

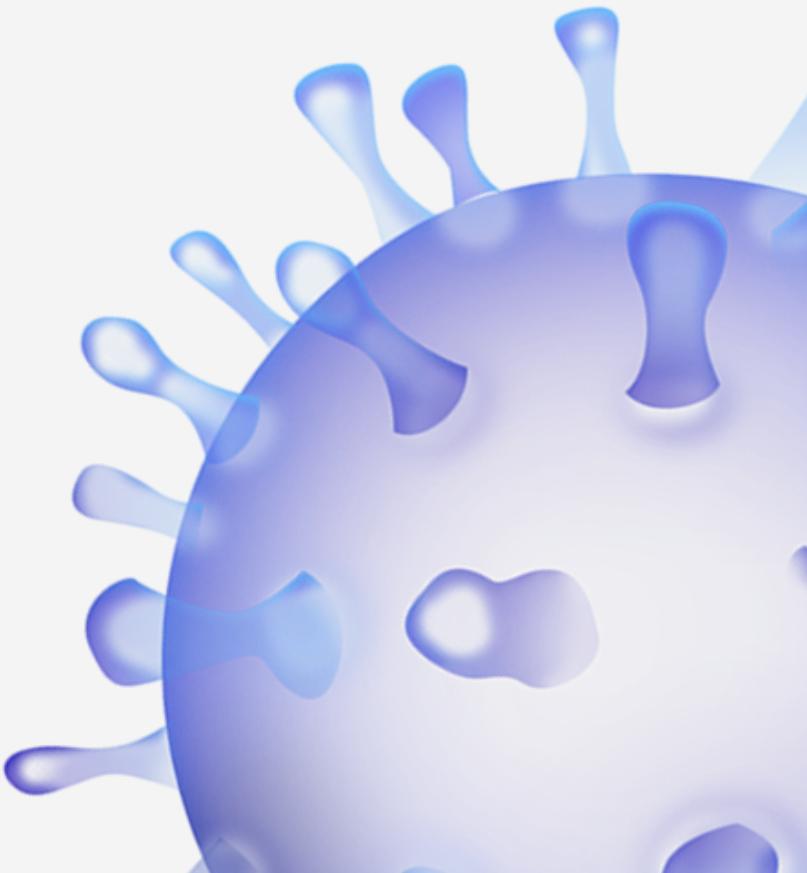


PROTEIN-BERT

A UNIVERSAL DEEP-LEARNING MODEL OF PROTEIN SEQUENCE AND FUNCTION

-Nadav Brandes et al

Proyecto Final IA
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

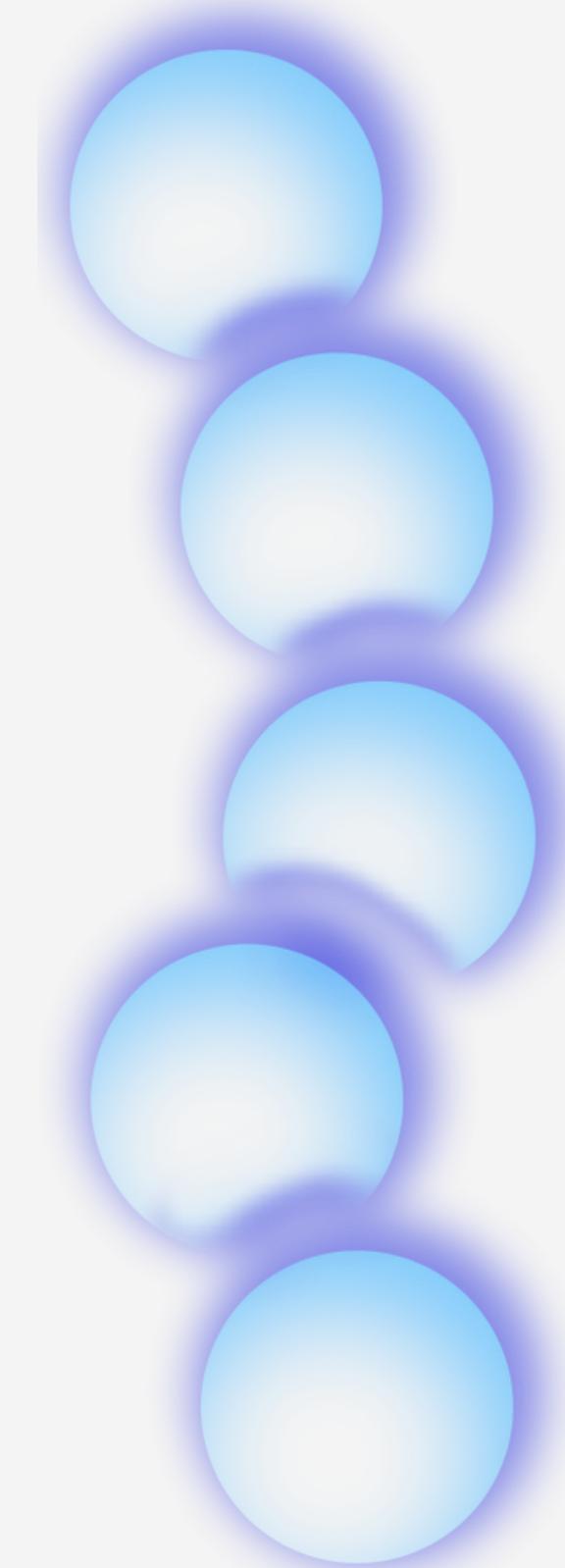
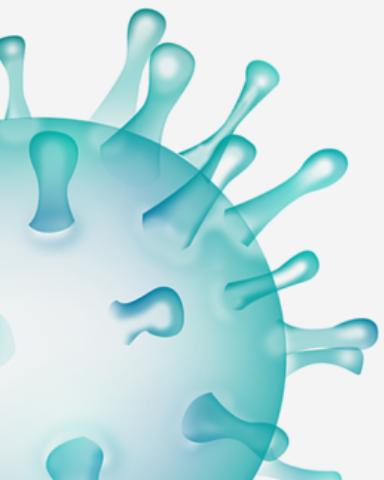
El paper basa en los avances recientes en modelos de lenguaje y su aplicación en la predicción de propiedades de péptidos.

03 Modelo ProteinBERT:

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

04 Resultados

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

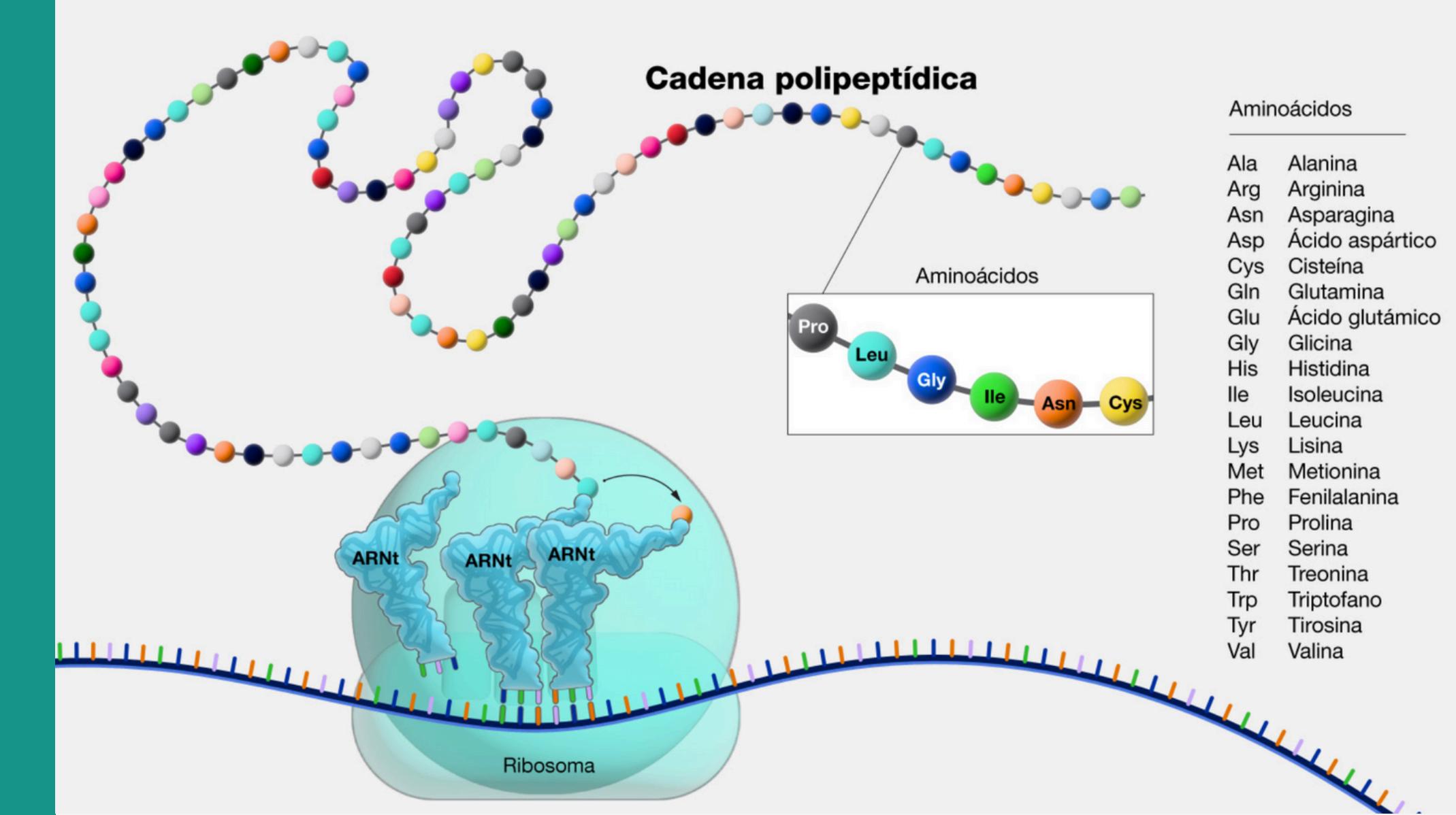


Aminoácido

- 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.

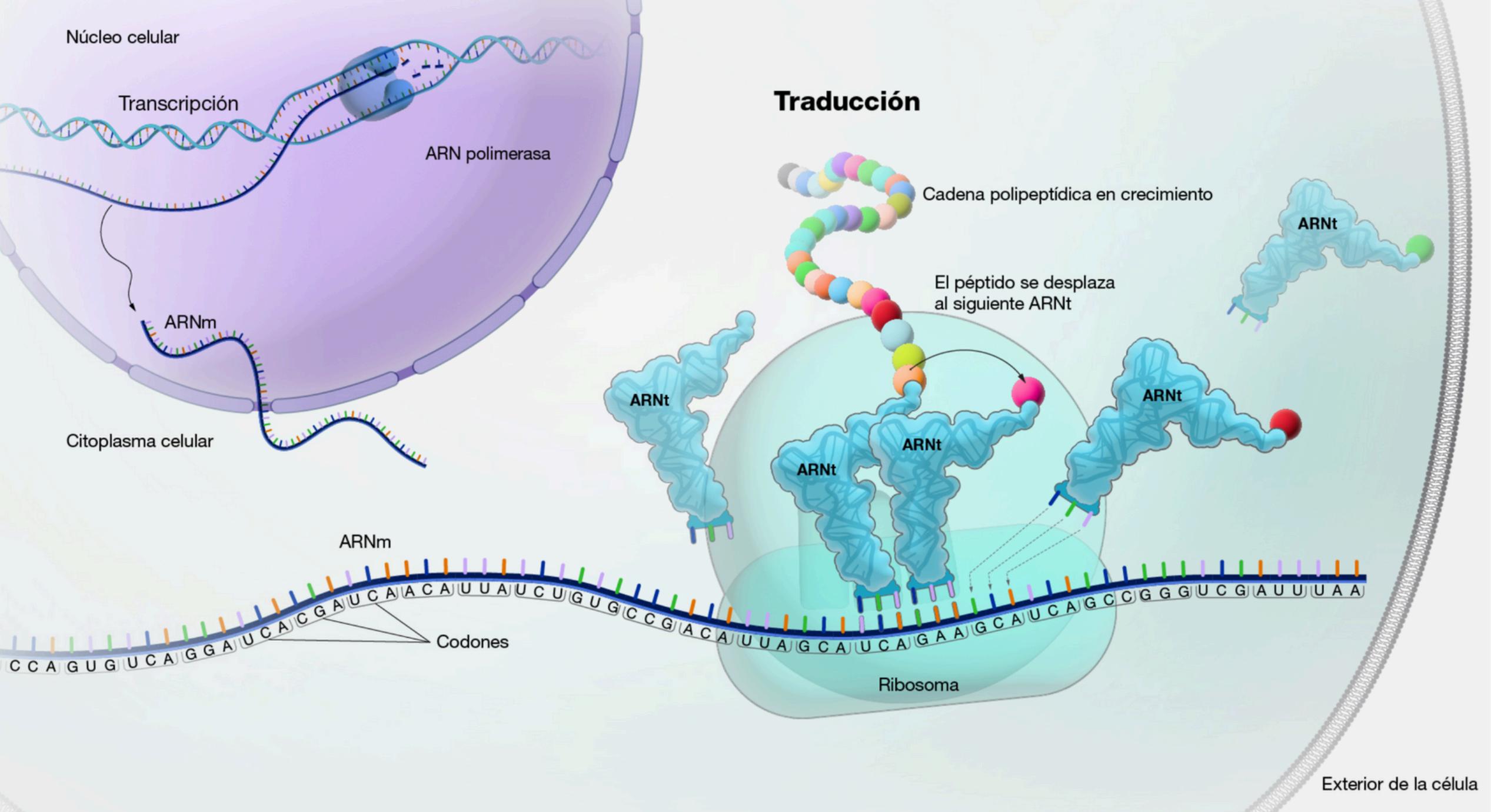
Péptido

- 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

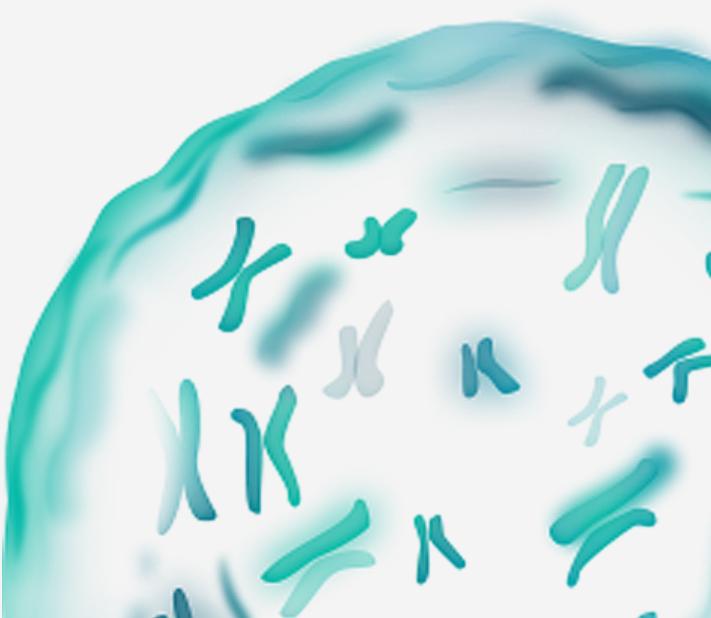
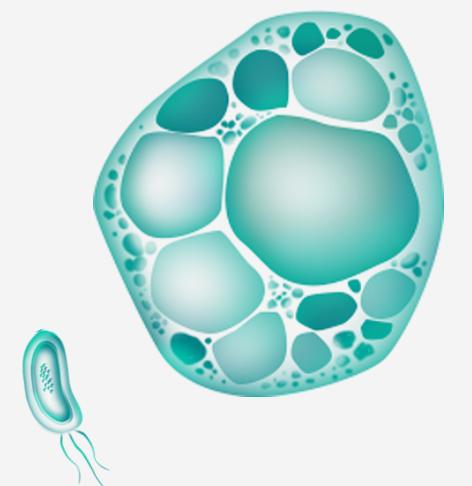
- Las proteínas son moléculas grandes, vitales en la mayoría de trabajos que realizan las células y necesarias para mantener la estructura, función y regulación de los tejidos. Están formada por una o más cadenas largas, plegadas de aminoácidos, cuyas secuencias están determinadas por la secuencia de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



¿Polipéptidos=Texto?

- Secuencias Lineales
- Contexto y Dependencia
- Mecanismo de Atención

¿Cómo se relaciona esto con el procesamiento del lenguaje?





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.

TOKENIZACIÓN

Symbols for Amino Acids

Amino acid	Three letter	One letter
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamine	Gln	Q
Glutamic acid	Glu	E
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

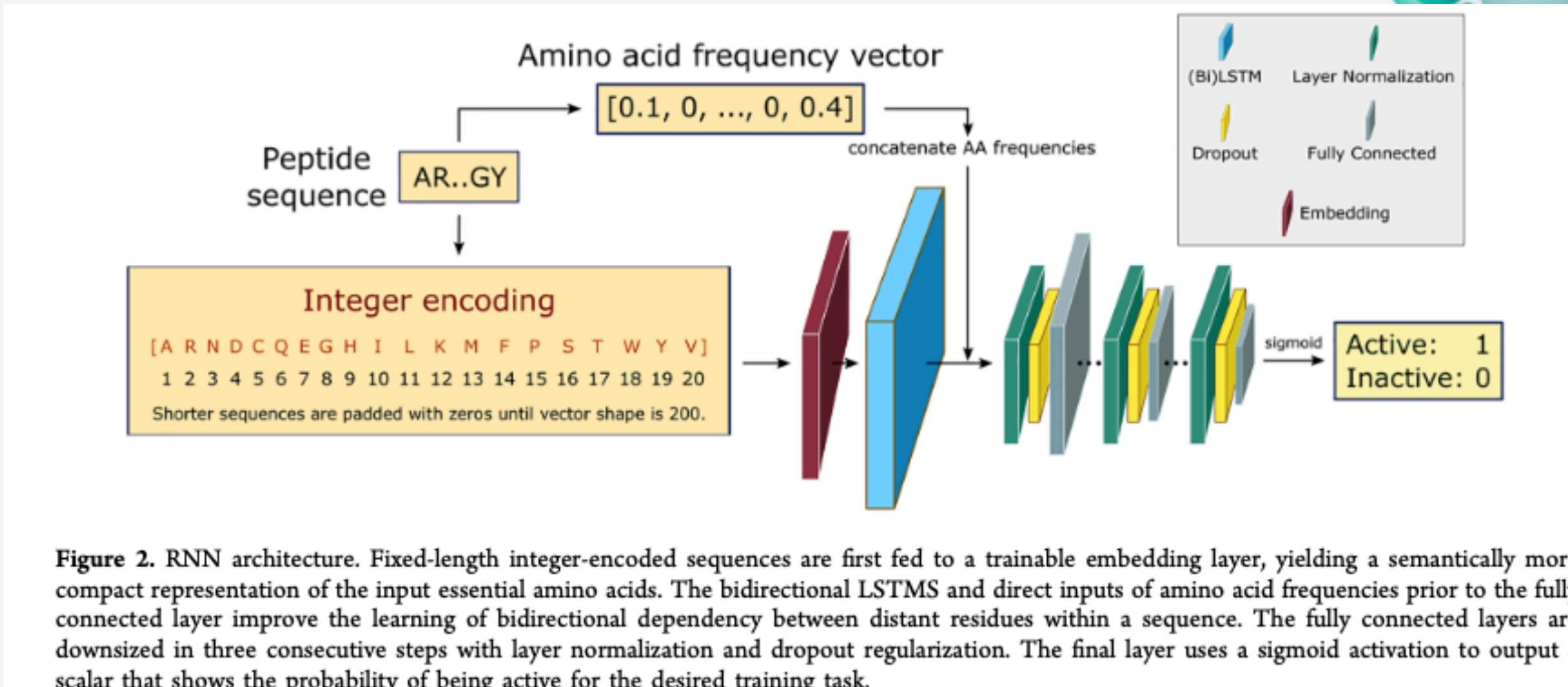


Figure 2. RNN architecture. Fixed-length integer-encoded sequences are first fed to a trainable embedding layer, yielding a semantically more compact representation of the input essential amino acids. The bidirectional LSTMs and direct inputs of amino acid frequencies prior to the fully connected layer improve the learning of bidirectional dependency between distant residues within a sequence. The fully connected layers are downsized in three consecutive steps with layer normalization and dropout regularization. The final layer uses a sigmoid activation to output a scalar that shows the probability of being active for the desired training task.

ProtBert

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:

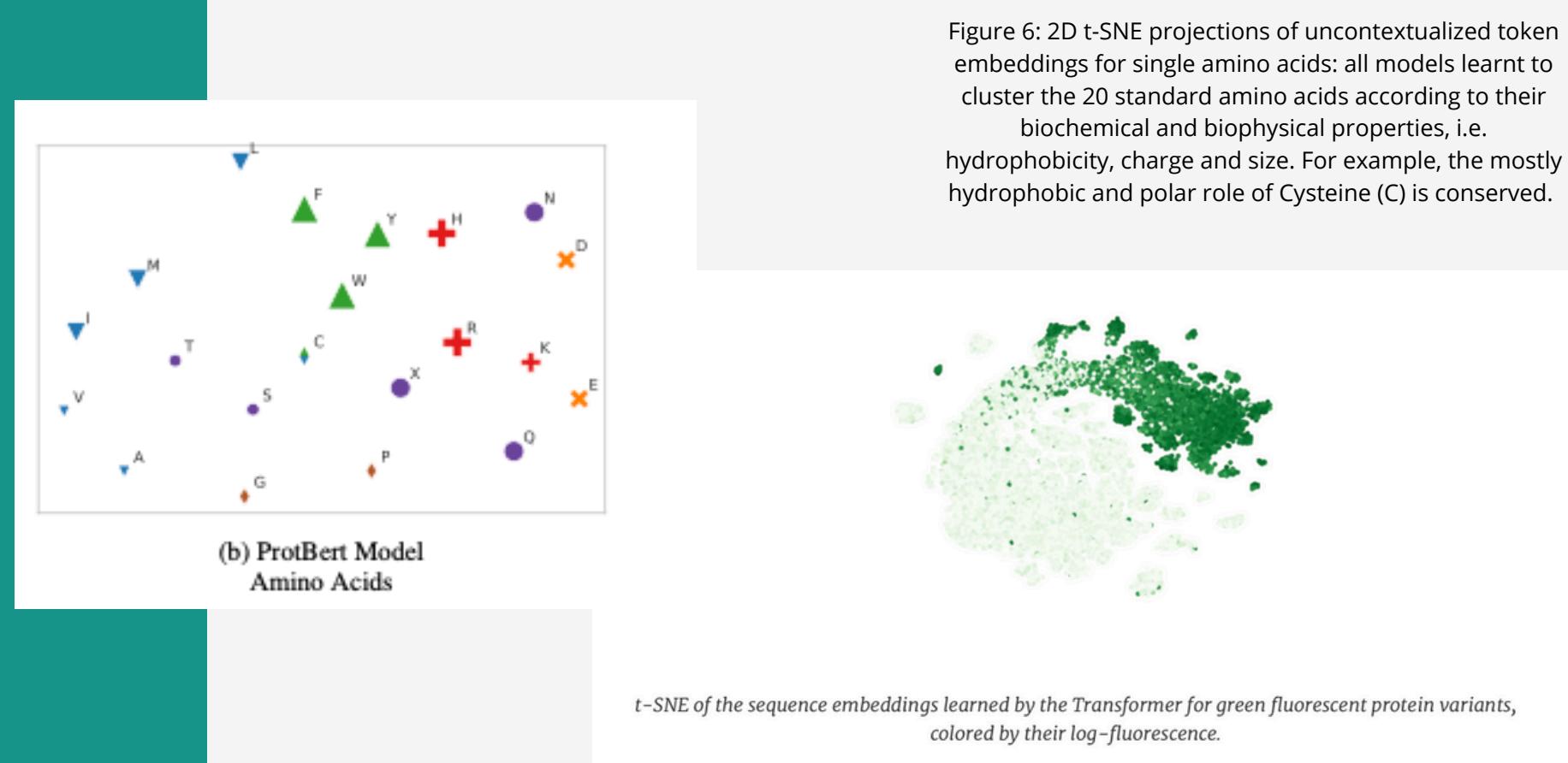
- El 15% de los aminoácidos están enmascarados.
- En el 80% de los casos, los aminoácidos enmascarados se reemplazan por [MASK].

Datasets:

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

Preproceso de los datos:

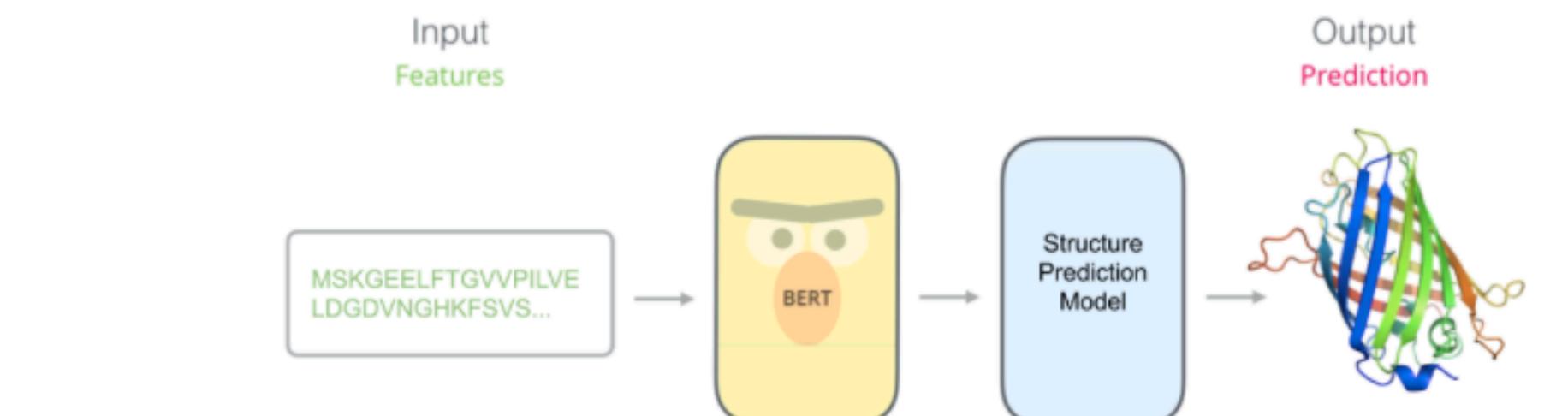
- Cada aminoácido es un token
- Cada secuencia aminoácidos (polipéptido) se almacenó en una línea separada:
[CLS] Secuencia de Proteína A [SEP] Secuencia de Proteína B [SEP]
- [UNK] para aminoácidos no protéicos



Mask out a random portion,

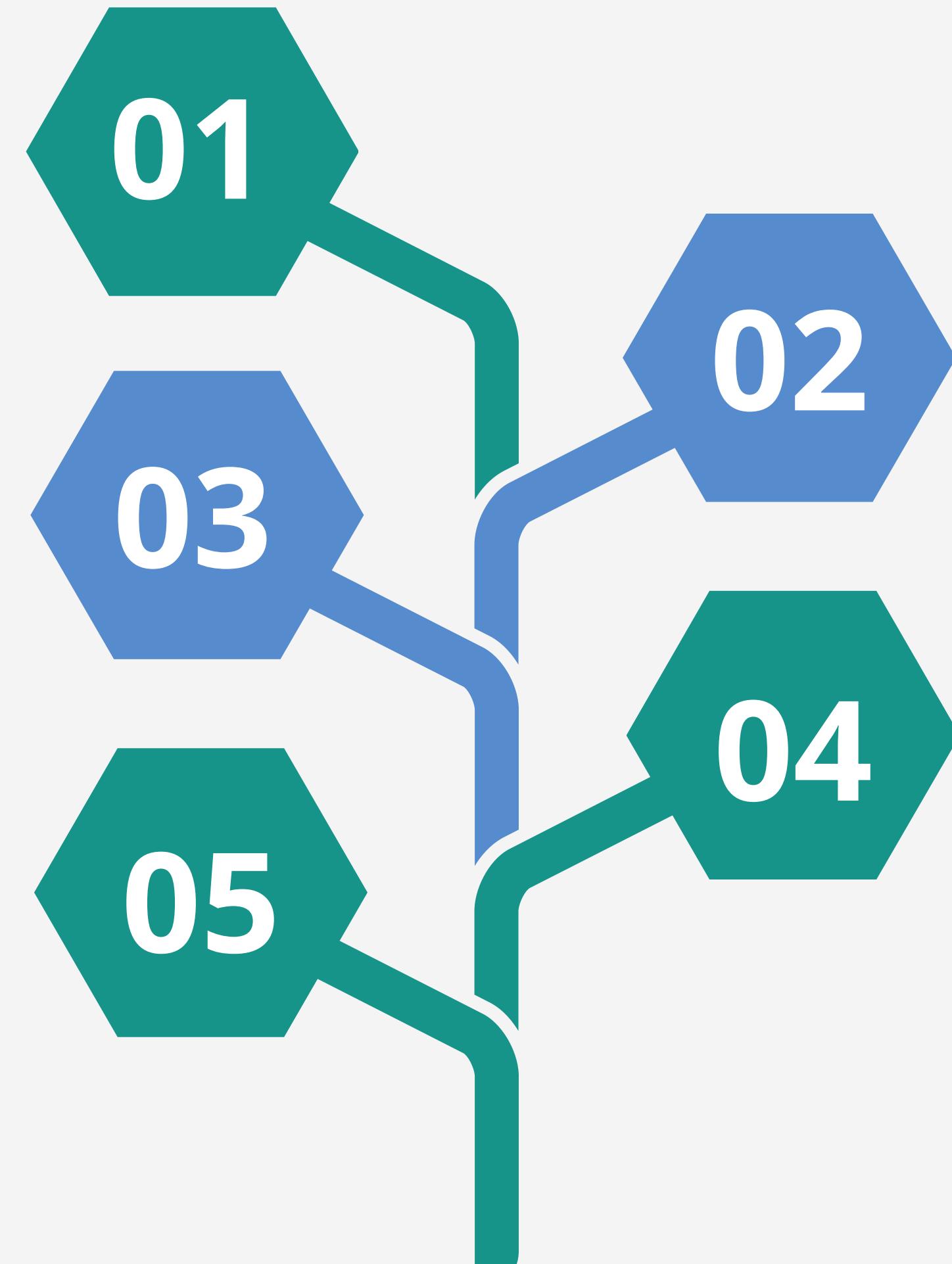
MSKGE?LFT?VVP?ILVELDGDV?GHKFVSVS...

and ask the model to fill in the rest.



Problemas

- 01 LLM son diseñados para lenguaje Natural
- 02 Las proteínas no tienen bloques de construcción multiletra claros
- 03 Existen interacciones entre posiciones MUY distantes
- 04 Las proteínas son Estructuras 3D
- 05 Los LLM no aprovechar las características únicas de las proteínas



Funciones Moleculares:

Actividades Bioquímicas

“Actividad unión ATP”

Componentes Celulares:

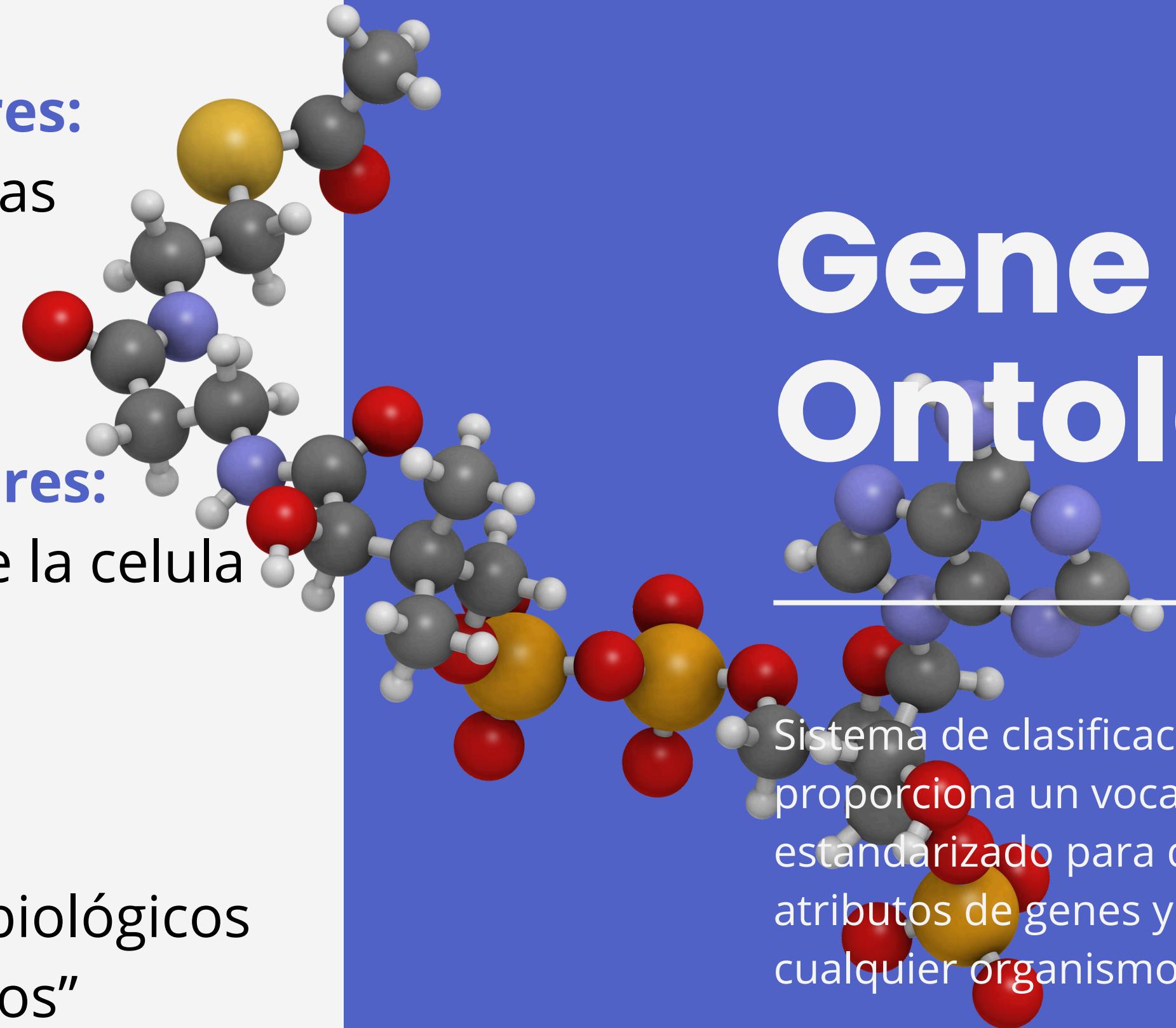
Ubicaciones dentro de la celula

“mitocondria”

Procesos Biológicos:

Conjunto de eventos biológicos

“metabolismo de lípidos”



Gene Ontology

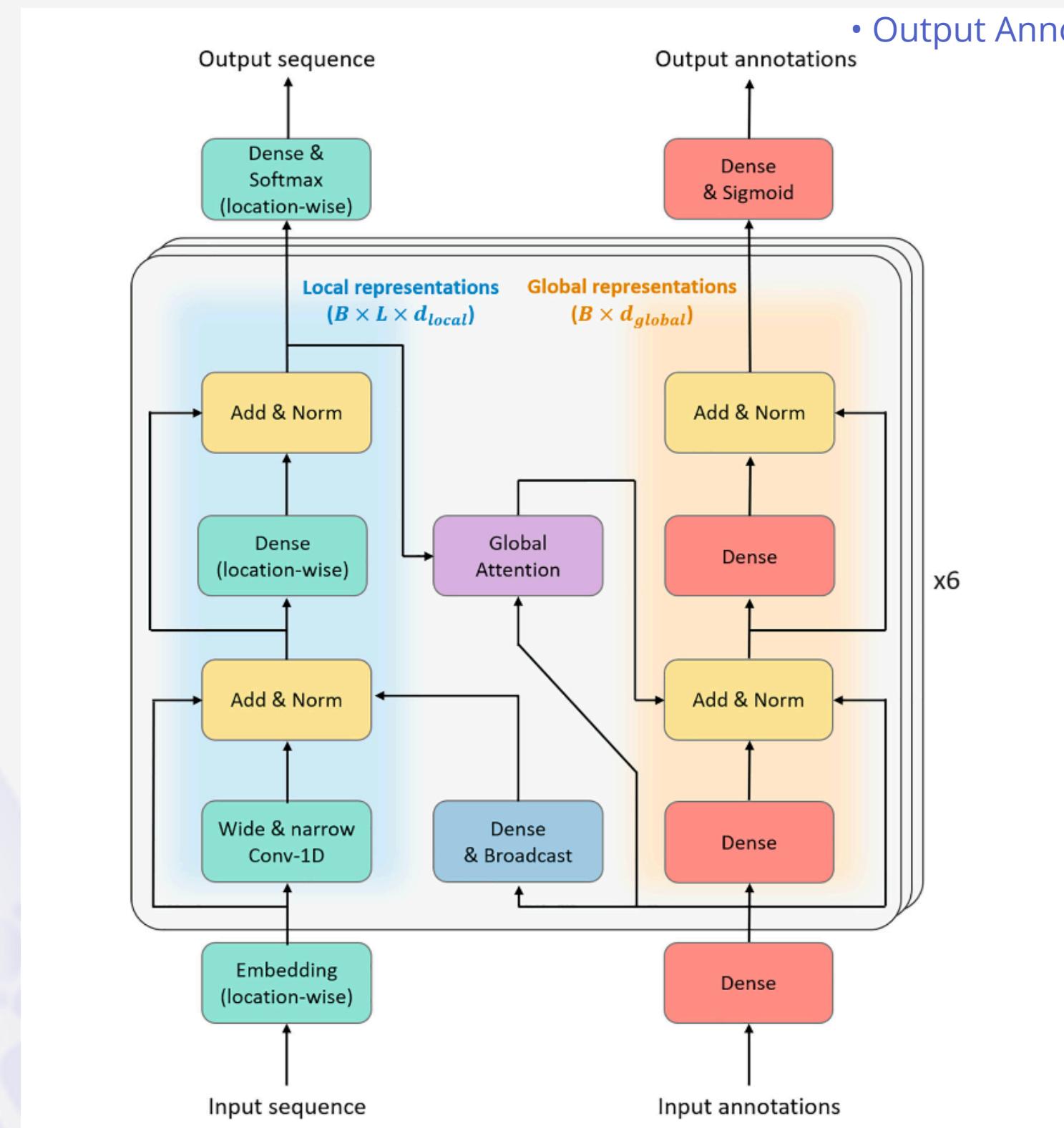
Sistema de clasificación que proporciona un vocabulario estandarizado para describir los atributos de genes y proteínas en cualquier organismo.

Beneficios: Estandarización, comprensión e integración

Arquitectura

Preprocesamiento

Atención global:
Interacción entre
representaciones locales y
globales, similar a la atención en
Transformers.



Salida del Modelo:

- Output Sequence: Secuencia de proteínas generada tras pasar por varias capas densas y activación softmax.
- Output Annotations: Anotaciones de GO predichas tras pasar por una capa densa y activación sigmoide.

Dense & Broadcast:
Difunde representaciones
globales a cada posición de
la secuencia.

Entrada de la Secuencia de
Proteínas (**Input Sequence**) y
Anotaciones (**Input Annotations**)

Atención Global vs. Auto-Regresiva



Transformers Tradicionales

Cada token en una secuencia puede considerar todos los otros tokens para calcular su representación



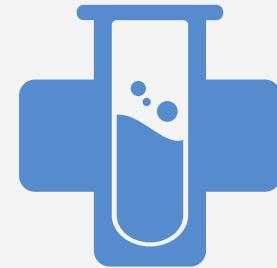
Atencion Global

Permite que las representaciones locales (nivel de aminoácido) interactúen con las representaciones globales (nivel de secuencia completa).



Interacción Local-Global:

Permite que la información a nivel de aminoácido influya en la comprensión general de la secuencia y viceversa



Difusión de Información

asegura que la información relevante a nivel local se propague a la representación global, permitiendo una visión más completa de la secuencia de proteínas

Fine-Tuning

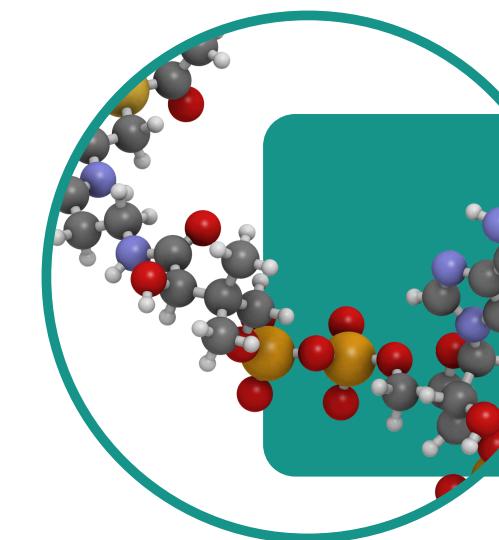
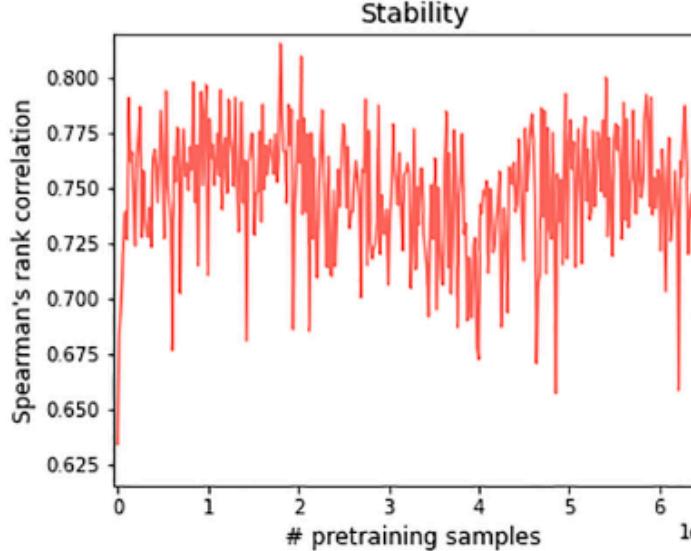
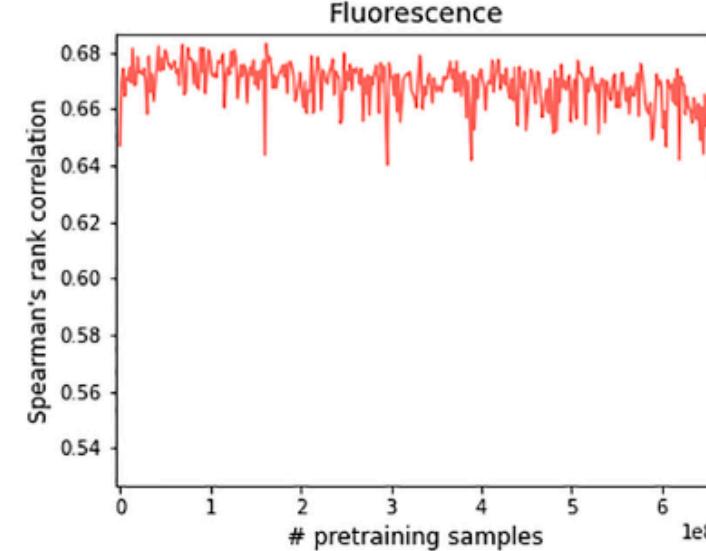
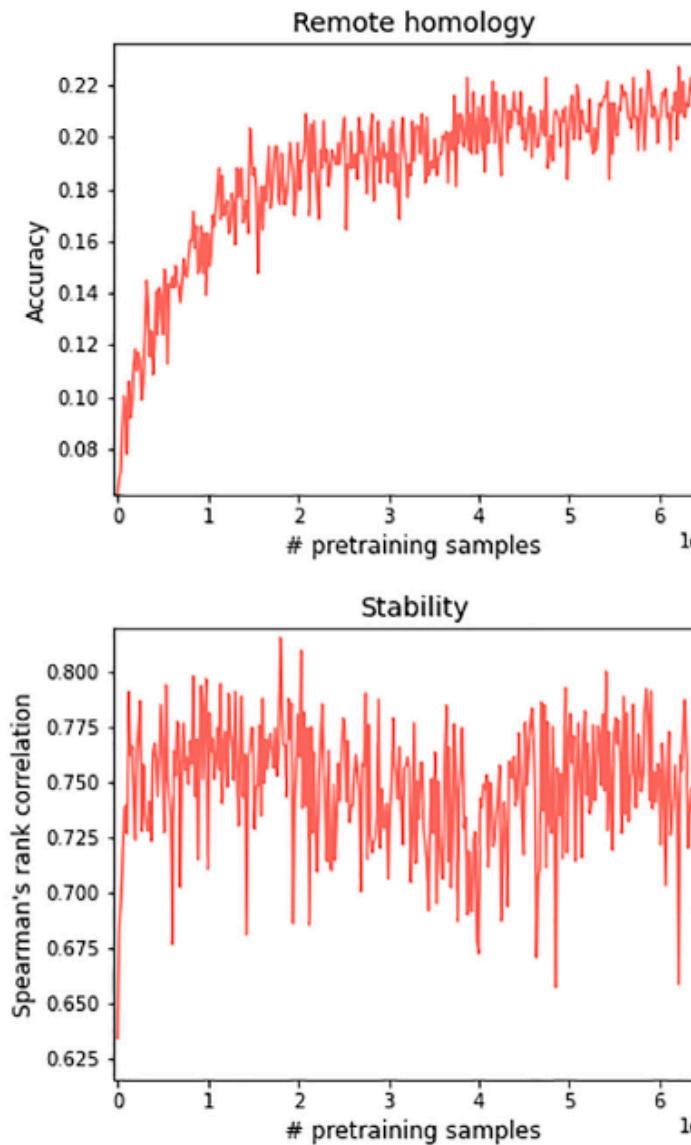
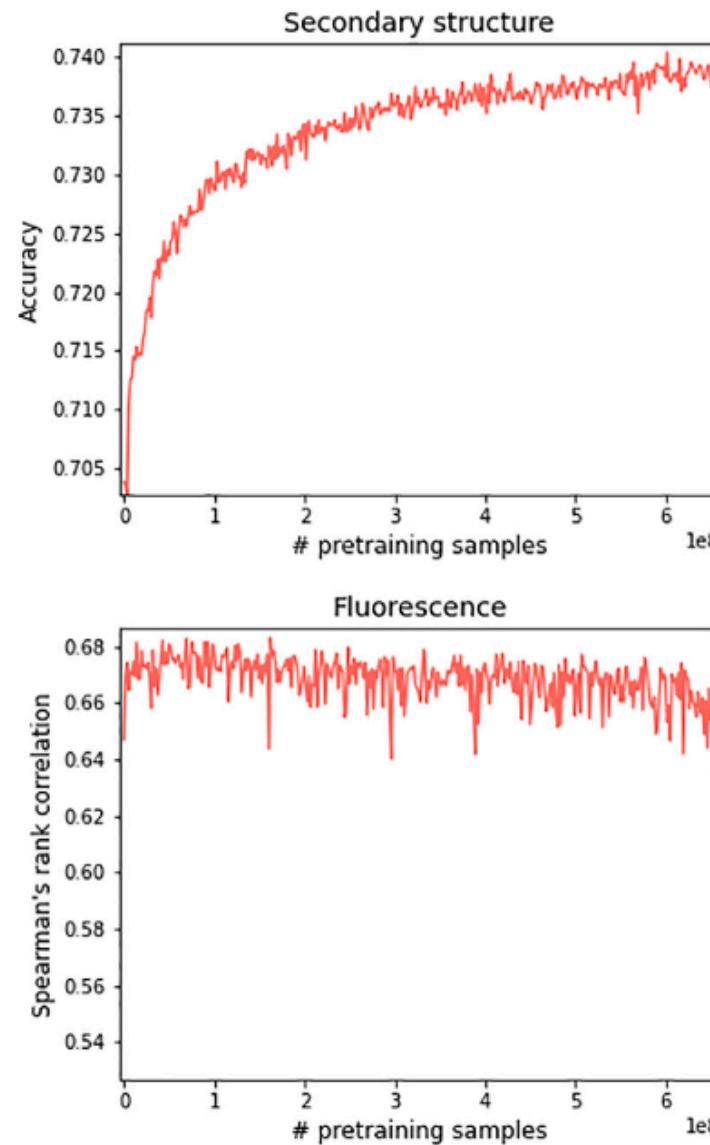
En nueve benchmarks que cubren todas las facetas principales de la investigación de proteínas.

Table 1. Protein benchmarks

Topic	Benchmark	Target type ^a	Resolution	# Training sequences	Source
Protein structure	Secondary structure	Categorical (3)	Local	8,678	(Moult <i>et al.</i> , 2018; Rao <i>et al.</i> , 2019)
	Disorder	Binary	Local	8,678	(Moult <i>et al.</i> , 2018)
	Remote homology	Categorical (1,195)	Global	12,312	(Andreeva <i>et al.</i> , 2014, 2020; Rao <i>et al.</i> , 2019)
	Fold classes	Categorical (7)	Global	15,680	(Andreeva <i>et al.</i> , 2014, 2020)
Post-translational modifications	Signal peptide	Binary	Global	16,606	(Armenteros <i>et al.</i> , 2019)
	Major PTMs	Binary	Local	43,356	(Hornbeck <i>et al.</i> , 2015)
	Neuropeptide cleavage	Binary	Local	2,727	(Ofer and Linial 2014, 2015; Brandes <i>et al.</i> , 2016)
Biophysical properties	Fluorescence	Continuous	Global	21,446	(Sarkisyan <i>et al.</i> , 2016; Rao <i>et al.</i> , 2019)
	Stability	Continuous	Global	53,679	(Rocklin <i>et al.</i> , 2017; Rao <i>et al.</i> , 2019)

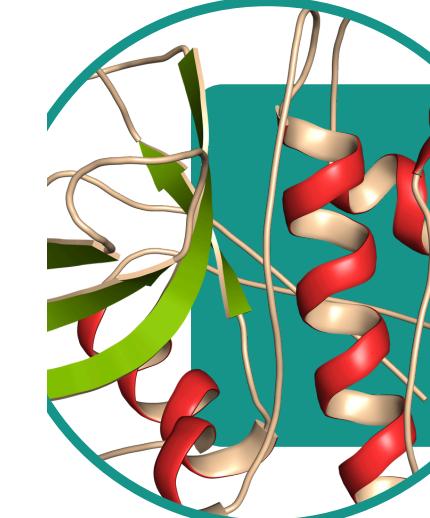
^aFor categorical targets, the number of classes appears in parentheses.

Resultados



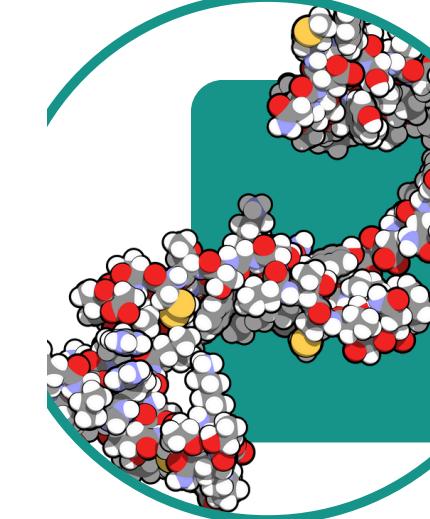
Protein Structure

74% en la predicción de la estructura secundaria, lo cual es comparable con modelos más grandes y complejos.



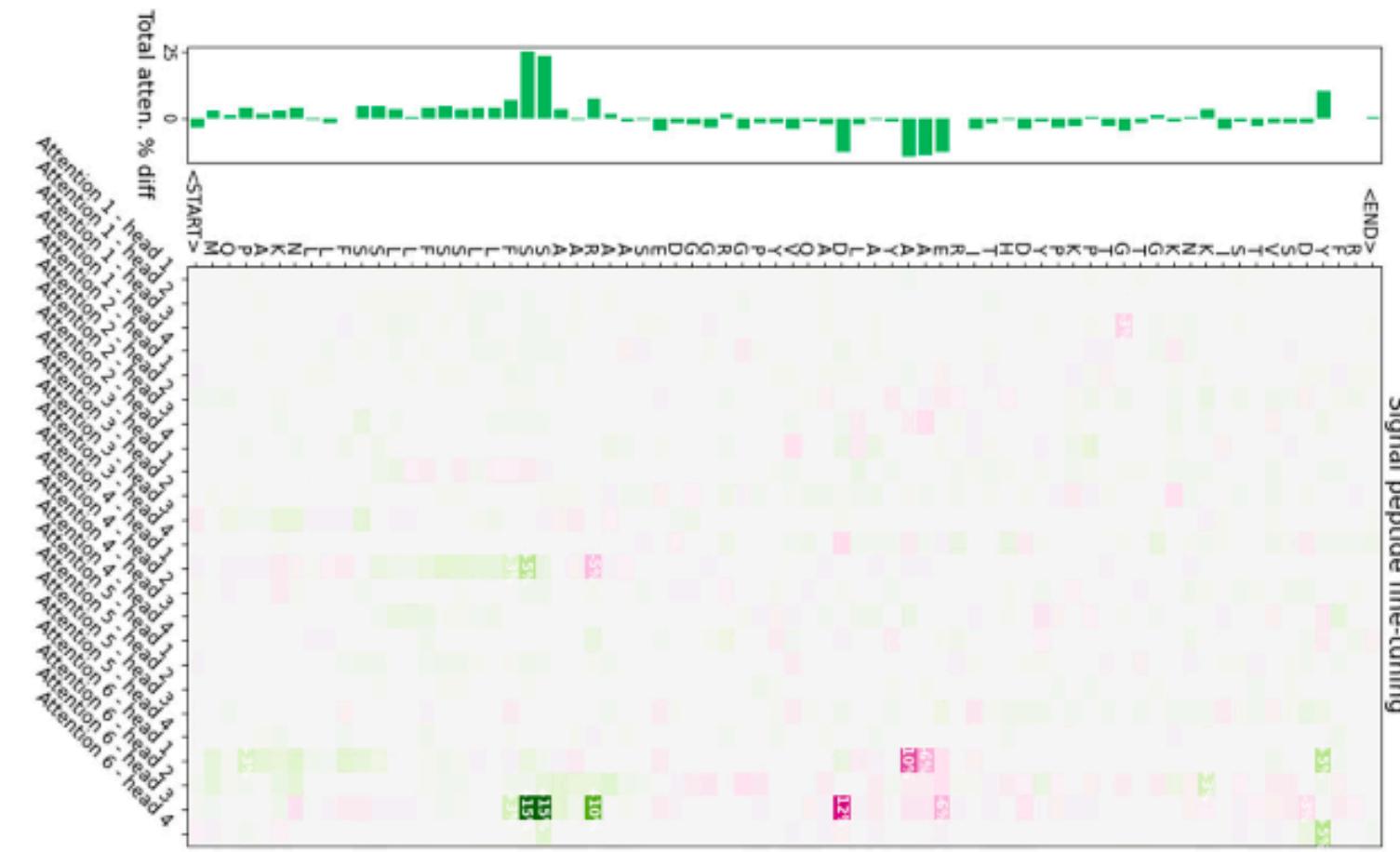
Propiedades Biofísicas

Alta correlación spearman para Fluorescencia y estabilidad

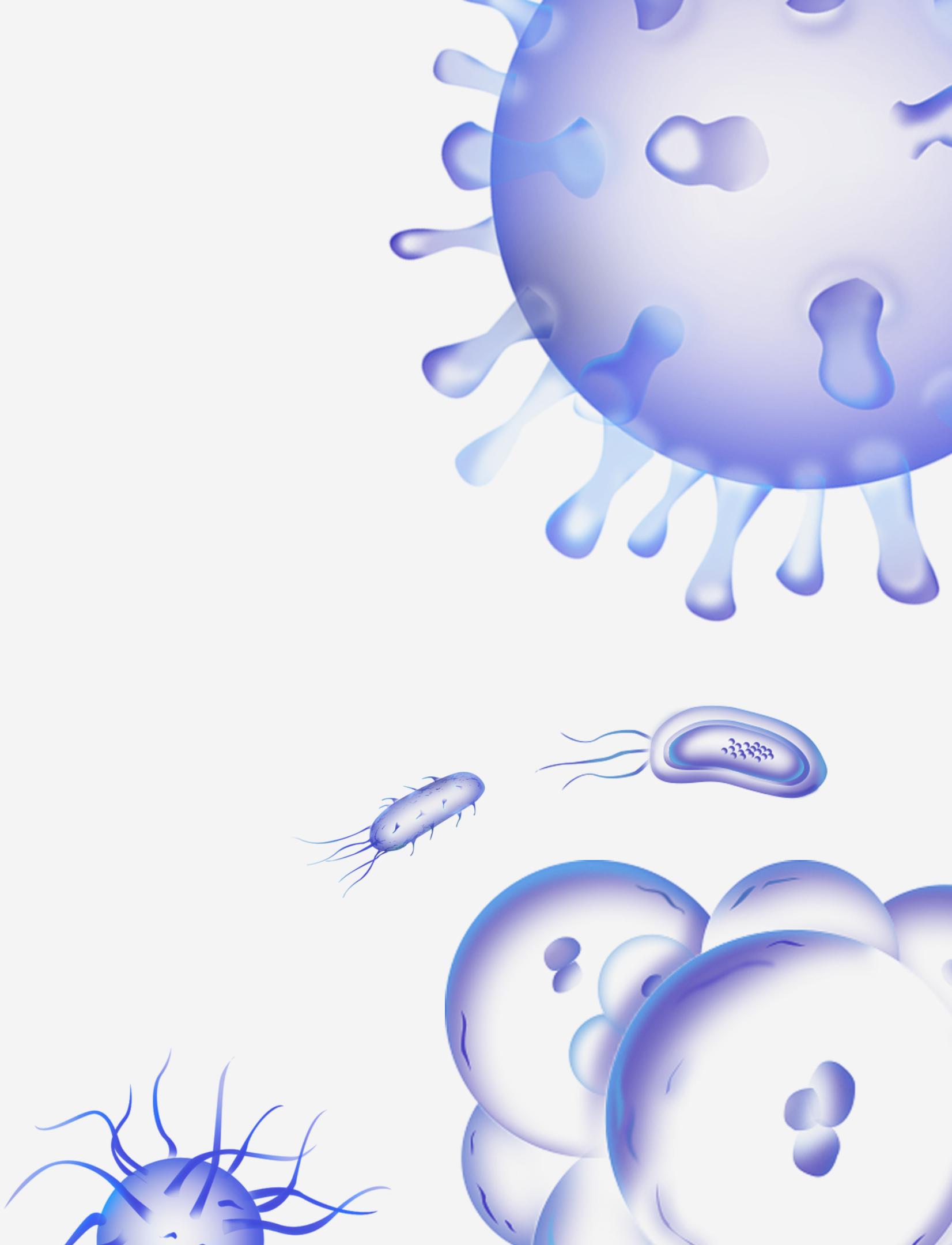


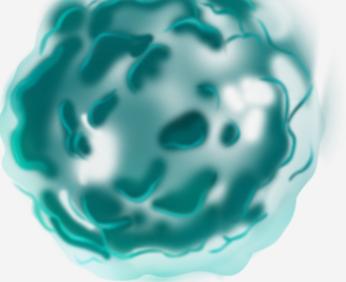
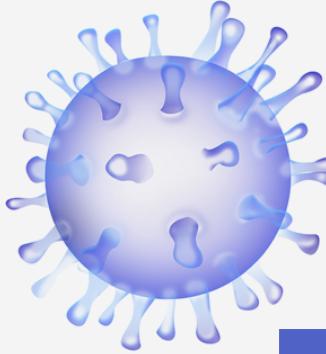
Modificaciones Postraduccionales

El modelo es capaz de predecir modificaciones postraduccionales importantes con alta precisión.



¿DUDAS?





Bibliografía

Artículos

Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., & Linial, M. (2022). ProteinBERT: A universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8), 2102-2110. <https://doi.org/10.1093/bioinformatics/btac020>

Chakradhar, G. et. al. (2023). PeptideBERT: A language model based on transformers for peptide property prediction. *bioRxiv*. <https://arxiv.org/pdf/2309.03099v1>

Elnaggar, A., et. al. (2020). ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing. *bioRxiv*. <https://doi.org/10.1101/2020.07.12.199554>

Ansari, M., & White, A. D. (2023). Serverless prediction of peptide properties with recurrent neural networks. *Journal of Chemical Information and Modeling*, 63(8), 2546-2553. <https://doi.org/10.1021/acs.jcim.2c01317>

Web

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

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

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

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

