha consolidado como una herramienta clave para la investigación científica. Bajo esta perspectiva, el Congreso Internacional de Supercómputo 2025, organizado por la UNAM bajo el lema "El supercómputo en los tiempos de la Inteligencia Artificial", destacó la estrecha relación entre ambas tecnologías y su capacidad para potenciar el modelado de sistemas complejos, optimizar procesos computacionales y fortalecer la innovación científica con el impacto social que ello conlleva. En este espacio, se expusieron los resultados de investigaciones de vanguardia obtenidos mediante la integración de diversas herramientas y metodologías orientadas a mejorar el desempeño de las infraestructuras de supercómputo, particularmente aquéllas basadas en inteligencia artificial.

En el presente número de la revista TIES, damos a conocer un primer volumen con cinco artículos que dan fe de lo ocurrido en el congreso, mostrando cómo el cómputo de alto desempeño y la inteligencia artificial se articulan como herramientas fundamentales para enfrentar desafíos ambientales, biomédicos y tecnológicos desde las instituciones de educación superior.

Estas tecnologías, más allá de representar únicamente capacidad de procesamiento, permiten enfrentar problemas complejos a través del análisis de grandes volúmenes de datos, la simulación avanzada y la construcción de modelos predictivos de fenómenos ambientales, de salud, biología computacional y diseño de fármacos.

El número inicia con el artículo "Caracterización de tormentas severas usando imágenes satelitales GOES-16 y procesamiento en paralelo", escrito por Jimena Ortiz Villalva, Erika Danaé López Espinoza, Dulce Rosario Herrera Moro, Jorge Zavala Hidalgo y Juan Manuel De Santiago Rodríguez. Éste aborda uno de los retos más relevantes para el monitoreo atmosférico y la prevención de riesgos hidrometeorológicos. A partir del uso de imágenes satelitales y esquemas de procesamiento paralelo implementados en el clúster Ometeotl, los

autores desarrollan un modelo capaz de optimizar el análisis de datos meteorológicos en tiempo casi real.

En esta misma línea, los autores Kevin Douglas Álvarez Segales, Alejandro Mendoza Reséndiz y Moisés Berezowsky Verduzco presentan, en el artículo "Aceleración de simulaciones fluviales computacionalmente intensivas mediante aprendizaje automático", un análisis de sistemas complejos a partir de un modelo híbrido que integra simulación numérica y redes neuronales para reducir los tiempos de procesamiento asociados a fenómenos hidromorfodinámicos. Esta propuesta plantea nuevas posibilidades para la exploración de escenarios de erosión y sedimentación.

La relación entre supercómputo, inteligencia artificial y ciencias biológicas adquiere una dimensión relevante en el artículo "Aplicación de inteligencia artificial en genómica y metagenómica viral para la clasificación taxonómica y el descubrimiento de nuevas proteínas virales", de los autores Blanca Itzel Taboada Ramírez, Lorena Díaz González, Oscar Alejandro Uscanga Junco, Alida Esmeralda Zárate Jiménez y Edna Cruz-Flores, quienes exploran el uso de la inteligencia artificial para el procesamiento y clasificación de millones de secuencias generadas mediante metagenómica viral.

En el campo biomédico, el artículo "Desde la quinolina hasta el Alzheimer. Diseño computacional de fármacos multifuncionales usando cómputo de alto desempeño", de Luis Felipe Hernández Ayala, Eduardo Gabriel López Guzmán, Mario Prejanò, Tiziana Marino y Annia Galano, expone el potencial del diseño racional asistido por computadora para acelerar la búsqueda de tratamientos dirigidos a enfermedades complejas. Mediante el uso de los clústeres Miztli y Yoltla, los autores desarrollan un flujo de trabajo que permite explorar miles de compuestos químicos y priorizar candidatos farmacológicos viables. Este enfoque muestra cómo el supercómputo contribuye a reducir tiempos de investigación y fortalece la capacidad de las universidades para generar conocimiento aplicado con impacto potencial en salud pública.

Finalmente, los autores Blanca Itzel Taboada Ramírez, Lorena Díaz-González, Oscar Alejandro Uscanga Junco, Alida Esmeralda Zárate Jiménez y Edna Cruz-Flores cierran este volumen con "VirDetect-AI: nueva herramienta de inteligencia artificial para identificar proteínas virales eucariontes en datos metagenómicos", artículo que presenta una arquitectura basada en redes neuronales convolucionales capaz de clasificar proteínas virales, con altos niveles de precisión y velocidad, para el análisis automatizado de secuencias

metagenómicas, confirmando el potencial de la inteligencia artificial como apoyo para la investigación biológica avanzada.

En conjunto, este número 15 de la Revista TIES destaca el papel estratégico que desempeñan el supercómputo y la inteligencia artificial en la generación de conocimiento científico y su impacto social. Más allá de una dimensión técnica, los trabajos aquí presentados muestran cómo estas tecnologías permiten construir nuevas rutas para abordar problemas complejos y evidencian la necesidad de fortalecer, en las instituciones de educación superior, la formación especializada y el acceso a infraestructuras avanzadas que les permitan consolidarse como espacios de vanguardia en innovación científica y tecnológica.

Se demuestra la necesidad de fortalecer, en las instituciones de educación superior, la formación especializada y el acceso a infraestructuras avanzadas que les permitan consolidarse como espacios de vanguardia en innovación científica y tecnológica.

Dr. Héctor Benítez Pérez

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

Caracterización de tormentas severas usando imágenes satelitales GOES-16 y procesamiento en paralelo

Jimena Ortiz Villalva
Universidad Nacional Autónoma de México, Posgrado en Ciencias de la Tierra,
Ciudad de México, México.
ORCID: 0009-0007-8028-3192

Erika Danaé López Espinoza
Universidad Nacional Autónoma de México, Instituto de Ciencias de la Atmósfera y Cambio Climático,
Ciudad de México, México.
ORCID: 0000-0003-2387-0335

Dulce Rosario Herrera Moro
Universidad Nacional Autónoma de México, Instituto de Ciencias de la Atmósfera y Cambio Climático,
Ciudad de México, México.
ORCID: 0000-0002-7889-154X

Jorge Zavala Hidalgo
Universidad Nacional Autónoma de México, Instituto de Ciencias de la Atmósfera y Cambio Climático,
Ciudad de México, México.
ORCID: 0000-0002-2737-434X

Juan Manuel De Santiago Rodríguez
Universidad Nacional Autónoma de México, Escuela Nacional de Ciencias de la Tierra,
Ciudad de México, México.
ORCID: 0009-0002-2460-774X

Recepción: 26 de febrero de 2026.
Aceptación: 17 de abril de 2026.

Mayo 2026 • número de revista 15 • DOI: 10.22201/dgtic.26832968e.2026.15.146

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

Caracterización de tormentas severas usando imágenes satelitales GOES-16 y procesamiento en paralelo

Resumen

Las tormentas severas asociadas a convección profunda representan un riesgo importante en regiones urbanas como la Ciudad de México por su potencial para generar lluvia intensa, granizo y actividad eléctrica, fenómenos vinculados a inundaciones y daños urbanos. Este trabajo desarrolla un flujo computacional para la caracterización de tormentas severas a partir de observaciones del satélite GOES-16, integrando variables térmicas y eléctricas de los sensores *Advanced Baseline Imager* (ABI) y *Geostationary Lightning Mapper* (GLM). El análisis se aplicó a 169 eventos de tormenta severa ocurridos entre 2020 y 2023, utilizando esquemas de procesamiento en paralelo en infraestructura de cómputo de alto rendimiento. La metodología se basa en la estimación de la temperatura de brillo del tope de nube y la actividad eléctrica de las tormentas severas, con el fin de identificar los umbrales del ciclo de vida convectivo. Los umbrales estimados son: −33°C, asociado al inicio de la actividad eléctrica, y −55°C, vinculado con la transición hacia la fase madura de desarrollo. Asimismo, se implementó una adaptación regional de la composición RGB, desarrollada previamente para latitudes medias, permitiendo mejorar la identificación de la convección profunda para la Ciudad de México. Los productos resultantes operan actualmente para el monitoreo de tormentas severas en tiempo casi real.

Palabras Clave: tormentas severas, procesamiento paralelo, imágenes GOES-16, ciclo de vida convectivo, convección profunda.

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

Characterization of Severe Storms Using GOES-16 Satellite Imagery and Parallel Processing

Abstract

Severe storms associated with deep convection pose a significant hazard in urban regions such as Mexico City due to their potential to produce intense rainfall, hail, and electrical activity, phenomena linked to flooding and urban damage. This study develops a computational workflow for the characterization of severe storms using observations from the GOES-16 satellite, integrating thermal and electrical variables from the Advanced Baseline Imager (ABI) and the Geostationary Lightning Mapper (GLM). The analysis was applied to 169 severe storm events that occurred between 2020 and 2023, employing parallel processing schemes within a high-performance computing infrastructure.

The methodology is based on the estimation of cloud-top brightness temperature and storm electrical activity to identify thresholds associated with the convective life cycle. An estimated threshold of −33 °C is associated with the onset of electrical activity, while a threshold of −55 °C is linked to the transition toward the mature development phase. Additionally, a regional adaptation of an RGB composite originally developed for midlatitudes was implemented, improving the identification of deep convection under local atmospheric conditions for Mexico City. The resulting products are currently operating for near–real-time monitoring of severe storms.

Keywords: *severe storms, parallel processing, GOES-16 imagery, convective life cycle, deep convection.*

Introducción

Las tormentas severas asociadas a convección profunda representan un riesgo significativo en regiones urbanas densamente pobladas debido a su potencial para producir precipitación intensa, granizo, rachas de viento y elevada actividad eléctrica [1]-[5]. En la Ciudad de México (CDMX), este riesgo se ve amplificado por la alta concentración de la población e

infraestructura, así como por la limitada capacidad de infiltración del suelo urbano, lo que favorece inundaciones recurrentes durante la temporada de lluvias.

El seguimiento de la convección profunda se ha beneficiado del uso de satélites geoestacionarios, que proporcionan observaciones continuas con alta resolución temporal [10]. En particular, el satélite GOES-16 incorpora el sensor ABI, que permite caracterizar propiedades térmicas y espectrales de las nubes, y el GLM, que observa la actividad eléctrica total asociada a sistemas convectivos [11] y [14]. La temperatura de brillo en el infrarrojo, especialmente en torno a 10.3 µm, se ha utilizado ampliamente como indicador de la altura del tope de nube y de la intensidad de las corrientes ascendentes [4], [5], mientras que la actividad eléctrica refleja los procesos microfísicos y dinámicos internos.

Diversos estudios han mostrado que incrementos abruptos en la tasa de descargas eléctricas, conocidos como lightning jumps, pueden preceder a la intensificación convectiva y a la ocurrencia de fenómenos hidrometeorológicos severos [12], [13]. Sin embargo, muchos de los criterios operativos y umbrales reportados han sido desarrollados para latitudes medias, por lo que su aplicación en regiones tropicales y subtropicales no es óptima y resulta necesario estimarlos para estas regiones [15].

El análisis conjunto de variables térmicas y eléctricas, a partir de los sensores ABI y GLM, ofrece una vía para caracterizar de forma más completa el ciclo de vida convectivo de las tormentas severas. No obstante, la alta resolución temporal y espacial de los productos satelitales generan grandes volúmenes de datos para ser procesados cuando se analizan múltiples eventos, imponiendo retos computacionales para un procesamiento sistemático.

En este trabajo, se desarrolló e implementó un flujo computacional para la caracterización de tormentas severas a partir de datos ABI y GLM del satélite GOES-16. Se caracterizaron 169 eventos ocurridos en la región centro-oriental de México, con énfasis en los que afectaron a la CDMX. El análisis se centró en la estimación de umbrales regionales térmicos y eléctricos, asociados a distintas etapas del ciclo de vida convectivo de las tormentas severas; éste destacó el papel del procesamiento en paralelo como la herramienta para la reproducibilidad, el escalamiento y el análisis. Todo lo anterior tuvo el objetivo de tener una herramienta operativa y un visualizador, en tiempo casi real, del desarrollo y ciclo de vida de las tormentas severas.

Desarrollo

Zona de estudio

La zona de estudio comprende la Ciudad de México y sus alrededores, definida por un dominio geográfico rectangular entre 18.24° y 20.66° N de latitud, así como 100.19° y 97.58° O de longitud. Este dominio, en adelante denominado CDMX, cubre una superficie aproximada de 73,170 km² y permite capturar tanto los sistemas convectivos que se desarrollan sobre la ciudad, como aquellos que la afectan desde regiones circundantes.

Base de datos y selección de eventos

Se analizaron 169 eventos de tormenta severa ocurridos entre 2020 y 2023. Se emplearon datos satelitales GOES-16 de los sensores ABI y GLM en formato NetCDF, obtenidos del repositorio de Datos Abiertos en el *bucket* S3 de Amazon Web Services (AWS) [16].

Del sensor ABI, se utilizó el producto Cloud and Moisture Imagery (CMI), considerando ocho bandas (B02, B05, B07, B08, B10, B11, B13 y B15) seleccionadas por su relevancia para el análisis de la convección profunda y la construcción de composiciones RGB [11]. Del sensor GLM, se emplearon los productos de eventos, grupos y destellos para caracterizar la evolución de la actividad eléctrica [14].

Flujo computacional y procesamiento en paralelo

La combinación de múltiples bandas ABI con una alta frecuencia temporal de GLM implicó el manejo de un volumen de datos del orden de 10^6 archivos para el conjunto de eventos de tormenta severa analizados. Para abordar este reto, se implementó, en el clúster Ometeotl del Instituto de Ciencias de la Atmósfera y Cambio Climático de la UNAM, un flujo automatizado de descarga, reproyección, recorte y generación de productos (imágenes RGB).

Los archivos con datos ABI se proyectaron desde la proyección geoestacionaria nativa a una grilla regular en coordenadas geográficas y, posteriormente, se recortaron al dominio de la CDMX.

Para la descarga de datos, se ejecutaron cuatro trabajos de descarga en el administrador de trabajos Slurm. No se ejecutaron más trabajos de descarga para evitar la saturación del ancho de banda.

La proyección de la imágen se realizó con la función ImageContainerNeares de la biblioteca especializada en el manejo de datos geoespaciales y satelitales Pyresample, la cual dividió la imagen en bloques e hizo que se procesarán de forma independiente y paralela. Para reproyectar y recortar cada banda al dominio CDMX, se utilizaron 20 núcleos de un nodo de procesamiento del clúster Ometeotl. Los datos reproyectados y recortados se almacenaron en un archivo en formato NetCDF.

La generación de los productos satelitales se hizo con un programa que originalmente era serial, el cual leía las imágenes reproyectadas con sus diferentes bandas para cada hora de cada evento de estudio. Considerando que la generación de una imagen RGB es independiente de la otra, se utilizó la biblioteca Multiprocessing de Python, con la cual se refactorizó el programa serial para generar de manera independiente y paralela cada imagen RGB horaria haciendo uso de un núcleo del nodo de procesamiento. Con ello, se obtuvo un programa en paralelo que redujo el tiempo de procesamiento en aproximadamente un 88.6%, aprovechando los 44 núcleos del nodo asignado. La paralelización fue esencial porque, dependiendo del caso de estudio, se podían procesar hasta 2,880 imágenes y los eventos de tormenta severa analizados fueron 169 casos.

Integración térmica y eléctrica

El análisis se basó en la temperatura de brillo del tope de nube obtenida a partir de la banda 13 (10.3 µm) del sensor ABI, utilizada como referencia térmica principal. Para cada imagen infrarroja se identificaron las descargas eléctricas registradas por el sensor GLM dentro del dominio CDMX y se asoció a cada descarga la temperatura de brillo correspondiente. Esta integración permitió construir series temporales y distribuciones estadísticas que describen la relación entre el enfriamiento del tope de nube y la evolución de la actividad eléctrica [12]-[15].

A partir de este análisis conjunto, se identificaron umbrales regionales térmicos asociados a la convección profunda. En particular, se observó un umbral cercano a −33°C,

asociado al inicio de la actividad eléctrica, y un segundo umbral alrededor de −55°C, vinculado con un incremento marcado en dicha actividad y con la transición hacia la fase madura de la tormenta severa.

Productos satelitales y caracterización convectiva

Con base en los umbrales térmicos identificados, se desarrolló un producto combinado ABI–GLM, en el que la temperatura de brillo del tope de nube se representa de manera continua y la actividad eléctrica se superpone espacialmente. Este producto permite seguir la evolución espacio-temporal de los sistemas convectivos e identificar transiciones entre las fases de desarrollo, madurez y disipación de la tormenta severa.

De manera complementaria, se implementaron composiciones RGB orientadas a la caracterización de la convección. La RGB de convección diurna se utilizó para clasificar la intensidad convectiva, mientras que la RGB de fase de nubes fue ajustada regionalmente para adecuarla a las condiciones atmosféricas y radiativas de la región centro-oriental de México, siguiendo criterios similares a los reportados en la literatura [11] y [15]. Este ajuste permitió mejorar la identificación de nubes profundas con topes de hielo de manera más consistente y reproducible para contextos operativos regionales.

Para el análisis nocturno, se empleó la composición RGB de la microfísica nocturna, lo que permitió extender el seguimiento del ciclo convectivo a periodos sin información visible. Además, la evolución temporal de la actividad eléctrica se analizó mediante el cálculo de tasas de cambio en el número de descargas, permitiendo identificar incrementos abruptos asociados a procesos de intensificación convectiva.

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

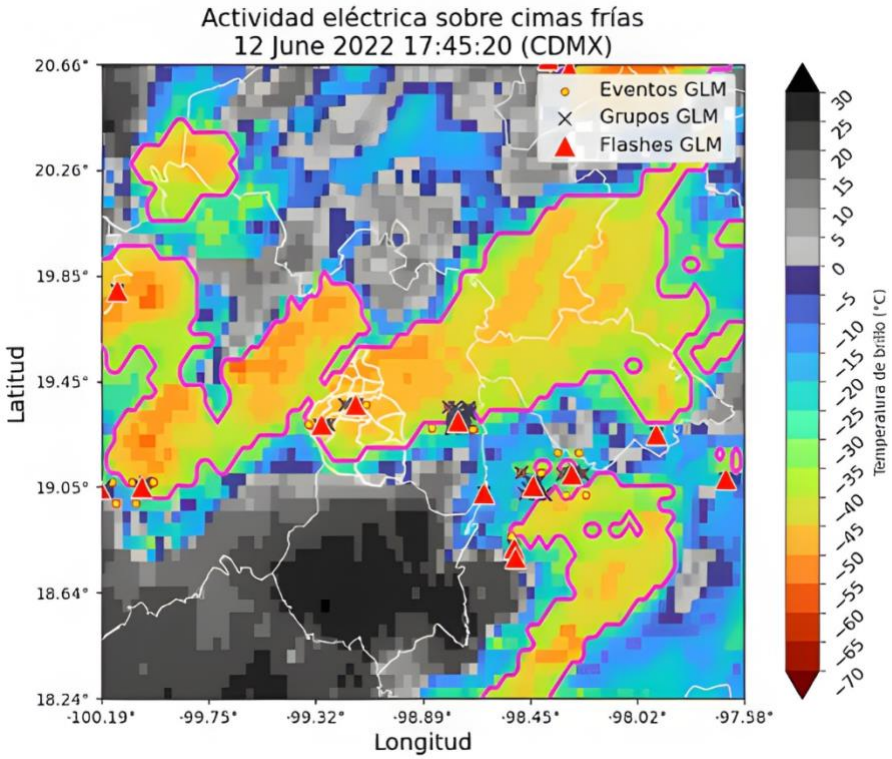


Fig.1. Temperatura de brillo del tope de nube (banda 13 del ABI) y actividad eléctrica del GLM durante la tormenta del 12 de junio de 2022 a las 17:45:20 h (hora local), poco antes de su punto más intenso. La escala muestra la temperatura (°C); los contornos en magenta indican la isoterma de −33 °C. Los símbolos señalan los eventos, grupos y flashes detectados dentro del área de estudio.

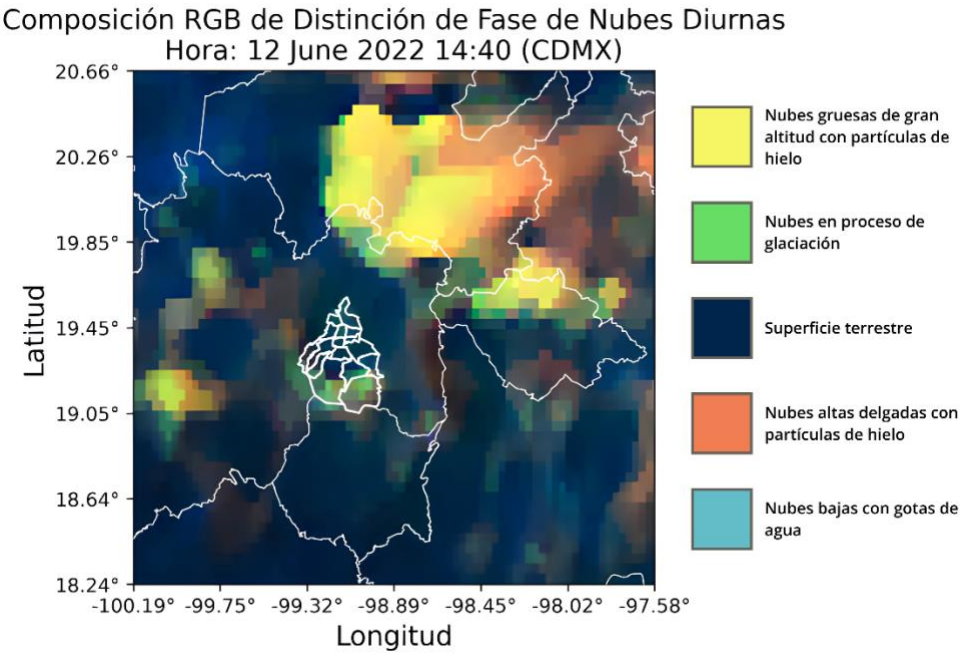


Fig. 2. Composición RGB de convección diurna, versión ajustada para el Valle de México, correspondiente al 12 de junio de 2022. Los colores permiten distinguir la fase de las cimas nubosas a partir de umbrales definidos para esta región.

Umbrales regionales operativos del ciclo de vida convectivo

El análisis integrado permitió definir tres rangos térmicos representativos del ciclo de vida convectivo de las tormentas severas. La etapa de desarrollo se asoció a temperaturas de tope de nube cercanas a −33°C, coincidiendo con el inicio de la actividad eléctrica. La transición hacia la fase madura se observó alrededor de −55°C, acompañada por un incremento abrupto en la actividad eléctrica y por señales espectrales consistentes con convección profunda. Finalmente, temperaturas cercanas o inferiores a −70°C se asociaron al inicio de la disipación, caracterizada por una disminución rápida de la actividad eléctrica. En la Tabla 1, se muestran los diferentes umbrales regionales estimados para diferentes variables que describen cada una de las etapas del ciclo de vida convectivo de las tormentas severas en la Ciudad de México.

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

I. Tabla

Características estimadas del ciclo de vida convectivo de las tormentas severas desarrolladas en la Ciudad de México para el periodo de 2020 a 2023

Etapa del ciclo de vida	Variable	Fuente	Rango numérico	Criterio de clasificación	Interpretación física
Desarrollo	Temperatura de tope de nube (Tb)	ABI Banda 13	**Tb ≤ - 33°C**	Umbral térmico	El tope de la nube comienza a enfriarse, indicando crecimiento vertical inicial
	Green (G) normalizado	RGB Convección	**0.27-0.45**	Convección débil	Señal RGB asociada a nubes un poco profundas
	ΔT (G)	RGB Convección	**-43 a -31°C**		Diferencias térmicas pequeñas, típicas de etapas iniciales
	Red (R) normalizado	RGB Convección	**.74 – 0.85**		Estructura térmica poco contrastada en la cima
	ΔT (R)	RGB Convección	**-5 a -1°C**		Señal térmica débil asociada al inicio del desarrollo
Madurez	Temperatura de tope de nube (Tb)	ABI Banda 13	**-55°C < Tb ≤ - 33°C**	Umbral térmico	El tope alcanza temperauras muy frías, asociadas a la máxima altura.
	Green (G) normalizado	RGB Convección	**0.47 – 0.63**	Convección moderada	Señal RGB intensa, asociada a convección activa

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

Etapa del ciclo de vida	Variable	Fuente	Rango numérico	Criterio de clasificación	Interpretación física
	ΔT (G)	RGB Convección	**-30 a -19°C**		Incremento del contraste térmico en la cima
	Red (R) normalizado	RGB Convección	**0.74 – 0.87**		Núcleo convectivo bien definido
	ΔT (R)	RGB Convección	**-5 a 0°C**		Mayor contraste térmico respecto al desarrollo.
	Green (G) normalizado	RGB Convección	**0.64 – 0.90**	Convección fuerte	Señal RGB muy intensa, asociada a nubes profundas
	ΔT (G)	RGB Convección	**-18 a -2°C**		Grandes diferencias térmicas, indicativas de convección profunda
	Red (R) normalizado	RGB Convección	**0.76 – 0.89**		Núcleo convectivo muy bien definido
	ΔT (R)	RGB Convección	**-5 a +1°C**		Máximo contraste térmico observado
Disipación	Temperatura de tope de nube (Tb)	ABI Banda 13	**Tb ≤ - 70°C**	Umbral térmico	Persisten topes fríos, pero la nube deja de intensificarse
	Green (G) normalizado	RGB Convección	**0.27 – 0.45**	Convección débil	Señal RGB débil asociada a nubes remanentes

Etapa del ciclo de vida	Variable	Fuente	Rango numérico	Criterio de clasificación	Interpretación física
	ΔT (G)	RGB Convección	**-43 a 31°C**		Reducción del contraste térmico respecto a la fase madura
	Red (R) normalizado	RGB Convección	**0.74 – 0.85**		Estructura térmica menos definida
	ΔT (R)	RGB Convección	**-5 a -1°C**		Señal térmica residual durante la disipación

Conclusiones

Este trabajo presentó una caracterización de nubes convectivas profundas asociadas a tormentas severas en la Ciudad de México a partir de imágenes satelitales del GOES-16, integrando variables térmicas y eléctricas derivadas de los sensores ABI y GLM. Al revisar la evolución de los casos se observó que los criterios definidos para latitudes medias no pueden aplicarse tal cual en la Ciudad de México. Las condiciones atmosféricas no son las mismas y eso se refleja en los valores de temperatura. Las recetas RGB desarrolladas para latitudes medias consideran topes de nube más fríos que los que normalmente se registran en nuestra región. Aquí los valores suelen ser más cálidos, y esa diferencia modifica los umbrales con los que se interpretan las distintas etapas de la tormenta.

Por ello fue necesario ajustar las recetas y trabajar con valores acordes a la región. El análisis permitió identificar temperaturas que se repiten a lo largo de los casos estudiados. En varios eventos, valores cercanos a −33 °C coinciden con el inicio de la actividad eléctrica. Alrededor de −55 °C se observa el paso hacia la fase madura, momento en el que los incrementos de actividad eléctrica son más notorios.Al integrar estos valores con productos combinados ABI–GLM y con composiciones RGB adaptadas a la región, se obtiene una base práctica para el seguimiento de la convección profunda en la Ciudad de México.

Los resultados de esta investigación se pueden consultar en el sitio https://pronosticos.atmosfera.unam.mx/operativo/index.php/sevstorm

Agradecimiento

Esta investigación ha sido apoyada por la Secretaría de Educación, Ciencia, Tecnología e Innovación (SECTEI) de la Ciudad de México bajo el proyecto "Sistema piloto de alertamientos hidrometeorológicos para la Ciudad de México" - SECTEI/145/2024. Asimismo, contó con el apoyo de la Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI), mediante la beca otorgada para la realización de estudios de maestría, registrada con CVU No. 1343182. Estos apoyos hicieron posible dedicar tiempo, atención y continuidad al desarrollo de este trabajo.

Referencias

[1] H. R. Byers and R. R. Braham, *The Thunderstorm*. Washington, DC, USA: U.S. Government Printing Office, 1949.

[2] C. A. Doswell III, "Severe convective storms—An overview," in *Severe Convective Storms*, Meteorological Monographs, vol. 28, no. 50. Boston, MA, USA: American Meteorological Society, 2001, pp. 1–26.

[3] J. M. Wallace and P. V. Hobbs, *Atmospheric Science: An Introductory Survey*, 2nd ed. Amsterdam, The Netherlands: Elsevier, 2006, doi: 10.1016/C2009-0-00034-8.

[4] W. R. Cotton, G. H. Bryan, and S. C. van den Heever, "Cumulonimbus clouds and severe convective storms," in *Storm and Cloud Dynamics*, Int. Geophys., vol. 99, pp. 315–454, 2011, doi: 10.1016/S0074-6142(10)09914-6.

[5] R. A. Houze Jr., "Cumulonimbus and severe storms," in *Cloud Dynamics*, 2nd ed., Int. Geophys., vol. 104. Oxford, U.K.: Academic Press, 2014, pp. 187–236, doi: 10.1016/B978-0-12-374266-7.00008-1.

[6] Instituto Nacional de Estadística y Geografía (INEGI), *Censo de Población y Vivienda 2020*. Mexico City, Mexico, 2020. [Online]. Disponible: http://www.inegi.org.mx

[7] Secretaría de Desarrollo Agrario, Territorial y Urbano (SEDATU), Consejo Nacional de Población (CONAPO), and Instituto Nacional de Estadística y Geografía (INEGI), *Delimitación de las zonas metropolitanas de México 2015*. Mexico City, Mexico, 2018. [Online]. Disponible: https://www.gob.mx/sedatu

[8] Y. Chen and A. M. Bilton, "Water stress, peri-urbanization, and community-based water systems: A reflective commentary on the metropolitan area of Mexico City," *Frontiers in Sustainable Cities*, vol. 4, p. 790633, 2022, doi: 10.3389/frsc.2022.790633.

[9] J. M. Montero-Martínez and M. Andrade-Velázquez, "Effects of urbanization on extreme climate indices in the Valley of Mexico Basin," *Atmosphere*, vol. 13, no. 5, p. 785, 2022, doi: 10.3390/atmos13050785.

[10] S. Q. Kidder and T. H. Vonder Haar, *Satellite Meteorology: An Introduction*. San Diego, CA, USA: Academic Press, 1995.

[11] T. J. Schmit *et al.*, "A closer look at the ABI on the GOES-R series," *Bull. Amer. Meteor. Soc.,* vol. 98, no. 4, pp. 681–698, Apr. 2017, doi: 10.1175/BAMS-D-15-00230.1.

[12] E. R. Williams *et al.*, "The behavior of total lightning activity in severe Florida thunderstorms," *Atmospheric Research*, vol. 51, no. 3–4, pp. 245–265, 1999.

[13] C. J. Schultz, W. A. Petersen, and L. D. Carey, "Lightning and severe weather: A comparison between total and cloud-to-ground lightning trends," *Weather and Forecasting*, vol. 26, no. 5, pp. 744–755, Oct. 2011, doi: 10.1175/WAF-D-10-05026.1.

[14] S. J. Goodman *et al*., "The GOES-R Geostationary Lightning Mapper (GLM)," *Atmospheric Research*, vol. 125–126, pp. 34–49, 2013, doi: 10.1016/j.atmosres.2013.01.006.

[15] K. C. Thiel, K. M. Calhoun, A. E. Reinhart, and D. R. MacGorman, "GLM and ABI characteristics of severe and convective storms," *J. Geophys. Res.: Atmospheres*, vol. 125, no. 17, p. e2020JD032858, 2020, doi: 10.1029/2020JD032858.

[16] National Oceanic and Atmospheric Administration, "GOES-16 Data on AWS," Amazon Web Services S3 Open Data Repository. [Online]. Available: https://noaa-goes16.s3.amazonaws.com/index.html

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

Aceleración de simulaciones fluviales computacionalmente intensivas mediante aprendizaje automático

Kevin Douglas Alvarez Segales
Universidad Nacional Autónoma de México, Instituto de Ingeniería, Ciudad de México, México.
ORCID: 0009-0009-2029-0547

Alejandro Mendoza Reséndiz
Universidad Nacional Autónoma de México, Instituto de Ingeniería, Ciudad de México, México.
ORCID: 0000-0002-2479-9799

Moisés Berezowsky Verduzco
Universidad Nacional Autónoma de México, Instituto de Ingeniería, Ciudad de México, México.
ORCID: 0000-0002-7675-3450

Recepción: 01 de marzo de 2026.
Aceptación: 22 de abril de 2026.

Mayo 2026 • número de revista 15 • DOI: 10.22201/dgtic.26832968e.2026.15.157

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

Aceleración de simulaciones fluviales computacionalmente intensivas mediante aprendizaje automático

Resumen

Predecir la evolución del lecho es clave para anticipar procesos de erosión y sedimentación que afectan la infraestructura. Aunque los modelos 2D resuelven la hidromorfodinámica mediante la ecuación de Exner, su costo computacional es elevado debido a restricciones de estabilidad, mallas finas y la sobrecarga del cómputo paralelo. Esta demanda técnica hace que la exploración de escenarios dependa habitualmente de plataformas de alto desempeño (HPC).

Se propone un esquema que combina simulación numérica y aprendizaje automático. Primero, se ejecutan modelos numéricos convencionales para generar bases de datos espacio-temporales. Luego, se entrenan redes neuronales tipo perceptrón multicapa (MLP) como modelos sustitutos, con el fin de predecir la evolución del fondo en pasos de tiempo sucesivos. El enfoque se evalúa con simulaciones 2D mediante un modelo de elementos finitos acoplado hidrodinámico-sedimentos, elaborado en TELEMAC-MASCARET. Los resultados muestran que el MLP reproduce la tendencia espacio-temporal del lecho y reduce de forma sustancial el tiempo de análisis, lo que habilita estudios iterativos y la exploración de escenarios. Se discuten las limitaciones por extrapolación y las recomendaciones para un uso robusto.

Palabras Clave: morfodinámica fluvial, transporte de sedimentos, modelación numérica, modelos sustitutos, perceptrón multicapa, TELEMAC–SISYPHE, aprendizaje de máquina.

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

Acceleration of computationally intensive fluvial simulations using machine learning

Abstract

Predicting riverbed evolution is essential in applied hydromorphology because it helps anticipate erosion and deposition processes that reshape channel geometry and affect hydraulic infrastructure. These dynamics are commonly simulated with 2D morphodynamic models that solve conservation equations for flow and sediment transport, including the Exner equation. However, computational cost can be high due to stability constraints, flow–sediment coupling, and, in 2D, mesh size and parallel communication overhead, motivating the use of high-performance computing (HPC) to explore scenarios more efficiently.

This paper proposes a hybrid workflow combining numerical simulation and machine learning: (i) running conventional high-fidelity models to generate spatiotemporal datasets and (ii) training multilayer perceptron (MLP) neural networks as surrogate models to predict bed evolution at successive time steps. The approach is evaluated using 2D simulations based on coupled hydrodynamics–sediment finite-element modeling in TELEMAC-MASCARET. Results indicate that the MLP reproduces the spatiotemporal trend of bed evolution while substantially reducing analysis time, enabling iterative studies. Limitations related to extrapolation are discussed, along with recommendations for robust operational use.

Keywords: *fluvial morphodynamics; sediment transport; numerical modeling; surrogate models; multilayer perceptron; TELEMAC–SISYPHE; machine learning.*

1 Introducción

La evolución del lecho fluvial surge de la interacción no lineal entre la hidrodinámica y el transporte de sedimentos, generando patrones de erosión y depósito que afectan la infraestructura hidráulica. En modelación clásica, este proceso se describe mediante las ecuaciones de flujo en lámina libre acopladas a la ecuación de Exner, la cual vincula la

divergencia del transporte con el cambio de elevación del fondo [1]. Para cerrar el sistema, las tasas de transporte se estiman mediante expresiones empíricas o semiempíricas ampliamente utilizadas (p. ej., Meyer-Peter y Müller; Engelund-Hansen; Van Rijn), derivadas de evidencia experimental y de campo [2], [3].

Este enfoque ofrece consistencia física, pero su alto costo computacional en mallas finas y simulaciones largas limita análisis iterativos (calibración e incertidumbre), incluso en sistemas HPC. Herramientas como TELEMAC-MASCARET y SISYPHE permiten modelar geometrías complejas y acoplamiento morfológico, pero mantienen una alta demanda de recursos en campañas extensivas [4], [5].

Los modelos de aprendizaje automático se han consolidado como una estrategia de aceleración al desplazar el costo computacional al entrenamiento, permitiendo inferencias rápidas. Específicamente, las redes MLP logran aproximar relaciones no lineales en procesos hidro-morfodinámicos [6], [7], [8]. Aplicaciones recientes confirman esta tendencia en la predicción de campos en ríos [9], los estudios que integran aprendizaje automático con modelación hidrodinámica para mejorar la gestión de recursos hídricos [10], las revisiones sobre predicción de caudales mediante inteligencia artificial [11] y los análisis de los desafíos y oportunidades del aprendizaje automático en hidrología de gran escala [12]. Este trabajo propone un flujo híbrido (modelos 2D y entrenamiento MLP) para emular la evolución del lecho, evaluando su precisión y ahorro computacional frente a métodos convencionales en HPC, así como sus riesgos de extrapolación.

2 Metodología y establecimiento del modelo

2.1 Modelación morfodinámica numérica de referencia

La evolución del lecho se modela mediante el acoplamiento entre la hidrodinámica de lámina libre y la continuidad de sedimentos. En 2D, se utiliza el sistema TELEMAC–SISYPHE sobre mallas no estructuradas, tomando sus salidas como referencia para entrenar modelos sustitutos [5].

$$\frac{\partial h}{\partial t}+\frac{\partial (uh)}{\partial x}=0 \qquad (1) \qquad \frac{\partial (uh)}{\partial x}+\frac{\partial}{\partial x}\left(uh^2+\frac{1}{2}gh^2\right)=-gh\frac{\partial z_b}{\partial x}-ghSf \qquad (2)$$

Donde h es la profundidad del agua, u es la velocidad en dirección del flujo, z_b es la elevación de fondo, y S_f es la velocidad de fricción. Estas ecuaciones corresponden a la aproximación bidimensional de las ecuaciones de Navier–Stokes para flujo a superficie libre, adoptadas siguiendo la formulación clásica descrita en [1] y [9]. La evolución de fondo, representada por la ecuación de Exner [1], se muestra a continuación.

$$\frac{\partial zb}{\partial t}+\left(\frac{1}{1-\gamma_p}\right)\frac{\partial qs}{\partial x}=0\ (1D) \qquad (3) \qquad \frac{\partial z}{\partial t}+\left(\frac{1}{1-p_0}\right)\left(\frac{\partial qx}{\partial x}+\frac{\partial qy}{\partial y}\right)=\frac{\partial (c\rho s)}{\partial t}(2D) \qquad (4)$$

Con γ_p/p_0 como la porosidad y q como el componente del flujo de sedimentos. En TELEMAC-SISYPHE, la hidrodinámica es resuelta con las ecuaciones de flujo superficial y de turbulencia de (κ–ε), con aplicación de Manning y transporte de sedimentos en su forma semi empírica [4].

2.2 Morfodinámica en dos dimensiones en TELEMAC

Tres casos de referencia bidimensionales a escala de laboratorio se utilizaron para generar los conjuntos de datos de entrenamiento y validación de la ANN: (i) un canal recto con un obstáculo localizado en el lecho, tomado de ejemplos de TELEMAC–MASCARET descritos por Hervouet [13]; (ii) una curva de 180° basada en Yen y Lee [14], también derivada de ejemplos de TELEMAC; y (iii) un canal meándrico basado en Moghaddassi *et al.* [15]. Los casos presentan una complejidad creciente, desde ajustes en canales rectos hasta redistribuciones por curvatura y corrientes secundarias en meandros. En todos los escenarios, el acoplamiento TELEMAC-2D/SISYPHE generó las series temporales utilizadas para el entrenamiento del modelo.

2.3 Conformación de datos de entrenamiento de la red neuronal

El conjunto de datos se genera exclusivamente a partir de simulaciones numéricas con el enfoque "t → t+1", donde el objetivo es predecir z_b^{t+1} a partir del estado hidráulico–morfodinámico en t. Para capturar la dependencia espacial local de la morfodinámica, se emplea un esquema de vecindad (esténcil o grafos) alrededor de cada nodo, con ventanas 37×37, apilando variables como profundidad h, velocidad u y elevación del lecho z_b en t para predecir z_b^{t+1} en el nodo central . Este diseño equivale a aprender un operador morfodinámico local coherente con el rol de los gradientes y divergencias de transporte que controlan erosión/deposición en 2D.

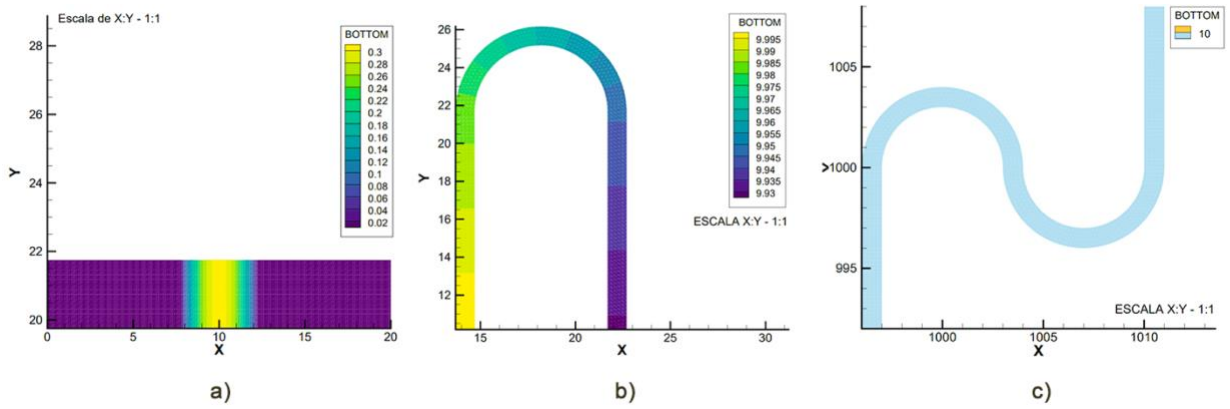


Fig. 1. Geometría inicial: a) canal recto con obstáculo, b) del canal con curva de 180° y c) del canal meándrico.

2.4 Combinaciones de entrada–salida de la RNA y reformulación adimensional

Una contribución clave es la construcción y evaluación sistemática de combinaciones de variables de entrada (C1–C7) y la reinterpretación de combinaciones bidimensionales previas de Kaveh *et al.* [4] en grupos adimensionales para mejorar la generalización física [5].

$$C4 = \begin{matrix}[\left(\frac{Z}{y_n}\right)_{i,j}^{t}, \left(\frac{Z}{y_n}\right)_{i-1,j}^{t}, \left(\frac{Z}{y_n}\right)_{i,j-1}^{t}, \left(\frac{U}{u_*}\right)_{i,j}^{t}, \left(\frac{U}{u_*}\right)_{i-1,j}^{t}, \left(\frac{U}{u_*}\right)_{i,j-1}^{t}, \left(\frac{H}{B}\right)_{i,j}^{t}, \left(\frac{H}{B}\right)_{i-1,j}^{t}, \left(\frac{H}{B}\right)_{i,j-1}^{t} \\ \left(\frac{Z}{y_n}\right)_{i,j}^{t-1}, \left(\frac{Z}{y_n}\right)_{i-1,j}^{t-1}, \left(\frac{Z}{y_n}\right)_{i,j-1}^{t-1}, \left(\frac{U}{u_*}\right)_{i,j}^{t-1}, \left(\frac{U}{u_*}\right)_{i-1,j}^{t-1}, \left(\frac{U}{u_*}\right)_{i,j-1}^{t-1}, \left(\frac{H}{B}\right)_{i,j}^{t-1}, \left(\frac{H}{B}\right)_{i-1,j}^{t-1}, \left(\frac{H}{B}\right)_{i,j-1}^{t-1}]\end{matrix} \quad (5)$$

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

$$C5 = \begin{matrix}[\left(\frac{Z}{y_n}\right)^{t}_{i,j}, \left(\frac{Z}{y_n}\right)^{t}_{i-1,j}, \left(\frac{Z}{y_n}\right)^{t}_{i,j-1}, \left(\frac{U}{u_*}\right)^{t}_{i,j}, \left(\frac{U}{u_*}\right)^{t}_{i-1,j}, \left(\frac{U}{u_*}\right)^{t}_{i,j-1}, \left(\frac{H}{B}\right)^{t}_{i,j}, \left(\frac{H}{B}\right)^{t}_{i-1,j}, \left(\frac{H}{B}\right)^{t}_{i,j-1}, \left(\frac{Z}{y_n}\right)^{t-1}_{i,j}, \\ \left(\frac{Z}{y_n}\right)^{t-1}_{i-1,j}, \left(\frac{Z}{y_n}\right)^{t-1}_{i,j-1}, \left(\frac{U}{u_*}\right)^{t+1}_{i,j}, \left(\frac{U}{u_*}\right)^{t+1}_{i-1,j}, \left(\frac{U}{u_*}\right)^{t+1}_{i,j-1}, \left(\frac{H}{B}\right)^{t-1}_{i,j}, \left(\frac{H}{B}\right)^{t-1}_{i-1,j}, \left(\frac{H}{B}\right)^{t-1}_{i,j-1}]\end{matrix} \quad (6)$$

2.5 Modelos MLP y protocolo de entrenamiento

Los modelos sustitutos se implementan mediante redes MLP, evaluando funciones de activación (ReLU, tanh, Radbas) y el optimizador AdamW. Para asegurar la estabilidad, se aplica normalización Min–Max y particiones de entrenamiento/validación/prueba. En 2D, se emplea una arquitectura compacta de dos capas ocultas que prioriza la velocidad de inferencia, trasladando el costo computacional a la generación de datos y al entrenamiento.

Formalmente, la red aprende una función no lineal f(x; W, b) que aproxima el operador morfodinámico local a partir del estado hidráulico–morfológico x. Para una red con L capas ocultas, la propagación hacia adelante se expresa como se muestra en la ecuación (5), donde $h^{(l)}$ denota la activación de la capa l, $W^{(l)}$ y $b^{(l)}$ son la matriz de pesos y el vector de sesgos asociados, y σ es la función de activación no lineal aplicada elemento a elemento [8].

$$h^{(l)} = \sigma(W^{(l)} h^{(l-1)} + b^{(l)}), \quad l = 1, \ldots, L \quad (7)$$

Los parámetros $\{W^{(l)}, b^{(l)}\}$ se ajustan minimizando la función de pérdida cuadrática media sobre el conjunto de entrenamiento, mediante descenso de gradiente estocástico con la variante AdamW.

2.6 Métricas de evaluación del modelo

La validación estadística del emulador se realiza comparando los valores predichos $\hat{y}_i$ con los de referencia y_i obtenidos de TELEMAC–SISYPHE. Se utilizan tres indicadores ampliamente aceptados en la modelación hidrológica [8]: el error cuadrático medio (RMSE), el error absoluto medio (MAE) y el coeficiente de determinación R^2. El RMSE penaliza errores grandes

en unidades físicas, el MAE mide el error promedio y R^2 la varianza explicada. Estos índices se evalúan en el conjunto de prueba y se complementan con mapas de diferencia espacial para identificar posibles sesgos locales.

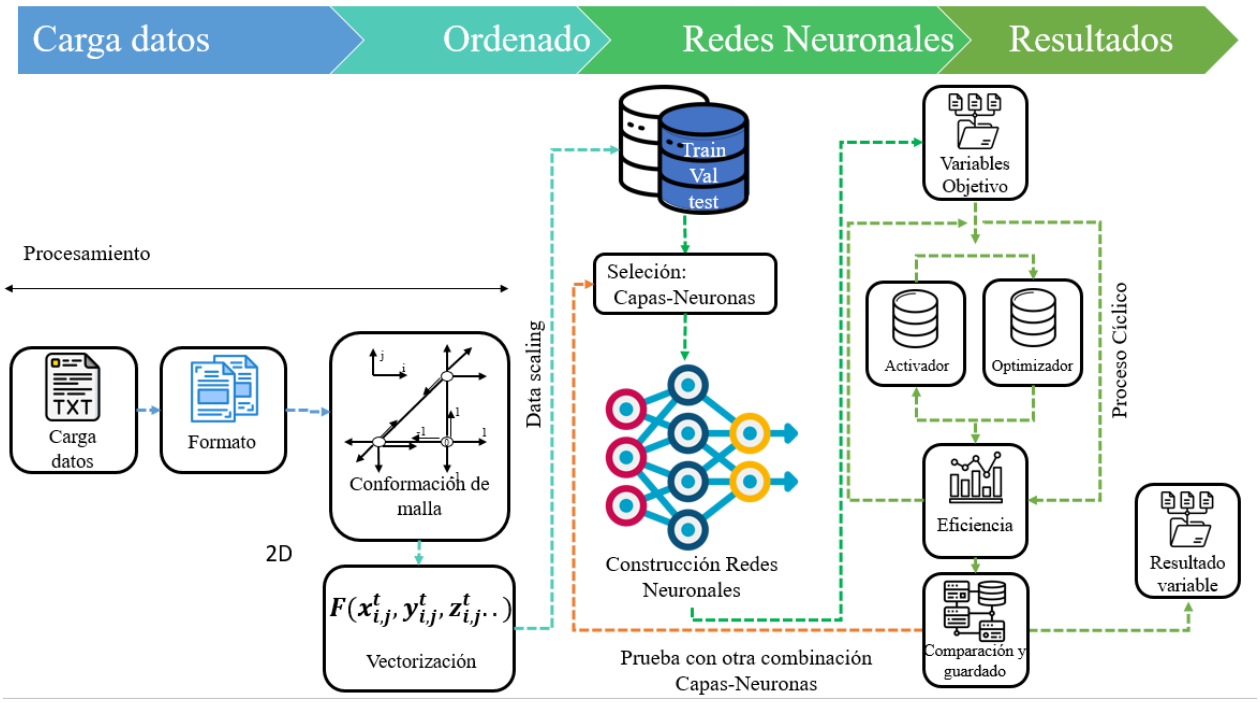


Fig. 2. Diagrama de flujo de red neuronal artificial para predicción de evolución en dos dimensiones.

Los resultados reportan que el caso del río meándrico de referencia 2D evaluó 7 combinaciones con diferentes funciones de activación y optimizadores (Tabla). La combinación con mejor desempeño (C4 y C5) utiliza la activación Radbas y el optimizador AdamW, alcanzando R = 0.999391 con RMSE = 8.249×10^{-4} en 22 épocas. El mejor desempeño estadístico no es proporcional a la velocidad de convergencia, lo que subraya la necesidad de comparar arquitecturas de forma controlada.

Tabla I. Resultados de evaluación de combinación C4 para caso de rio Meándrico

Activador	Optimizador	Épocas	R	MAE	RMSE	MAE
sigmoid	adamw	34	0.999315	0.004835	0.009995	0.001042

tanh	amsgrad	18	0.999364	0.003882	0.009453	0.000837
tanh	adam	20	0.999363	0.004788	0.009592	0.001032
relu	adam	15	0.99937	0.00478	0.009651	0.00103
relu	adamw	19	0.999365	0.004727	0.009757	0.001019
relu	amsgrad	38	0.999323	0.003911	0.009807	0.000843
relu	adamw	16	0.999354	0.005022	0.009833	0.001082
radbas	amsgrad	16	0.999303	0.00387	0.009892	0.000834
sigmoid	amsgrad	81	0.999311	0.004229	0.009947	0.000911
radbas	adamw	22	0.999391	0.003415	0.008249	0.000736
sigmoid	adam	24	0.999335	0.004791	0.010043	0.001033
radbas	adam	16	0.999266	0.004887	0.010809	0.001053
tanh	sgd	200	0.998573	0.007007	0.014454	0.00151
radbas	sgd	151	0.998388	0.006837	0.015086	0.001474
relu	sgd	200	0.998176	0.008059	0.016916	0.001737
sigmoid	sgd	200	0.990905	0.023056	0.037767	0.004969

3. Evaluación y resultados: precisión, patrones morfológicos y consistencia

En 2D, los mapas de elevación del lecho predichos reproducen los patrones de erosión y deposición obtenidos con TELEMAC–SISYPHE, incluyendo asimetrías asociadas a la geometría (p. ej., en curvatura). Por ejemplo, para el rio meándrico (último tiempo 432,000 s), el estudio reporta RMSE≈4.90×10^{-4}–4.91×10^{-4} m, con correlación cercana a 1, y diferencias espaciales típicamente del orden de milésimas [5]. Más allá del valor global de RMSE, el análisis de mapas de diferencia es crítico: permite verificar ausencia de sesgos sistemáticos (p. ej., sobre erosión en márgenes o depósito artificial), lo cual es el criterio práctico de consistencia morfológica frente al modelo físico.

Desde el punto de vista computacional, la MLP actúa como un acelerador del proceso morfológico al sustituir los cálculos iterativos del transporte de sedimentos por una

aproximación aprendida, reduciendo de forma importante los tiempos de simulación asociados a la morfodinámica del lecho.

En este trabajo, TELEMAC y el módulo SISYPHE actualizan el fondo cada 3Δt, dado que la evolución morfológica ocurre en escalas temporales más lentas que las hidráulicas. Una vez entrenada la red, cada actualización del campo morfológico puede obtenerse en segundos, con mayor eficiencia al ejecutar TensorFlow en GPU, frente a una simulación bidimensional convencional.

La MLP optimiza el cálculo morfodinámico mediante operaciones matriciales simples, manteniendo la consistencia con los resultados de TELEMAC sin requerir las iteraciones del modelo numérico. Según la Figura 4, el tiempo en CPU para el canal meándrico (1 h 17 min) disminuye drásticamente al aplicar la red neuronal, ya que el costo principal se desplaza al entrenamiento previo y no a la fase de evaluación.

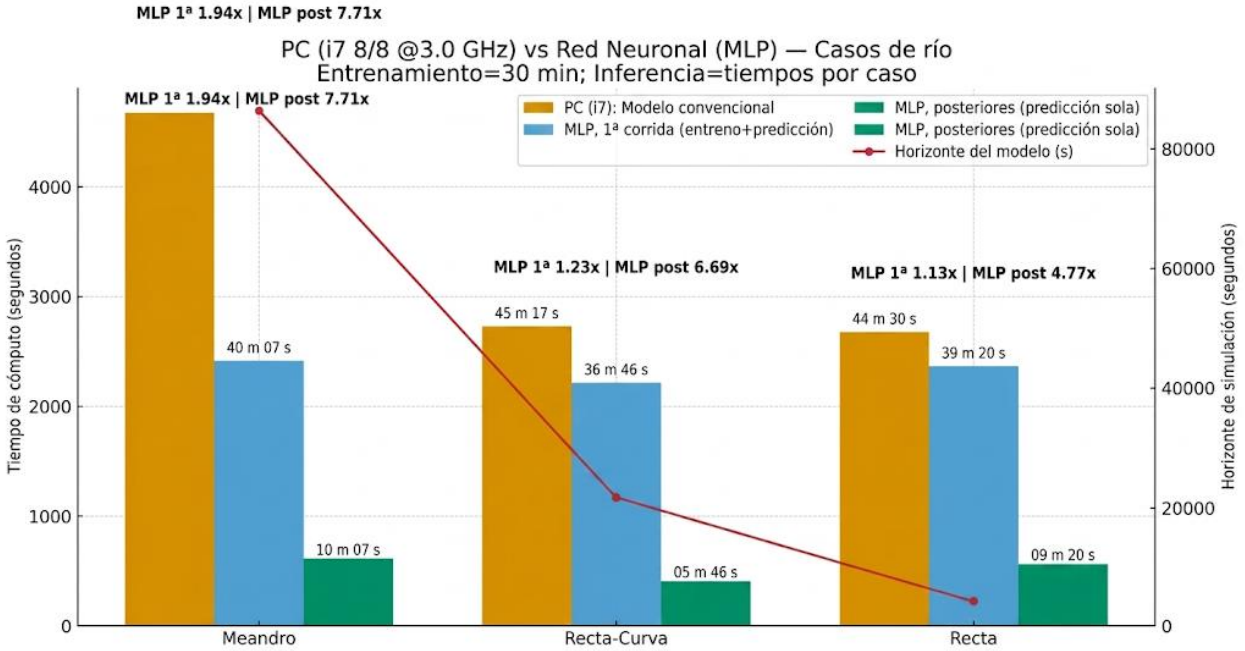


Fig. 4. Comparación de eficiencia computacional de los casos de estudio.

4. Conclusión

Una estrategia híbrida, como la modelación morfodinámica 2D para generar datos y MLP como sustituto, preserva la estructura espaciotemporal del lecho con RMSE ≈ 4.9×10^{-4} m y reduce el tiempo de evaluación de horas a minutos. La comparación TELEMAC–MLP debe interpretarse como emulación del operador numérico dentro del dominio de entrenamiento: la RNA no impone explícitamente conservación tipo Exner, pero, al aprender de soluciones físicas, reproduce patrones de erosión y deposición sin sesgos evidentes. Aun así, para una aplicación en escenarios reales, el método actual presenta limitaciones concretas. El modelo se desarrolló bajo geometrías y caudales controlados; por tanto, carece de pruebas en condiciones extremas o de flujo no permanente propios de la naturaleza. Además, el cálculo se restringe al transporte de carga de fondo y omite el transporte en suspensión. Estructuralmente, la arquitectura MLP puede ser insuficiente para captar dependencias espaciotemporales complejas, donde redes convolucionales o recurrentes serían más aptas. El esquema actual carece de paralelización o implementación en lenguajes óptimos (como C o plataformas HPC) para dominios extensos. Además, la precisión puede degradarse fuera del rango entrenado, por lo que se recomienda diversificar el conjunto de datos e integrar restricciones físicas para mejorar la generalización.

Referencias

[1] M. Garcia, Ed., *Sedimentation Engineering: Processes, Measurements, Modeling, and Practice*. Reston, VA, USA: American Society of Civil Engineers, 2008. doi: 10.1061/9780784408148.

[2] E. Meyer-Peter and R. Muller, “Formulas for Bed-Load transport,” in *Proceedings of 2nd meeting of the International Association for Hydraulic Structures Research*, 1948. Accessed: Oct. 20, 2025. [Online]. Available: https://repository.tudelft.nl/record/uuid:4fda9b61-be28-4703-ab06-43cdc2a21bd7

[3] L. C. van Rijn, *Principles of Sediment Transport in Rivers, Estuaries and Coastal Seas*. Aqua Publications, 1993. [Online]. Available: https://books.google.com.mx/books?id=gGIYAQAAIAAJ

[4] K. Kaveh, "Development of Data Driven Models for Hydromorphology and Sediment Transport," Tesis doctoral. Technische Universität München, 2019. Accessed: Oct. 20, 2025. [Online]. Available: https://mediatum.ub.tum.de/node?id=1451898

[5] K. D. A. Segales, "Aplicación De Técnicas De Aprendizaje De Máquina Para El Análisis Morfológico De Ríos," Tesis de licenciatura, Universidad Nacional Autónoma de México, México, 2025.

[6] A. W. Minns, "Hydrological Modelling in a Hydroinformatics Context," in *Distributed Hydrological Modelling*, M. B. Abbott and J. C. Refsgaard, Eds., Dordrecht: Springer Netherlands, 1996, pp. 297–312. doi: 10.1007/978-94-009-0257-2_16.

[7] C. W. Dawson and R. L. Wilby, "A comparison of artificial neural networks used for river forecasting," *Hydrology and Earth System Sciences Discussions*, vol. 3, no. 4, pp. 529–540, 1999.

[8] R. Abrahart, P. E. Kneale, and L. M. See, Eds., *Neural Networks for Hydrological Modeling*, CRC Press, 2004. doi: 10.1201/9780203024119.

[9] X. Yan, F. Du, T. Zhang, Q. Cui, Z. Zhu, and Z. Song, "Predicting the Flow Fields in Meandering Rivers with a Deep Super-Resolution Convolutional Neural Network," *Water*, vol. 16, no. 3, p. 425, Jan. 2024, doi: 10.3390/w16030425.

[10] S. R. O. Marshall, T.-N.-D. Tran, M. R. Tapas, and B. Q. Nguyen, "Integrating artificial intelligence and machine learning in hydrological modeling for sustainable resource management," *International Journal of River Basin Management*, pp. 1–17, Mar. 2025, doi: 10.1080/15715124.2025.2478280.

[11] J. G. Gacu, C. E. F. Monjardin, R. G. T. Mangulabnan, and J. C. F. Mendez, "Application of Artificial Intelligence in Hydrological Modeling for Streamflow Prediction in Ungauged Watersheds: A Review," *Water*, vol. 17, no. 18, p. 2722, Sep. 2025, doi: 10.3390/w17182722.

[12] L. Slater *et al.*, "Challenges and opportunities of ML and explainable AI in large-sample hydrology," *Phil. Trans. R. Soc. A.*, vol. 383, no. 2302, p. 20240287, Jul. 2025, doi: 10.1098/rsta.2024.0287.

[13] J. Hervouet, *Hydrodynamics of Free Surface Flows: Modelling with the finite element method*, 1st ed. Wiley, 2007. doi: 10.1002/9780470319628.

[14] C. Yen and K. T. Lee, "Bed Topography and Sediment Sorting in Channel Bend with Unsteady Flow," *J. Hydraul. Eng.*, vol. 121, no. 8, pp. 591–599, Aug. 1995, doi: 10.1061/(ASCE)0733-9429(1995)121:8(591).

[15] N. Moghaddassi, S. H. Musavi-Jahromi, M. Vaghefi, and A. Khosrojerdi, "Effect of Mean Velocity-to-Critical Velocity Ratios on Bed Topography and Incipient Motion in a Meandering Channel: Experimental Investigation," *Water*, vol. 13, no. 7, p. 883, Mar. 2021, doi: 10.3390/w13070883.

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

Aplicación de inteligencia artificial en genómica y metagenómica viral para la clasificación taxonómica y el descubrimiento de nuevas proteínas virales

Blanca Itzel Taboada Ramírez
Universidad Nacional Autónoma de México, Instituto de Biotecnología, Cuernavaca, Morelos, México.
ORCID: 0000-0003-1896-5962

Lorena Díaz-González
Universidad Autónoma del Estado de Morelos, Centro de Investigación en Ciencias,
Cuernavaca, Morelos, México.
ORCID: 0000-0003-1577-5629

Oscar Alejandro Uscanga Junco
Universidad Autónoma del Estado de Morelos, Instituto de Investigación en Ciencias Básicas Aplicadas (IICBA),
Cuernavaca, Morelos, México.
ORCID: 0000-0002-4179-6725

Alida Esmeralda Zárate Jiménez
Universidad Nacional Autónoma de México, Instituto de Biotecnología, Cuernavaca, Morelos, México.
ORCID: 0009-0006-5407-6598

Edna Cruz-Flores
Universidad Autónoma del Estado de Morelos, Instituto de Investigación en Ciencias Básicas Aplicadas (IICBA),
Cuernavaca, Morelos, México.
ORCID: 0000-0002-4187-3428

Recepción: 01 de marzo de 2026.
Aceptación: 28 de abril de 2026.

Mayo 2026 • número de revista 15 • DOI: 10.22201/dgtic.26832968e.2026.15.158

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

Aplicación de inteligencia artificial en genómica y metagenómica viral para la clasificación taxonómica y el descubrimiento de nuevas proteínas virales

Resumen

La metagenómica viral permite caracterizar comunidades de virus en muestras clínicas o ambientales a partir de millones de secuencias de ADN generadas por tecnologías de próxima generación. Este trabajo describe tres aplicaciones computacionales que, apoyadas en el uso de supercómputo, buscan mejorar el análisis de datos metagenómicos: i) una metodología que elimina la redundancia de las bases de datos de secuencias de referencia de genomas de virus mediante la construcción de pangenomas, conservando secuencias específicas de cada especie y las compartidas a nivel de género, lo que permite identificar virus de manera más precisa; ii) una herramienta que usa inteligencia artificial para identificar secuencias de virus eucariontes a nivel de proteínas, facilitando la detección de virus nuevos o con baja similitud a los anotados; y iii) una herramienta, también basada en inteligencia artificial, para identificar arreglos CRISPR en genomas bacterianos, lo que favorece el estudio de las interacciones fago-bacteria en datos metagenómicos. Estas aplicaciones apoyan el análisis de datos metagenómicos, contribuyendo a comprender mejor la diversidad viral y las relaciones virus-bacteria.

Palabras Clave: metagenómica viral, clasificación taxonómica, proteínas virales, bacteriófagos, CRISPR, aprendizaje profundo, supercómputo.

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

Application of Artificial Intelligence in Viral Genomics and Metagenomics for Taxonomic Classification and the Discovery of Novel Viral Proteins

Abstract

Viral metagenomics enables the characterization of viral communities in clinical or environmental samples based on millions of DNA sequences generated by next-generation sequencing technologies. This work describes three computational applications that leverage high-performance computing to improve metagenomic data analysis: (i) a methodology that reduces redundancy in viral reference genome databases through the construction of pangenomes, preserving species-specific sequences as well as sequences shared at the genus level, enabling more accurate virus identification; (ii) an artificial intelligence–based tool for identifying eukaryotic viral sequences at the protein level, facilitating the detection of novel viruses or those with low similarity to annotated references; and (iii) an artificial intelligence–based tool for identifying CRISPR arrays in bacterial genomes, which supports the study of phage–bacteria interactions in metagenomic datasets. Together, these applications enhance metagenomic data analysis and contribute to a better understanding of viral diversity and virus–bacteria relationships.

Keywords: *viral metagenomics, taxonomic classification, viral proteins, bacteriophages, CRISPR, deep learning, high-performance computing.*

Introducción

La metagenómica es una herramienta que permite estudiar las comunidades microbianas presentes en una muestra clínica o ambiental mediante el análisis de millones de secuencias de ADN o ARN, siguiendo el flujo de trabajo descrito en la Fig.1. Esto ha permitido identificar virus en océanos, suelos, animales, plantas y humanos, entre otros, mostrando que la diversidad de virus es mucho más extensa de lo que se había pensado.

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

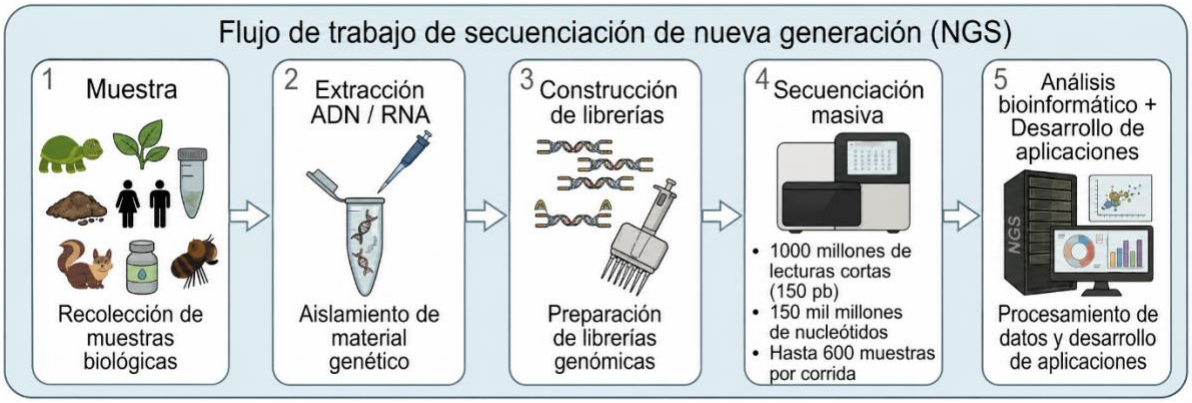


Fig. 1. Flujo de trabajo general para el análisis metagenómico mediante NGS.

Los virus se pueden dividir en dos grupos: los bacteriófagos, que infectan bacterias, y los eucariontes, que infectan a plantas, animales, hongos, insectos y humanos. Estos últimos incluyen virus de importancia médica, veterinaria y agrícola, por lo que conocer su diversidad y mejorar su detección es crucial para el desarrollo de estrategias de vigilancia, prevención y control. Por otra parte, los bacteriófagos son los virus más abundantes del planeta, con más de 10^{31} partículas en la biosfera, y desempeñan un papel importante al modular a las comunidades bacterianas. Sin embargo, la mayoría de estos virus permanece sin descubrirse [1]. Estudiar su diversidad y sus interacciones es clave para explicar su efecto en distintos ecosistemas.

Los estudios metagenómicos generan un inmenso volumen de secuencias cortas de ADN llamadas lecturas (Fig.1). Sin embargo, entre el 40 % y el 90 % de éstas no puede asignarse a ninguna especie viral, conociéndose como “materia oscura viral”. Esto es debido al uso de métodos basados en alineamientos y/o búsquedas por similitud contra bases de datos de genomas de referencia, las cuales no reflejan toda la diversidad viral. Además, la redundancia de estas bases de datos puede generar asignaciones no específicas. Todo esto afecta la vigilancia genómica, el estudio de la diversidad viral y la comprensión del papel de los bacteriófagos en diferentes ecosistemas.

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

En este contexto, el desarrollo de nuevas herramientas computacionales basadas en inteligencia artificial y la mejora de las bases de datos de referencia abren nuevas posibilidades para abordar esta problemática. El objetivo de este trabajo es describir un conjunto de aplicaciones (Fig. 2), que se apoyaron en el uso de cómputo de alto desempeño para (i) reducir y enriquecer bases de datos de secuencias virales de referencia [2], (ii) clasificar taxonómicamente proteínas de virus eucariontes [3] y (iii) apoyar el estudio de interacciones fago-bacteria mediante la identificación de arreglos CRISPR.

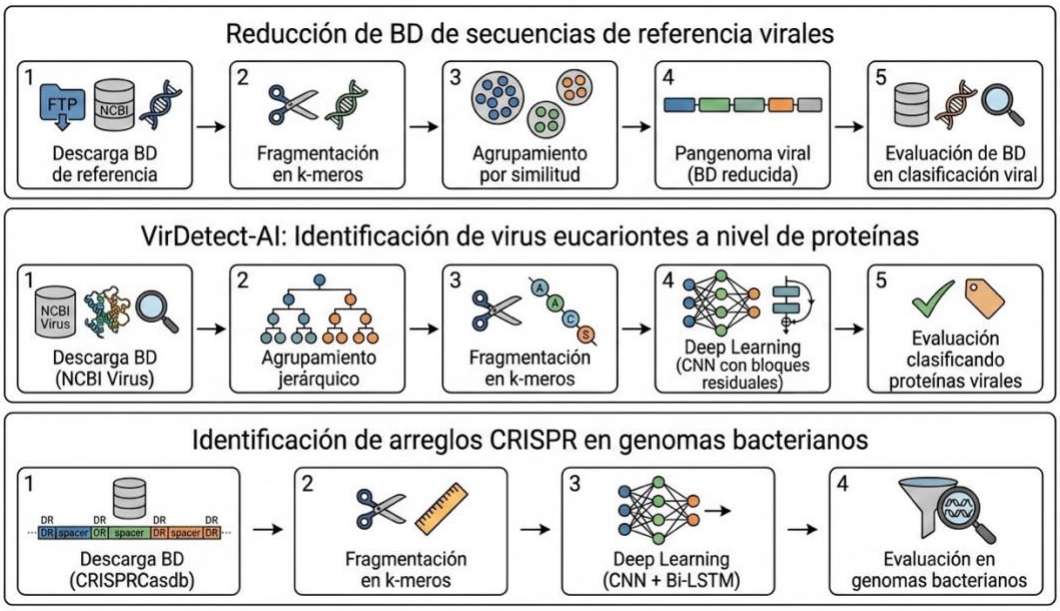


Fig. 2. Desarrollo de aplicaciones en metagenómica viral.

Este artículo tiene un enfoque de divulgación científica, en el que las aplicaciones (i) y (ii) se basan en desarrollos previamente validados, mientras que la aplicación (iii) corresponde a una línea de trabajo en desarrollo.

Marco metodológico para el análisis metagenómico de la diversidad viral

Como ya se mencionó, la metagenómica viral produce grandes volúmenes de información en forma de secuencias denominadas lecturas que son, de manera general, de una longitud

de 150 nucleótidos. A partir de estos datos, el reto consiste en encontrarles un sentido biológico, identificando virus, mejorando su clasificación y, cuando sea posible, proponiendo funciones o relaciones dentro del microbioma.

Este trabajo integra el desarrollo de aplicaciones computacionales que operan en distintos niveles. Por un lado, se emplean estrategias a nivel de nucleótido y k-meros, que trabajan directamente sobre secuencias de ADN para reducir las de bases de datos de genomas de referencia, enriquecer regiones informativas y favorecer identificaciones más específicas; dentro de esta misma línea, se ubica el método de identificación de arreglos CRISPR, útiles como evidencia para explorar interacciones fago-bacteria.

Por otro lado, cuando las secuencias presentan baja similitud con lo ya anotado, se recurre al análisis a nivel de proteínas, con el objetivo de capturar patrones más allá de la similitud inmediata y permitir la clasificación taxonómica de proteínas virales eucariontes que sean muy diferentes a lo conocido. Esto permitirá clasificar proteínas virales y descubrir nuevas, ampliando el repertorio funcional del viroma.

Dado el tamaño de los conjuntos de datos y la complejidad de los análisis, incluyendo el entrenamiento de modelos con millones de secuencias y la ejecución de procesos a gran escala, el desarrollo y evaluación de estas aplicaciones se realizaron con el apoyo de cómputo de alto rendimiento, incluyendo el uso de GPUs, en la supercomputadora Miztli (UNAM). Esta infraestructura permite entrenar modelos en tiempos razonables y ejecutar pruebas comparativas de manera sistemática, obteniendo resultados escalables.

Aplicación para reducir la base de datos de secuencias de referencia virales

De acuerdo con la International Committee on Taxonomy of Viruses (ICTV), los virus se organizan en 15 rangos taxonómicos, desde dominio hasta especie [4]. Las especies distintas de un mismo género comparten regiones genómicas conservadas, generando información redundante. En estudios metagenómicos, esto puede provocar que las lecturas obtenidas se asignen indistintamente a varias referencias, generando asignaciones poco específicas que no permiten diferenciar especies concretas. Por lo anterior, es necesario realizar análisis

adicionales que requieren un tiempo de cómputo considerable, tales como el análisis de ancestro común [5].

Para resolver esta problemática, diseñamos K-FluDB, una metodología computacional basada en el agrupamiento de subsecuencias de genomas virales [2]. Esto permite identificar subsecuencias que son específicas de cada especie y otras que son comunes entre varias especies. Estas últimas se denominan "piezas genómicas dispensables", pues son secuencias que permiten clasificar a nivel de género. Al conjunto total de secuencias específicas y dispensables, se le conoce como pangenoma viral.

Resumen de la metodología

La base de datos de virus de referencia se descargó de NCBI [6], obteniendo 14,331,967 secuencias de nucleótidos de referencia correspondientes a 31,573 especies y 3,328 géneros.

La metodología de reducción, ilustrada en la Fig. 3, consiste en obtener, para cada conjunto de secuencias de una especie s, todas las subsecuencias o k-meros de tamaño k = 200 con un sobrelape(o) de 150 nucleótidos. Esta configuración de k y de o garantiza que cualquier subsecuencia de 150 nucleótidos, como las lecturas generadas por las tecnologías de secuenciación masiva en estudios genómicos o metagenómicos, se encuentre dentro de, al menos, un k-mero.

Los k-meros obtenidos de cada especie se agrupan por similitud al 94%, utilizando la herramienta CD-HIT [7], y se selecciona una secuencia representativa para cada grupo. A partir de esto, se definen dos grupos:

a) $CUEs$ (Clústeres de un elemento): grupos que tienen una sola secuencia específica de la especie, sin similitud con secuencias de otras especies.
b) $CMEs$ (Clústeres de múltiples elementos): grupos de múltiples secuencias compartidas entre distintas especies, por lo que representan información redundante que debería estar relacionada a un nivel taxonómico superior, en este caso género.

Las secuencias representativas de los $CUEs$ y los $CMEs$ se concatenan si son consecutivas y provienen de la misma secuencia. El conjunto final de secuencias representa el pangenoma viral de la especie s.

Este procedimiento se repite utilizando todos los pangenomas virales generados de todas las especies asociadas al género G. De este modo, se obtienen secuencias específicas que permiten identificar especies virales con precisión, derivadas de los $CUEs$, y secuencias compartidas a nivel de género que son derivadas de los $CMEs$. Estas últimas secuencias compartidas representan la información que se hubiera obtenido de un análisis LCA. En el pangenoma a nivel de género, los $CUEs$ y los $CMEs$ se denominan CGs y CSs, respectivamente, por las secuencias representativas de géneros y especies.

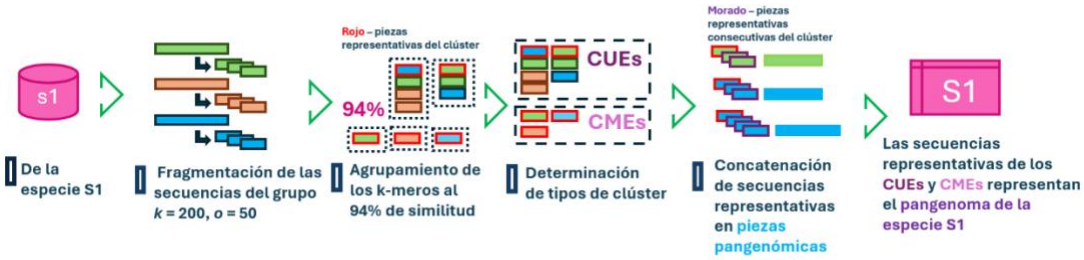


Fig. 3. Esquema de metodología para obtención de pangenomas virales.

Resultados de la compresión y validación del pangenoma

Para evaluar la eficacia de la metodología, se construyó un índice de búsqueda rápida del pangenoma viral. La validación se realizó mediante la simulación de más de 105 millones de lecturas de 150 nucleótidos, pertenecientes a las 31,573 especies virales, las cuales se mapearon contra este índice para encontrar su anotación a nivel de especie o género.

Los resultados mostraron un desempeño sobresaliente, con una recuperación promedio del 99.86% de las secuencias simuladas, lo que indica una alta precisión y sensibilidad. Estos valores confirman que la reducción de la base de datos a pangenomas no compromete la capacidad de identificación, permitiendo una clasificación eficiente.

En cuanto a la eficiencia de la compresión, se observa que, de más de 9 mil millones de nucleótidos iniciales, el pangenoma final representa solo el 12.8% del total. Esto permite realizar búsquedas eficientes, utilizando poco más de 1.1 mil millones de nucleótidos, sin sacrificar resolución a nivel de especie y género.

VirDetect-AI: una nueva forma de ver la diversidad viral eucarionte

En la última década, se han desarrollado herramientas computacionales para identificar virus en datos metagenómicos, como VirFinder, DeepVirFinder y VirSorter2, entre otras [8], [9], [10]. Sin embargo, la mayoría analiza secuencias a nivel de ADN, lo que limita estos métodos en identificar virus con baja homología con secuencias de referencia. Además, muchas funcionan como clasificadores binarios, distinguiendo sólo entre secuencias virales y no virales, sin aportar información sobre la familia, género o especies virales identificadas, la función de las proteínas virales o el posible hospedero. Además, estas herramientas suelen ser específicas en un ámbito, por ejemplo, la identificación sólo de bacteriófagos o de muestras humanas, lo que limita su desempeño en la detección de virus eucariontes y deja una fracción considerable de la diversidad viral sin explorar.

Para enfrentar esta problemática, desarrollamos VirDetect-AI [3], cuyo proceso se muestra en la Fig. 4. VirDetect-AI es una herramienta innovadora, basada en inteligencia artificial, diseñada específicamente para la identificación de virus eucariontes en muestras metagenómicas.

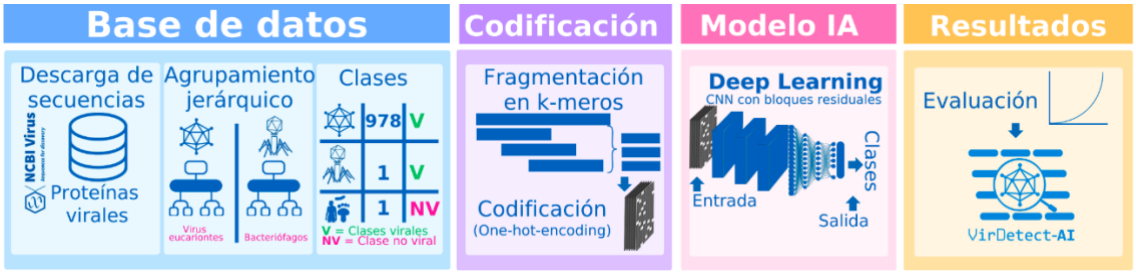


Fig. 4. Metodología detallada para el desarrollo de la herramienta VirDetect-AI para la identificación de secuencias de proteínas virales eucariontes en datos metagenómicos.

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

VirDetect-AI trabaja a nivel de aminoácidos (aa) para identificar proteínas virales o fragmentos de éstas. Este cambio es clave, ya que las proteínas suelen conservar motivos funcionales, incluso cuando sus secuencias de nucleótidos cambian considerablemente. De este modo, se puede identificar relaciones evolutivas aun cuando la similitud entre las secuencias sea baja. Esto permite identificar virus poco parecidos a los ya descritos en las bases de datos actuales [11].

Para poder construir esta herramienta, se descargaron secuencias de proteínas virales de NCBI [6]. A continuación, estas secuencias fueron agrupadas de manera jerárquica usando umbrales de similitud desde el 80% hasta el 30%. Posteriormente, a partir de cada secuencia, se generaron k-meros de 300 aa de longitud con saltos de 20 aa.

VirDetect-AI está basado en un modelo de deep learning que combina arquitecturas de redes neuronales convolucionales (CNN) [12] con bloques residuales (ResNet) [13]. Éste clasifica secuencias en 978 clases de proteínas virales eucariontes, abarcando decenas de familias virales y géneros que representan diversas funciones y tipos de hospederos. De esta forma, no sólo identifica si una secuencia es viral, sino que ofrece una clasificación mucho más rica y detallada, lo que permite interpretar mejor los resultados biológicos.

Asimismo, VirDetect-AI incluyó una clase viral negativa que identifica bacteriófagos y una clase negativa para identificar secuencias de humano, bacterias, hongos y arqueas. El total de datos se dividió en 80% para entrenamiento, 10% para validación y 10% como conjunto de prueba.

Identificación de arreglos CRISPR en genomas bacterianos

Como ya mencionamos, los bacteriófagos o fagos son impulsores clave de la evolución bacteriana, moldeando la composición, diversidad y dinámica de sus comunidades. Comprender estas interacciones fago-bacteria es fundamental para desentrañar la dinámica ecológica microbiana [14].

Para contrarrestar los ataques virales, las bacterias han desarrollado diversos sistemas de defensa, entre ellos CRISPR-Cas (*Clustered Regularly Interspaced Short Palindromic Repeats*

- CRISPR-*associated protein*), el cual funciona como un sistema inmune adaptativo, ya que almacena fragmentos del ADN fago invasor como memoria molecular. Cuando una bacteria sobrevive a una infección, integra fragmentos del genoma del fago invasor (conocidos como espaciadores) en un arreglo CRISPR. Esto le permite reconocer y neutralizar al mismo virus en infecciones posteriores [15].

Los arreglos CRISPR-Cas tienen una estructura característica, secuencias de repetidores CRISPR (*Direct Repeats* o DR) separadas por espaciadores únicos. Estas repeticiones suelen tener entre 23 y 55 nucleótidos de longitud y están separadas por espaciadores o distancias específicas que van de los 16 a 109 nucleótidos.

Actualmente, no existe un método computacional que permita predecir con alta confiabilidad las interacciones fago-bacteria. No obstante, la identificación precisa de sistemas CRISPR, junto con la extracción y el análisis comparativo de sus espaciadores, representan una aproximación prometedora, al ser un registro molecular de encuentros históricos entre bacterias y fagos [16].

A diferencia de los métodos tradicionales, basados en conservación de secuencias, nuestro enfoque, ilustrado en la Fig. 5, utiliza aprendizaje profundo capaz de identificar arreglos CRISPR poco conservados. El conjunto positivo se construyó con los sistemas CRISPR-Cas validados experimentalmente de la base de datos CRISPRCasdb [17]. El conjunto negativo combinó secuencias de nucleótidos de baja complejidad y secuencias sintéticas del trabajo de Wang & Liang en 2017 [18]. El conjunto de datos final incluyó 84,065 ejemplos positivos (DR) y 85,949 negativos (no-DR), dividido en 80% para entrenamiento y 20% para prueba/validación. Todas las secuencias se fragmentaron en k-meros de 30 nucleótidos, usados como entrada del modelo.

La arquitectura del modelo integra una capa de *embedding* para codificar numéricamente los nucleótidos de las secuencias, capas convolucionales (Conv1D) para detectar patrones locales, capas Bi-LSTM para modelar dependencias de mayor alcance en ambas direcciones de la secuencia y una capa densa para generar una salida binaria (DR/no-DR). Se evaluaron diferentes combinaciones de hiperparámetros incluyendo capas Conv1D, número de filtros y tamaño del *kernel* en cada Conv1D; número de unidades y profundidad en Bi-LSTM; y valores de regularización *dropout*.

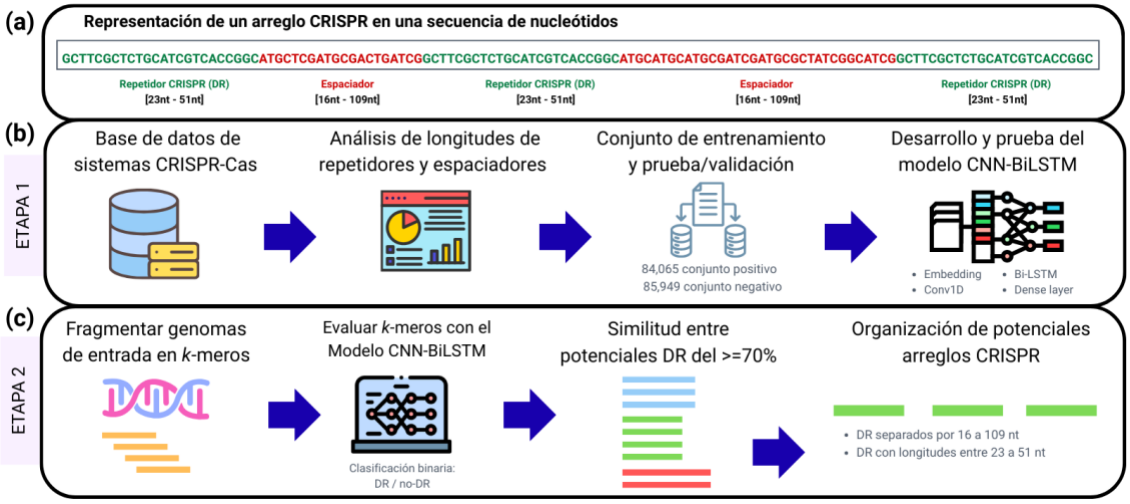


Fig. 5. (a) Representación de un arreglo CRISPR en una secuencia de nucleótidos. (b) Etapa 1 del desarrollo del modelo CNN-BiLSTM. (C) Etapa 2 del análisis bioinformático para validar los candidatos a arreglos CRISPR.

La arquitectura del modelo con mejor desempeño consistió en una capa de *embedding* seguida de dos capas convolucionales Conv1D de 64 y 32 filtros de tamaño 7, normalización por lotes (*batch normalization*), reducción de dimensionalidad mediante *max pooling* y *dropout* de 0.2. Posteriormente, se incorporó una capa Bi-LSTM de 64 unidades, así como una capa densa de 128 neuronas y, finalmente, la capa de salida con una única unidad.

Este modelo alcanzó una exactitud y precisión del 98.4% sobre el conjunto de prueba, lo cual muestra la capacidad del enfoque para discriminar de manera robusta entre repetidores directos (DR) y secuencias no-DR, incluso en escenarios con variación en los patrones de secuencia.

En la segunda etapa, implementamos un proceso bioinformático para validar los DR identificados por el modelo y pertenecientes al mismo sistema CRISPR, usando tres criterios: i) distancia espacial entre los DR, que es de 16 a 109 nucleótidos (rango típico de espaciadores); ii) similitud entre DR, que debía ser ≥70%; y iii) organización característica, que consiste en el patrón repetitivo de arreglos CRISPR auténticos.

En conjunto, este enfoque combina un modelo de aprendizaje profundo para la detección de repeticiones directas y un proceso bioinformático para delimitar DRs y extraer

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

los espaciadores, ofreciendo dos ventajas: mayor flexibilidad para identificar arreglos atípicos, que no se detectan por homología, y la aplicabilidad a secuencias cortas típicas de metagenomas fragmentados, donde los métodos convencionales muestran sensibilidad limitada [19].

Conclusión

En este trabajo, se presentan tres aplicaciones basadas en la inteligencia artificial y el supercómputo que contribuyen a los estudios de metagenómica viral: (i) K-FluDB, una herramienta de optimización de la base de datos de secuencias de referencia virales, mediante la construcción de pangenomas virales a nivel especie y género, para mejorar la clasificación taxonómica de virus; (ii) VirDetect-AI, una nueva herramienta para clasificar proteínas virales y descubrir nuevas, ampliando el repertorio funcional del viroma; y (iii) la identificación de arreglos CRISPR en genomas bacterianos para ampliar el estudio de las interacciones fago-bacteria en datos metagenómicos.

Estas nuevas herramientas contribuyen al análisis de datos metagenómicos virales, teniendo un impacto en la vigilancia genómica, el estudio de la diversidad viral y la comprensión del papel de los bacteriófagos en diferentes ecosistemas.

Financiamiento

Este trabajo fue financiado por los proyectos PAPIIT-DGAPA-IN230523 y PAPIIT-DGAPA-IN225126 de la DGAPA-UNAM (a B.T.), así como por el proyecto CBF-2025-I-1026 de SECIHTI (a B.T.).

Agradecimientos

Agradecemos a Jérôme Verleyen, Roberto Bahena y Juan Manuel Hurtado por su invaluable apoyo en el ámbito computacional.

Referencias

[1] A. D. Rowan-Nash, B. J. Korry, E. Mylonakis, and P. Belenky, "Cross-Domain and Viral Interactions in the Microbiome," *Microbiology and Molecular Biology Reviews*, vol. 83, no. 1, Feb. 2019, doi: 10.1128/MMBR.00044-18.

[2] A. Uscanga Junco, L. Díaz-González, and B. Taboada, "K-FluDB: A Novel K-Mer Based Database for Enhanced Genomic Surveillance of Influenza A Viruses," *Bioinformatics Advances*, Oct. 2025, doi: 10.1093/bioadv/vbaf254.

[3] A. Zárate, L. Díaz-González, and B. Taboada, "VirDetect-AI: a residual and convolutional neural network–based metagenomic tool for eukaryotic viral protein identification," *Brief. Bioinform.*, vol. 26, no. 1, Nov. 2024, doi: 10.1093/bib/bbaf001.

[4] E. J. Black, C. S. Powell, D. M. Dempsey, R. C. Hendrickson, L. R. Mims, and E. J. Lefkowitz, "Virus taxonomy: the database of the International Committee on Taxonomy of Viruses," *Nucleic Acids Res.*, vol. 54, no. D1, pp. D776–D789, Jan. 2026, doi: 10.1093/nar/gkaf1159.

[5] Y. Wang, T. S. Korneliussen, L. E. Holman, A. Manica, and M. W. Pedersen, "ngs LCA - A toolkit for fast and flexible lowest common ancestor inference and taxonomic profiling of metagenomic data," Methods Ecol. Evol., vol. 13, no. 12, pp. 2699–2708, Dec. 2022, doi: 10.1111/2041-210X.14006.

[6] National Center for Biotechnology Information (NCBI), "Virus genomes – All nucleotide sequences," NCBI FTP Server. [Online]. Available: https://ftp.ncbi.nlm.nih.gov/genomes/Viruses/AllNucleotide/. [Accessed: Aug. 24, 2025].

[7] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, Dec. 2012, doi: 10.1093/bioinformatics/bts565.

[8] J. Ren, N. A. Ahlgren, Y. Y. Lu, J. A. Fuhrman, and F. Sun, "VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data," Microbiome, vol. 5, no. 1, p. 69, Dec. 2017, doi: 10.1186/s40168-017-0283-5.

[9] J. Ren, K. Song, C. Deng, N. A. Ahlgren, J. A. Fuhrman, Y. Li, X. Xie, R. Poplin, and F. Sun, "Identifying viruses from metagenomic data using deep learning," *Quantitative Biology*, vol. 8, no. 1, p. 64, 2020, doi: 10.1007/s40484-019-0187-4.

[10] J. Guo *et al.*, "VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses," *Microbiome*, vol. 9, no. 1, p. 37, Dec. 2021, doi: 10.1186/s40168-020-00990-y.

[11] C. Chothia and A. M. Lesk, "The relation between the divergence of sequence and structure in proteins.," *EMBO J.*, vol. 5, no. 4, pp. 823–826, Apr. 1986, doi: 10.1002/j.1460-2075.1986.tb04288.x.

[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv:1512.03385 [cs]*, Dec. 2015, [Online]. Available: http://arxiv.org/abs/1512.03385

[14] L. Beller and J. Matthijnssens, "What is (not) known about the dynamics of the human gut virome in health and disease," *Curr. Opin. Virol.*, vol. 37, pp. 52–57, Aug. 2019, doi: 10.1016/j.coviro.2019.05.013.

[15] E. V. Koonin and K. S. Makarova, "Origins and evolution of CRISPR-Cas systems," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 374, no. 1772, p. 20180087, May 2019, doi: 10.1098/rstb.2018.0087.

[16] D. J. Nasko, B. D. Ferrell, R. M. Moore, J. D. Bhavsar, S. W. Polson, and K. E. Wommack, "CRISPR Spacers Indicate Preferential Matching of Specific Virioplankton Genes," *mBio*, vol. 10, no. 2, Apr. 2019, doi: 10.1128/mBio.02651-18.

[17] C. Pourcel *et al.*, "CRISPRCasdb a successor of CRISPRdb containing CRISPR arrays and cas genes from complete genome sequences, and tools to download and query lists of repeats and spacers," *Nucleic Acids Res.*, Oct. 2019, doi: 10.1093/nar/gkz915.

[18] K. Wang and C. Liang, "CRF: detection of CRISPR arrays using random forest," *PeerJ*, vol. 5, p. e3219, Apr. 2017, doi: 10.7717/peerj.3219.

[19] C. Coclet and S. Roux, "Global overview and major challenges of host prediction methods for uncultivated phages," *Curr. Opin. Virol.*, vol. 49, pp. 117–126, Aug. 2021, doi: 10.1016/j.coviro.2021.05.003.

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

Desde la quinolina hasta el Alzheimer. Diseño computacional de fármacos multifuncionales usando cómputo de alto desempeño

Luis Felipe Hernández Ayala
Universidad Nacional Autónoma de México, Facultad de Química, Departamento de Química Inorgánica, Ciudad de México, México.
ORCID: 0000-0002-3387-763X

Eduardo Gabriel López Guzmán
Universidad Autónoma Metropolitana-Iztapalapa, Departamento de Química, Ciudad de México, México.
ORCID: 0000-0003-1443-4136

Mario Prejanò
Universidad de Calabria, Departamento de Química y Tecnologías Químicas, Arcavacata de Rende, Italia.
ORCID: 0000-0002-9140-6246

Tiziana Marino
Universidad de Calabria, Departamento de Química y Tecnologías Químicas, Arcavacata de Rende, Italia.
ORCID: 0009-0008-5856-0664

Annia Galano
Universidad Autónoma Metropolitana-Iztapalapa, Departamento de Química, Ciudad de México, México.
ORCID: 0000-0002-1470-3060

Recepción: 26 de febrero de 2026.
Aceptación: 16 de abril de 2026.

Mayo 2026 • número de revista 15 • DOI: 10.22201/dgtic.26832968e.2026.15.147

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

Desde la quinolina hasta el Alzheimer. Diseño computacional de fármacos multifuncionales usando cómputo de alto desempeño

Resumen

La enfermedad de Alzheimer (EA) representa un desafío biomédico actual debido a su naturaleza multifactorial y a la ausencia de terapias efectivas. Frente a este escenario, el diseño de fármacos multifuncionales surge como una estrategia prometedora, impulsada por el desarrollo tecnológico. En este trabajo, se describe la implementación del protocolo CADMA-Chem en infraestructuras de supercómputo (HPC) para la exploración masiva de un espacio químico de 8356 derivados de la quinolina en los clústeres Miztli y Yoltla (LANCAD-UNAM). Mediante el uso de quimioinformática y simulaciones de acoplamiento molecular, se priorizaron cinco candidatos con actividad antioxidante y afinidad por la acetilcolinesterasa (AChE), diana clave en la EA. Simulaciones de dinámica molecular y estimaciones energéticas MM/GBSA, permitieron identificar al derivado dQ1368 como el candidato más prometedor. Finalmente, la integración de inteligencia artificial para el análisis retrosintético permite sentar las bases para el trabajo experimental, demostrando cómo la sinergia entre el HPC universitario y la IA acelera la investigación de frontera.

Palabras Clave: diseño de fármacos, fármacos multifuncionales, Alzheimer, HPC, quimioinformática.

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

From quinoline to Alzheimer's: Computational design of multifunctional drugs using high-performance computing

Abstract

Alzheimer's disease (AD) represents one of the most significant current biomedical challenges due to its multifactorial nature and the lack of effective therapies. In this scenario, the design of multifunctional drugs (MFDs) emerges as a promising strategy, driven by technological development. This work describes the implementation of the CADMA-Chem protocol on High-Performance Computing (HPC) infrastructures for the massive exploration of a chemical space comprising 8,356 quinoline derivatives using the Miztli and Yoltla clusters (LANCAD-UNAM). Through chemoinformatics and molecular docking simulations, five candidates were prioritized based on their antioxidant activity and affinity for acetylcholinesterase (AChE), a key target in AD. Advanced molecular dynamics simulations and MM/GBSA energy estimations identified the derivative dQ1368 as the most promising candidate due to its stability and energetic profile. Finally, the integration of artificial intelligence for retrosynthetic analysis lays the groundwork for experimental work, demonstrating how the synergy between university-based HPC and AI accelerates frontier research.

Keywords: *drug design, multifunctional drugs, Alzheimer, HPC, chemoinformatics.*

Introducción

La enfermedad de Alzheimer (EA) constituye uno de los principales retos de salud pública. La ausencia de terapias efectivas y la progresión irreversible evidencian la necesidad de estrategias innovadoras que superen los enfoques terapéuticos tradicionales. La EA se caracteriza por una compleja red de mecanismos patológicos interconectados. La acumulación de proteínas mal plegadas, el desequilibrio en la homeostasis de metales, el incremento del estrés oxidativo y la alteración de la neurotransmisión actúan de manera simultánea, empujando a la neurona hacia el colapso [1]. Este carácter multifactorial limita la eficacia de fármacos dirigidos, cuyos efectos son paliativos y de alcance moderado.

Ante este escenario, los fármacos multifuncionales (FMF) surgen como una estrategia prometedora que integra, en una sola molécula, propiedades complementarias, tales como capacidad antioxidante, actividad quelante, inhibición enzimática y actividad reparadora con efectos sinérgicos, menor polifarmacia y reducción de reacciones adversas. Sin embargo, el diseño de FMF requiere la exploración sistemática de espacios químicos masivos y la estimación temprana de propiedades ADME (Administración, Distribución, Metabolismo y Excreción), toxicidad y, en función de la diana terapéutica, actividad antioxidante u otras propiedades específicas. Todas estas tareas demandan una alta capacidad computacional. La integración de infraestructuras de supercómputo (HPC), IA y quimioinformática dentro del entorno universitario permite realizar cribados masivos y simulaciones avanzadas a una escala sin precedentes. La disponibilidad de clústeres como Miztli y Yoltla permite realizar simulaciones moleculares con niveles de detalle que, hace unas décadas, eran inalcanzables. Lejos de sustituir la validación experimental, estas tecnologías permiten formular hipótesis mejor fundamentadas, optimizar recursos y reducir drásticamente los tiempos de investigación.

El diseño de fármacos ha evolucionado hacia el uso de modelos de aprendizaje profundo para explorar la diversidad química. Estos enfoques incluyen el uso de SMILES y redes neuronales para generar estructuras con propiedades biológicas específicas [2], [3], así como algoritmos evolutivos integrados con dinámica molecular para el diseño automatizado de inhibidores [4]. A diferencia de estas estrategias, enfocadas en un solo blanco o en la generación estocástica, el protocolo CADMA-Chem permite priorizar sistemáticamente moléculas multifuncionales. Al integrar simultáneamente actividad antioxidante, inhibición enzimática y viabilidad farmacocinética, junto con retrosíntesis asistida por IA, este protocolo de diseño establece un puente eficiente hacia la validación experimental.

Por todo lo anterior, el presente trabajo busca explorar cómo la combinación de HPC, quimioinformática e IA puede transformar la manera en que diseñamos fármacos multifuncionales, tomando como ejemplo la identificación y priorización de derivados quinolínicos (dQu) con potencial neuroprotector mediante el protocolo CADMA-Chem.

En este contexto, se describe una aproximación integral que combina infraestructura HPC, IA y quimioinformática para el diseño de potenciales FMF derivados de la quinolina

(dQu). Se destaca el papel de estos candidatos como antioxidantes y como inhibidores de la acetilcolinesterasa (AChE), diana terapéutica por excelencia en la EA [5], demostrando cómo el ecosistema tecnológico potencia la investigación académica hacia soluciones con impacto social.

Infraestructura de supercómputo e integración tecnológica en instituciones de educación superior

El diseño de fármacos asistido por computadora ha evolucionado en las últimas décadas. Lo que antes se limitaba a estudios de acoplamiento (*docking*) con bibliotecas pequeñas, hoy implica la exploración masiva de espacios químicos de hasta incluso millones de estructuras virtuales. Esta expansión ha sido posible gracias al desarrollo de arquitecturas HPC, incorporación de unidades de procesamiento gráfico (GPU) y el desarrollo de software especializado.

Dichas tecnologías permiten ejecutar cálculos intensivos, tales como los basados en la teoría de los funcionales de la densidad (DFT), cribados virtuales masivos *(High-Throughput Virtual Screening)*, simulaciones de dinámica molecular (MD) y estimaciones energéticas con métodos eficientes como MM/GBSA (mecánica molecular / área superficial de Born generalizada). La integración de estos recursos ha incrementado la velocidad de cálculo en varios órdenes de magnitud, reduciendo tiempos de ejecución de semanas o días a sólo unas horas o incluso minutos. No obstante, es vital reconocer que el incremento en potencia computacional no sustituye la evidencia biológica. El valor del supercómputo reside en la capacidad de priorizar candidatos de manera más inteligente y acelerada, disminuyendo la incertidumbre antes de la validación experimental que continúa siendo indispensable.

Para las instituciones de educación superior, la disponibilidad de centros con capacidades HPC constituye un activo estratégico. Además de acelerar proyectos específicos, estas plataformas promueven la formación de recursos humanos en áreas interdisciplinarias y fortalecen la autonomía tecnológica que es necesaria para sostener la investigación de frontera. En este trabajo, el uso de infraestructura de Miztli y Yoltla (LANCAD-UNAM) fue determinante para implementar el protocolo CADMA-Chem a escala, procesando más de ocho mil estructuras y ejecutando simulaciones de dinámica molecular que habrían sido

inviables sin dicha infraestructura. Los recursos computacionales ascendieron a aproximadamente 13,500 h-CPU y 1,300 h-GPU, excluyendo aquellas ejecuciones que no convergieron o presentaron problemas técnicos.

Protocolo CADMA-Chem para el diseño racional de quinolinas multifuncionales

Con el objetivo de diseñar compuestos con potencial actividad multifuncional frente a la EA, se implementó el protocolo computacional CADMA-Chem [6], que permite trasladarse de una molécula progenitora, hacia derivados priorizados de manera sistemática y reproducible.

El punto de partida es la selección de un marco molecular con actividad biológica conocida, en este caso, la quinolina (Fig. 1). Este marco posee gran versatilidad biológica y rutas sintéticas establecidas, lo que lo convierte en un candidato ideal para la generación de derivados con potencial farmacológico [7]. A partir de este núcleo, se realizaron sustituciones sistemáticas con grupos funcionales de interés químico (–OH, –CHO, $–COCH_3$, $–COOCH_3$, –SH, $–NH_2$), utilizando el software SmileIt desarrollado en el grupo de investigación [8]. Este proceso generó un espacio químico estructuralmente diverso, pero controlado, de poco más de 8000 derivados.

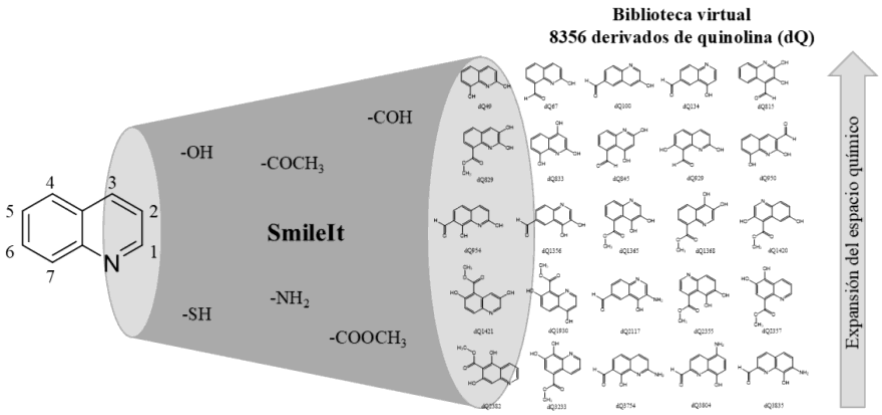


Fig. 1. Generación de la biblioteca virtual de dQu.

El protocolo contempla una primera etapa de filtrado basada en propiedades farmacocinéticas, toxicológicas y de factibilidad sintética. Se estimaron parámetros ADME (absorción, distribución, metabolismo y excreción), toxicidad (dosis letal media, mutagenicidad y toxicidad del desarrollo) y accesibilidad sintética con herramientas quimioinformáticas como RDKit, T.E.S.T., Ambit-SA, entre otros [9]-[11]. Estos criterios permiten descartar compuestos con bajo potencial de desarrollo. La elección de los candidatos con las propiedades farmacológicas más prometedoras se realizó mediante índices de selección y eliminación, que combinan los descriptores mencionados en una métrica cuantitativa que compara cada derivado contra un conjunto de referencia de fármacos neuroprotectores comerciales, descartando automáticamente aquellos con perfil farmacocinético inadecuado o con valores de toxicidad inaceptables.

En una segunda etapa, se evaluaron la energía de ionización (IP) y la energía de disociación de enlace mediante cálculos DFT implementados en Gaussian 16 [12]. Estos parámetros están relacionados con mecanismos de donación de electrones (SET) o de transferencias de átomo de hidrógeno (HAT), procesos que, a su vez, dan cuenta de su actividad antioxidante. Los resultados se presentan en la Fig. 2, en un mapa conocido como eH-DAMA (*electron and hydrogen donating ability maps*), una herramienta visual que permite identificar de forma intuitiva qué candidatos superan en eficacia a antioxidantes estándar como la vitamina E, vitamina C o al Trolox frente a especies oxidantes (como el radical hidroperoxilo, $OOH^{\bullet}$).

Posteriormente, se incorporó la evaluación de actividad poligénica frente a enzimas relevantes en la neurotransmisión (Catecol-o-metiltransferasa, "COMT", Monoaminooxidasa-B "MAO-B" y AChE) y en la fisiopatología de enfermedades neurodegenerativas como EA y Parkinson mediante simulaciones de *docking* molecular de los 20 mejores candidatos implementados con el software Autodock Vina [13]. Se utilizaron las estructuras de origen humano de COMT (PDB Id: 3S68 en complejo con Mg^{2+}, S-adenosilmetionina y tolcapone), MAO-B (PDB Id: 2V5Z en complejo con flavin-adenin dinucleótido y safinamida) y AChE (PDB Id: 4EY7 en complejo con donepezilo).

Finalmente, tras la priorización de candidatos, se encontró que los compuestos diseñados tienen mayor potencial para inhibir a la AChE en comparación con las otras enzimas.

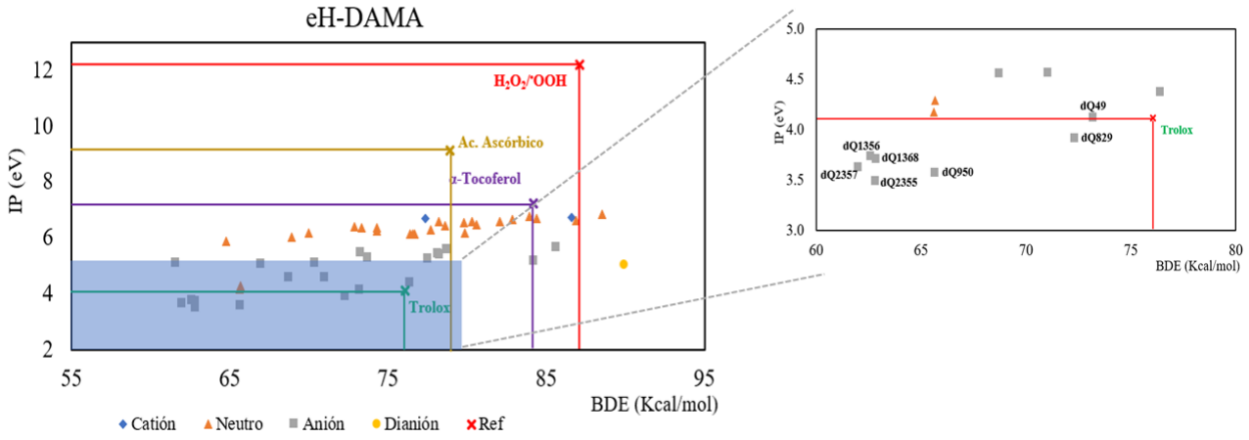


Fig. 2. Mapa eH-DAMA de los mejores 20 candidatos. Los ejes representan los potenciales de ionización (IP) y las energías de disociación de enlace (BDE). El recuadro destaca los derivados con un perfil antioxidante superior a los estándares comerciales.

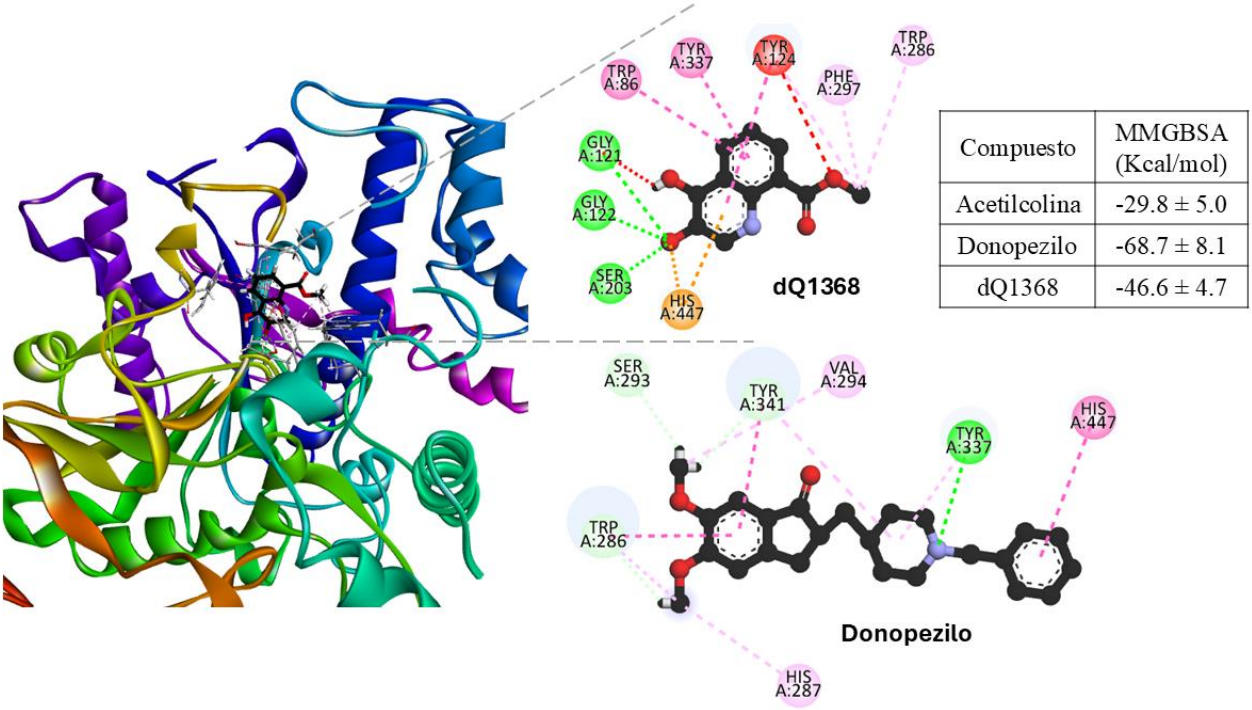
La integración secuencial de estas etapas logra reducir el conjunto de 8356 derivados a 5 compuestos que equilibran viabilidad farmacocinética, potencial antioxidante y actividad inhibitoria de AChE, lo que los convierte en candidatos a FMF contra la EA [14].

Validación dinámica en HPC: simulaciones de dinámica molecular y análisis MM/GBSA

Utilizando la suite Schroedinger®, se realizaron simulaciones MD de los complejos quinolina-AChE seleccionados para evaluar su estabilidad, conformación e interacciones a lo largo del tiempo (200 ns). Las simulaciones de estos sistemas de miles de átomos implican una alta demanda computacional. La infraestructura HPC, en particular el uso de GPU, permite ejecutarlas, optimizando tiempos de procesamiento y posibilitando el análisis de múltiples sistemas bajo condiciones comparables. La incorporación de GPU fue determinante para alcanzar los tiempos de simulación sin comprometer el tamaño del sistema ni la calidad del muestreo conformacional. Esta etapa resulta crítica, ya que la dinámica molecular permite analizar el sistema en condiciones más cercanas al entorno fisiológico.

El análisis se centró en la entrada y la región de "la garganta" del sitio activo de la AChE, regiones que controlan el acceso al centro catalítico [15]. Algunos de los derivados

mantuvieron interacciones persistentes mediante enlaces de hidrógeno, apilamientos aromáticos y distintos contactos hidrofóbicos, sugiriendo una posible oclusión funcional similar a la observada en inhibidores de referencia como el donepezilo (Fig. 3) [15].



Compuesto	MMGBSA (Kcal/mol)
Acetilcolina	-29.8 ± 5.0
Donopezilo	-68.7 ± 8.1
dQ1368	-46.6 ± 4.7

Fig. 3. Diagramas 3D (izquierda) y 2D (centro arriba) de la AChE ocluida con dQ1368. Diagrama de 2D donepezilo (centro abajo) y tabla con energías MMGBSA (derecha).

Para complementar el análisis, se estimaron energías libres de unión mediante el método MM/GBSA. Las afinidades calculadas mostraron correlación con los puntajes obtenidos en el docking del protocolo, validando la aplicabilidad y consistencia interna del flujo computacional. Entre los compuestos evaluados, dQ1368 evidenció mayor estabilidad posicional y mejor perfil energético, permaneciendo cercano a su configuración inicial durante la mayor parte de la trayectoria simulada y respaldando su potencial actividad como inhibidor de la AChE [16]. Este derivado es el candidato más prometedor debido a su notable estabilidad en el sitio activo de la AChE. Destaca su mimetismo con donepezilo, estableciendo interacciones aromáticas con residuos clave de la enzima (Trp86 y Tyr337), permitiendo una oclusión efectiva del sitio activo. Además, su capacidad para formar enlaces de hidrógeno

cerca de la tríada catalítica y su diseño como ligando multifuncional lo posicionan como un potencial FMF contra la EA.

Aunque la MD aporta un nivel de precisión superior al *docking*, los resultados continúan siendo predicciones. El valor del supercómputo radica en permitir la priorización con más información de los candidatos, disminuyendo la incertidumbre antes de la validación experimental. En un entorno académico, esta capacidad permite optimizar recursos, enfocar esfuerzos sintéticos y fortalecer la toma de decisiones.

Evaluación de viabilidad sintética y análisis de retrosíntesis asistida por IA

El diseño racional de compuestos con propiedades farmacológicas favorables posee poca utilidad si no se acompaña de un análisis de factibilidad sintética. El protocolo incorpora la evaluación de accesibilidad sintética como criterio estructural de priorización y fue estimada con descriptores que consideran complejidad estructural, fragmentos reactivos y disponibilidad comercial de los reactivos. Este análisis permite descartar estructuras con alta dificultad sintética, favoreciendo derivados compatibles con rutas orgánicas convencionales y disminuyendo los costos experimentales.

Para los 5 candidatos elegidos, se realizó un análisis de retrosíntesis asistida por IA con la plataforma IBM RXN for Chemistry [17], la cual emplea modelos de aprendizaje profundo entrenados en grandes bases de datos de reacciones químicas. Se identificaron rutas sintéticas cortas, del orden de cuatro etapas, que emplearon reactivos, disolventes y catalizadores de uso comercial, incluyendo alternativas compatibles con procedimientos tradicionales y con esquemas automatizables. La incorporación de retrosíntesis representa un puente entre el diseño virtual y la química experimental. A modo de ejemplo, se presenta la retrosíntesis de dQ1368 (Fig. 4). En un entorno académico, esta integración permite cerrar el ciclo “diseño-evaluación-síntesis” de manera coherente, reduciendo la brecha entre modelado molecular y validación en laboratorio. Este ciclo permitirá la síntesis y evaluación biológica de los 5 mejores candidatos.

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

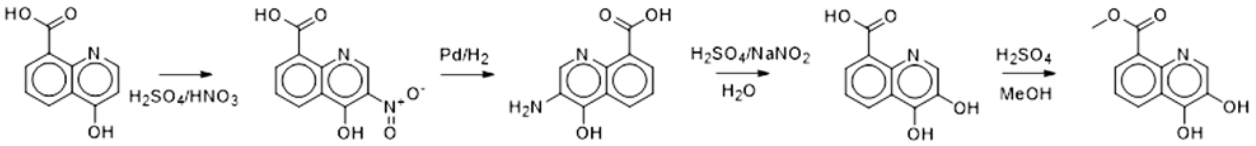


Fig. 4. Retrosíntesis del compuesto dQ1368.

Impacto del supercómputo y retos de validación

La integración de HPC, IA, quimioinformática y modelado molecular en el diseño de FMF permite explorar espacios químicos extensos, evaluar múltiples propiedades simultáneamente y analizar sistemas complejos con mayor profundidad y eficiencia. En el entorno académico, esta aproximación fomenta la multidisciplinariedad, ya que requiere la convergencia de química orgánica, química teórica, bioquímica, modelado molecular, ciencia de datos e inteligencia artificial, fortaleciendo la formación de recursos humanos capaces de integrar el entorno computacional y la validación experimental.

Por otro lado, es indispensable reconocer los límites inherentes a los modelos computacionales. Las predicciones de *docking*, MD y MM/GBSA, aunque robustas, siguen siendo estimaciones: dependen de supuestos fisicoquímicos y simplificaciones del entorno biológico. Por ello, y en el contexto de este trabajo, el supercómputo debe concebirse como una herramienta de priorización y de generación de hipótesis, no como sustituto de la validación experimental. El reto inmediato es trasladar los candidatos seleccionados, como el dQ1368, a etapas experimentales que incluyan síntesis y evaluación biológica. De esta manera, se consolida en las universidades un entorno donde infraestructura tecnológica, investigación interdisciplinaria y formación académica converjan para enfrentar problemas de salud pública.

Conclusiones

La combinación de quimioinformática, evaluación farmacológica temprana, modelado molecular y simulaciones MD permitió estructurar un flujo de trabajo coherente para la priorización de derivados quinolínicos con potencial como FMF.

El protocolo CADMA-Chem demostró su utilidad para reducir un espacio químico de 8356 derivados a 5 candidatos con un perfil equilibrado de viabilidad farmacocinética, accesibilidad sintética, actividad antioxidante y potencial inhibitorio contra la AChE.

La validación mediante MD y estimaciones MM/GBSA refinó y respaldó los resultados del *docking*, permitiendo la identificación del derivado dQ1368 como el candidato más prometedor a FMF contra la EA.

El análisis retrosintético posibilita enlazar los resultados computacionales con el ámbito experimental al proponer rutas de síntesis plausibles mediante el uso de herramientas asistidas por IA.

En conjunto, la integración de supercómputo, inteligencia artificial y diseño racional de fármacos demuestra que las IES mexicanas cuentan con las capacidades tecnológicas y humanas para producir ciencia de frontera con impacto social, formación de recursos humanos y desarrollo de soluciones terapéuticas innovadoras.

Agradecimientos

Al proyecto SECIHTI CBF2023-2024-1141, por el financiamiento, y al proyecto LANCAD-DGTIC-UNAM-352, por los recursos de supercómputo. LFHA (CVU:385666) agradece a la SECIHTI por la beca de Estancias Posdoctorales por México 2022(1). A EGGL (CVU 649128), por la beca de posgrado 762905.

Referencias

[1] M. C. Carreiras, E. Mendes, M. J. Perry, A. P. Francisco, and J. Marco-Contelles, "The multifactorial nature of Alzheimer's disease for developing potential therapeutics," Curr. Top. Med. Chem., vol. 13, no. 15, pp. 1745-1770, 2013, doi: 10.2174/15680266113139990135.

[2] R. Özçelik, H. Brinkmann, E. Criscuolo, and F. Grisoni, "Generative deep learning for de novo drug design—A chemical space odyssey," J. Chem. Inf. Model., vol. 65, no. 14, pp. 7352-7372, 2025, doi: 10.1021/acs.jcim.5c00641.

[3] F. Frausto-Parada, I. Vargas-Rodríguez, I. Mercado-Sánchez, A. Bazán-Jiménez, E. Díaz-Cervantes, M. A. Sotelo-Figueroa, and M. A. García-Revilla, "Grammatical evolution-based design of SARS-CoV-2 main protease inhibitors," Phys. Chem. Chem. Phys., vol. 24, pp. 5233-5245, 2022, doi: 10.1039/d1cp04159b.

[4] W. Kong, Y. Hu, J. Zhang, and Q. Tan, "Application of SMILES-based molecular generative model in new drug design," Front. Pharmacol., vol. 13, Art. no. 1046524, 2022, doi: 10.3389/fphar.2022.1046524.

[5] H. Hampel, M. M. Mesulam, A. C. Cuello, M. R. Farlow, E. Giacobini, G. T. Grossberg, A. S. Khachaturian, A. Vergallo, E. Cavedo, P. J. Snyder, and Z. S. Khachaturian, "The cholinergic system in the pathophysiology and treatment of Alzheimer's disease," Brain, vol. 141, no. 7, pp. 1917-1933, 2018, doi: 10.1093/brain/awy132.

[6] E. G. Guzmán-López, M. Reina, A. Pérez-González, M. Francisco-Márquez, L. F. Hernández-Ayala, R. Castañeda-Arriaga, and A. Galano, "CADMA-Chem: A Computational Protocol Based on Chemical Properties Aimed to Design Multifunctional Antioxidants," Int. J. Mol. Sci., vol. 23, no. 21, Art. no. 13246, 2022, doi: 10.3390/ijms232113246.

[7] L. F. E. Moor, T. R. A. Vasconcelos, R. da R. Reis, L. S. S. Pinto, and T. M. da Costa, "Quinoline: An Attractive Scaffold in Drug Design," Mini Rev. Med. Chem., vol. 21, no. 16, pp. 2209-2226, 2021, doi: 10.2174/1389557521666210210155908.

[8] A. Galano, "SmileIt 1.0 JAVA." [Online]. Available: https://agalano.com/smile-it/ [Accessed: Mar. 2024].

[9] G. Landrum, "RDKit: Open-source cheminformatics software." [Online]. Available: https://www.rdkit.org/ [Accessed: Mar. 2024].

[10] T. Martin, Toxicity Estimation Software Tool (TEST). Washington, DC, USA: U.S. Environmental Protection Agency, 2016.

[11] N. Kochev, S. Avramova, P. Angelov, and N. Jeliazkova, "Computational Prediction of Synthetic Accessibility of Organic Molecules with Ambit-Synthetic Accessibility Tool," Org. Chem. Ind. J., vol. 12, no. 2, Art. no. 123, 2018.

[12] M. J. Frisch et al., Gaussian 16, Revision C.01. Wallingford, CT, USA: Gaussian, Inc., 2016.

[13] Eberhardt, D. Santos-Martins, A. F. Tillack, and S. Forli, "AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings," J. Chem. Inf. Model., vol. 61, pp. 3891-3898, 2021, doi: 10.1021/acs.jcim.1c00203.

[14] L. F. Hernández-Ayala, E. G. Guzmán-López, and A. Galano, "Quinoline Derivatives: Promising Antioxidants with Neuroprotective Potential," Antioxidants (Basel), vol. 12, no. 10, Art. no. 1853, 2023, doi: 10.3390/antiox12101853.

[15] M. A. Silva, A. S. Kiametis, and W. Treptow, "Donepezil inhibits acetylcholinesterase via multiple binding modes at room temperature," J. Chem. Inf. Model., vol. 60, no. 7, pp. 3463-3471, 2020, doi: 10.1021/acs.jcim.9b01073.

[16] M. Prejanò, I. Romeo, L. F. Hernández-Ayala, G. E. Guzmán-López, S. Alcaro, A. Galano, and T. Marino, "Evaluating Quinolines: Molecular Dynamics Approach to Assess Their Potential as Acetylcholinesterase Inhibitors for Alzheimer's Disease," ChemPhysChem., vol. 26, no. 1, Art. no. e202400653, 2025, doi: 10.1002/cphc.202400653.

[17] IBM RXN for Chemistry [Online]. Available: https://rxn.app.accelerate.science/rxn/ [Accessed: Nov. 8, 2024).

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

VirDetect-AI: nueva herramienta de inteligencia artificial para identificar proteínas virales eucariontes en datos metagenómicos

Alida Esmeralda Zárate Jiménez
Universidad Nacional Autónoma de México, Instituto de Biotecnología, Cuernavaca, Morelos, México.
ORCID: 0009-0006-5407-6598

Blanca Itzel Taboada Ramírez
Universidad Nacional Autónoma de México, Instituto de Biotecnología, Cuernavaca, Morelos, México.
ORCID: 0000-0003-1896-5962

Lorena Díaz-González
Universidad Autónoma del Estado de Morelos, Centro de Investigación en Ciencias, Cuernavaca, Morelos, México.
ORCID: 0000-0003-1577-5629

Recepción: 01 de marzo de 2026.
Aceptación: 07 de mayo de 2026.

Mayo 2026 • número de revista 15 • DOI: 10.22201/dgtic.26832968e.2025.15.156

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

VirDetect-AI: nueva herramienta de inteligencia artificial para identificar proteínas virales eucariontes en datos metagenómicos

Resumen

Identificar secuencias de virus en muestras metagenómicas de origen ambiental, animal o humano representa un gran reto científico. Los virus son extremadamente diversos, evolucionan rápidamente; además, muchos de ellos no cuentan con genomas de referencia en las bases de datos actuales, lo que dificulta su detección mediante métodos tradicionales. Si bien, en años recientes, se han incorporado herramientas basadas en inteligencia artificial, la mayoría operan de forma binaria, distinguen sólo entre secuencias virales y no virales, y se basan en información genética a nivel de nucleótidos.

En respuesta a este escenario, se desarrolló VirDetect-AI, una herramienta basada en inteligencia artificial diseñada para identificar secuencias de virus eucariontes a partir de sus secuencias proteicas. En este trabajo, se describe su diseño y flujo de trabajo, desde la construcción del conjunto de datos y el procesamiento de secuencias hasta la implementación del modelo de aprendizaje profundo utilizando redes neuronales convolucionales y bloques residuales. El sistema reconoce patrones discriminativos asociados con dominios y motivos en las secuencias de proteínas virales. Asimismo, esta herramienta logra clasificar secuencias metagenómicas en 979 clases de proteínas virales con alta precisión, lo que amplía las posibilidades para explorar la diversidad viral, descubrir virus previamente no descritos y fortalecer la vigilancia ecológica y de salud pública.

Palabras Clave: VirDetect-AI, inteligencia artificial, metagenómica viral, redes neuronales convolucionales, redes neuronales residuales, clasificación viral.

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

VirDetect-AI: A Novel Artificial Intelligence Tool for Identifying Eukaryotic Viral Proteins in Metagenomic Data

Abstract

Identifying viral sequences in metagenomic samples from environmental, animal, or human sources represents a major scientific challenge. Viruses are extremely diverse and evolve rapidly; moreover, many lack reference genomes in current databases, which hinders their detection using traditional methods. Although artificial intelligence-based tools have been introduced in recent years, most operate in a binary manner, distinguishing only between viral and non-viral sequences, and rely primarily on nucleotide-level genetic information.

In response to this scenario, VirDetect-AI was developed, an artificial intelligence-based tool designed to identify eukaryotic viral sequences from protein sequences. In this work, its design and workflow are described, from dataset construction and sequence processing to the implementation of the deep learning model. By leveraging convolutional neural networks and residual blocks, VirDetect-AI recognizes discriminative patterns associated with domains and motifs in viral protein sequences. The tool is capable of classifying metagenomic sequences into 979 viral protein classes with high accuracy, expanding the possibilities for exploring viral diversity, discovering previously undescribed viruses, and strengthening ecological and public health surveillance.

Keywords: *VirDetect-AI, artificial intelligence, viral metagenomics, convolutional neural networks, residual neural networks, viral classification.*

La diversidad viral invisible: un reto para la ciencia moderna

Los virus están presentes en prácticamente todos los ecosistemas del planeta, incluyendo océanos, suelos, plantas, animales y el cuerpo humano. Sin embargo, a pesar de su abundancia y relevancia ecológica, médica y evolutiva, una gran fracción de los virus que nos rodean sigue siendo desconocida. Esta brecha de conocimiento se ha vuelto especialmente evidente con el auge de la metagenómica, una disciplina que permite estudiar el material

genético directamente desde muestras ambientales, clínicas o biológicas, sin necesidad de cultivar los organismos en el laboratorio [1].

Los virus pueden clasificarse de acuerdo con el tipo de hospedero que invadan. Por ejemplo, los bacteriófagos infectan bacterias, mientras que otros virus pueden encontrarse en organismos eucariontes, como plantas, animales, hongos, insectos y humanos [2]. En este trabajo, y con fines de claridad terminológica, éstos últimos se denominarán en adelante "virus eucariontes". Debido a su vasta diversidad genética y a su relevancia en los ámbitos médico, agrícola y veterinario, los virus eucariontes representan uno de los mayores retos para los sistemas de clasificación y análisis computacional [3].

En el contexto de la metagenómica, esta diversidad se traduce en un obstáculo crítico. Aunque esta tecnología permite secuenciar millones de fragmentos genéticos prevenientes de un ambiente o de un hospedero, identificar a qué organismos pertenecen esas secuencias sigue siendo un desafío mayor. Actualmente, se estima que entre el 40% y el 90% de las secuencias metagenómicas obtenidas no logran ser asignadas taxonómicamente. Esto impide conocer la naturaleza biológica de una parte importante de los datos. A este conjunto de secuencias que no son clasificadas por los métodos convencionales, se le conoce como materia oscura viral (viral dark matter), la cual conforma un universo de virus potencialmente nuevos aún por descubrir [4].

¿Por qué es difícil identificar virus?

Se estima que existen del orden de 10^{31} viriones en el planeta Tierra, una cifra que representa la totalidad de las partículas virales individuales presentes en distintos ambientes y correspondientes a una gran diversidad de virus. Sin embargo, sólo una fracción mínima de esta diversidad ha sido caracterizada hasta la fecha [5]. Dicha brecha en el conocimiento se manifiesta claramente en los estudios metagenómicos, donde la enorme proporción de materia oscura viral evidencia que muchos virus carecen de genomas de referencia en las bases de datos públicas. Como consecuencia, la ausencia de estos genomas reduce la efectividad de métodos computacionales tradicionales, particularmente aquellos basados en la comparación directa de secuencias. Los enfoques de alineamiento, que incluyen herramientas como BLAST [6], Bowtie [7], SOAP [8] y BWA [9], dependen en gran medida de la similitud con secuencias analizadas y aquellas previamente anotadas. Esta dependencia

limita su capacidad para detectar virus altamente divergentes. Además, herramientas como BLAST, aunque ampliamente utilizadas, pueden requerir tiempos de ejecución elevados cuando se aplican a grandes volúmenes de datos metagenómicos.

Por otro lado, los enfoques basados en perfiles, como los modelos ocultos de Markov (HMM), permiten detectar identidades más distantes mediante modelos probabilísticos construidos a partir de alineamientos múltiples de proteínas. Herramientas como Pfam [10] y HMMER [11] utilizan este tipo de estrategia. No obstante, su desempeño también depende de la disponibilidad, calidad y cobertura de las bases de datos de referencia, lo que limita la identificación de virus previamente no descritos.

El auge de la inteligencia artificial y el análisis proteico

Ante estas limitaciones, los enfoques basados en inteligencia artificial han surgido como una alternativa para analizar patrones en secuencias biológicas sin depender exclusivamente de la similitud directa con referencias conocidas. No obstante, gran parte de las herramientas actuales operan como clasificadores binarios, es decir, distinguen únicamente entre secuencias virales y no virales, o se limitan a un número reducido de clases, menos de 6 [12]. Asimismo, muchas de estas herramientas se enfocan principalmente en bacteriófagos [13], [14], [15] o en muestras humanas [12], [16], [17], [18], [19], [20], y suelen trabajar a nivel de la secuencia de nucleótidos, lo que reduce su capacidad para identificar virus con baja similitud genómica.

Para solventar este problema, el análisis de proteínas ofrece una ventaja importante. A nivel de aminoácidos, es posible detectar señales evolutivas más conservadas, como dominios funcionales y motivos característicos, que suelen mantenerse incluso cuando la identidad nucleotídica es baja. En este contexto, el análisis proteico permite extender la detección hacia virus más divergentes y fortalecer la clasificación taxonómica de proteínas virales eucariontes.

En este contexto, VirDetect-AI se propone como una herramienta computacional avanzada, diseñada para superar algunas de estas limitaciones. El modelo combina redes neuronales convolucionales (CNNs) [21] y bloques residuales (ResNets) [22] para extraer características jerárquicas y patrones complejos a partir de secuencias proteicas. Aunque el

potencial predictivo de VirDetect-AI ya fue reportado previamente [23], el presente artículo tiene un enfoque distinto, el explicar su diseño y flujo de trabajo de una manera aplicada. Se describen sus principales etapas, desde la construcción del conjunto de datos hasta su aplicación en datos metagenómicos reales, con el objetivo de facilitar la transferencia de conocimiento hacia audiencias interesadas en inteligencia artificial aplicada y fomentar el desarrollo de nuevas herramientas de análisis de secuencias genómicas.

Construyendo VirDetect-AI: datos, biología e inteligencia artificial

La metodología de VirDetect-AI, ilustrada en la Fig. 1, implicó trabajar con volúmenes masivos de datos biológicos. Inicialmente, se descargaron 9.9 millones de secuencias de proteínas virales de la base de datos de NCBI (National Center for Biotechnology Information). Después de eliminar las secuencias redundantes con un umbral de 98% de identidad, se obtuvo un conjunto de más de 2.2 millones de proteínas virales, las cuales fueron divididas en 2 conjuntos principales: uno correspondiente a bacteriófagos y otro a virus eucariontes, con un total de 1,013,722 secuencias (Fig. 1, panel superior).

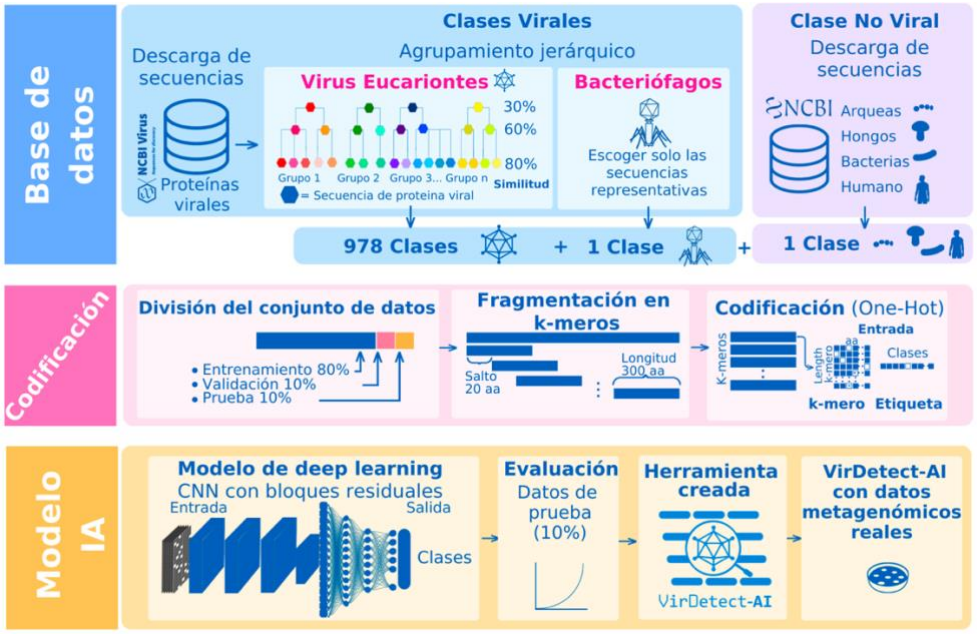


Fig. 1. Flujo de trabajo de VirDetect-AI para la identificación de proteínas virales eucariontes en datos metagenómicos. El *pipeline* comprende: (i) construcción de una base de datos a partir de secuencias virales y no virales; (ii) agrupamiento jerárquico (*hierarchical clustering*) para definir 978 clases virales eucariontes, una clase viral correspondiente a bacteriófagos y una clase negativa; (iii) división del conjunto de datos en entrenamiento, validación y prueba; (iv) fragmentación de secuencias en k-meros y su codificación mediante *one-hot encoding*; y (v) entrenamiento y evaluación de un modelo de *deep learning* basado en redes CNN con bloques residuales, el cual puede aplicarse a datos metagenómicos reales para la detección de proteínas virales.

Este conjunto de virus eucariontes se procesó mediante un enfoque de agrupamiento jerárquico (hierarchical clustering), aplicando umbrales sucesivos de similitud del 80%, 60% y 30%. La estrategia permitió definir grupos con una identidad mínima del 30% entre secuencias. Con este procedimiento, se definieron 978 clases de proteínas virales eucariontes asociadas con información de familia, género, función proteica y hospedero. Estas clases abarcan 86 diferentes familias virales y una amplia diversidad de hospederos, incluyendo humanos, animales, y plantas, entre otros.

No obstante, se observa una mayor representación de clases asociadas a Homo sapiens (38.4%) y a familias virales ampliamente estudiadas como Orthomyxoviridae, Retroviridae, Flaviviridae y Coronaviridae. Este sesgo no se atribuye al método de clasificación, sino a la distribución de las secuencias disponibles en las bases de datos públicas, donde los virus de relevancia médica suelen estar mejor representados. La definición de estas clases permite una clasificación más detallada que facilita la interpretación biológica de los resultados. Adicionalmente, se incluyó una clase para la identificación de bacteriófagos, integrada por 6,464 secuencias, muestreadas aleatoriamente por familia del conjunto original y una clase negativa no viral, compuesta por 7,551 secuencias de origen humano, bacteriano, fúngico y arqueas. Finalmente, para el entrenamiento y evaluación del modelo, los datos se dividieron en tres conjuntos: 80% para entrenamiento, 10% para validación y 10% para prueba.

Fragmentación de secuencias y codificación de los datos

Para preparar los datos de entrada de VirDetect-AI, las secuencias proteicas se fragmentaron en k-meros de 300 aminoácidos (aa), con un desplazamiento de 20 aa entre fragmentos consecutivos. Esta estrategia generó más de 19 millones de fragmentos para el entrenamiento y evaluación del modelo. El tamaño del k-mero y su desplazamiento se seleccionaron con base en pruebas experimentales, considerando tanto la necesidad de conservar información biológica suficiente como la utilidad del modelo en datos metagenómicos reales.

Los k-meros de mayor longitud permiten capturar patrones distribuidos en regiones más amplias de la secuencia e incorporar múltiples regiones funcionales dentro de un mismo fragmento. Sin embargo, utilizar fragmentos demasiados largos, como aquellos mayores a

1,000 nucleótidos empleados en algunos trabajos previos de identificación viral basada en IA [13,15], puede limitar la aplicación del modelo a secuencias más cortas que son frecuentes en datos mutagénicos reales. Esto puede ocasionar que una parte importante de la información disponible no sea utilizada. Así, el uso de k-meros de 300 aminoácidos permite conservar contexto biológico suficiente sin excluir una proporción importante de los fragmentos cortos presentes en datos metagenómicos reales.

Por otra parte, el solapamiento entre fragmentos ayuda a reducir los efectos de borde, es decir, disminuye la posibilidad de que motivos biológicos relevantes queden divididos entre los extremos de fragmentos consecutivos. Esto es particularmente importante en arquitecturas CNN, dado que los filtros detectan patrones locales dentro de regiones específicas de la matriz de entrada. Además, el solapamiento incrementa el número de ejemplos de entrenamiento y contribuye a mejorar la capacidad de generalización del modelo.

Posteriormente, cada k-mero se transformó en una matriz binaria de 300 × 26 mediante codificación one-hot, convirtiendo la información biológica en una representación numérica adecuada para el aprendizaje profundo (Fig.1, panel intermedio). Aunque existen representaciones más complejas, como embeddings aprendidos o preentrenados, la codificación one-hot no es una elección arbitraria. Se trata de una estrategia ampliamente utilizada para transformar secuencias de proteínas en representaciones adecuadas para modelos de aprendizaje profundo, ya que no impone relaciones previas entre aminoácidos y permite que el modelo aprenda directamente a partir de los patrones presentes en las secuencias de referencia utilizadas. Además, su simplicidad la hace adecuada para arquitecturas basadas en CNN, ya que éstas explotan relaciones locales mediante filtros aplicados sobre campos receptivos definidos [24].

Arquitectura del modelo VirDetect-AI

La arquitectura de VirDetect-AI se ilustra en la Fig. 2. El modelo fue entrenado para realizar una clasificación multiclase compuesta por 978 clases de virus eucariontes, una clase correspondiente a bacteriófagos y una clase negativa no viral. Para cada predicción, el modelo produce un valor de probabilidad que refleja la confianza de la asignación. Este diseño permite no sólo identificar secuencias virales conocidas, sino también señalar

secuencias potencialmente novedosas cuando presentan patrones compatibles con proteínas virales, aun cuando su similitud con referencias conocidas sea baja.

La arquitectura híbrida, basada en redes CNN con bloques residuales, fue seleccionada con base en criterios teóricos y evidencia experimental. Las CNN permiten capturar patrones locales y motivos conservados en secuencias biológicas, mientras que las conexiones residuales favorecen el flujo de información entre capas y ayudan a prevenir la degradación del desempeño en modelos profundos. Esto permite construir redes más complejas sin perder capacidad de aprendizaje.

Además, en el estudio de ablación reportado por Zárate *et al*. [23], se observó que la reducción o eliminación de componentes clave del modelo disminuyó su desempeño. En particular, la eliminación de los bloques residuales redujo la precisión y la sensibilidad, mientras que su eliminación completa ocasionó la mayor degradación del modelo. De manera similar, la eliminación de la clase negativa no viral y la reducción del número o tamaño de los filtros también afectaron negativamente el desempeño. En contraste, modificar las capas completamente conectadas tuvo un impacto menor, principalmente sobre la certeza de las predicciones. En conjunto, estos resultados respaldan la selección de la arquitectura final de VirDetect-AI.

El papel del supercómputo

El entrenamiento de VirDetect-AI representó un desafío computacional considerable. El modelo ajustó más de 27 millones de parámetros y fue entrenado durante más de cinco días, procesando millones de k-meros a lo largo de múltiples épocas de aprendizaje (Fig. 2).

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

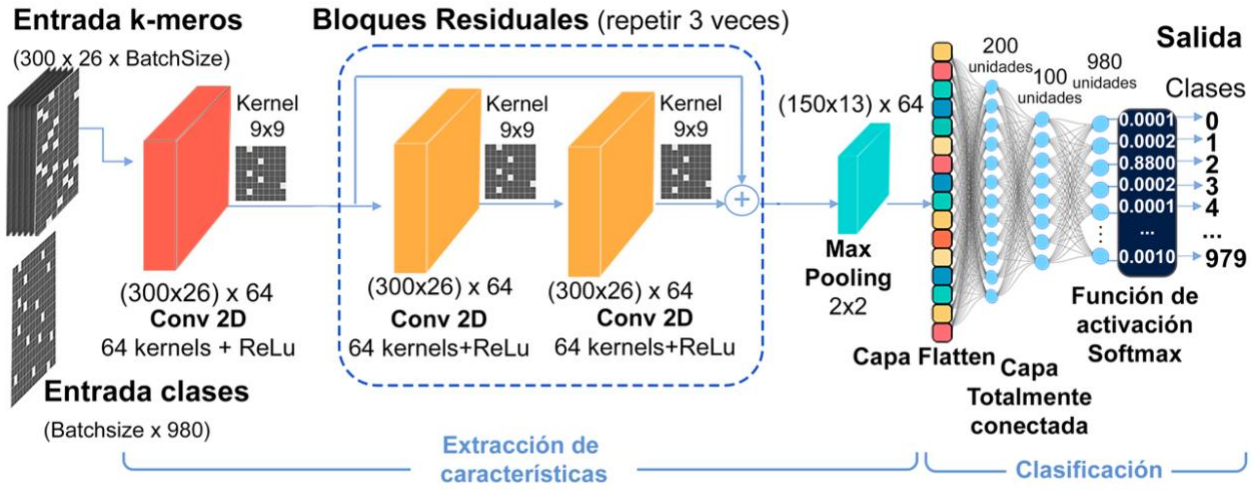


Fig. 2. Estructura del modelo de aprendizaje profundo VirDetect-AI, que integra una arquitectura de CNN con bloques residuales.

Debido a la magnitud del conjunto de datos y a la complejidad de la arquitectura, los cálculos se realizaron utilizando la supercomputadora Miztli, operada por la Dirección General de Cómputo y de Tecnologías de Información y Comunicación (DGTIC) de la UNAM. El entrenamiento fue acelerado mediante unidades de procesamiento gráfico (GPU NVIDIA Tesla V100-16 GB), lo que permitió ejecutar de manera eficiente las operaciones matriciales necesarias para optimizar los parámetros del modelo.

Este trabajo ejemplifica cómo el supercómputo académico se ha convertido en un componente esencial para la investigación científica moderna. En particular, permite abordar problemas biológicos de gran escala que serían impracticables con recursos computacionales convencionales. Así, VirDetect-AI no sólo representa una herramienta de inteligencia artificial aplicada a la virología, sino también un ejemplo de cómo la infraestructura tecnológica universitaria puede impulsar el desarrollo de soluciones innovadoras para el análisis de datos biológicos complejos.

TiES Revista de Tecnología e Innovación en Educación Superior
www.ties.unam.mx

Resultados

Resultados del entrenamiento y del conjunto de prueba

El modelo VirDetect-AI mostró un desempeño alto durante el entrenamiento y la validación. En el conjunto de entrenamiento, obtuvo una exactitud de 0.99 y una pérdida de 0.02, mientras que, en el conjunto de validación, alcanzó una exactitud de 0.99 y una pérdida de 0.03 (Fig. 3). Estos resultados indican que el modelo aprendió de manera adecuada los patrones presentes en las secuencias utilizadas para su entrenamiento, sin mostrar una pérdida importante de desempeño al evaluarse con datos no vistos durante esta etapa.

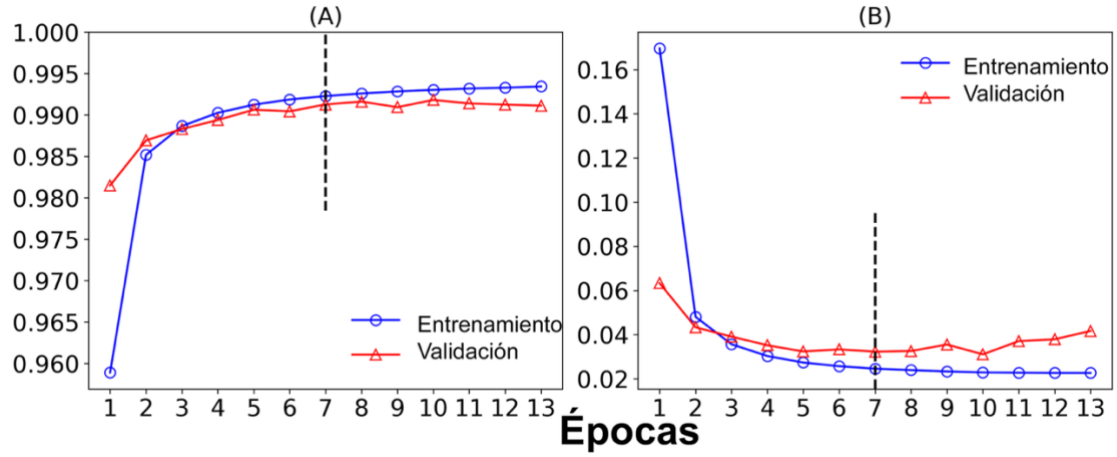


Fig. 3. Rendimiento del entrenamiento y validación de VirDetect-AI en el conjunto de datos de prueba (*Test*). (A) Exactitud del modelo (B) Comportamiento de la función de perdida (*loss*).

Posteriormente, el modelo fue evaluado con un conjunto de pruebas independiente que no fue utilizado durante el entrenamiento. En esta evaluación, VirDetect-AI alcanzó una precisión de 0.98, una sensibilidad de 0.97 y un F1-score de 0.98. Además, el coeficiente de correlación de Matthews (MCC) alcanzó un valor de 0.99, lo que indica un desempeño robusto incluso en un escenario con clases desbalanceadas (Fig. 4).

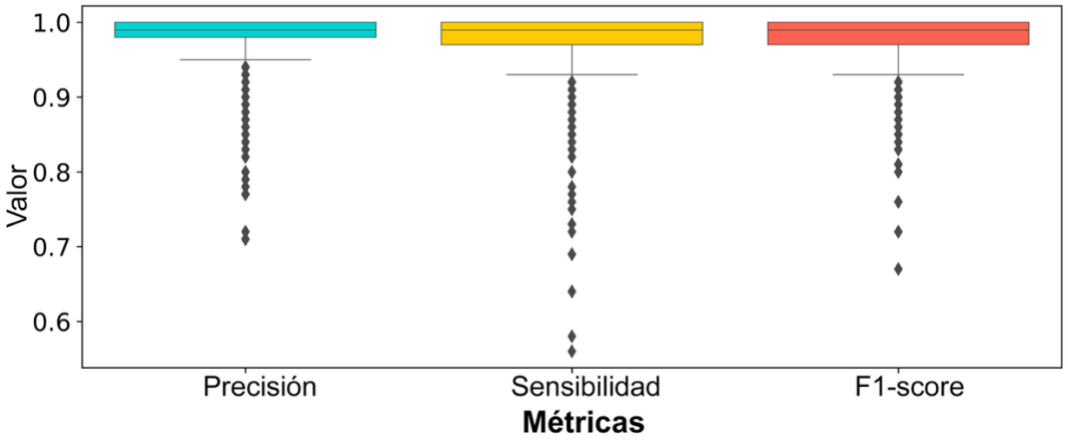


Fig. 4. Rendimiento de VirDetect-AI en el conjunto de datos de Prueba (*Test*) para las 980 clases evaluadas.

De manera adicional, el 94.4% de las clases presentó una sensibilidad superior a 0.9, mientras que sólo el 1.3% presentó valores por debajo de 0.7. Esto sugiere que VirDetect-AI no sólo tuvo un buen desempeño global, sino que también mantuvo una capacidad de calificación consistente para la mayoría de las clases evaluadas.

Desarrollo de la herramienta VirDetect-AI

A partir del modelo entrenado, se desarrolló la herramienta VirDetect-AI, diseñada para facilitar la identificación de proteínas virales eucariontes en secuencias metagenómicas (Fig. 5). La herramienta fue implementada en un repositorio público de GitHub [25], el cual incluye una versión local instalable y una *notebook* que permite realizar el análisis de manera guiada, desde la carga de archivos en formato FASTA hasta la obtención de reportes finales de clasificación. Esta implementación busca que el modelo pueda ser utilizado no sólo por especialistas en aprendizaje profundo, sino también por usuarios interesados en el análisis de datos metagenómicos virales.

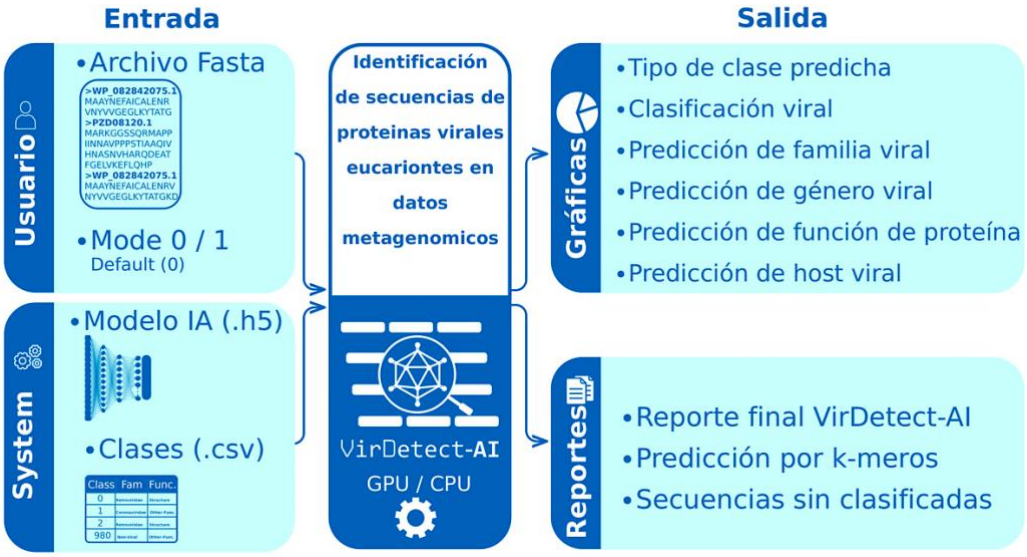


Fig. 5. Flujo general de predicción de la herramienta VirDetect-AI.

Evaluación de VirDetect-AI en datos metagenómicos reales

Con el objetivo de evaluar el desempeño de VirDetect-AI en escenarios más cercanos a su aplicación práctica, el modelo se probó con conjuntos de datos metagenómicos derivados de estudios clínicos humanos, así como con conjuntos negativos de origen humano y bacteriano (Tabla I). Estos datos permitieron valorar si el modelo podía identificar proteínas virales en muestras reales y, al mismo tiempo, estimar su comportamiento frente a secuencias no virales. Adicionalmente, los resultados de VirDetect-AI fueron comparados con los obtenidos mediante BLASTp, una herramienta ampliamente utilizada que permite identificar similitudes entre secuencias de proteínas y empleada comúnmente como referencia en análisis bioinformáticos.

Tabla I. Información de los conjuntos de muestras metagenómicas reales evaluadas con VirDetect-AI (formato aminoácidos).

Conjunto de datos	**Descripción (muestras metagenómicas)**		**Estadística de longitud de secuencias (aa)**		
	#	Origen	Min	Max	Mediana
Meta-EukVirus-set	703	orofaríngeas	300	4,557	606
Meta-Unknown-set	113	orofaríngeas	300	516	332
Meta-Human-set	1,280	orofaríngeas	300	1,720	369
Meta-Bacteria-set	2,428	fecal de infante	300	3,172	434

Se analizaron 120 conjuntos de datos metagenómicos provenientes de muestras orofaríngeas humanas de pacientes positivos a COVID-19 [26]. Después del preprocesamiento, ensamblado y predicción de marcos abiertos de lecturas (ORFs), que permite pasar de DNA a aminoácidos, se obtuvieron 703 proteínas con similitud con proteínas de virus eucariontes y 113 secuencias sin homólogos conocidos. Estos conjuntos fueron denominados Meta-EukVirus-set y Meta-Unknown-set, respectivamente. También, se incluyeron dos conjuntos negativos: Meta-Human-set, compuesto por 1,280 ORFs de origen humano, y Meta-Bacteria-set, integrado por 2,428 proteínas bacterianas de *Bifidobacterium* provenientes de una muestra fecal de infante [27].

En el conjunto Meta-EukVirus-set, VirDetect-AI clasificó el 94.6% de las secuencias como virales, el 3.3% como no virales y el 2.1% como desconocidas. De las secuencias virales identificadas, el 98.2% correspondió a virus eucariontes. A nivel taxonómico, las familias más frecuentes fueron *Coronaviridae* (69.2%), *Papillomaviridae* (13.3%), *Anelloviridae* (2.9%) y *Orthoherpesviridae* (2.3%), con predominancia del género *Betacoronavirus*. Este resultado es consistente con el origen de las muestras, provenientes de pacientes positivos a COVID-19.

La comparación con BLASTp mostró resultados concordantes en el conjunto Meta-EukVirus-set. BLASTp clasificó el 96.1% de las secuencias como virales, mientras que VirDetect-AI identificó el 94.6%. Aunque BLASTp presentó una sensibilidad ligeramente mayor, VirDetect-AI mostró ventajas en términos de eficiencia computacional y permitió asignar algunas secuencias con baja similitud respecto a las referencias disponibles. Este resultado sugiere que VirDetect-AI puede ser útil como herramienta complementaria para explorar proteínas virales difíciles de detectar mediante enfoques basados exclusivamente en alineamiento.

En el conjunto Meta-Unknown-set, VirDetect-AI clasificó el 29.2% de las secuencias como virales, el 46.9% como no virales y el 23.9% como desconocidas. Las proteínas virales predichas incluyeron familias como *Potyviridae*, *Orthoherpesviridae* y *Nodaviridae*, asociadas con hospederos diversos, incluyendo plantas, humanos, animales e insectos. Estos resultados sugieren que VirDetect-AI puede contribuir a la identificación inicial de secuencias con características compatibles con proteínas virales, incluso cuando no presentan una similitud evidente con secuencias previamente descritas.

Evaluación con conjuntos negativos

Para estimar la especificidad del modelo, VirDetect-AI también fue evaluado con conjuntos de datos negativos de origen humano y bacteriano. En el conjunto Meta-Human-set, VirDetect-AI clasificó el 69.8% de las secuencias como no virales y el 10.1% como desconocidas, mientras que el 16.5% fue asignado a clases virales eucariontes. Estas asignaciones virales se distribuyeron en múltiples clases con baja abundancia, sugiriendo una proporción limitada de posibles falsos positivos.

En el Meta-Bacteria-set, el 65.5% de las secuencias se clasificó como no viral, el 21.1% como viral procarionte y sólo el 7% como viral eucarionte. Este resultado indica que VirDetect-AI mantiene una baja proporción de clasificación errónea hacia virus eucariontes al analizar proteínas bacterianas.

Al comparar estos resultados con BLASTp, se observó que ambas herramientas presentaron proporciones globales similares; sin embargo, el solapamiento entre los posibles falsos positivos fue limitado. En el conjunto humano, el 21.8% de los falsos positivos de BLASTp coincidió con los de VirDetect-AI, mientras que, en el conjunto bacteriano, la coincidencia fue de 24%. Esto indica que los errores de clasificación no necesariamente

ocurren sobre las mismas secuencias, lo cual es esperable debido a que ambos métodos se basan en principios distintos: BLASTp depende de la similitud por alineamiento, mientras que VirDetect-AI identifica patrones aprendidos a partir de las secuencias de entrenamiento.

Comparación de eficiencia computacional

Finalmente, se comparó el rendimiento computacional de VirDetect-AI frente a BLASTp en los conjuntos metagenómicos evaluados. En todos los casos, VirDetect-AI mostró una reducción importante en el tiempo de análisis. Utilizando GPU, la herramienta fue entre 2,120 y 4,221 veces más rápido que BLASTp, mientras que, en CPU, fue entre 14 y 33 veces más rápido (Tabla II).

Tabla II. Comparación del tiempo de ejecución entre VirDetect-AI y BLASTp en los conjuntos de datos metagenómicos evaluados.

Conjunto de datos	# Secuencias	Tiempo de ejecución (segundos)		
		VirDetect-AI		BLASTp 2.11.10
		GPU	CPU	CPU
Meta-EukVirus-set	703	170.4	35,762	656,862
Meta-Unknown-set	113	6.1	384.1	12,976
Meta-Human-set	1,280	42.9	8,862	181,088
Meta-Bacteria-set	2,428	126.3	27,677	391,607

Estos resultados muestran que VirDetect-AI combina un alto desempeño de clasificación con una mayor eficiencia computacional. Por ello, representa una alternativa útil para el análisis de grandes volúmenes de datos metagenómicos, especialmente en contextos donde se

requiere procesar información de manera rápida, como estudios de vigilancia genómica o exploración de diversidad viral.

Conclusión y discusión

VirDetect-AI mostró un alto desempeño en la identificación y clasificación de secuencias metagenómicas, tanto en los conjuntos de entrenamiento y prueba como en datos reales. Estos resultados indican que el modelo puede reconocer patrones asociados con proteínas virales eucariontes y distinguirlas de secuencias no virales con alta precisión. Su aplicación en datos metagenómicos reales sugiere que puede utilizarse como una herramienta complementaria para explorar la diversidad viral en muestras biológicas complejas.

Sin embargo, VirDetect-AI presenta algunas limitaciones. Se observó una fracción reducida de posibles falsos positivos, principalmente en secuencias negativas de origen humano y bacteriano. Esto puede explicarse por la complejidad biológica de las secuencias analizadas, ya que el genoma humano contiene regiones derivadas de virus endógenos y algunas bacterias presentan elementos genéticos relacionados con virus, como profagos, secuencias móviles o dominios conservados. Estos elementos pueden compartir características con proteínas virales y dificultar su clasificación precisa.

Por ello, los resultados de VirDetect-AI deben interpretarse como una primera aproximación computacional para priorizar secuencias de interés. Las secuencias clasificadas como virales, especialmente aquellas con baja similitud frente a bases de datos de referencia, requieren análisis complementarios para confirmar su origen, función y relevancia biológica. En este sentido, VirDetect-AI no sustituye a herramientas tradicionales como BLASTp, sino que las complementa al ofrecer una estrategia rápida y escalable para analizar grandes volúmenes de datos metagenómicos.

La principal aportación de VirDetect-AI es su capacidad para apoyar la exploración de la diversidad viral eucarionte, permitiendo identificar tanto secuencias virales conocidas como candidatas asociadas con virus poco caracterizados o potencialmente nuevos.

La integración de metagenómica, inteligencia artificial y supercómputo abre nuevas posibilidades para el estudio de los virus y el análisis sistemático de datos virales a gran escala. En un contexto marcado por la emergencia de nuevas enfermedades, los cambios

ambientales y la necesidad de fortalecer la vigilancia epidemiológica, herramientas como VirDetect-AI pueden contribuir a la investigación en salud pública, biodiversidad viral y desarrollo de tecnología bioinformática en México.

Financiamiento

Este trabajo fue financiado por los proyectos PAPIIT-DGAPA-IN230523 y PAPIIT-DGAPA-IN225126 de la DGAPA-UNAM (a B.T.), así como por el proyecto CBF-2025-I-1026 de SECIHTI (a B.T.).

Agradecimientos

Extendemos nuestro agradecimiento a la Universidad Nacional Autónoma de México (UNAM) y a la Dirección General de Cómputo y Tecnologías de Información y Comunicación (DGTIC) por otorgar acceso a la supercomputadora Miztli, mediante el proyecto LANCAD-UNAM-DGTIC-350. Finalmente, agradecemos a Jerome Verleyen, Juan Manuel Hurtado y Roberto Bahena, del Instituto de Biotecnología de la UNAM, por su invaluable apoyo en tareas de cómputo.

Referencias

[1] N. Nam, H. Do, K. L. Trinh, and N. Lee, “Metagenomics: an effective approach for exploring microbial diversity and functions,” *Foods*, vol. 12, no. 11, p. 2140, May 2023, doi: 10.3390/foods12112140.

[2] E. V. Koonin, V. V. Dolja, M. Krupovic, A. M. Varsani, Y. I. Wolf, and N. Yutin, *et al.*, “Global organization and proposed megataxonomy of the virus world,” *Microbiology and Molecular Biology Reviews*, vol. 84, no. 2, pp. e00061-19, 2020, doi: 10.1128/MMBR.00061-19.

[3] R. K. Sales, J. Oraño, R. D. Estanislao, A. J. Ballesteros, and M. I. F. Gomez, “Research priority-setting for human, plant, and animal virology: an online experience for the

Virology Institute of the Philippines," *Health Res Policy Sys,* vol. 19, no. 1, Apr. 2021, doi: 10.1186/s12961-021-00723-z.

[4] S. R. Krishnamurthy and D. Wang, "Origins and challenges of viral dark matter," *Virus Res.*, vol. 239, pp. 136–142, Jul. 2017, doi: 10.1016/j.virusres.2017.02.002.

[5] A. R. Mushegian, "Are there 10^31 virus particles on earth, or more, or fewer?," *J. Bacteriol.*, vol. 202, no. 9, Apr. 2020, doi: 10.1128/JB.00052-20.

[6] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden, "BLAST+: architecture and applications," *BMC Bioinformatics*, vol. 10, no. 421, 2009, doi:10.1186/1471-2105-10-421.

[7] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biology*, vol. 10, no. 3, p. R25, 2009, doi: 10.1186/gb-2009-10-3-r25.

[8] R. Li, Y. Li, K. Kristiansen, and J. Wang, "SOAP: short oligonucleotide alignment program," *Bioinformatics*, vol. 24, no. 5, pp. 713–714, 2008, doi: 10.1093/bioinformatics/btn025.

[9] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows–Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009, doi: 10.1093/bioinformatics/btp324.

[10] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. L. Sonnhammer, J. Tate, and M. Punta, "Pfam: the protein families database," *Nucleic Acids Research*, vol. 42, no. 1, pp. 222–230, 2014, doi: 10.1093/nar/gkt1223.

[11] R. D. Finn, J. Clements, and S. R. Eddy, "HMMER web server: interactive sequence similarity searching," *Nucleic Acids Research*, vol. 39, no. Suppl. 2, pp. 29–37, 2011, doi: 10.1093/nar/gkr367.

[12] J. Guo, B. Bolduc, A. A. Zayed, A. Varsani, G. Dominguez-Huerta, and T. O. Delmont, *et al*., "VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA

and RNA viruses," *Microbiome*, vol. 9, no. 1, p. 37, Feb. 2021, doi: 10.1186/s40168-020-00990-y.

[13] J. Ren, N. A. Ahlgren, Y. Y. Lu, J. A. Fuhrman, and F. Sun, "VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data," *Microbiome*, vol. 5, no. 1, p. 69, Dec. 2017, doi: 10.1186/s40168-017-0283-5.

[14] J. Ren, K. Song, C. Deng, *et al*., "Identifying viruses from metagenomic data using deep learning," *Quantitative Biology*, vol. 8, no. 1, p. 64, 2020, doi: 10.1007/s40484-019-0187-4.

[15] Y. Miao, F. Liu, T. Hou, and Y. Liu, "Virtifier: a deep learning-based identifier for viral sequences from metagenomes" *Bioinformatics*, vol. 38, no. 5, pp. 1216–1222, Feb. 2022, doi: 10.1093/bioinformatics/btab845.

[16] Z. Bzhalava, A. Tampuu, P. Bała, R. Vicente, and J. Dillner, "Machine Learning for detection of viral sequences in human metagenomic datasets," *BMC Bioinformatics*, vol. 19, no. 1, p. 336, Dec. 2018, doi: 10.1186/s12859-018-2340-x.

[17] M. H. Alshayeji, S. C. Sindhu, and S. Abed, "Viral genome prediction from raw human DNA sequence samples by combining natural language processing and machine learning techniques," *Expert Syst. Appl.*, vol. 218, p. 119641, May 2023, doi: 10.1016/j.eswa.2023.119641.

[18] A. Tampuu, Z. Bzhalava, J. Dillner, and R. Vicente, "ViraMiner: deep learning on raw DNA sequences for identifying viral genomes in human samples," Apr. 2019. doi: 10.1101/602656.

[19] C. M. Dasari and R. Bhukya, "Explainable deep neural networks for novel viral genome prediction," *Applied Intelligence*, vol. 52, no. 3, pp. 3002–3017, Feb. 2022, doi: 10.1007/s10489-021-02572-3.

[20] Y. Zhang, C. Li, H. Feng, and D. Zhu, "DLmeta: a deep learning method for metagenomic identification," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, Dec. 2022, pp. 303–308. doi: 10.1109/BIBM55620.2022.9995231.

[21] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv:1512.03385 [cs]*, Dec. 2015, [Online]. Available: http://arxiv.org/abs/1512.03385

[23] A. Zárate, L. Díaz-González, and B. Taboada, "VirDetect-AI: a residual and convolutional neural network–based metagenomic tool for eukaryotic viral protein identification," *Brief. Bioinform*., vol. 26, no. 1, Jan. 2025, doi: 10.1093/bib/bbaf001.

[24] D. Harding-Larsen, J. Funk, N. G. Madsen, H. Gharabli, C. G. Acevedo-Rocha, S. Mazurenko, "Protein representations: encoding biological information for machine learning in biocatalysis," *Biotechnology Advances*, vol. 77, p. 108459, 2024. doi: 10.1016/j.biotechadv.2024.108459.

[25] Alyzart22, *VirDetect-AI*, GitHub repository. [Online]. Available: https://github.com/alyzart22/VirDetect-AI

[26] P. Iša, B. Taboada, R. García-López, C. Boukadida, J. E. Ramírez-González, J. A. Vázquez-Pérez, *et al.*, "Metagenomic analysis reveals differences in the co-occurrence and abundance of viral species in SARS-CoV-2 patients with different severity of disease," *BMC Infect. Dis.*, vol. 22, no. 1, p. 792, Oct. 2022, doi: 10.1186/s12879-022-07783-8.

[27] X. Rivera-Gutiérrez, P. Morán, B. Taboada, A. Serrano-Vázquez, P. Isa, L. Rojas-Velázquez, *et al.*, "The fecal and oropharyngeal eukaryotic viromes of healthy infants during the first year of life are personal," *Sci. Rep.*, vol. 13, no. 1, p. 938, Jan. 2023, doi: 10.1038/s41598-022-26707-9.