



Original article

**EVALUACIÓN DE EMBEDDINGS DEL MODELO DE LENGUAJE DE
PROTEÍNAS ESM-2 PARA LA PREDICCIÓN DE PÉPTIDOS
ANTIMICROBIANOS DE INTERÉS VETERINARIO: ESTUDIO
COMPUTACIONAL PRELIMINAR**

Arnulfo Villanueva-Castillo^{1,2*}, Ruby S. Moreno-Mejía¹, César F. Pastelín-Rojas^{1,2},
Claudia Mancilla-Simbro³, Briseida L. Castro-Bautista^{1,2}, Erick C. Fernández-Meneses^{1,2},
Miguel Á. Zambrano-González^{1,2}, Fabiola Rodríguez-Andrade¹, Juan R. Cruz-Aviña¹,
Jesús J. Hinojosa-Moya⁴, Hermilo Lucio-Castillo⁵

¹Facultad de Medicina Veterinaria y Zootecnia, Benemérita Universidad Autónoma de Puebla, El Salado,
km 7.5 carr. Cañada-Morelos, C.P. 75487, Tecamachalco, Puebla, México.

²Cuerpo Académico de Enfermedades Emergentes, Bioinformática y Dinámica Molecular BUAP-CA-274,
Laboratorio de Biología Molecular y Biotecnología.

³HybridLab Fisiología y Biología Molecular de Células Excitables, Instituto de Fisiología, BUAP, 14 Sur 6301,
CU San Manuel, 72570, Puebla, Pue., México.

⁴Facultad de Ingeniería Química y Complejo Regional Centro, Benemérita Universidad Autónoma de
Puebla.

⁵Universidad Autónoma de Tamaulipas, Ciudad Mante, Tamaulipas, México.

*Autor de correspondencia: arnulfo.villanueva@correo.buap.mx

Recibido 02-04-2026; Formulario revisado 20-05-2026; Aceptado 24-06-2026

RESUMEN: La resistencia antimicrobiana (RAM) constituye una amenaza prioritaria de salud única (One Health) que compromete el tratamiento de infecciones en animales de producción, compañía y fauna silvestre. Los péptidos antimicrobianos (AMP) representan una alternativa terapéutica prometedora y los modelos de lenguaje de proteínas ofrecen herramientas novedosas para su predicción. El presente estudio evalúa, como prueba de concepto, la capacidad del modelo ESM-2 (facebook/esm2_t6_8M_UR50D) para discriminar AMP de controles negativos. Se creó un conjunto de 24 AMP de referencia (APD3/DBAASP) y 24 controles negativos generados mediante permutación aleatoria, que mantiene la composición de aminoácidos. Se calcularon los embeddings (media de las representaciones de residuos, 320 dimensiones) y se entrenaron clasificadores (regresión logística y SVM) con validación cruzada de 5 pliegues. La regresión logística alcanzó AUC = 0,74 (exactitud 0,71; F1 0,71) y la SVM AUC = 0,72. Los embeddings de ESM-2 separaron en parte AMP y no-AMP en el espacio t-SNE, incluso usando el modelo más pequeño, lo que muestra que son útiles como representaciones para predecir AMP. Se discuten las limitaciones (tamaño del modelo y de la muestra) y se plantea la proyección hacia modelos ESM-2 mayores y la validación in vitro. Este trabajo establece



bases informáticas para encontrar AMP contra los patógenos veterinarios más importantes (*Staphylococcus aureus* mecC, *Escherichia coli*, *Salmonella* spp.).

Palabras clave: péptidos antimicrobianos; ESM-2; modelos de lenguaje de proteínas; resistencia antimicrobiana; bioinformática veterinaria.

ABSTRACT Antimicrobial resistance (AMR) is a priority One Health threat that compromises the treatment of infections in livestock, companion animals and wildlife. Antimicrobial peptides (AMP) are a promising therapeutic alternative and protein language models offer novel tools for their prediction. As a proof of concept, this study evaluates the capacity of the ESM-2 model (facebook/esm2_t6_8M_UR50D) to discriminate AMPs from negative controls. A set of 24 reference AMPs (APD3/DBAASP) and 24 negative controls generated by random permutation (preserving amino acid composition) was assembled. Embeddings (mean of residue representations, 320 dimensions) were computed and classifiers (logistic regression and SVM) were trained with 5-fold cross-validation. Logistic regression reached AUC = 0.74 (accuracy 0.71; F1 0.71) and SVM AUC = 0.72. ESM-2 embeddings partially separated AMPs and non-AMPs in t-SNE space, even with the smallest model, evidencing their usefulness as representations for AMP prediction. Limitations (model and sample size) and the projection toward larger ESM-2 models and in vitro validation are discussed. This work provides a computational foundation for discovering antimicrobial peptides (AMPs) to combat important animal health threats like mecC *Staphylococcus aureus*, *Escherichia coli*, and *Salmonella* species.

Keywords: antimicrobial peptides; ESM-2; protein language models; antimicrobial resistance; veterinary bioinformatics.

INTRODUCCIÓN

La resistencia antimicrobiana (RAM) es una de las principales amenazas para la salud humana y animal del siglo XXI y se aborda desde un enfoque One Health que reconoce la interconexión entre salud humana, animal y ambiental (ONeill, 2016; OMS, 2014). En medicina veterinaria, la difusión de patógenos que son resistentes a muchos medicamentos, como *Staphylococcus*

aureus con el gen mecA (Villa-Díaz et al., 2025), *Escherichia coli* y *Salmonella* spp., disminuye las opciones de tratamiento y afecta la salud de los animales y la seguridad alimentaria.

Los péptidos antimicrobianos (AMP) son moléculas de la inmunidad innata con actividad contra bacterias, hongos y virus, que actúan principalmente sobre la membrana celular y cuya inducción de resistencia es baja, por lo que se



consideran alternativas prometedoras a los antibióticos convencionales (Hancock y Sahl, 2006; Peschel y Sahl, 2006; Magaña et al., 2020).

El descubrimiento de nuevos AMP se ha visto transformado por el aprendizaje profundo. Los modelos de lenguaje de proteínas, entrenados sobre millones de secuencias, generan representaciones (embeddings) que capturan propiedades estructurales y funcionales y permiten predecir actividad antimicrobiana a partir de la secuencia (Rives et al., 2021; Lin et al., 2023; Veltri et al., 2018). ESM-2, uno de los modelos de lenguaje de proteínas más utilizados, ha mostrado un desempeño destacado en tareas de predicción de estructura y función (Lin et al., 2023).

El objetivo de este trabajo, como un estudio computacional inicial, fue comprobar si los embeddings del modelo ESM-2 (versión ligera t6_8M) pueden diferenciar AMP de referencia de controles negativos generados por permutación. Esto establece las bases para descubrir AMP enfocados en patógenos veterinarios importantes.

MATERIALES Y MÉTODOS

Conjunto de datos. Se seleccionaron 24 AMP de referencia que están bien descritos, como magaininas, cecropinas, melitina, LL-37, indolicidina, protegrina, taquiplésina, histatina, pleurocidina, thanatin, mastoparanes, aureinas,

temporinas, lactoferricina y bactenecina. Estos se reportaron en las bases APD3 (Wang et al., 2016) y DBAASP (Pirtskhalava et al., 2016). Como controles negativos, se generó, para cada AMP, una secuencia permutada aleatoriamente que preserva la composición de aminoácidos pero pierde la actividad, estrategia ampliamente usada en la predicción de AMP (Veltri et al., 2018). El conjunto total fue de 48 secuencias.

Embeddings de ESM-2. Se empleó el modelo ESM-2 (facebook/esm2_t6_8M_UR50D, 7,5 millones de parámetros, 320 dimensiones) mediante la biblioteca Transformers (Hugging Face). Para cada secuencia se calculó el embedding como la media de las representaciones de los residuos, excluyendo los tokens especiales.

Clasificación y validación. Se normalizaron las representaciones y se entrenaron dos modelos (regresión logística y máquina de vectores de soporte con núcleo RBF) utilizando validación cruzada estratificada de 5 pliegues. Se reportaron exactitud, F1 y el área bajo la curva ROC (AUC).

Software. Python 3, PyTorch, scikit-learn y matplotlib. El análisis es reproducible y los datos/secuencias se depositarán en un repositorio público.

RESULTADOS



Separación en el espacio de embeddings. Los embeddings de ESM-2 (t6_8M) que se observaron con t-SNE mostraron una separación parcial entre los AMP y los controles mezclados (Fig. 1). Aun

tratándose del modelo más compacto de la familia ESM-2, las representaciones capturan información que distingue a los AMP de secuencias de misma composición pero sin actividad.

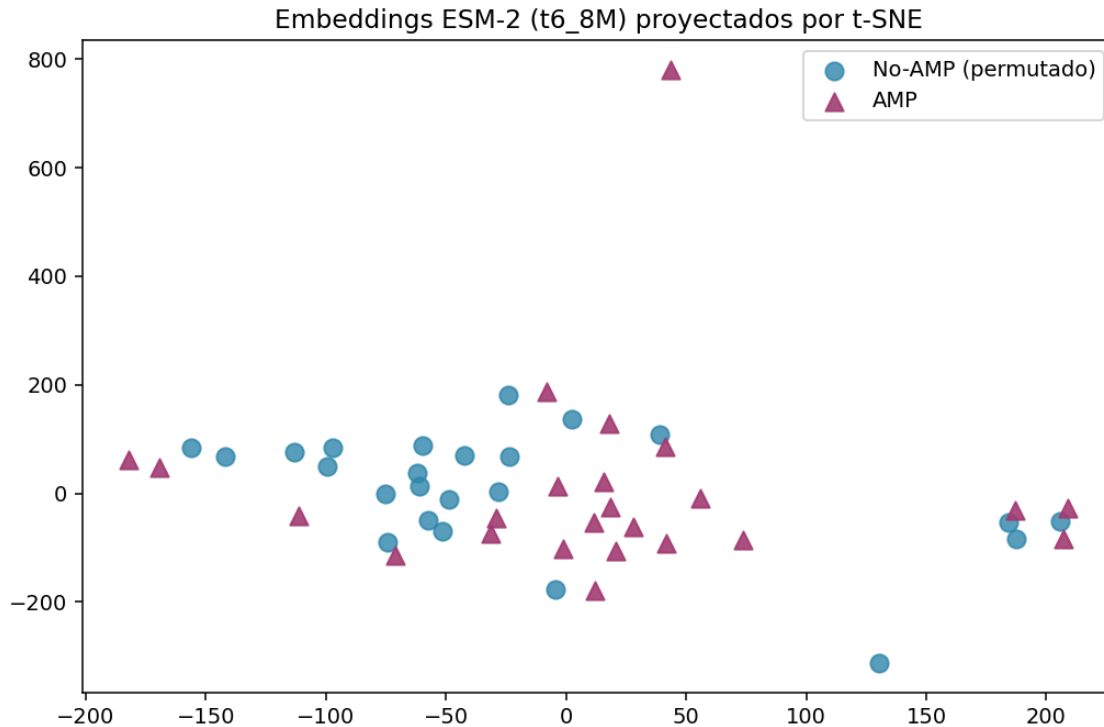


Figura 1. Proyección t-SNE de los embeddings de ESM-2 (t6_8M). Triángulos púrpura: AMP; círculos azules: controles permutados.

Desempeño de los clasificadores. La regresión logística alcanzó AUC = 0,74 (exactitud 0,71; F1 0,71), mientras que la SVM obtuvo AUC = 0,72 (exactitud 0,65; F1 0,64) en validación cruzada de

5 pliegues (Fig. 2; Tabla I). Estos valores, aunque moderados, indican que los embeddings contienen señal discriminativa útil para la predicción de AMP.

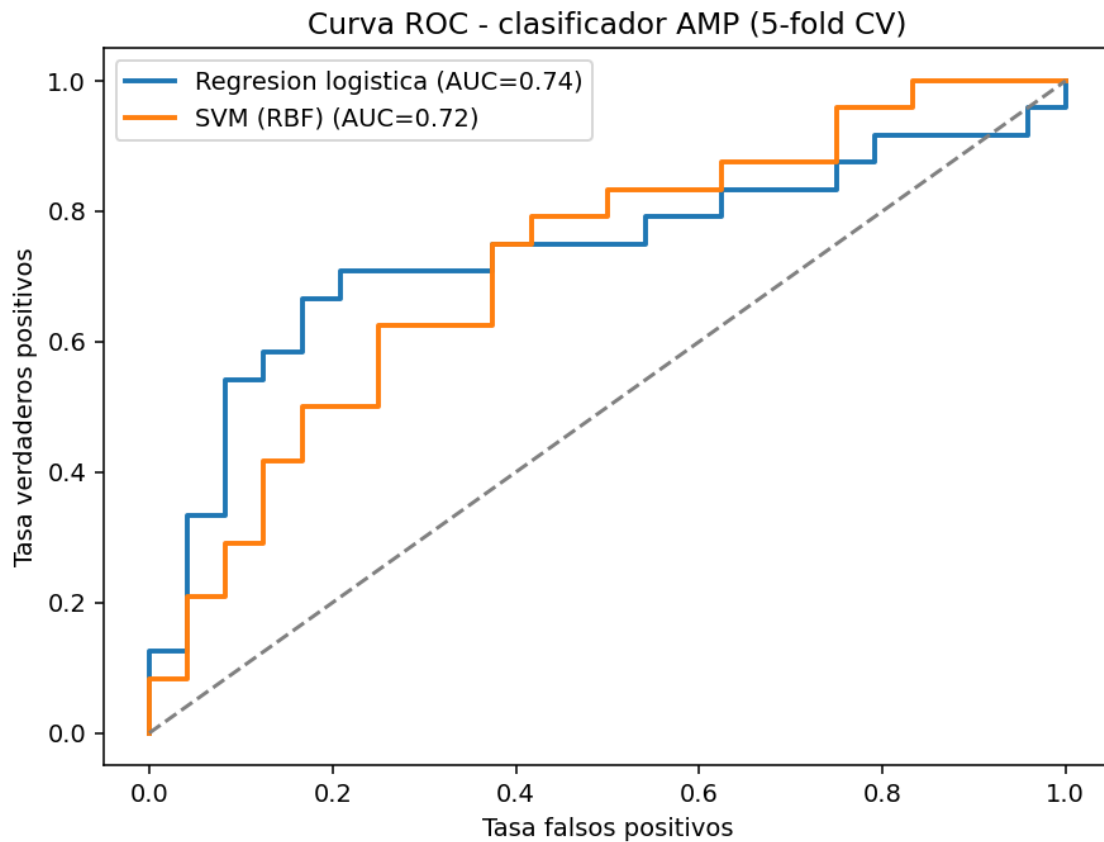


Figura 2. Curvas ROC de los clasificadores (5-fold CV). Regresión logística AUC = 0,74; SVM AUC = 0,72.

Propiedades fisicoquímicas. Los AMP que se evaluaron tenían longitudes de 12 a 37 residuos y una carga neta positiva,

como es típico en los AMP catiónicos y anfipáticos (Fig. 3).

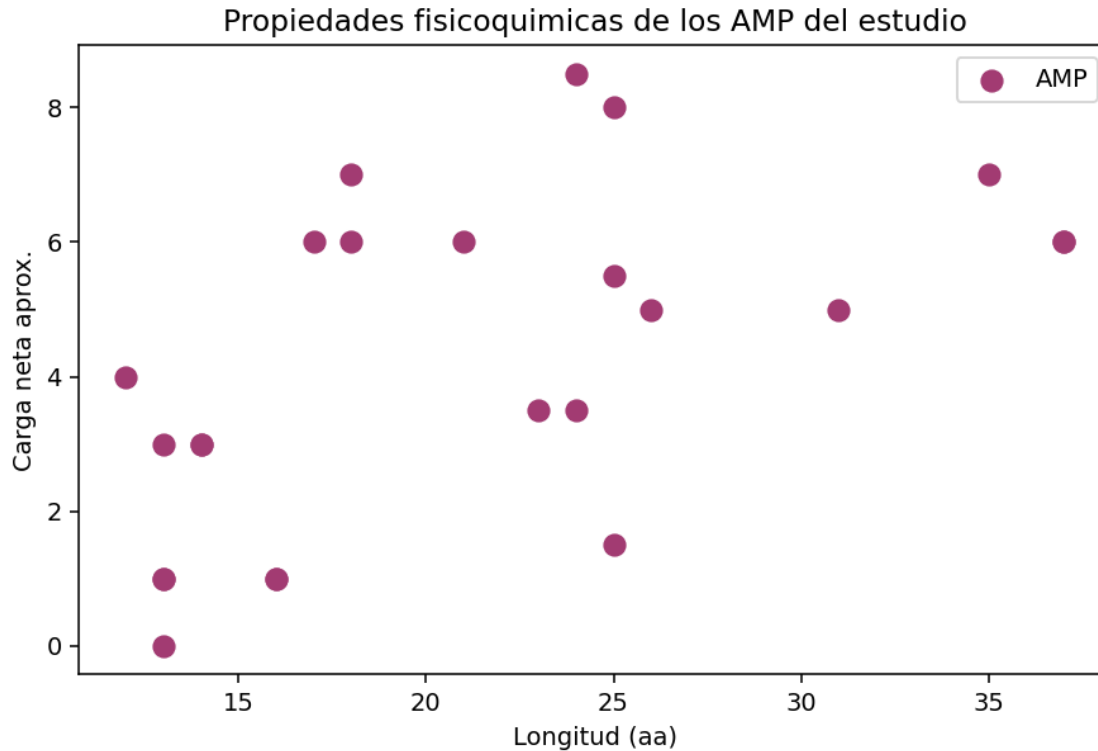


Figura 3. Distribución de longitud y carga neta aproximada de los AMP del estudio.

Tabla I. Desempeño de los clasificadores (5-fold CV) sobre embeddings ESM-2 t6_8M.

Clasificador	Exactitud	F1	AUC
Regresión logística	0,71	0,71	0,74
SVM (RBF)	0,65	0,64	0,72

DISCUSIÓN

Los resultados indican que, incluso en su versión más sencilla, ESM-2 produce representaciones útiles para diferenciar AMP de controles con la misma composición. Esta tarea es compleja porque la permutación mantiene la mezcla de aminoácidos. El desempeño moderado (AUC 0,72-0,74) es coherente con el tamaño reducido del modelo (7,5

millones de parámetros) y de la muestra, y se espera que estructuras ESM-2 mayores (p. ej., t33_650M o t36_3B) incrementen de forma sustancial la precisión, como se ha reportado en tareas análogas (Lin et al., 2023; Yan et al., 2020).

La metodología se puede reproducir y ampliar fácilmente: aumentar el conjunto con miles de AMP seleccionados



(DBAASP, APD3) y usar modelos más grandes ayudará a crear un predictor fuerte y, luego, descubrir nuevos candidatos. La integración de la validación de toxicidad/hemólisis (como HemoPI) y la predicción de estructuras (AlphaFold) completaría el proceso de descubrimiento in silico.

Las aplicaciones veterinarias son directas: AMP es efectivo contra *S. aureus* (incluyendo cepas mecA; Villa-Díaz et al., 2025), *E. coli* y *Salmonella* en la producción de animales, y contra *Batrachochytrium dendrobatidis* en la conservación de ranas. No obstante, los candidatos requieren validación in vitro (CIM, hemólisis) antes de cualquier afirmación de eficacia; el presente trabajo se limita al alcance computacional.

Entre las limitaciones destacan el tamaño del modelo y del conjunto, el uso de controles permutados (que pueden subestimar el desempeño frente a péptidos reales no AMP) y la ausencia de validación experimental.

CONCLUSIONES

Los embeddings del modelo de lenguaje de proteínas ESM-2, incluso en su versión más simple, contienen información valiosa para distinguir péptidos antimicrobianos de referencia. Esto establece bases computacionales para encontrar AMP relevantes en veterinaria. Se sugiere utilizar modelos ESM-2 más grandes,

aumentar el conjunto de datos y avanzar en la creación de candidatos y su validación en laboratorio como una estrategia contra la resistencia a los antimicrobianos.

CONTRIBUCIÓN DE LOS AUTORES

AVC: Conceptualización, metodología, redacción – borrador original, supervisión, administración del proyecto, adquisición de fondos. RSMM: Investigación, curaduría de datos. CFPR: Metodología, Investigación, Redacción – revisión y edición. CMS: Investigación, Análisis formal. BLCB: Investigación, curaduría de datos. ECFM: Investigación, Metodología. MAZG: Investigación, redacción, revisión y edición. FRA: Investigación, curaduría de datos. JRCA: Investigación (necropsia), metodología, recursos JJHM: Investigación, recursos.

HLC: Investigación, Recursos. Todos los autores leyeron y aprobaron el manuscrito final.

RECONOCIMIENTOS

Los autores agradecen a la Facultad de Medicina Veterinaria y Zootecnia (FMVZ) y al Cuerpo Académico de Enfermedades Emergentes, Bioinformática y Dinámica Molecular BUAP-CA-274 por las facilidades otorgadas para la realización de este trabajo.

BIBLIOGRAFÍA



- Hancock, R. E. W., & Sahl, H. G. (2006). Antimicrobial and host-defense peptides as new anti-infective therapeutic strategies. *Nature Biotechnology*, 24(12), 1551–1557.
- Kumar, P., Kizhakedathu, V. N., & Straus, S. K. (2018). Antimicrobial peptides: diversity, mechanism of action and strategies to improve the activity and biocompatibility in vivo. *Biomolecules*, 8(1), E4.
- Lin, Z., Akin, H., Rao, R., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637), 1123–1130.
- Magaña, M., Pushpanathan, M., Santos, A. L., et al. (2020). The value of antimicrobial peptides in the age of resistance. *Frontiers in Microbiology*, 11, 582779.
- O'Neill, J. (2016). Tackling drug-resistant infections globally: final report and recommendations. *The Review on Antimicrobial Resistance*.
- OMS. (2014). Antimicrobial resistance: global report on surveillance. Organización Mundial de la Salud.
- Peschel, A., & Sahl, H. G. (2006). The co-evolution of host cationic antimicrobial peptides and microbial resistance. *Nature Reviews Microbiology*, 4(7), 529–536.
- Pirtskhalava, M., Gabrielian, A., Cruz, P., et al. (2016). DBAASP v.2: an enhanced database of structures and annotations for antimicrobial peptides. *Nucleic Acids Research*, 44(D1), D1104–D1112.
- Rives, A., Meier, S., Sercu, T., et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS*, 118(15), e2016239118.
- Veltri, D., Kamath, U., & Shehu, A. (2018). Deep learning improves antimicrobial peptide recognition. *Briefings in Bioinformatics*, 19(6), 1346–1362.
- Villa-Díaz, F., Ramírez-Almaraz, P. S., Ramírez-Mata, A., et al. (2025). Bioinformatic analysis of the mecC gene in *Staphylococcus aureus*. *Revista Ciencia Veterinaria y Biotecnología*, 2(1), 1–11.
- Wang, G., Li, X., & Wang, Z. (2016). APD3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Research*, 44(D1), D1087–D1093.
- Yan, J., Bhadra, P., Li, A., et al. (2020). Deep-AmPEP30: improve short antimicrobial peptides prediction with deep learning. *ACS Omega*, 5(14), 7948–7954.



Copyright: © 2026 Arnulfo Villanueva-Castillo. This is an open-access article published under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. Competing Interests: The authors have declared that no competing interests exist. Please cite this article as: Arnulfo Villanueva-Castillo, Ruby S. Moreno-Mejía, César F. Pastelín-Rojas, Claudia Mancilla-Simbro, Briseida L. Castro-Bautista, Erick C. Fernández-Meneses, Miguel Á. Zambrano-González, Fabiola Rodríguez-Andrade, Juan R. Cruz-Aviña, Jesús J. Hinojosa-Moya, Hermilo Lucio-Castillo. EVALUACIÓN DE EMBEDDINGS DEL MODELO DE LENGUAJE DE PROTEÍNAS ESM-2 PARA LA PREDICCIÓN DE PÉPTIDOS ANTIMICROBIANOS DE INTERÉS VETERINARIO: ESTUDIO COMPUTACIONAL PRELIMINAR. Revista Ciencia Veterinaria y Biotecnología. Rev Cien Vet Bio 2026; 3 (1): 102-110.