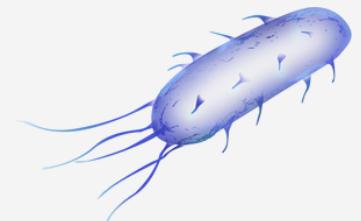


PEPTIDE-BERT

A LANGUAGE MODEL BASED ON TRANSFORMERS FOR PEPTIDE PROPERTY PREDICTION

Proyecto Final NLP
Y. Sarahi García González



Contents

01 Conceptos biológicos

Aminoácidos, péptidos y propiedades cruciales de éstos como la hemólisis, solubilidad y características antifouling

02 Introducción y contexto

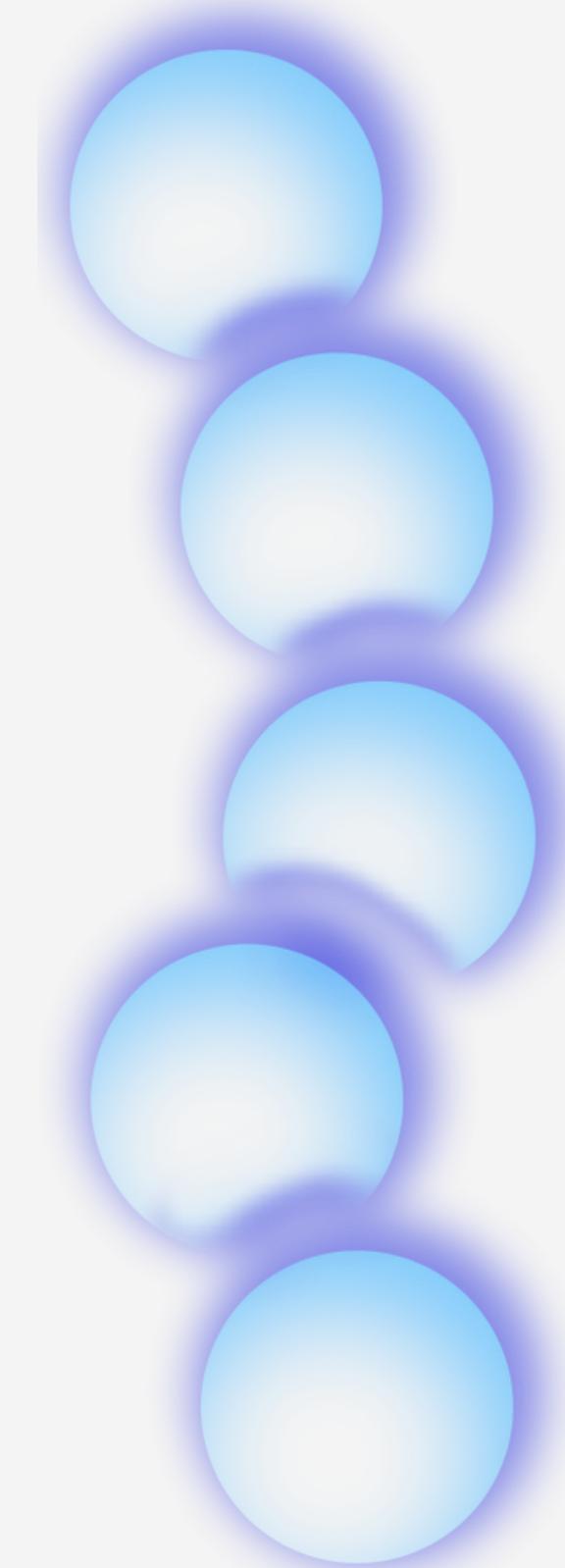
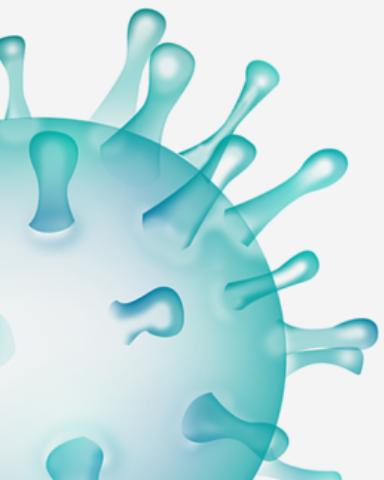
El paper basa en los avances recientes en modelos de lenguaje y su aplicación en la predicción de propiedades de péptidos.

03 Modelo PeptideBERT:

Arquitectura del modelo: Tokenización, ProtBert
Fine-Tunning y entrenamiento

04 Resultados

Resultados para cada tarea específica y visualización de embedding mediante t-sne

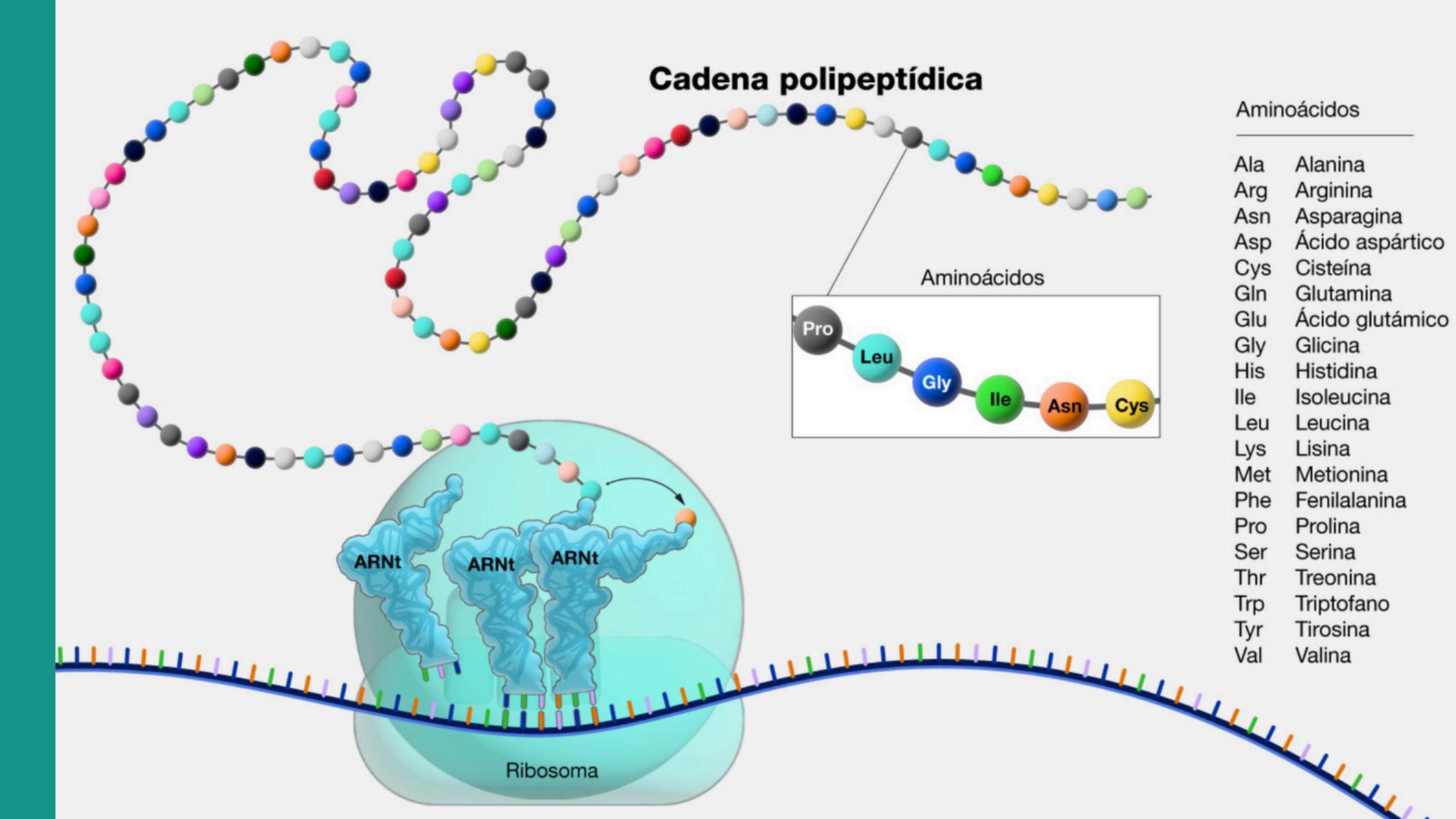


Aminoácido

- Un aminoácido son moléculas es la unidad base que actúa como estructura fundamental de los péptidos y proteínas. En los organismos vivos, existen 20 aminoácidos estándar y se conocen como los aminoácidos proteicos.

Péptido

- Un péptido es una cadena corta de aminoácidos (habitualmente de 2 a 50) vinculados por uniones químicas (denominados enlaces peptídicos). Una cadena más larga de aminoácidos unidos (51 o más) es un polipéptido. Los péptidos se organizan en estructuras más complejas, que se llaman proteínas. Las proteínas son los bloques de construcción de la célula.



Proteina

- Las proteínas son moléculas grandes, vitales en la mayoría de trabajos que realizan las células y necesarias para mantener la estructura, función y regulación de los tejidos. Están formada por una o más cadenas largas, plegadas de aminoácidos, cuyas secuencias están determinadas por la secuencia de ADN del gen que codifica la proteína. En el genoma humano, hay aproximadamente 20.000 genes que codifican para la producción de proteínas.



Linealidad

Los modelos de lenguaje están diseñados para captar patrones y relaciones en secuencias de texto, lo que es directamente aplicable a las secuencias de aminoácidos en péptidos.

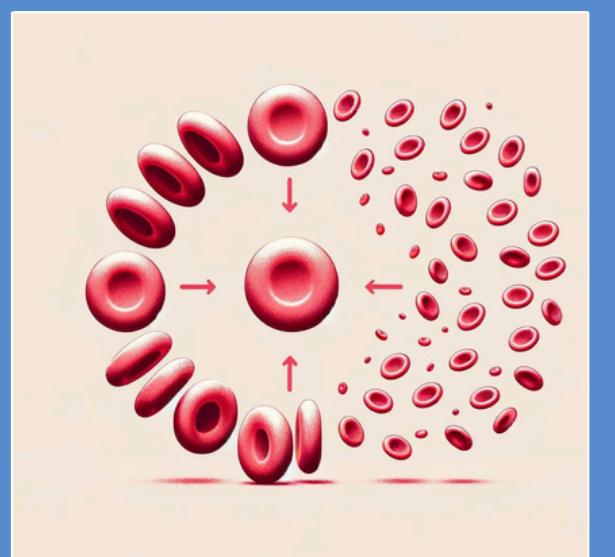
Dependencia

La función y las propiedades de un péptido dependen no solo de los aminoácidos individuales sino también de su contexto y su posición en la secuencia.

Atención

Los modelos de lenguaje utilizan mecanismos de atención para identificar las relaciones entre diferentes partes de la secuencia de texto.

Propiedades de interés



Hemólisis

La hemólisis es la ruptura de los glóbulos rojos, liberando hemoglobina en el entorno. En el contexto de los péptidos, la hemólisis se refiere a la capacidad de un péptido para inducir esta ruptura.



Capacidad de un péptido para disolverse en un solvente, generalmente agua.

Péptidos insolubles pueden ser difíciles de administrar y menos eficaces.



Capacidad de un péptido para resistir interacciones no específicas, evitando la adhesión de moléculas no deseadas a su superficie. Importantes para evitar la acumulación de bacterias y otros microorganismos en superficies médicas.

Datasets

Hemólisis: Database of antimicrobial Activity and structure of peptides [DBAASPv3](#)

Train: 9316 muestras. 19.6% + y 80.4% -

Solubilidad: Protein Data Bank [PDB](#)

Train: 18453 muestras. 47.6% + y 52.4% -

Non-fouling: Database of antimicrobial Activity and structure of peptides [DBAASPv3](#)

Train: 17185 muestras. 20.9% + y 69.1% -

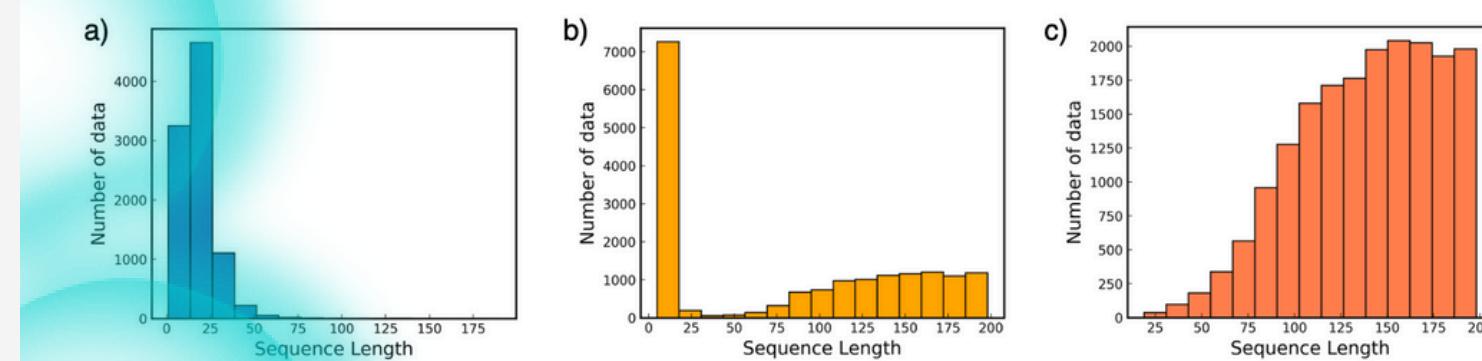
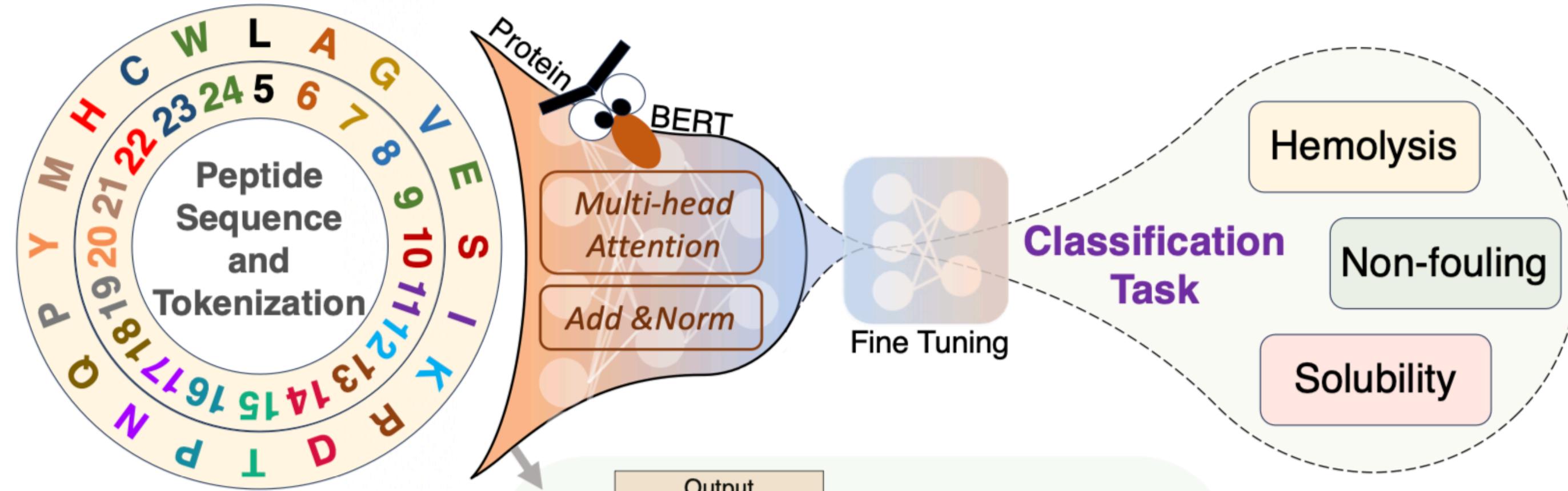


Figure 2: Sequence Length of each Peptide property dataset (a) Hemolysis, (b) Non-fouling and (c) Solubility.

Arquitectura de PeptideBert



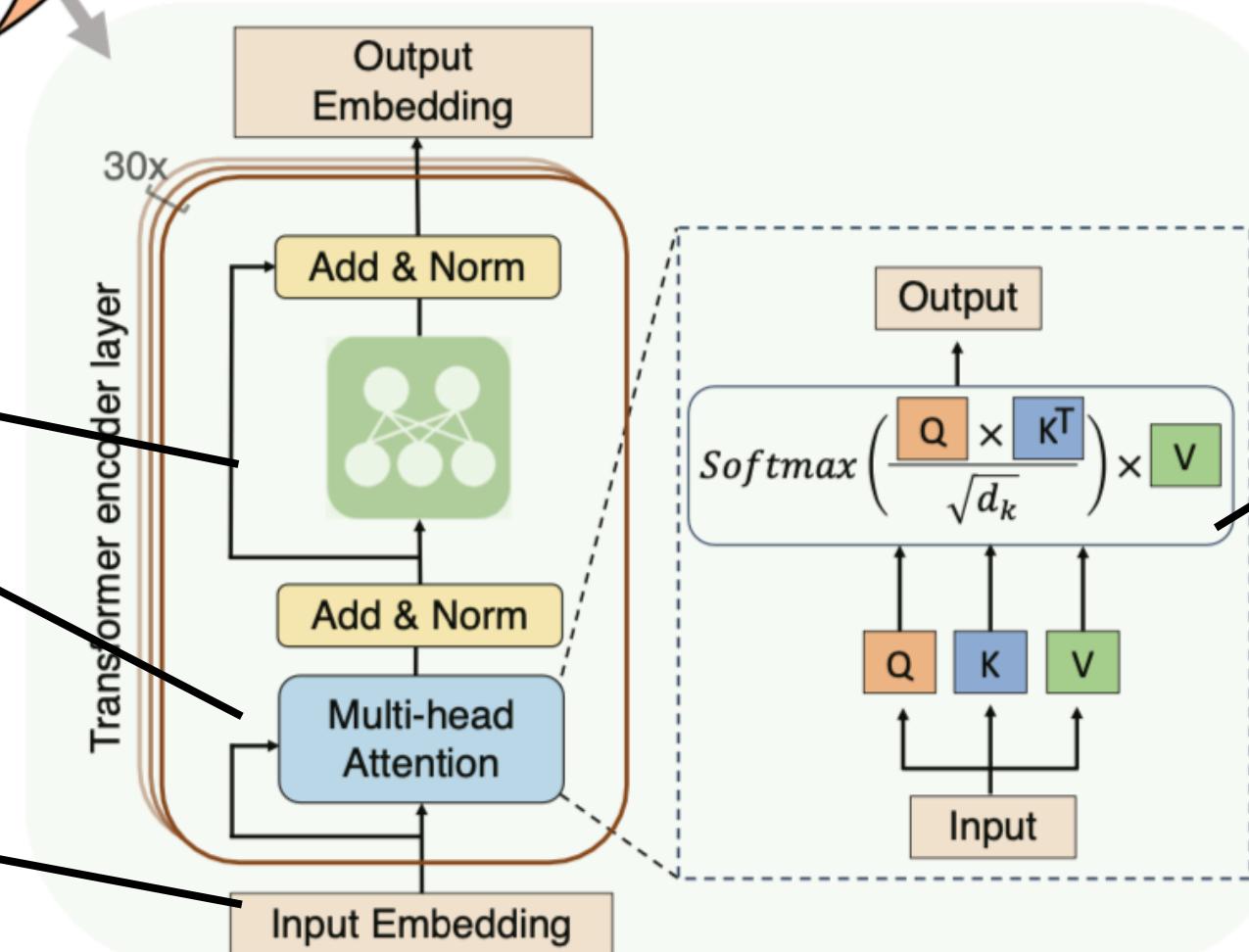
30 capas de codificación:

*Feed-Forward

Network

*Multihead attention

Las secuencias de aminoácidos
se convierten en vectores de alta
dimensión



Se obtienen los pesos de
atención

Protein-Bert

ProtBERT tiene la misma arquitectura que BERT, pero su preentrenamiento se llevó a cabo utilizando secuencias de proteínas en lugar de texto natural. De igual manera, sigue la misma arquitectura bidireccional, incluye capas de transformers, mecanismos de atención, y la capacidad de procesar contextos a la izquierda y derecha simultáneamente usando un Masked LM.

El enmascaramiento sigue el entrenamiento original de Bert:

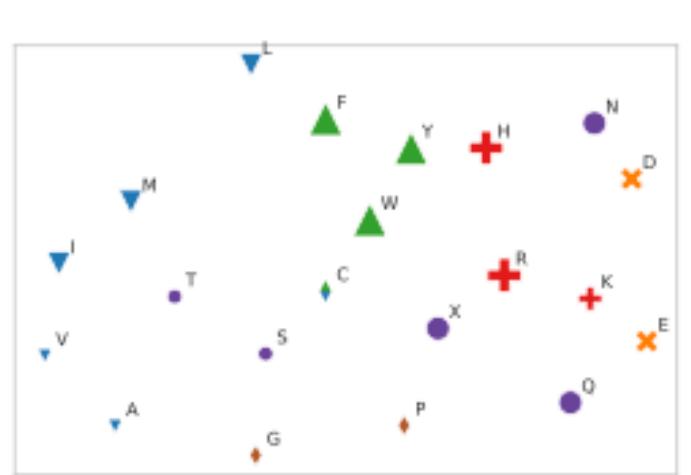
- El 15% de los aminoácidos están enmascarados.
- En el 80% de los casos, los aminoácidos enmascarados se reemplazan por [MASK].

Datasets:

- UniRef100: 216 millones de secuencias de proteínas.
- BFD: 2.1 mil millones de secuencias de proteínas.

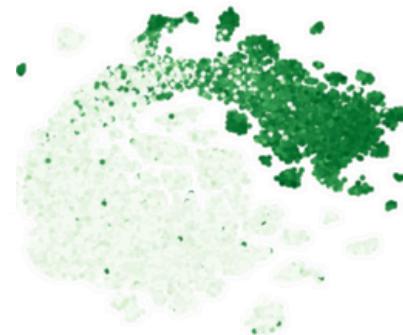
Preproceso de los datos:

- Cada aminoácido es un token
- Cada secuencia aminoácidos (polipéptido) se almacenó en una línea separada:
[CLS] Secuencia de Proteína A [SEP] Secuencia de Proteína B [SEP]
- [UNK] para aminoácidos no protéicos



(b) ProtBert Model
Amino Acids

t-SNE of the sequence embeddings learned by the Transformer for green fluorescent protein variants, colored by their log-fluorescence.



Mask out a random portion,

MSKGE?LFT?VVP?ILVELDGDV?GHKFVSVS...

and ask the model to fill in the rest.

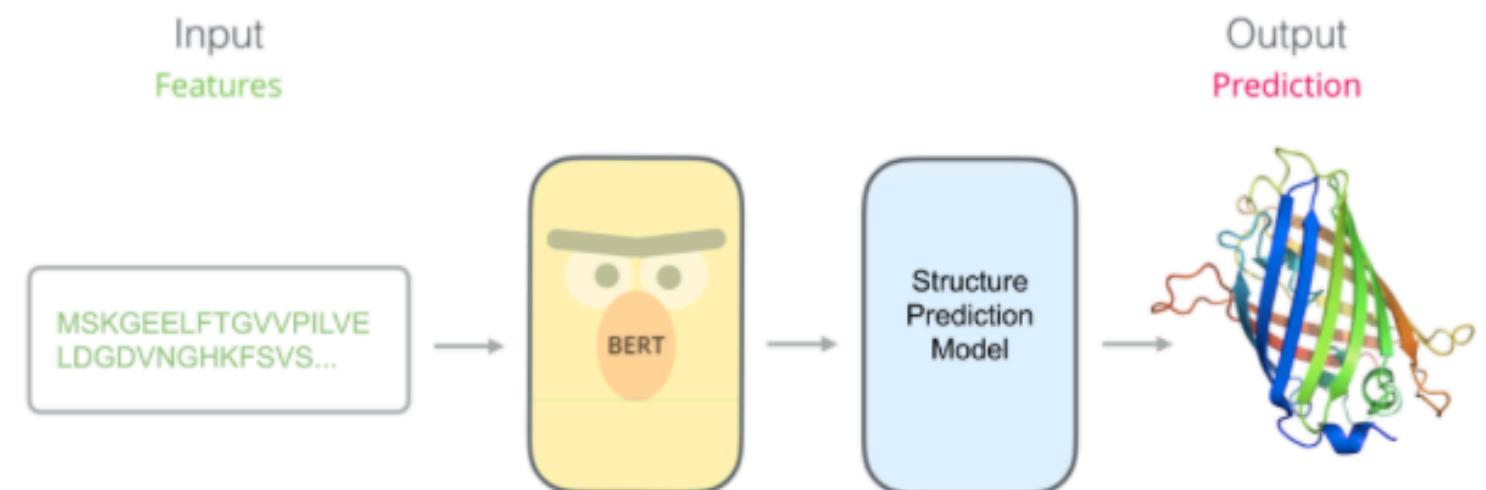


Figure 6: 2D t-SNE projections of uncontextualized token embeddings for single amino acids: all models learnt to cluster the 20 standard amino acids according to their biochemical and biophysical properties, i.e. hydrophobicity, charge and size. For example, the mostly hydrophobic and polar role of Cysteine (C) is conserved.

Fine-Tuning

Regression

Head: Después de ProtBERT, se añade una capa de regresión, una red neuronal totalmente conectada de 480 nodos. Su función es tomar las salidas de ProtBERT y mapearlas a un valor continuo.

Función de activación:

La salida de la capa de regresión se pasa a través de una función Sísmoide, que convierte el valor continuo en un valor entre 0 y 1. El valor resultante de la función Sísmoide se umbraliza (threshold) a 0.5 para obtener la predicción binaria final

Entrenamiento:

Para cada tarea específica, se afinó un modelo utilizando los siguientes parámetros:

- Optimizador: AdamW.
- Función de Pérdida: binary cross-entropy
- Learning-rate: 0.00001
- Batch-size: 32
- Número de Épocas: 30

Hyperparameter	Optimal value
Initial LR	$1.0 * 10^{-5}$
Batch Size	32
Number of Attention Heads	12
Number of Hidden Layers	12
Hidden Size	480
Hidden Layer Dropout	0.15
LR Scheduler (factor)	0.1
LR Scheduler (Patience)	4

Task	Training time (minutes)
Nonfouling	58.28
Hemolysis	69.28
Solubility	116.42

El proceso descrito combina la potencia de ProtBERT para extraer características ricas de las secuencias de proteínas con una capa de regresión sencilla pero efectiva para realizar predicciones binarias. La elección de la arquitectura de la capa de regresión se optimiza mediante experimentos para asegurar el mejor rendimiento posible del modelo.

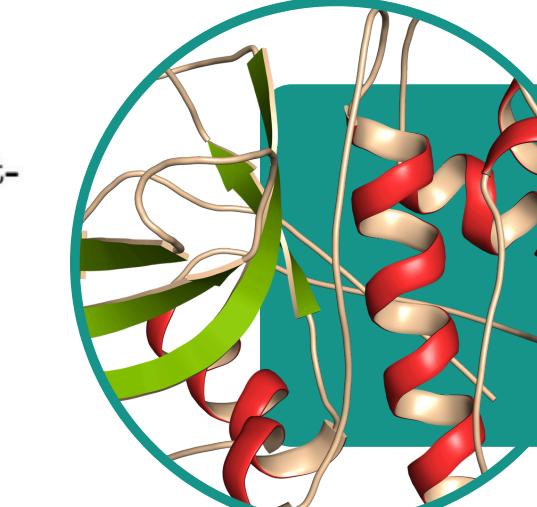
Resultados

Table 2: Classification Accuracy Comparison of previous methods and our Prot-BERT based approach on each of the 3 prediction tasks

Approach	Task	Accuracy(%)
PeptideBERT (Ours)	Non-fouling	88.365
Embedding + LSTM	Non-fouling	82.0
PeptideBERT (Ours)	Hemolysis	86.051
Embedding + Bi-LSTM	Hemolysis	84.0
UniRep + Logistic Regression	Hemolysis	82.0
UniRep + Random Forests	Hemolysis	84.0
HAPPENN ³⁴	Hemolysis	85.7
HLPpred-Fuse ⁴⁹	Hemolysis	-
one-hots + RNN ⁵⁰	Hemolysis	76.0
PeptideBERT (Ours) (With Augmentation)	Solubility	70.018
PeptideBERT (Ours) (Without Augmentation)	Solubility	69.175
Embedding + Bi-LSTM	Solubility	70.0
PROSO II ⁴²	Solubility	71.0
DSResSol (1) ²⁹	Solubility	75.1

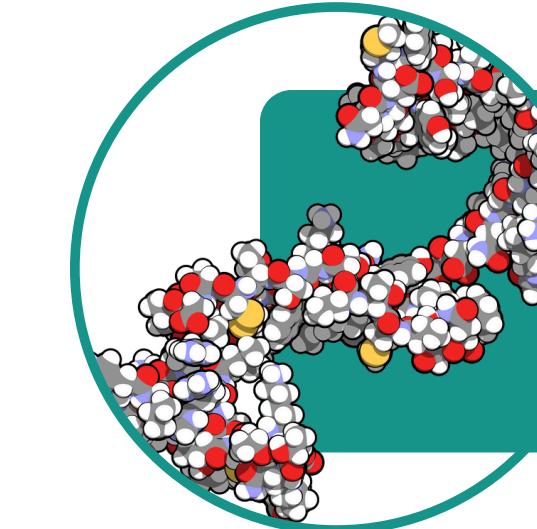
Predictión de Hemólisis

alcanzó una precisión del 86.051% en la predicción de hemólisis, superando a otros modelos existentes.



Predictión de Solubilidad

logró una precisión del 70.018% en la predicción de solubilidad de los péptidos. El aumento de datos contribuyó a mejorar la capacidad del modelo para predecir la solubilidad aunque no significativamente.



Predictión de Antifouling

obtuvo una precisión del 88.365% superior en comparación con otros modelos, demostrando la eficacia de PeptideBERT.

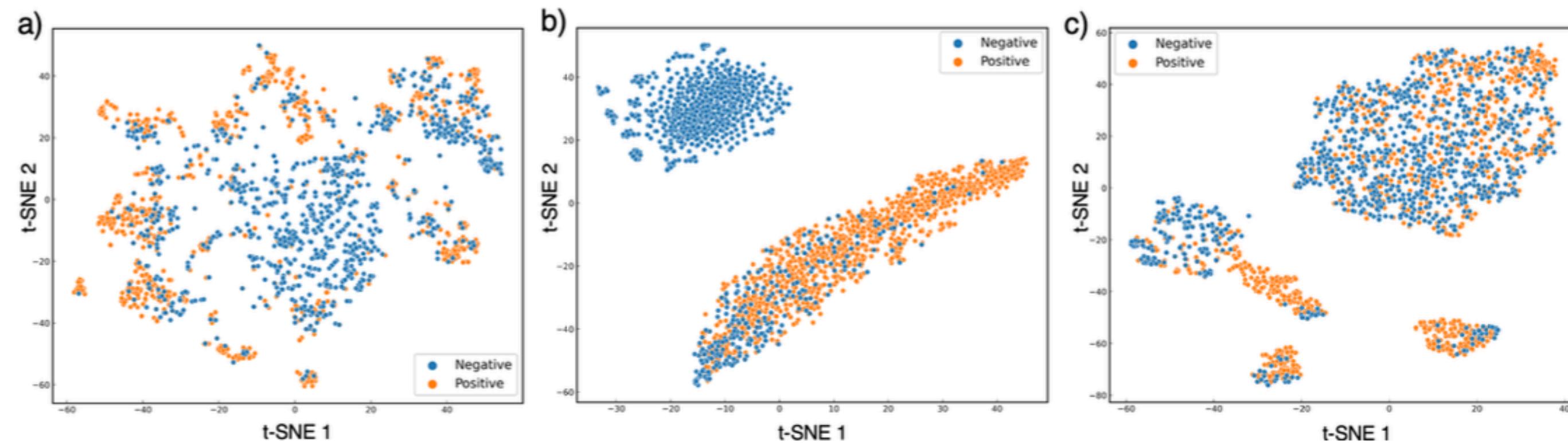
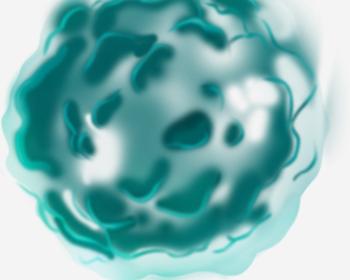
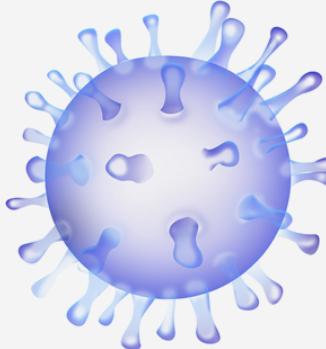


Figure 3: t-SNE visualization of peptide properties (a) Hemolysis, (b) Non-fouling and (c) Solubility. The [CLS] token embedding from the last hidden state of PeptideBERT is visualized after dimensionality reduction.

Mejoramiento de Atención: permite que cada embedding de token en el codificador capture la información de toda la secuencia de entrada.

El tokenizador de ProtBERT: agrega un token [CLS] al inicio de cada secuencia. Diseñado para contener información de todos los embeddings de tokens en la secuencia.

Extracción: Para visualizar la comprensión de PeptideBERT sobre diversas secuencias de péptidos se extrajeron los tokens [CLS]



Bibliografía

Artículos

<https://arxiv.org/pdf/2309.03099v1> PeptideBert

<https://pubmed.ncbi.nlm.nih.gov/34232869/> ProtTrans

https://huggingface.co/Rostlab/prot_bert/blob/main/README.md

<https://pubs.acs.org/doi/epdf/10.1021/acs.jcim.2c01317> RNN

Web

<https://www.genome.gov/genetics-glossary> NHRGI

<https://bair.berkeley.edu/blog/2019/11/04/proteins/>
[Language of Proteins \(Blog\)](#)

<https://www.uniprot.org/help/uniref>

