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).

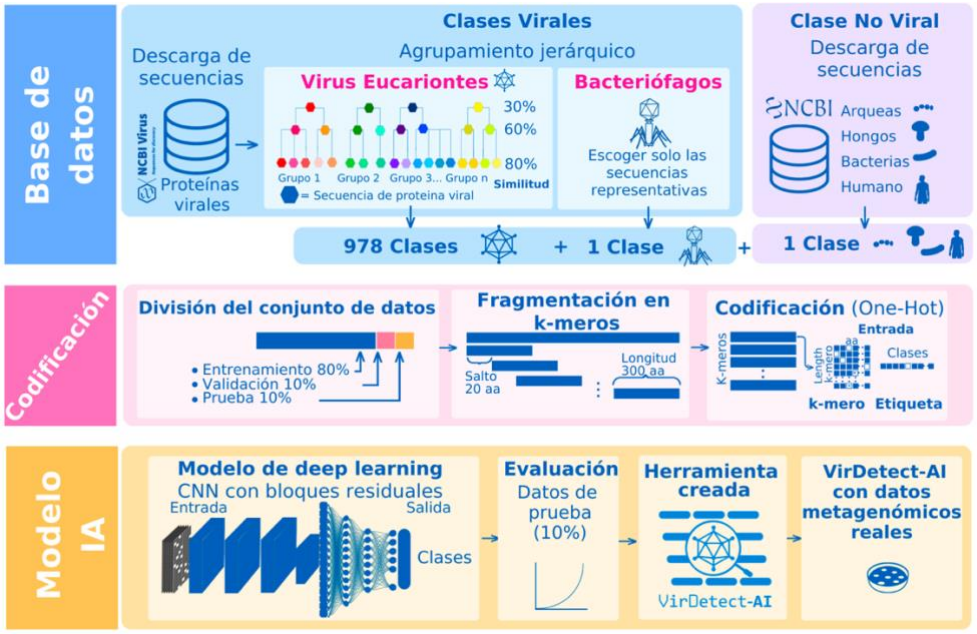

Fig. 1. Flujo de trabajo de VirDetect-AI para la identificación de proteínas virales eucariontes en datos metagenómicos. El *pipeline* comprende: (i) construcción de una base de datos a partir de secuencias virales y no virales; (ii) agrupamiento jerárquico (*hierarchical clustering*) para definir 978 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

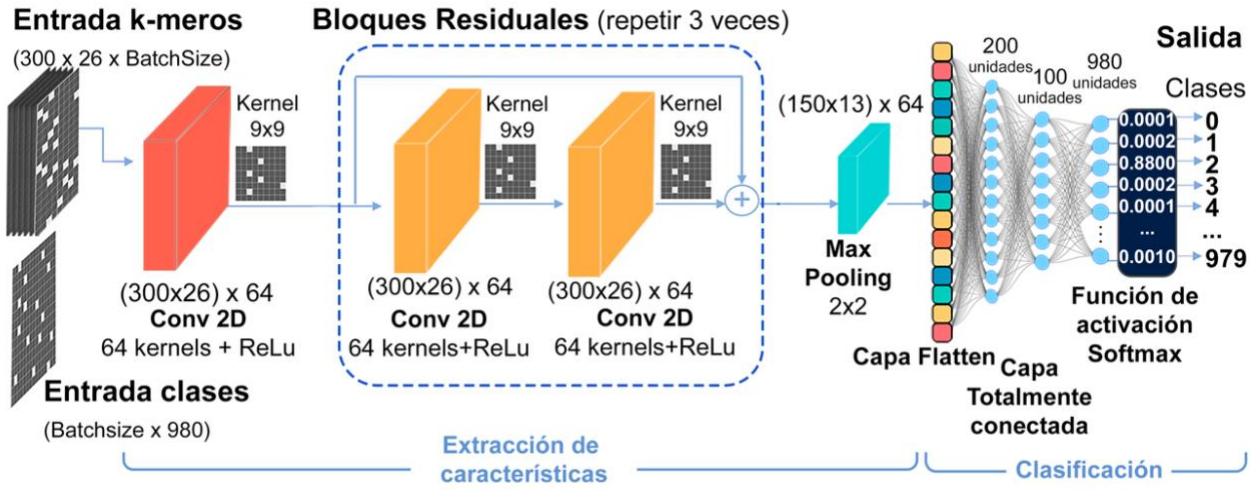

Fig. 2. Estructura del modelo de aprendizaje profundo VirDetect-AI, que integra una arquitectura de CNN con bloques residuales.

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.

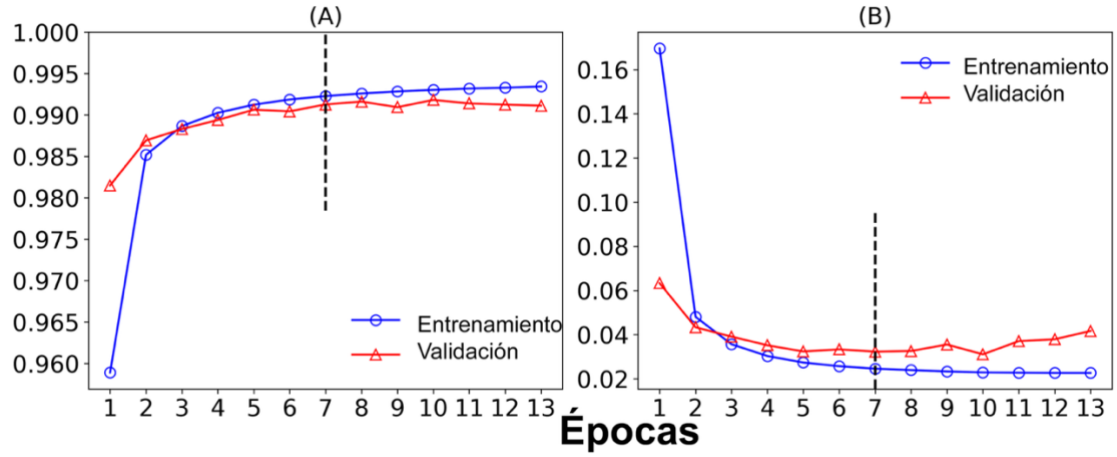

Fig. 3. Rendimiento del entrenamiento y validación de VirDetect-AI en el conjunto de datos de prueba (*Test*). (A) Exactitud del modelo (B) Comportamiento de la función de perdida (*loss*).

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).

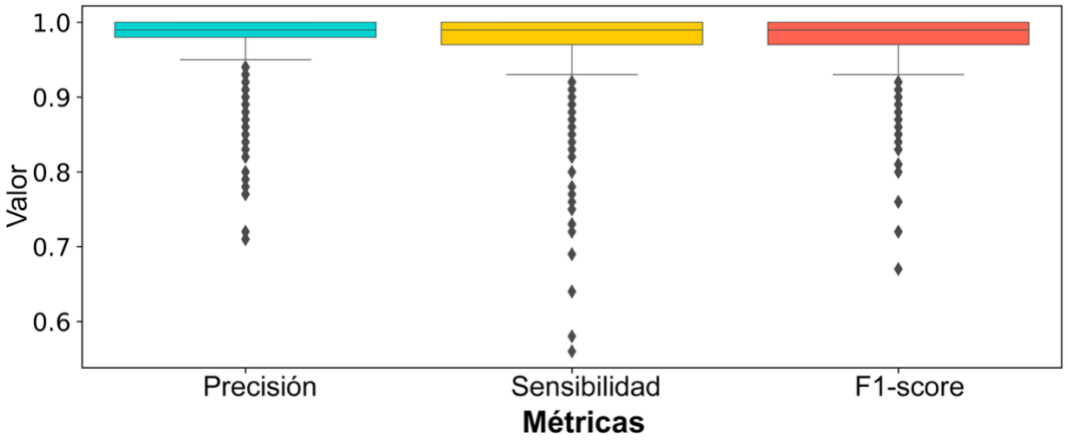

Fig. 4. Rendimiento de VirDetect-AI en el conjunto de datos de Prueba (*Test*) para las 980 clases evaluadas.

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.

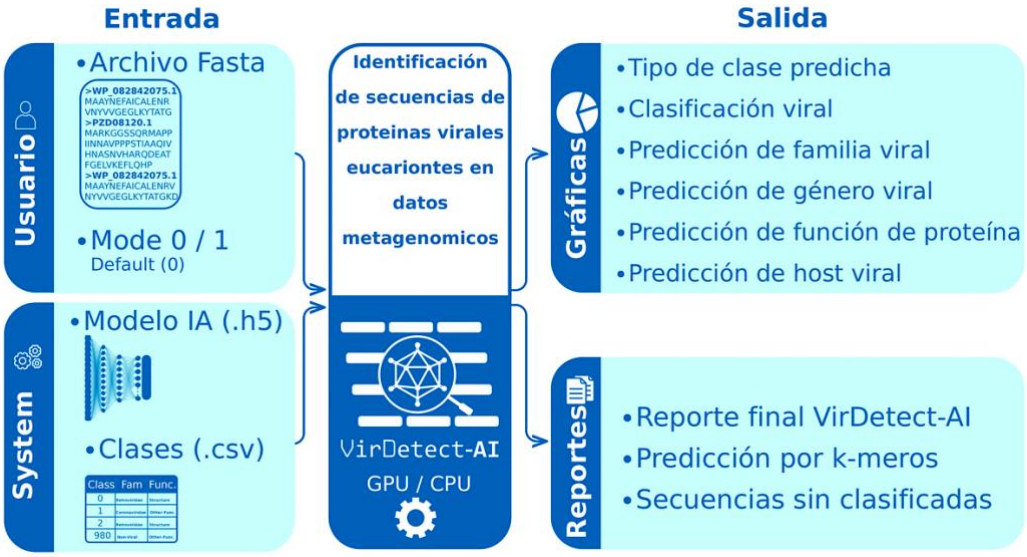

Fig. 5. Flujo general de predicción de la herramienta VirDetect-AI.

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.