Taller SAJIB 9:
Modelos fundacionales para secuencias biológicas

Leandro Bugnon
lbugnon.github.io

UNL CONICET sinc(i)

https://github.com/lbugnon/foundation_models_bioinfo

UNL CONICET

s i n c (*i*)

https://sinc.unl.edu.ar/

https://x.com/sinc_i

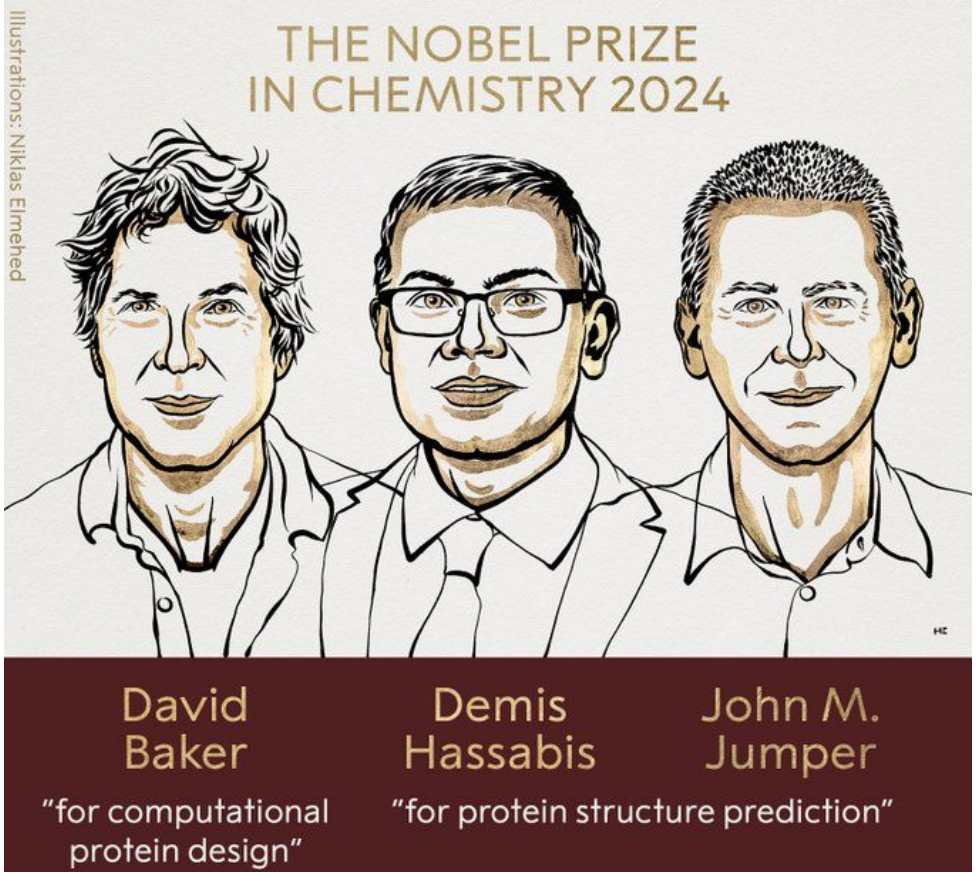
UNL CONICET

sinc(i)

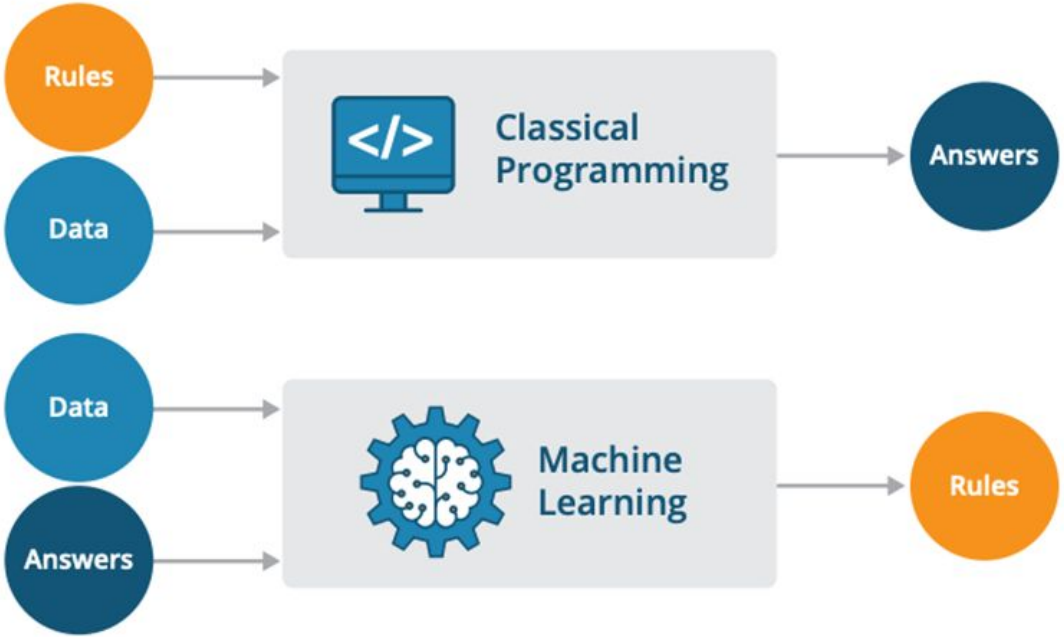
Objetivos

- Incorporar bases de modelos fundacionales y cómo funcionan
- Interpretar las representaciones (embeddings) de secuencias
- Utilización de herramientas de código abierto
- Revisar cuestiones metodológicas
- Resumir algunas de sus aplicaciones en bioinformática

Impacto de la IA en biología computacional

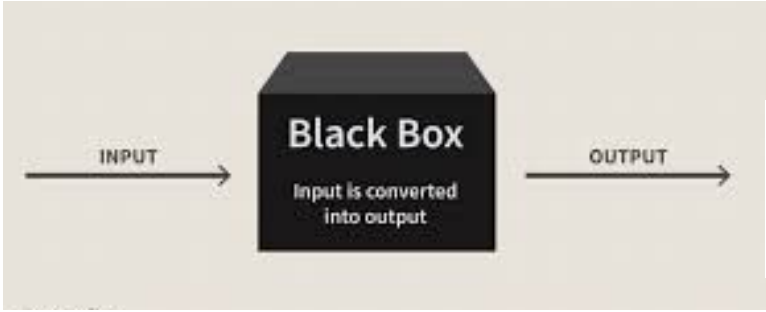


De programación clásica a “modelos fundacionales”



De programación clásica a “modelos fundacionales”

$$X = \{x_i \in \mathbb{R}^M\}$$

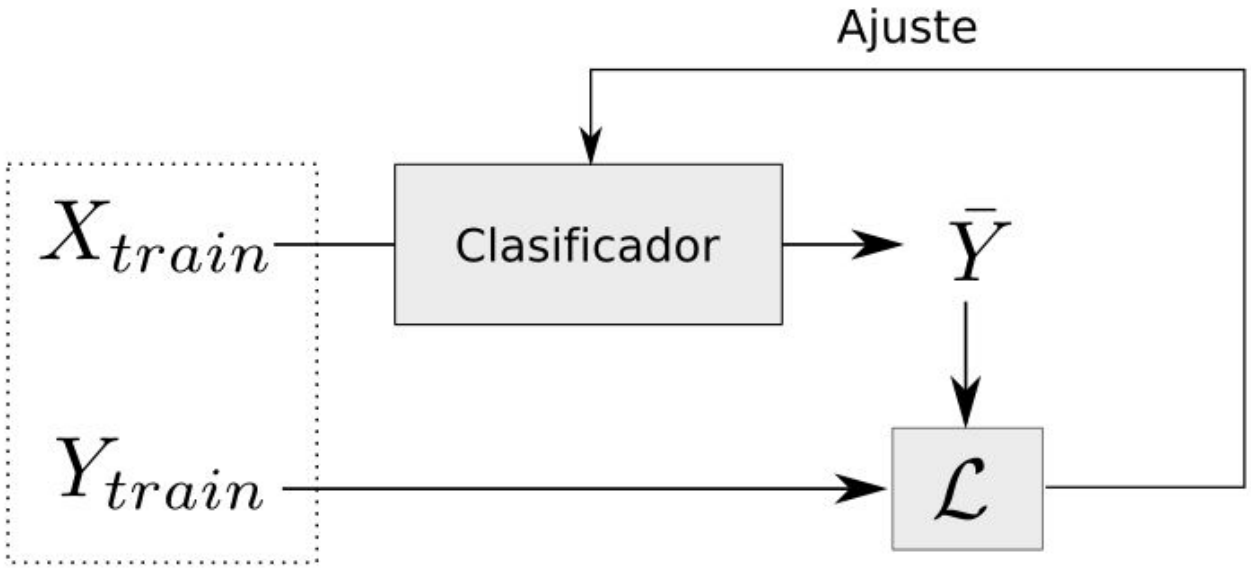


$$y = \{c_1, c_2, ...\}$$
$$y \in \mathbb{R}$$

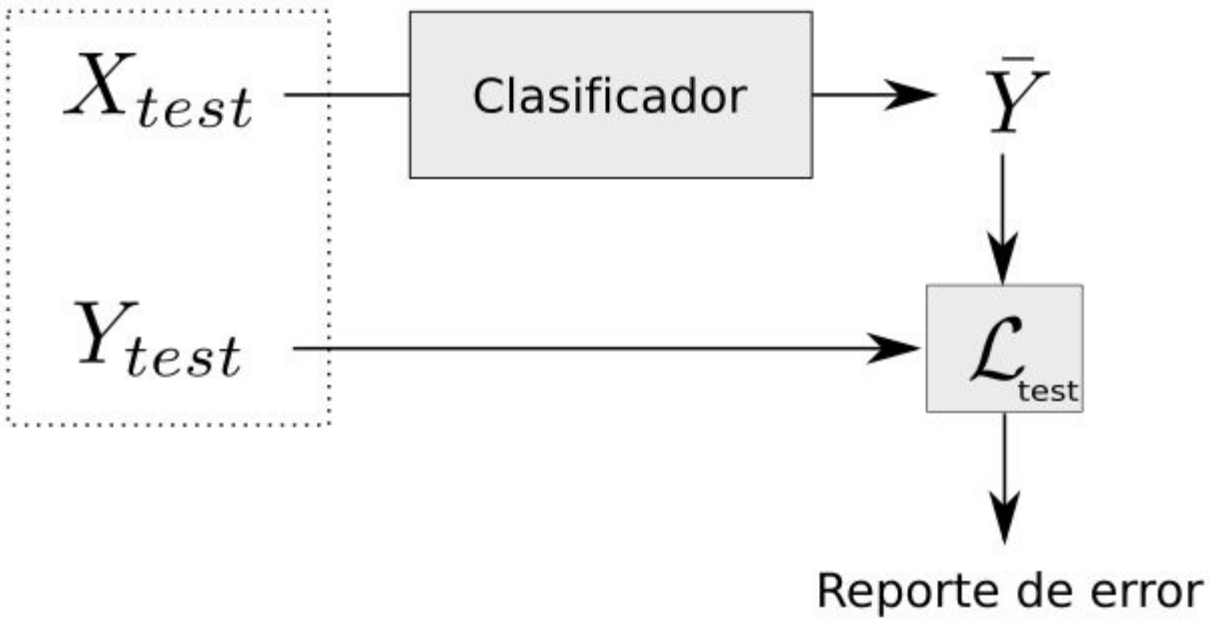
La tarea de **clasificación** implica poner etiquetas a cada punto del dataset.

La tarea de **regresión** implica que el modelo pueda estimar una variable (contínua) a partir de otras.

De programación clásica a “modelos fundacionales”

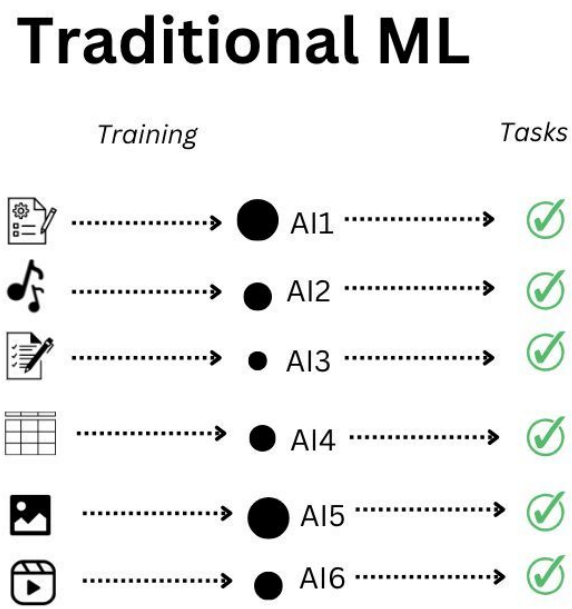


De programación clásica a “modelos fundacionales”

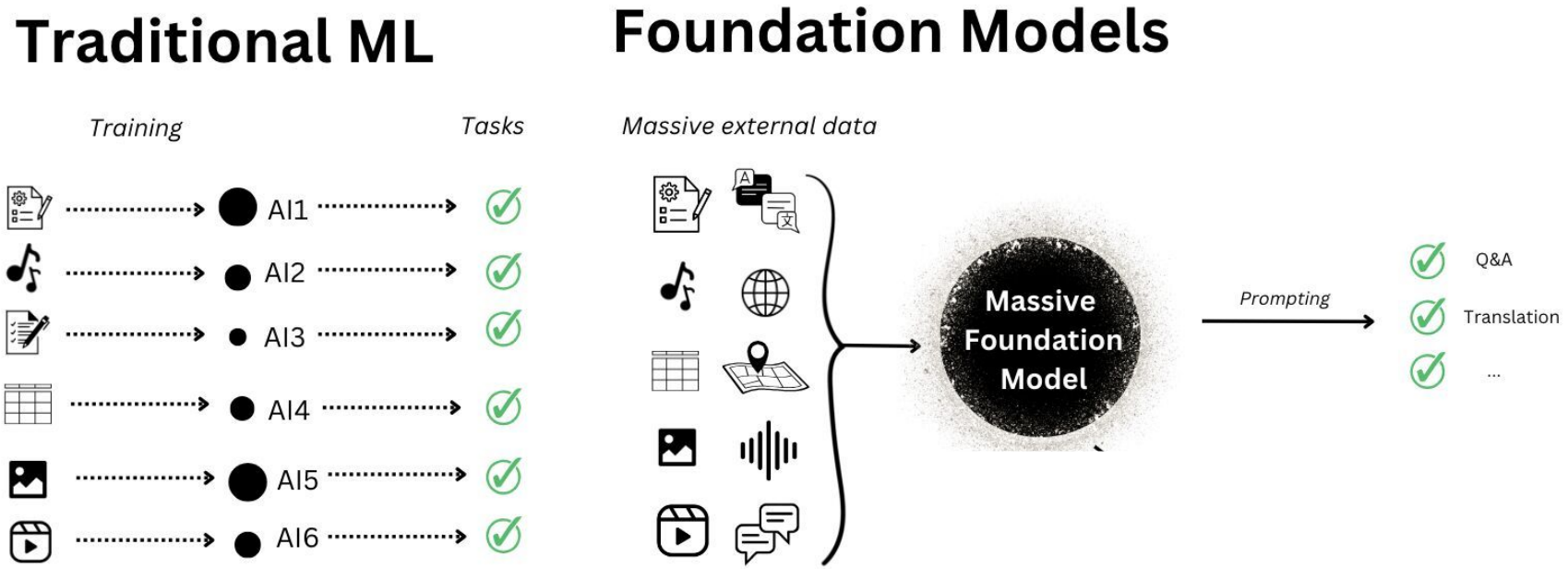


De programación clásica a “modelos fundacionales”

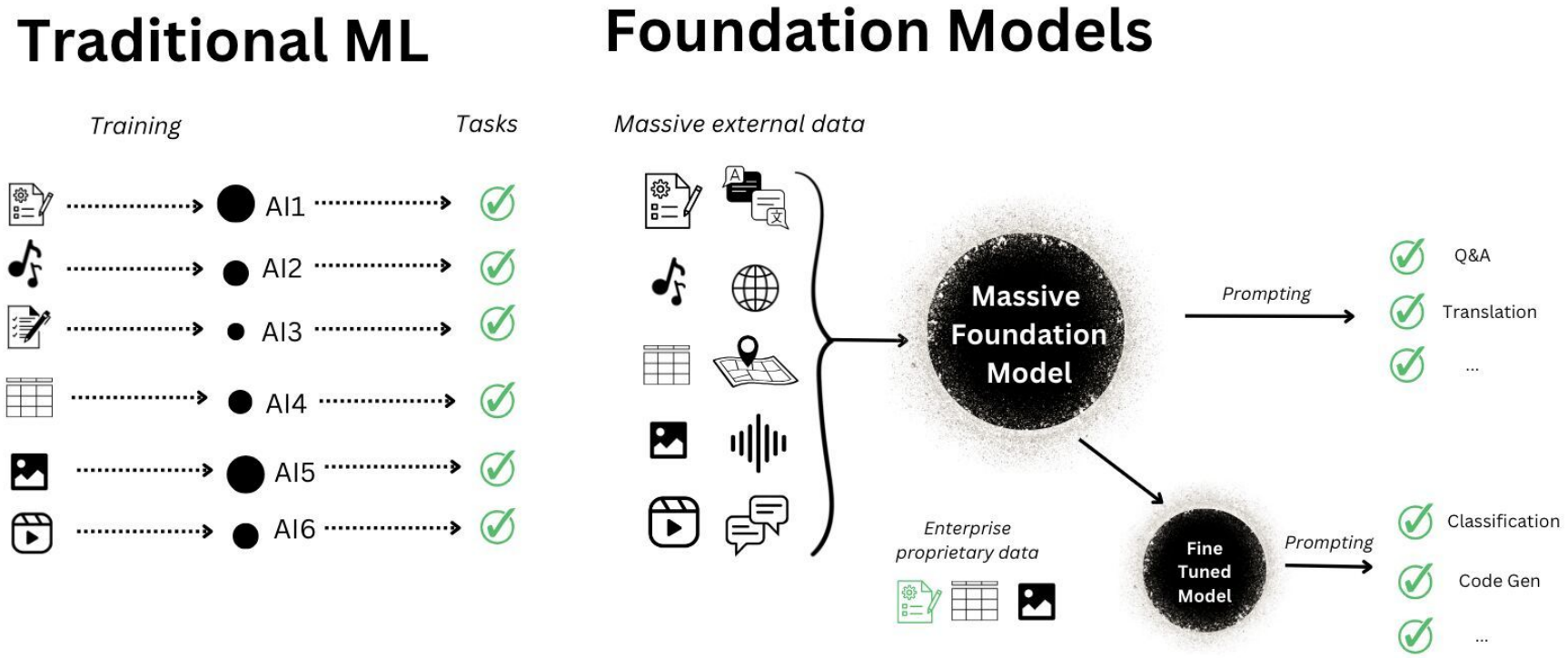
De programación clásica a “modelos fundacionales”



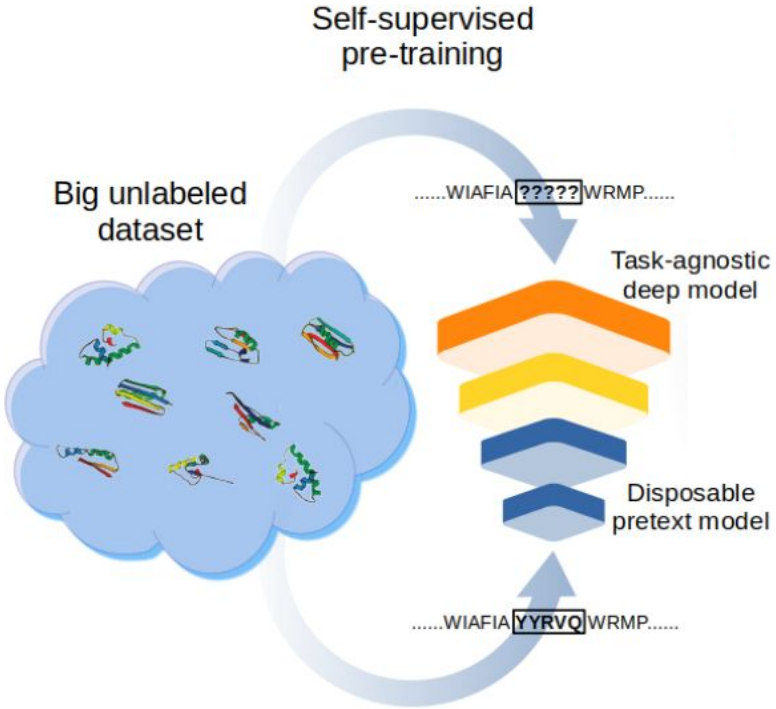
De programación clásica a “modelos fundacionales”



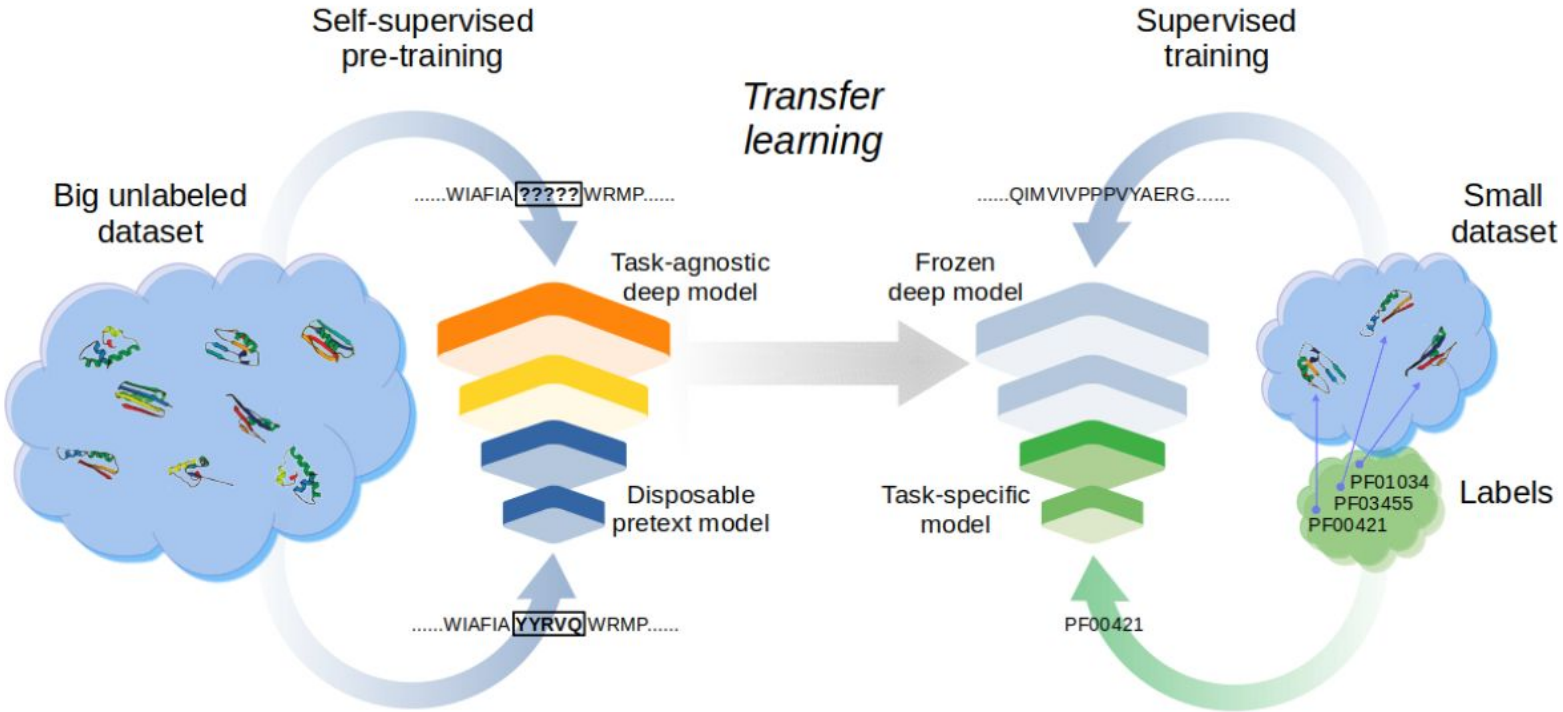
De programación clásica a “modelos fundacionales”



Transfiriendo el aprendizaje

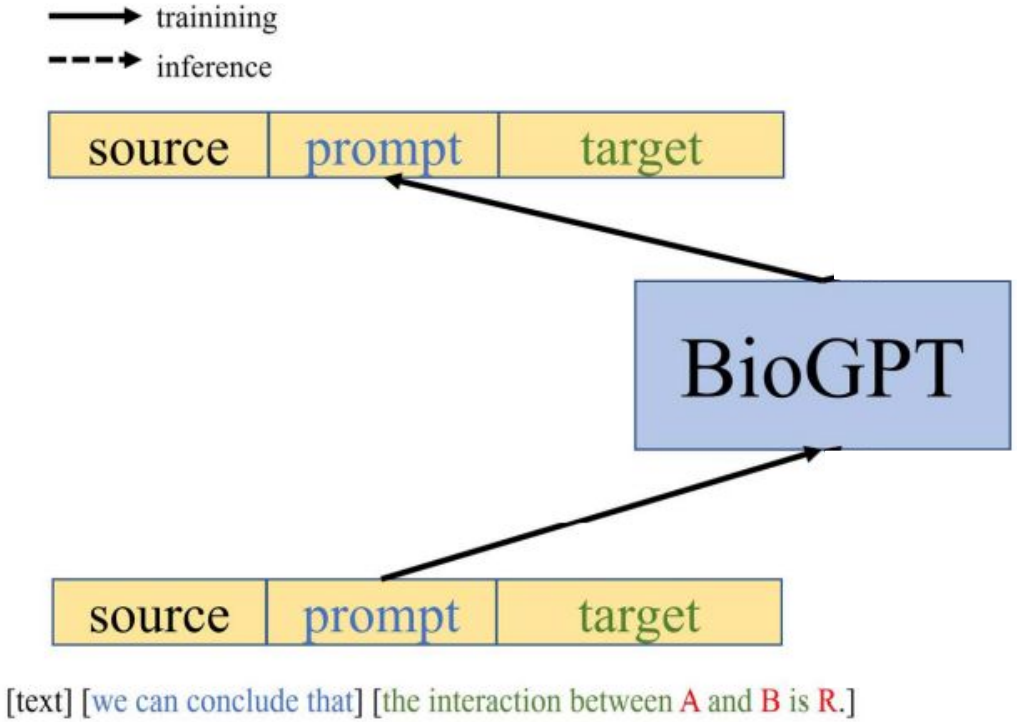


Transfiriendo el aprendizaje



Modelos fundacionales en biología computacional

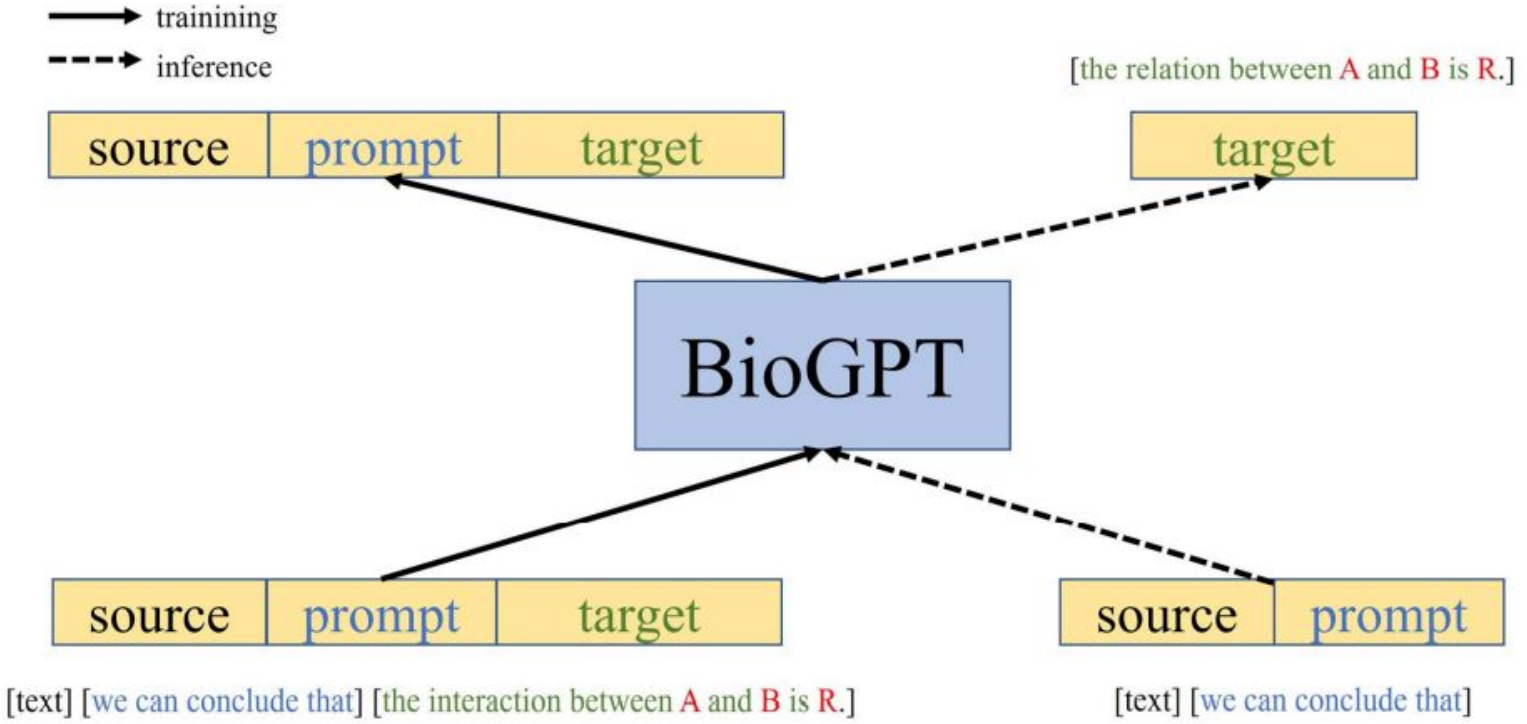
GPT para ciencias biológicas



Usa grandes datasets de tareas en documentos específicos de ciencias biológicas (**PubMedQA, BioASQ** y **SciDocs).**

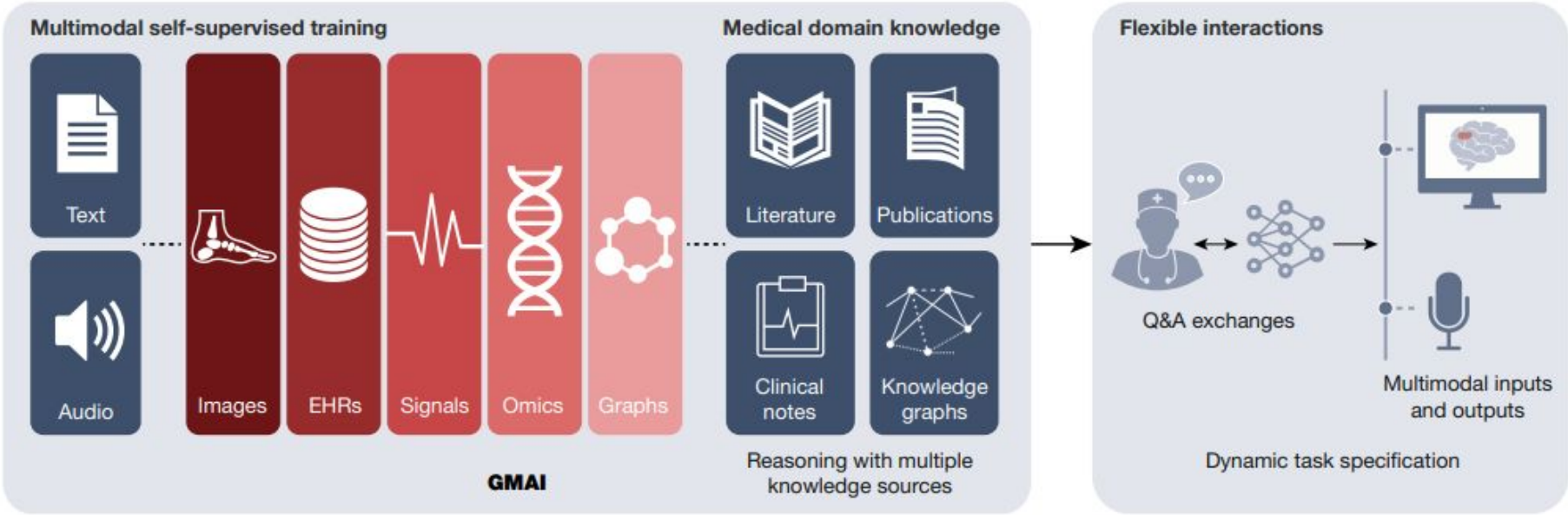
BioGPT: generative pre-trained transformer for biomedical text generation and mining, Renqian Luo et al., 2022, Briefings in Bioinformatics.

GPT para ciencias biológicas



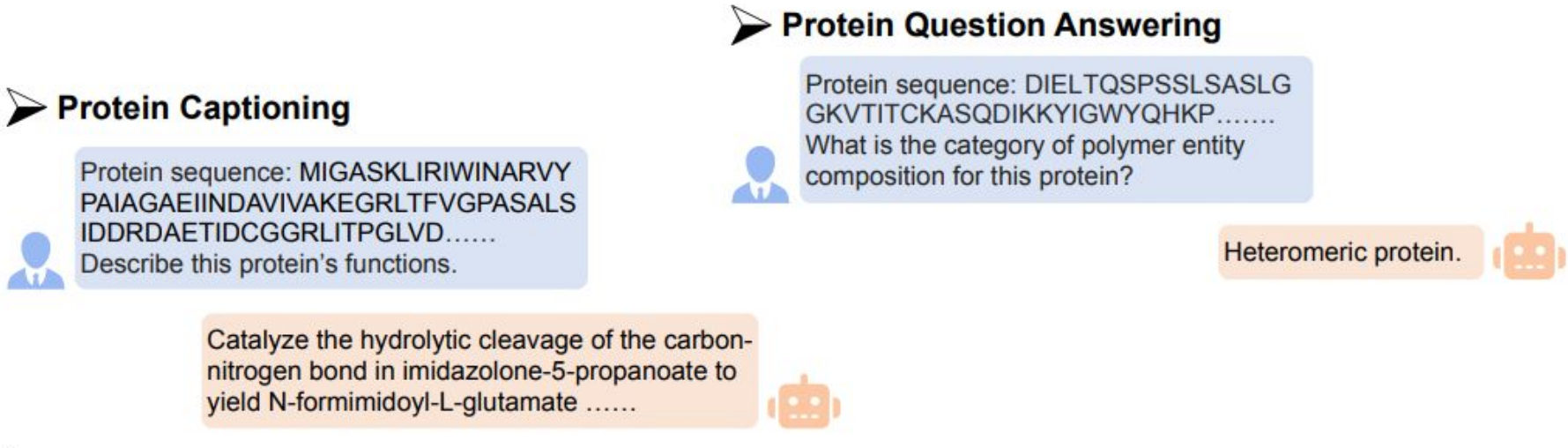
BioGPT: generative pre-trained transformer for biomedical text generation and mining, Renqian Luo et al., 2022, Briefings in Bioinformatics.

GPT para ciencias biológicas



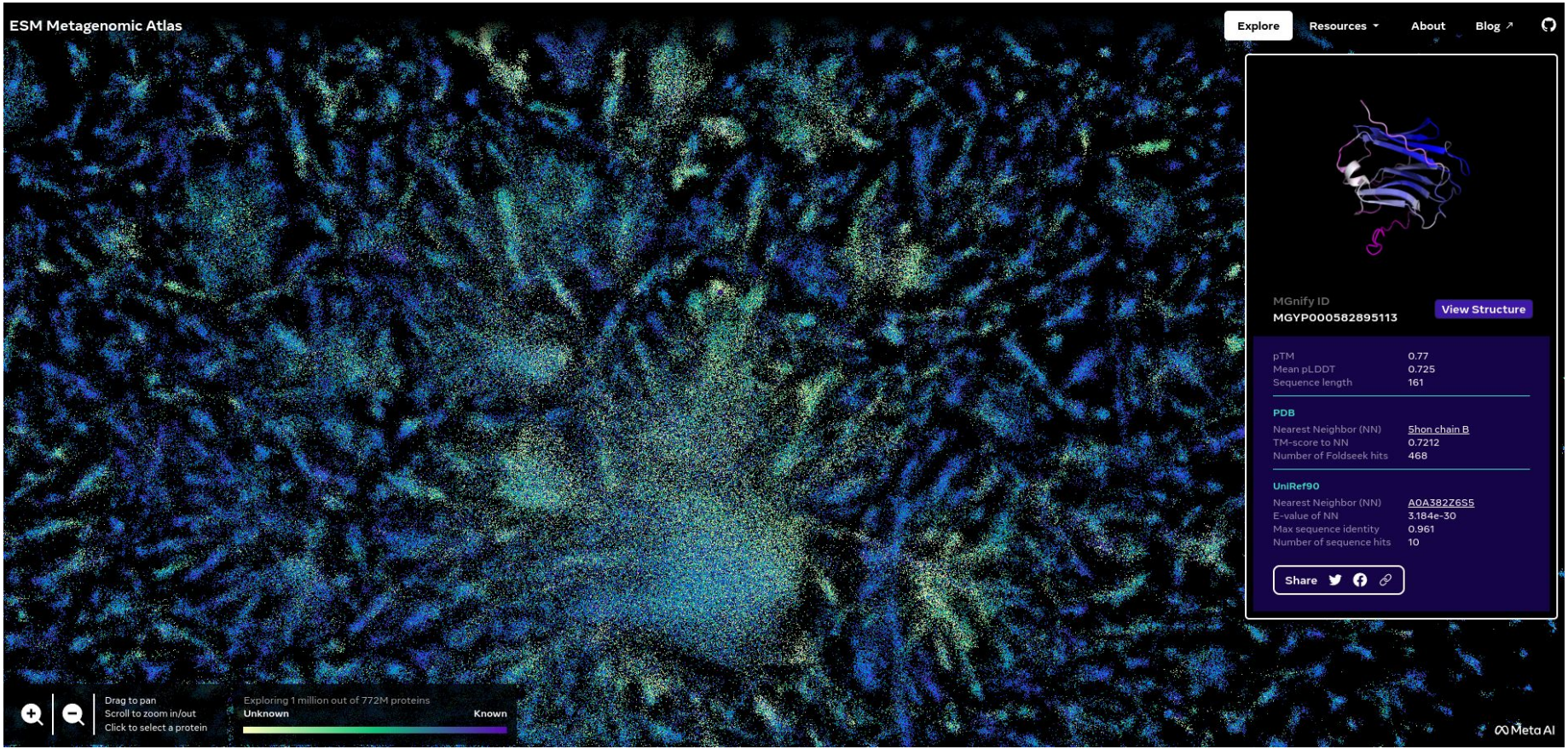
Foundation models for generalist medical artificial intelligence, Moor, M., et al. Nature 2023.

GPT para ciencias biológicas

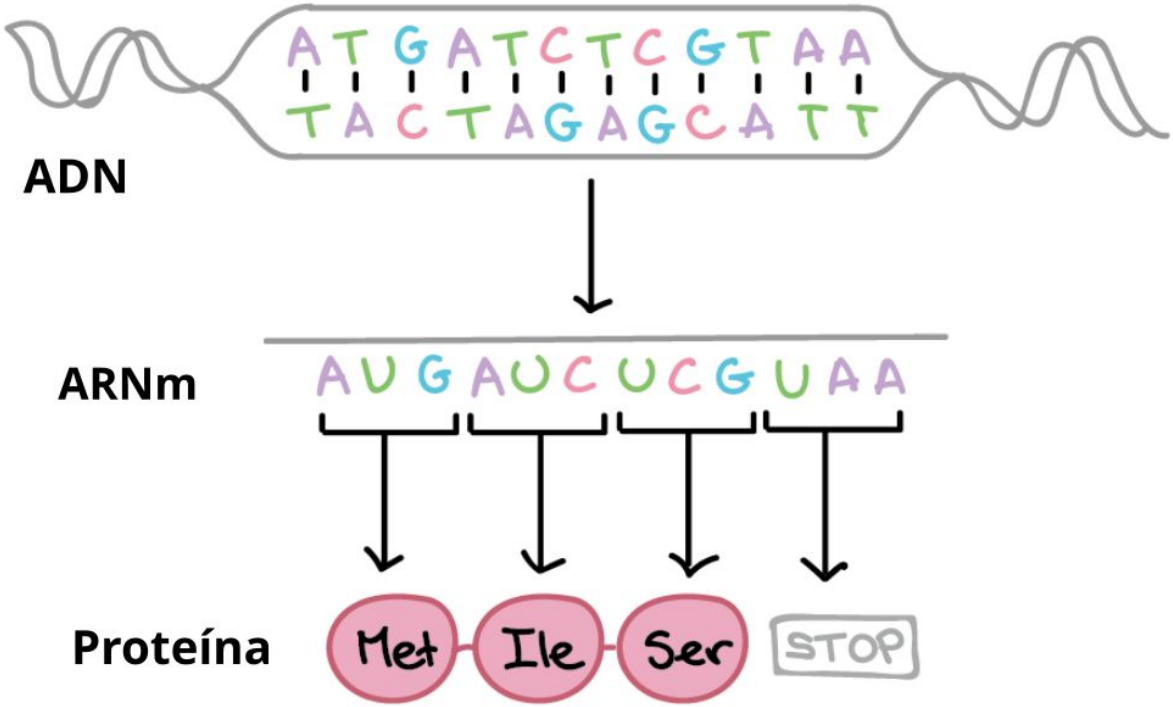


ProtT3: Protein-to-Text Generation for Text-based Protein Understanding, Z. Liu et al., arXiv 2024

Modelos fundacionales para secuencias biológicas



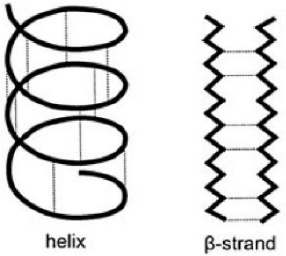
Aprendiendo el “lenguaje de la vida”

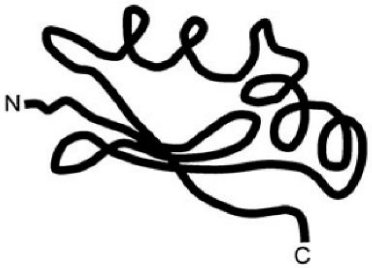


Aprendiendo el “lenguaje de la vida”

Proteina

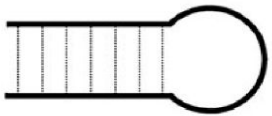
–M-D-Y-E-K-T-L-L-M–

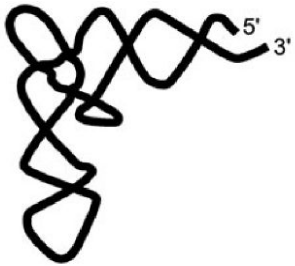


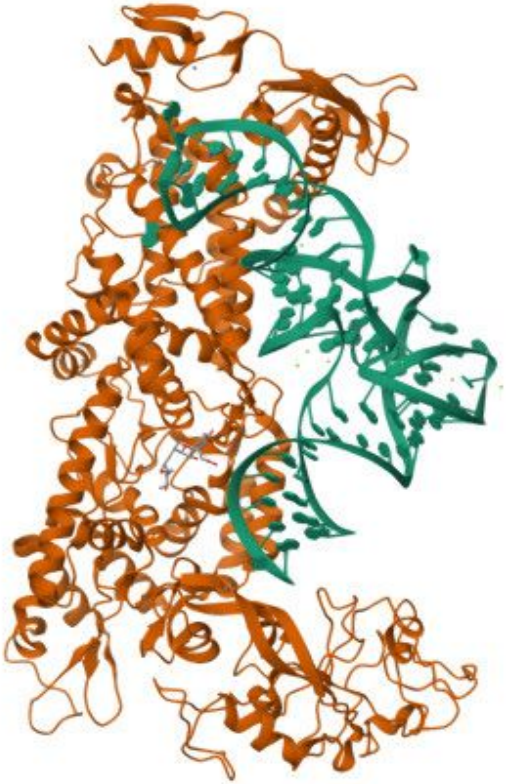


ARN

–G-C-G-G-A-U-U-U-A–







Aprendiendo la semántica

Semántica distribucional

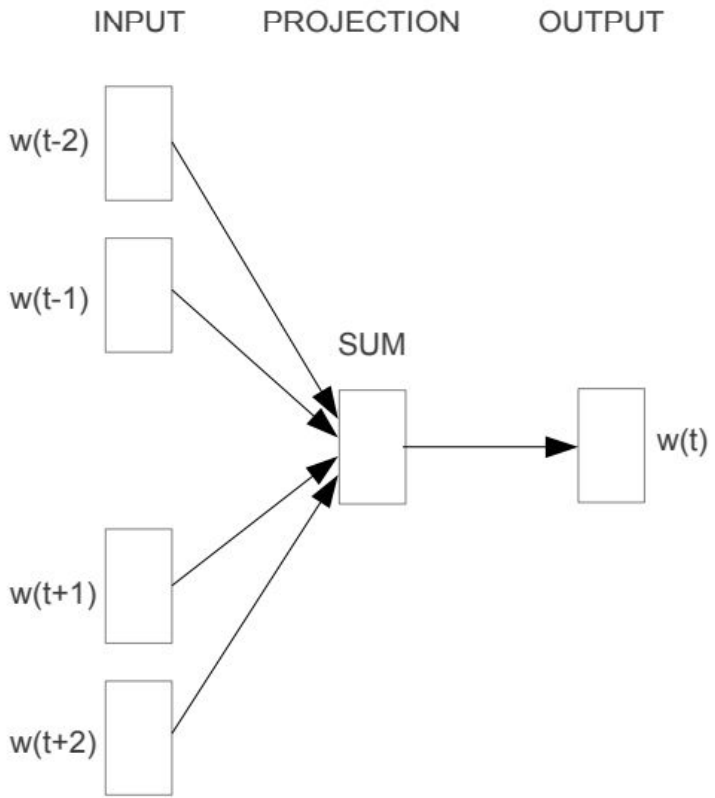
"You shall know a word by the company it keeps" (J. Firth, 1957)

	the	red	dog	cat	eats	food
1. the red dog →	1	1	1	0	0	0
2. cat eats dog →	0	0	1	1	1	0
3. dog eats food →	0	0	1	0	1	1
4. red cat eats →	0	1	0	1	1	0

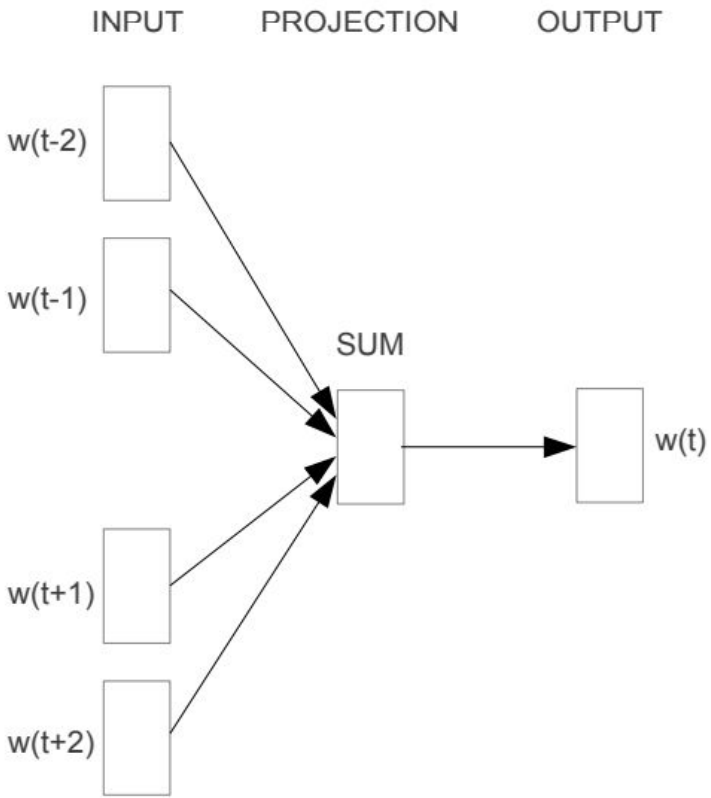
Aprendiendo la semántica

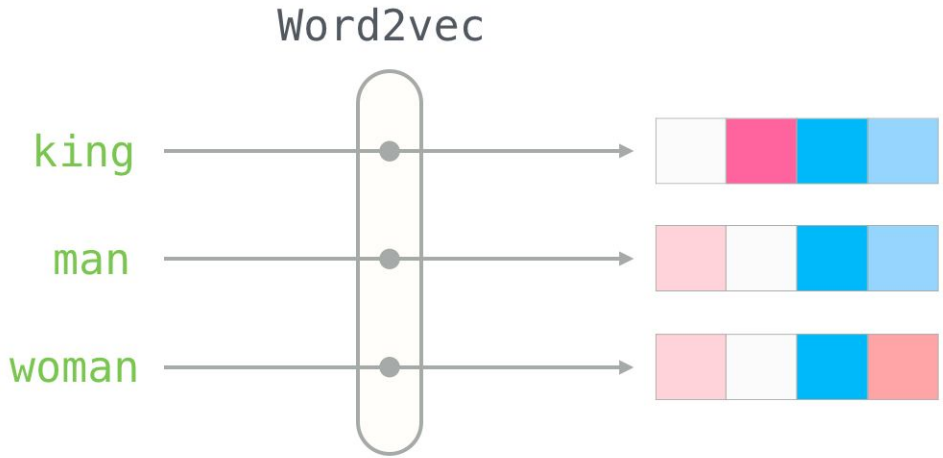
Semántica distribucional

"You shall know a word by the company it keeps" (J. Firth, 1957)

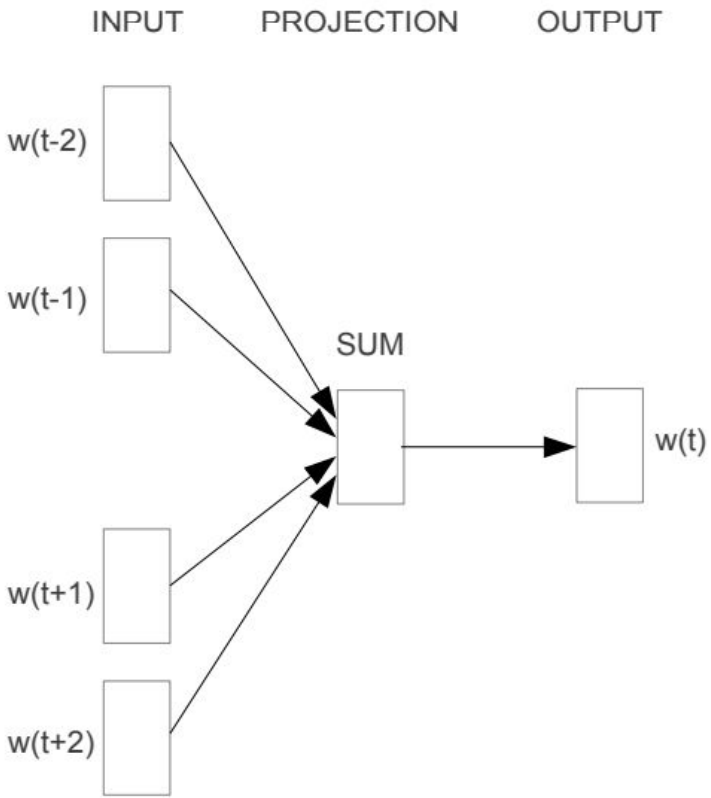


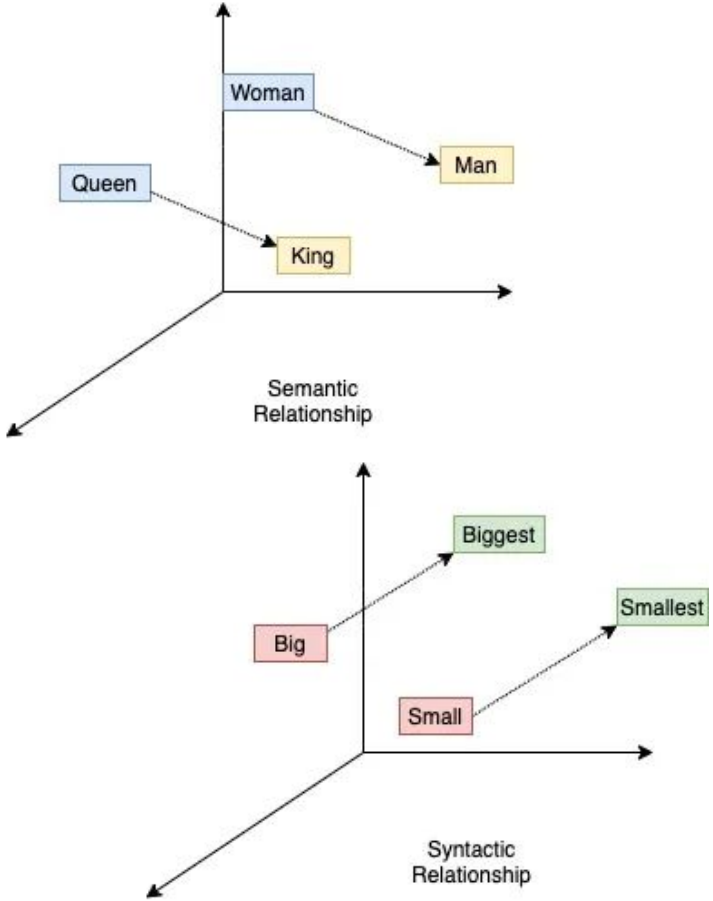
Aprendiendo la semántica





Aprendiendo la semántica





Aprendiendo la semántica

Original Sequence

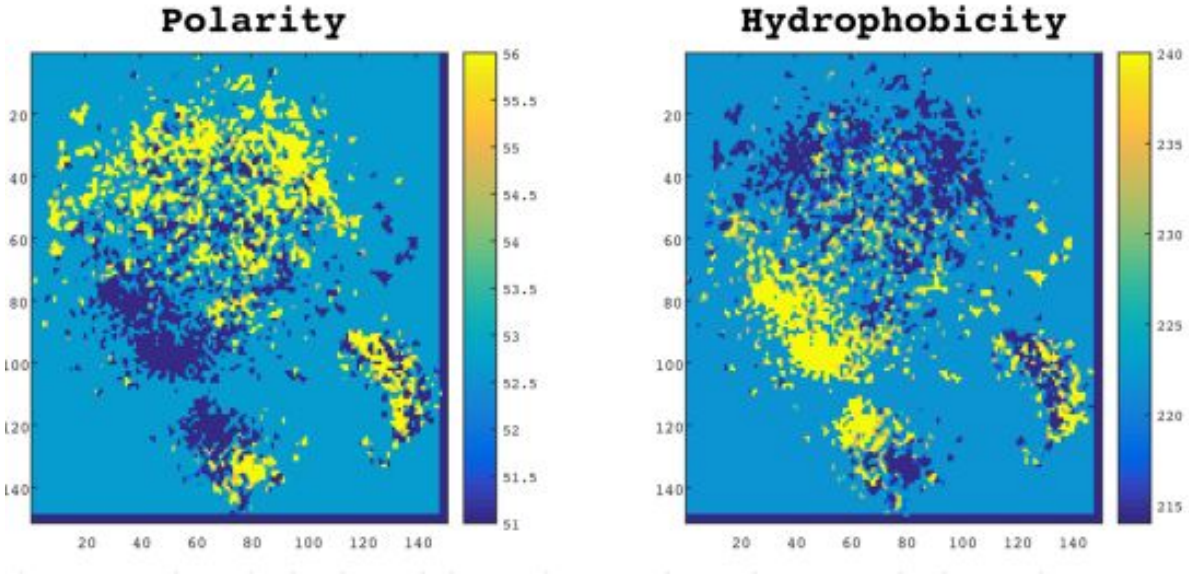
$$^{(1)}\vec{M}\,^{(2)}\vec{A}\,^{(3)}\vec{F}\,SAEDVLKEYDRRRRMEAL..$$

Splittings

$$\begin{cases} 1) & \text{MAF, SAE, DVL, KEY, DRR, RRM, ..} \\ 2) & \text{AFS, AED, VLK, EYD, RRR, RME, ..} \\ 3) & \text{FSA ,EDV, LKE, YDR, RRR, MEA, ..} \end{cases}$$

Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics
E. Asgari et al., Plos One 2015

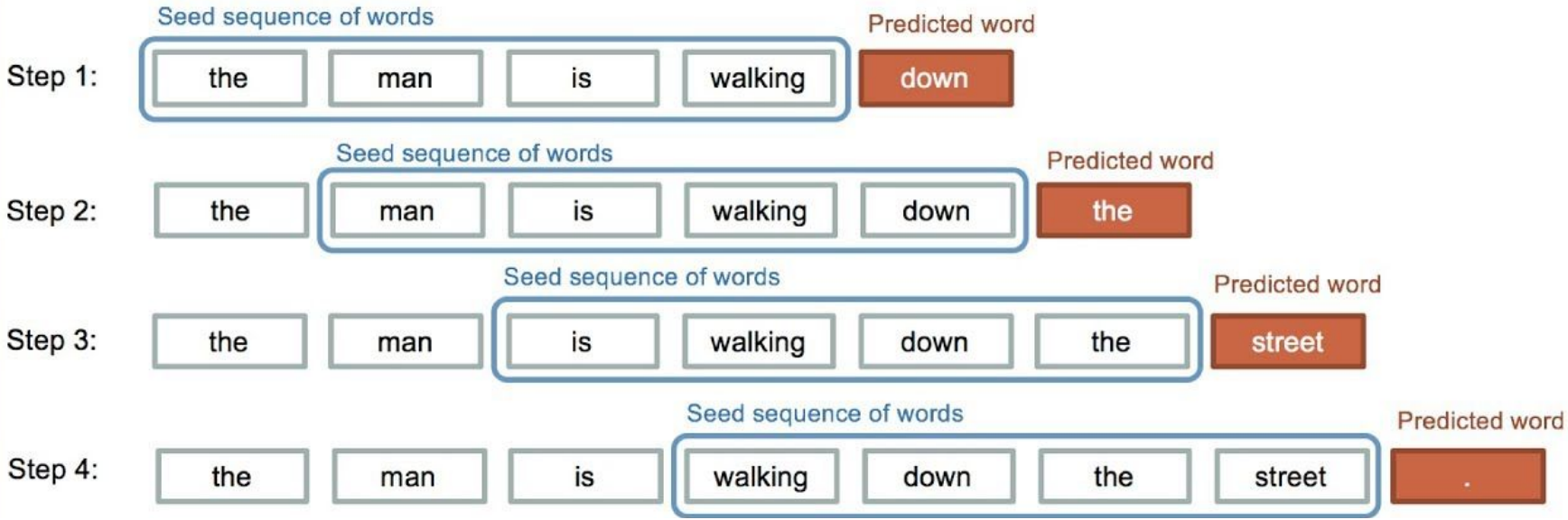
Aprendiendo la semántica



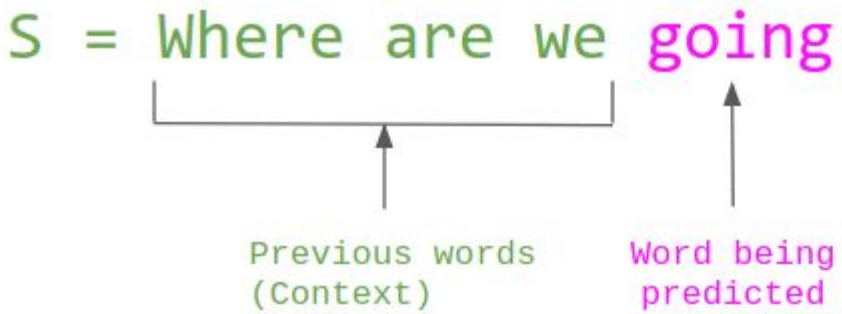
Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics
E. Asgari et al., Plos One 2015

Modelos de lenguaje

Modelos de lenguaje

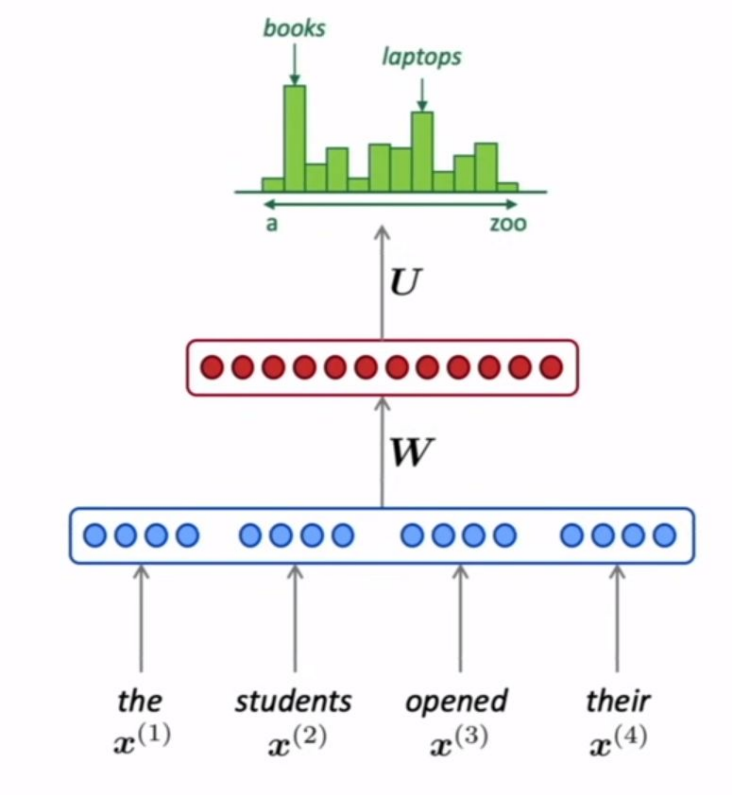


Modelos de lenguaje

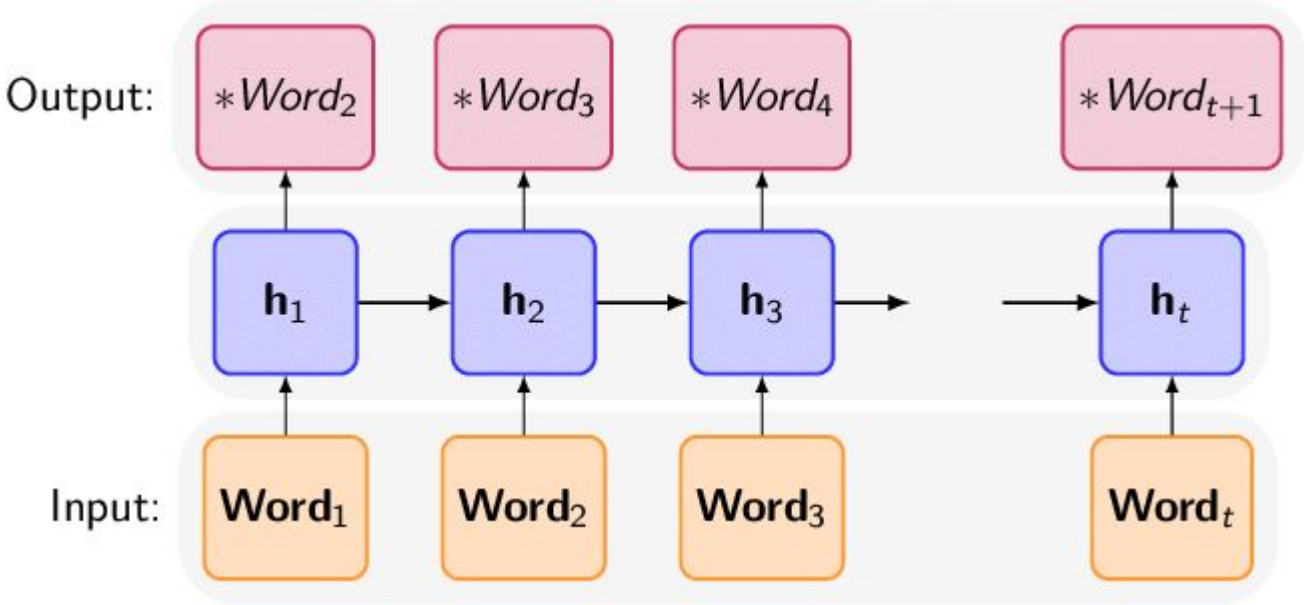


$$P(S) = P(\text{Where}) \times P(\text{are} \mid \text{Where}) \times P(\text{we} \mid \text{Where are}) \times P(\text{going} \mid \text{Where are we})$$

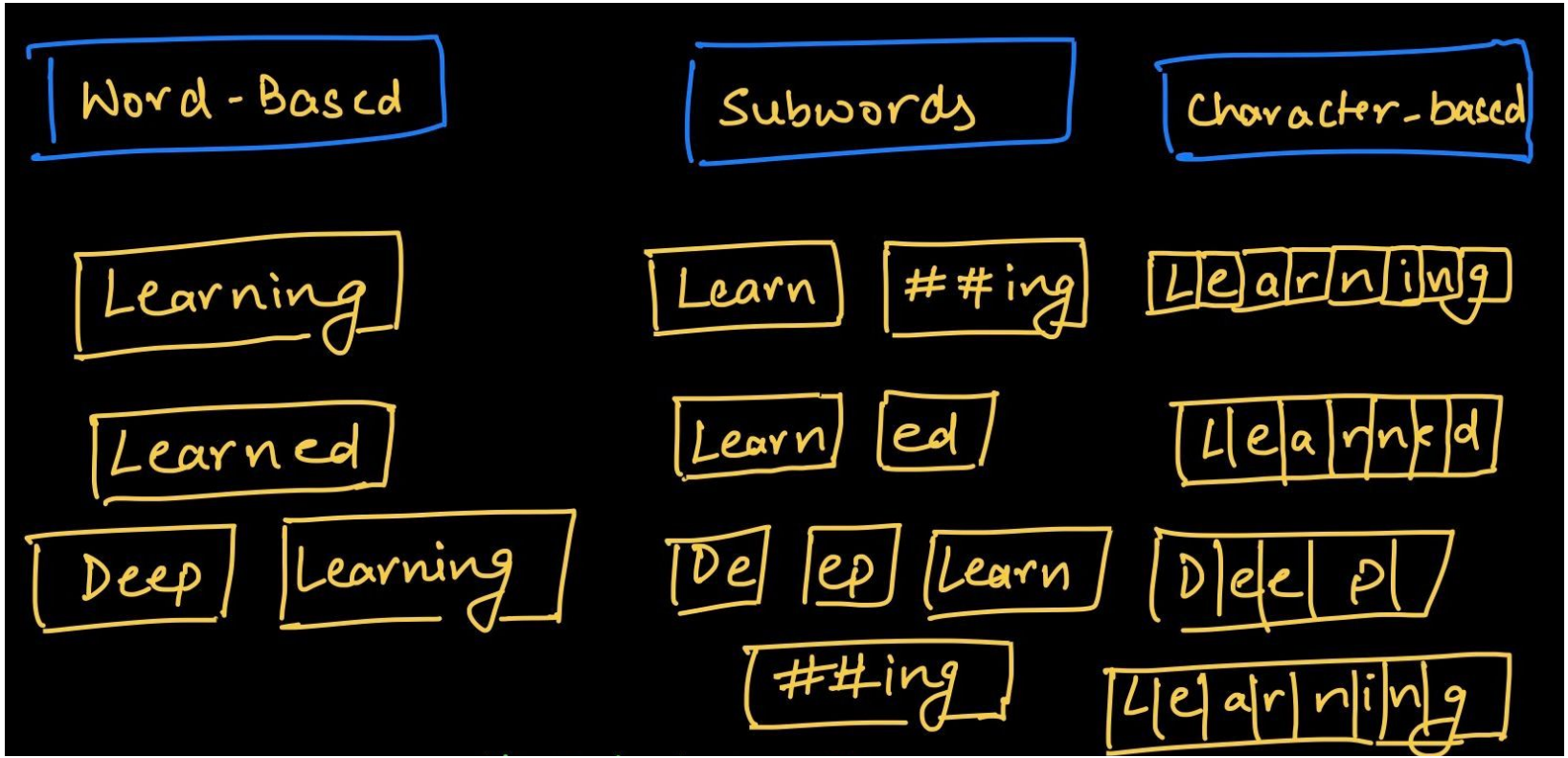
Modelos de lenguaje con redes neuronales



Modelos de lenguaje con redes recurrentes



¿Cómo representamos el texto?

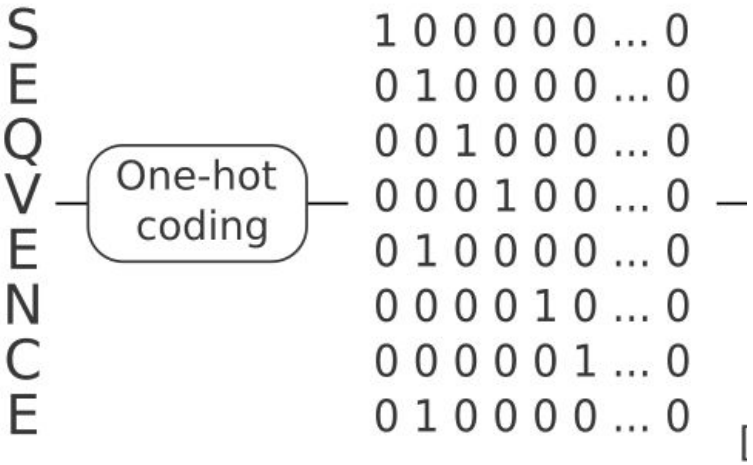


¿Cómo representamos los elementos de una secuencia?

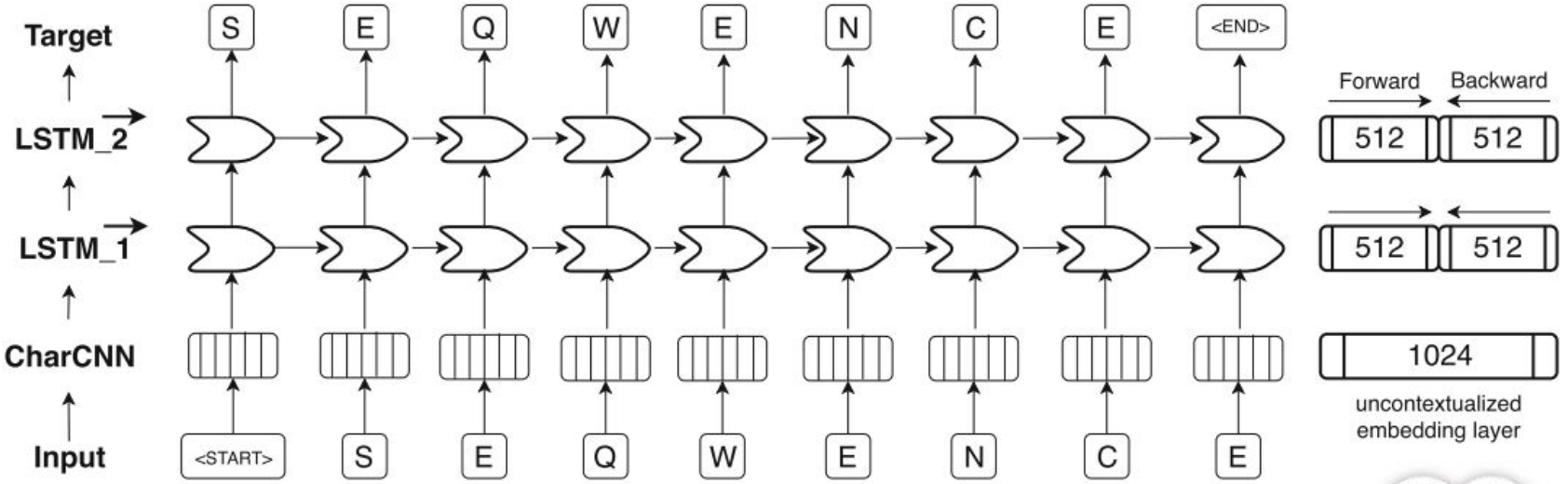
Podríamos armar un diccionario, en el cual cada residuo (AA) tiene asignado un ID: 0, 1, … 20.

Cada ID puede apuntar a un vector one-hot

Pero: Limita lo que podemos representar

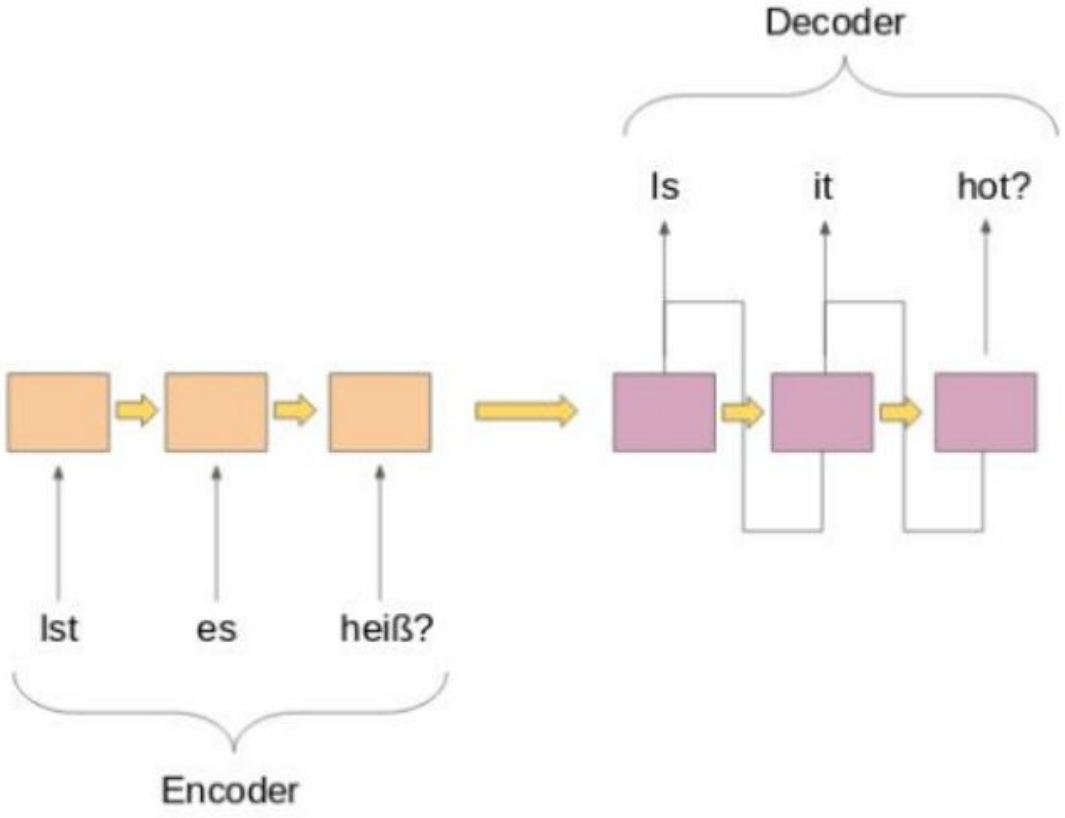


Embeddings from Language Models (ELMo)

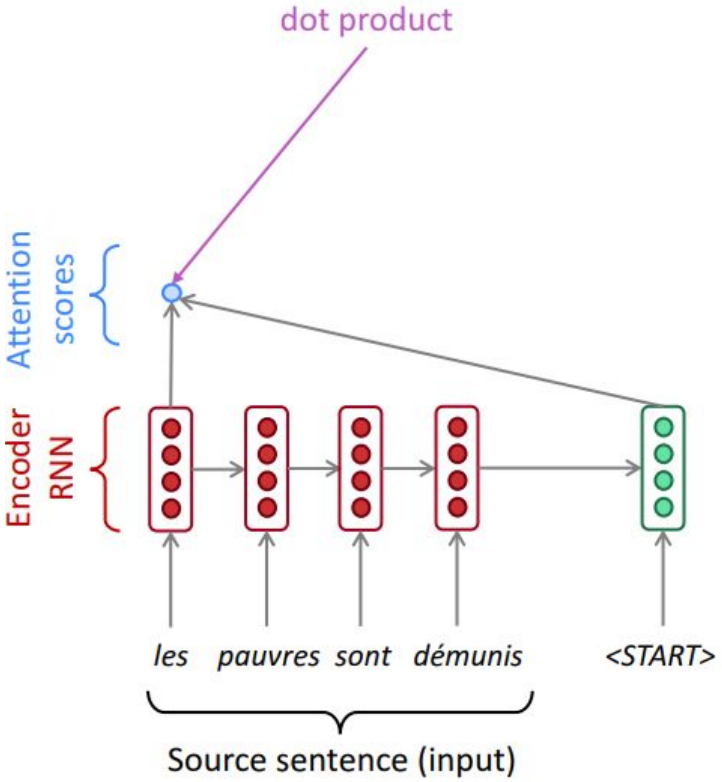


Modeling aspects of the language of life through transfer-learning protein sequences, M. Heinzinger et al., BMC Bioinformatics 2019

Limitaciones de las redes recurrentes

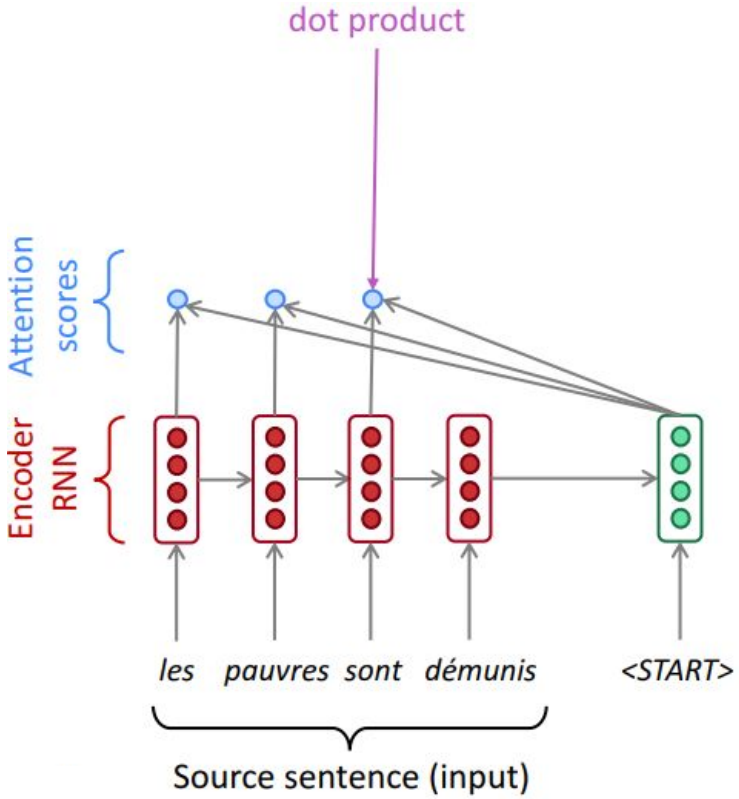


Mecanismo de atención



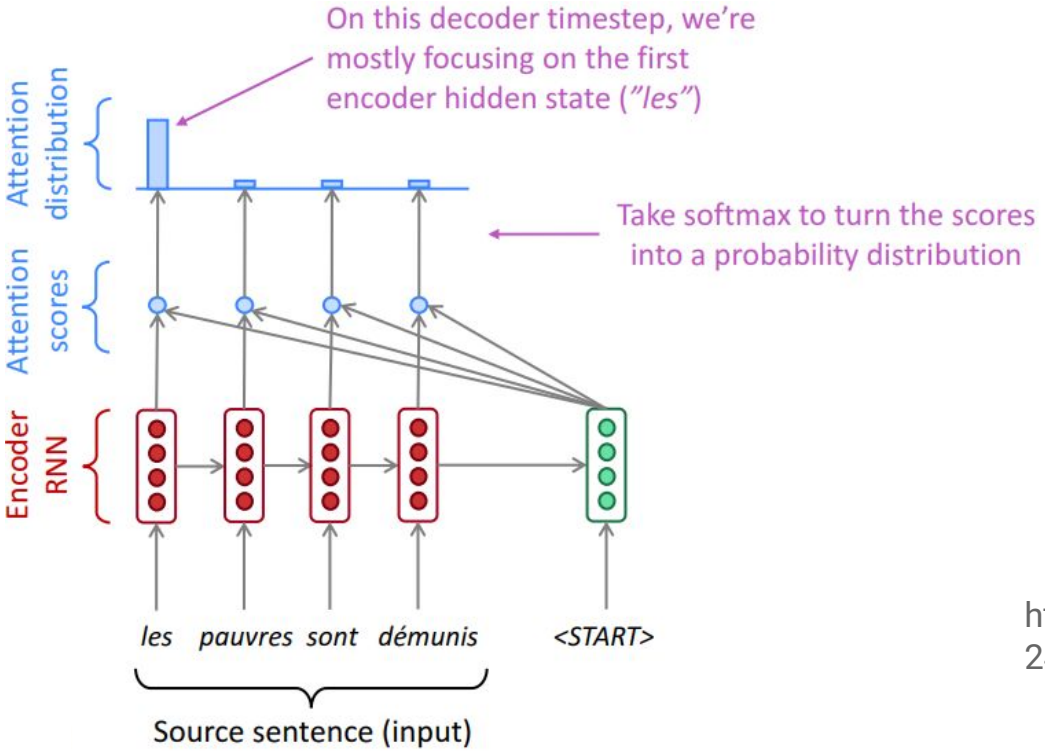
https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1184/lectures/lecture11.pdf

Mecanismo de atención



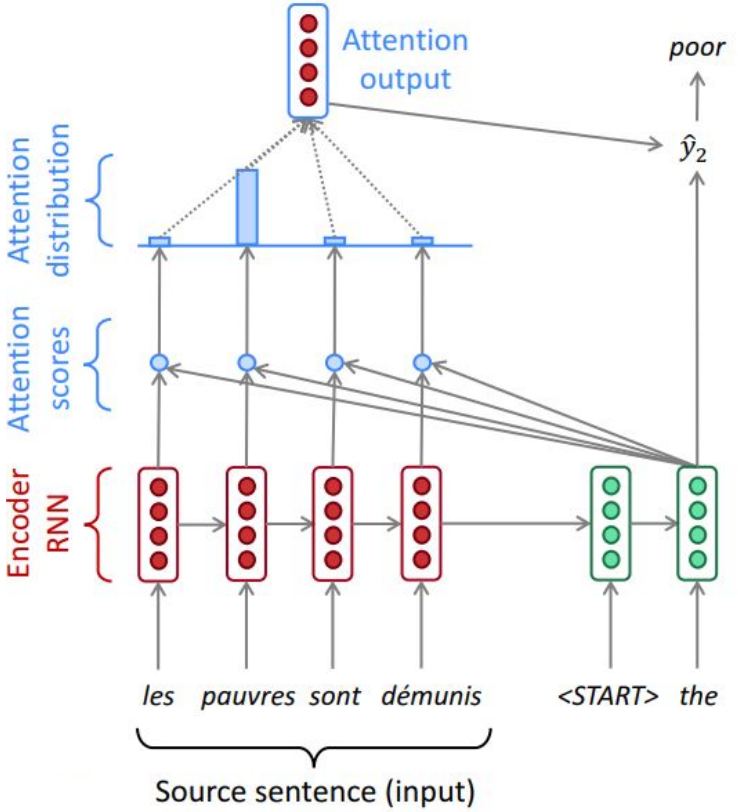
https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1184/lectures/lecture11.pdf

Mecanismo de atención



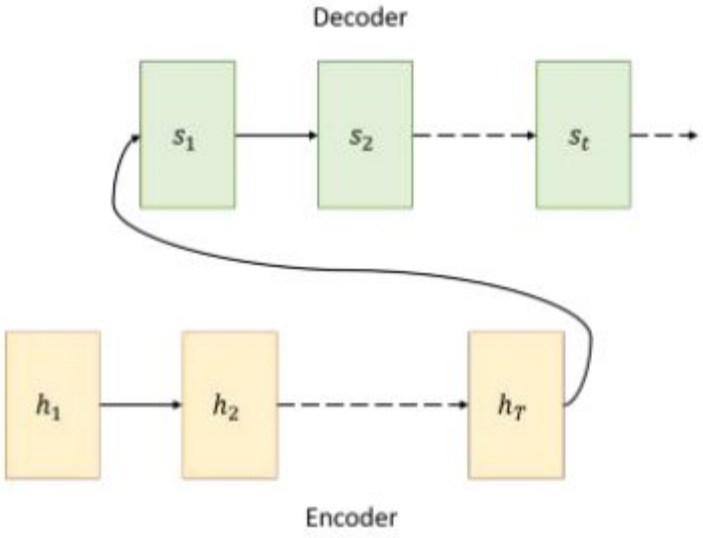
https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1184/lectures/lecture11.pdf

Mecanismo de atención

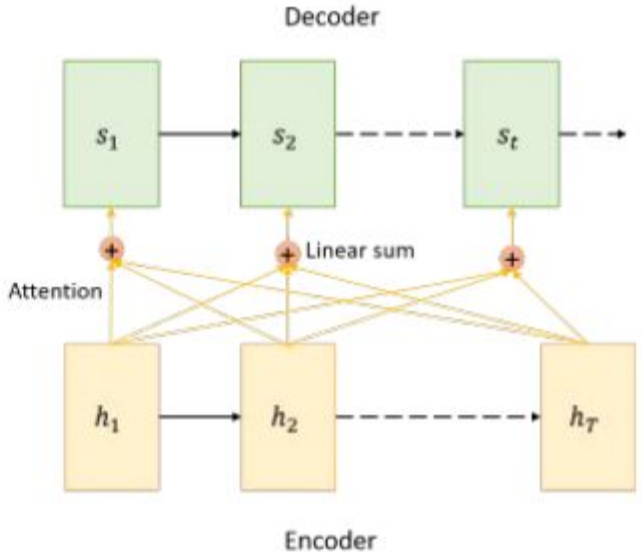


https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1184/lectures/lecture11.pdf

Mecanismo de atención

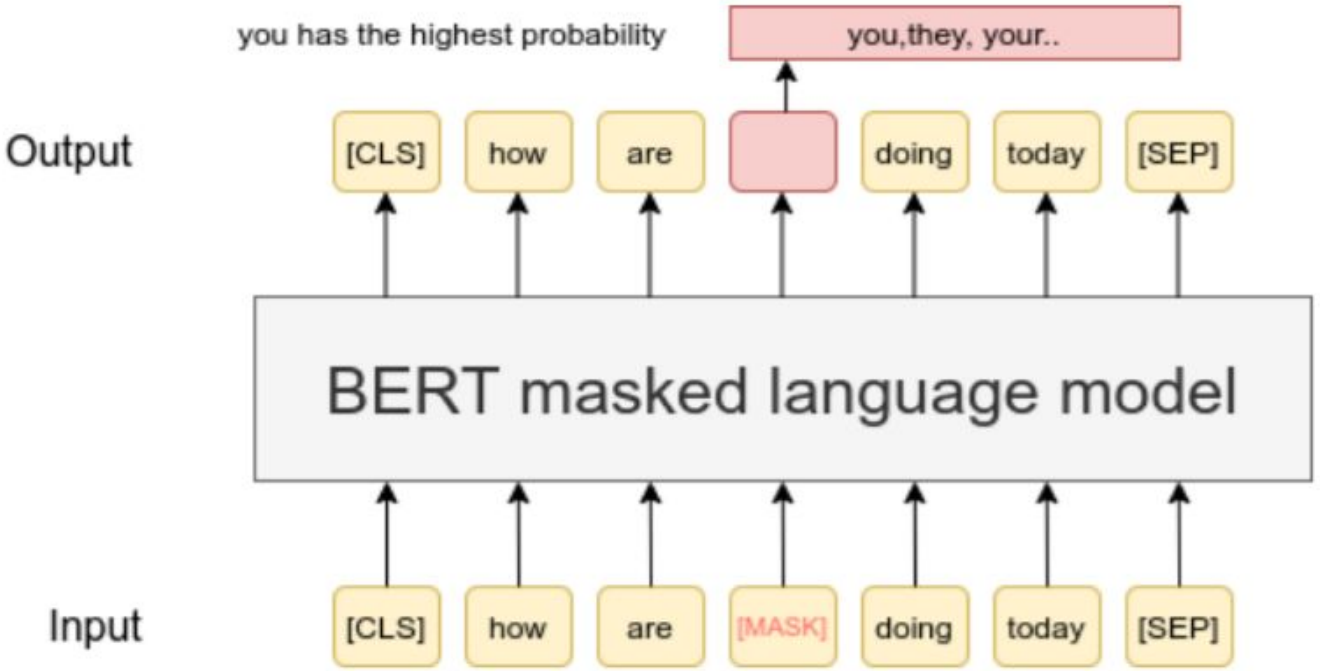


(a) Vanilla Encoder Decoder Architecture



(b) Attention Mechanism

Bidirectional Encoder Representations from Transformers (BERT)

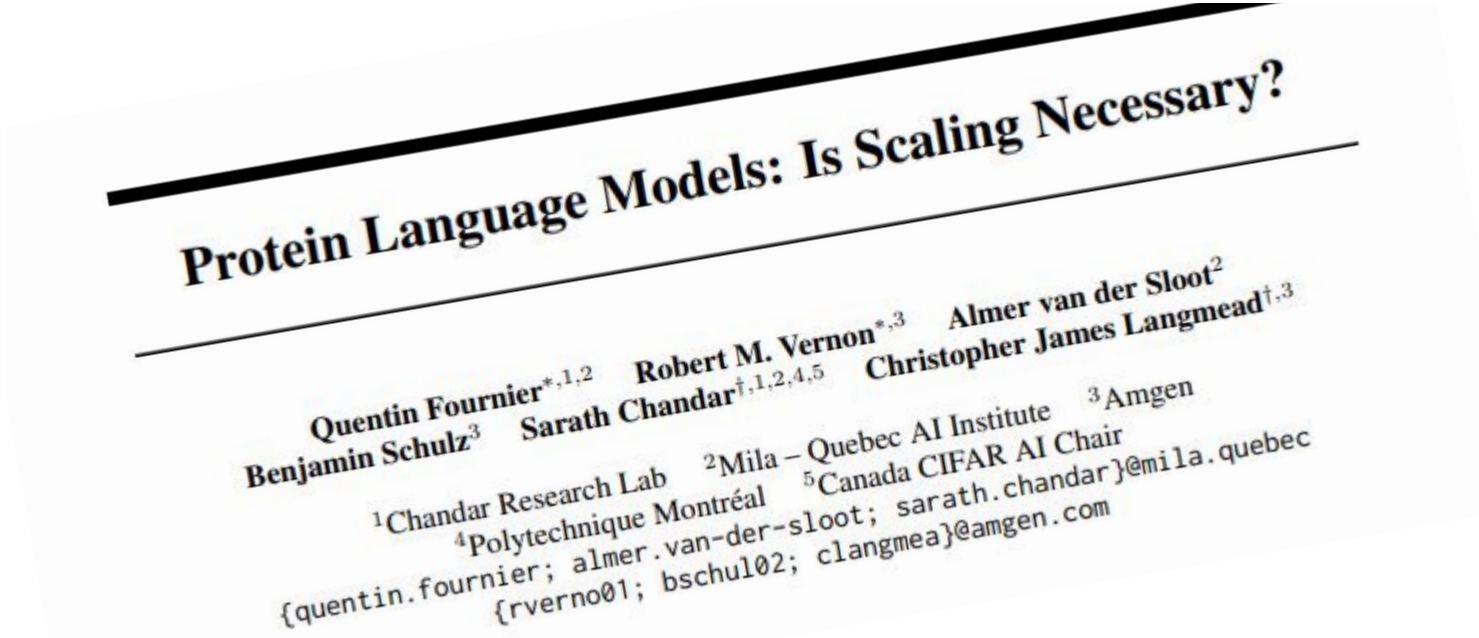


Evolutionary Scale Modeling (ESM)

	Model		Params	Training	ECE
(a)	Oracle				1
	Uniform Random				25
(b)	n-gram	4-gram		UR50/S	17.18
(c)	LSTM	Small	28.4M	UR50/S	14.42
	LSTM	Large	113.4M	UR50/S	13.54
(d)	Transformer	6-layer	42.6M	UR50/S	11.79
	Transformer	12-layer	85.1M	UR50/S	10.45
(e)	Transformer	34-layer	669.2M	UR100	10.32
	Transformer	34-layer	669.2M	UR50/S	8.54
	Transformer	34-layer	669.2M	UR50/D	**8.46**
(f)	Transformer	10% data	669.2M	UR50/S	10.99
	Transformer	1% data	669.2M	UR50/S	15.01
	Transformer	0.1% data	669.2M	UR50/S	17.50

Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, A. Rives, PNAS 2021,

¿Scaling vs diseño experimental?

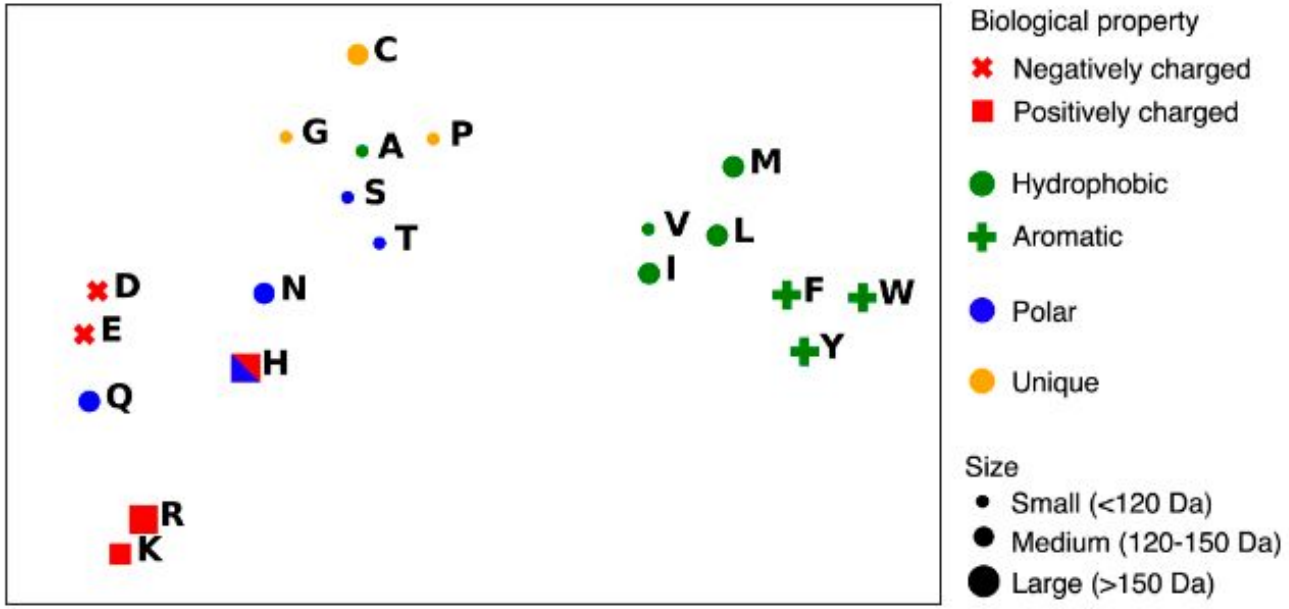


Protein Language Models: Is Scaling Necessary?

Quentin Fournier[*,1,2] **Robert M. Vernon**[*,3] **Almer van der Sloot**[2]
Benjamin Schulz[3] **Sarath Chandar**[†,1,2,4,5] **Christopher James Langmead**[†,3]

[1]Chandar Research Lab [2]Mila – Quebec AI Institute [3]Amgen
[4]Polytechnique Montréal [5]Canada CIFAR AI Chair
{quentin.fournier; almer.van-der-sloot; sarath.chandar}@mila.quebec
{rverno01; bschul02; clangmea}@amgen.com

Representaciones de aminoácidos



Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, A. Rives, PNAS 2021,

Herramientas: HuggingFace

https://huggingface.co/

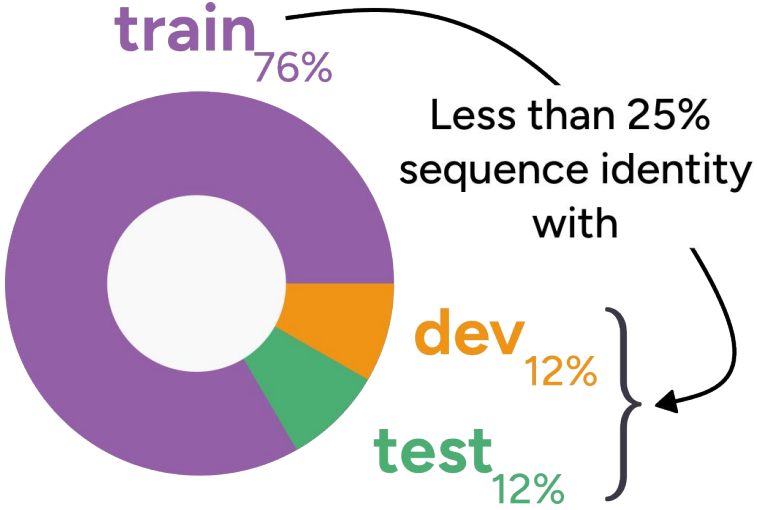
Herramientas: PyTorch & Google Colab

Tutorial 1

Aplicación: predecir dominio de proteínas

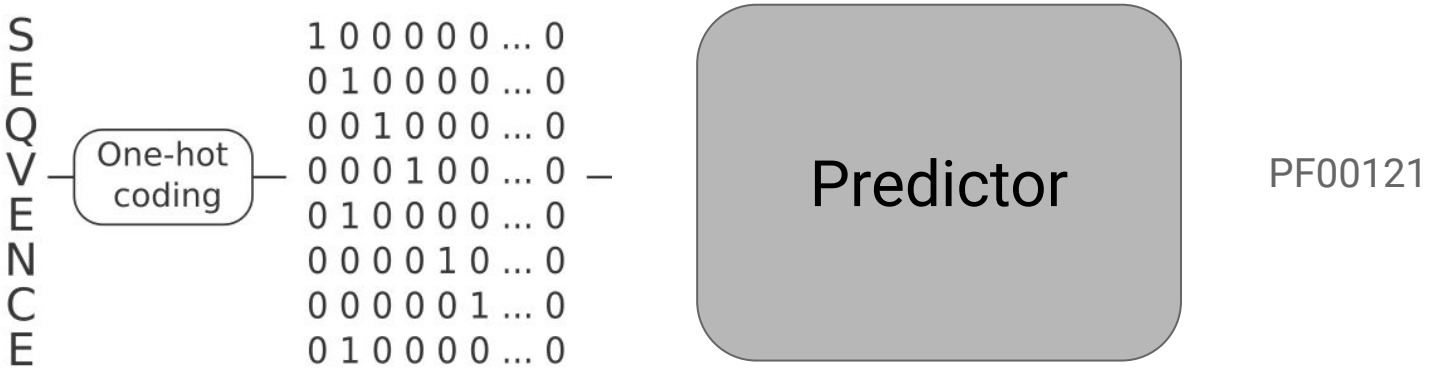
Pfam
v. 32.0

subset
- 395 families
- 131 clans
- 74 719 proteins



Aplicación: predecir dominio de proteínas

(clipped domains)

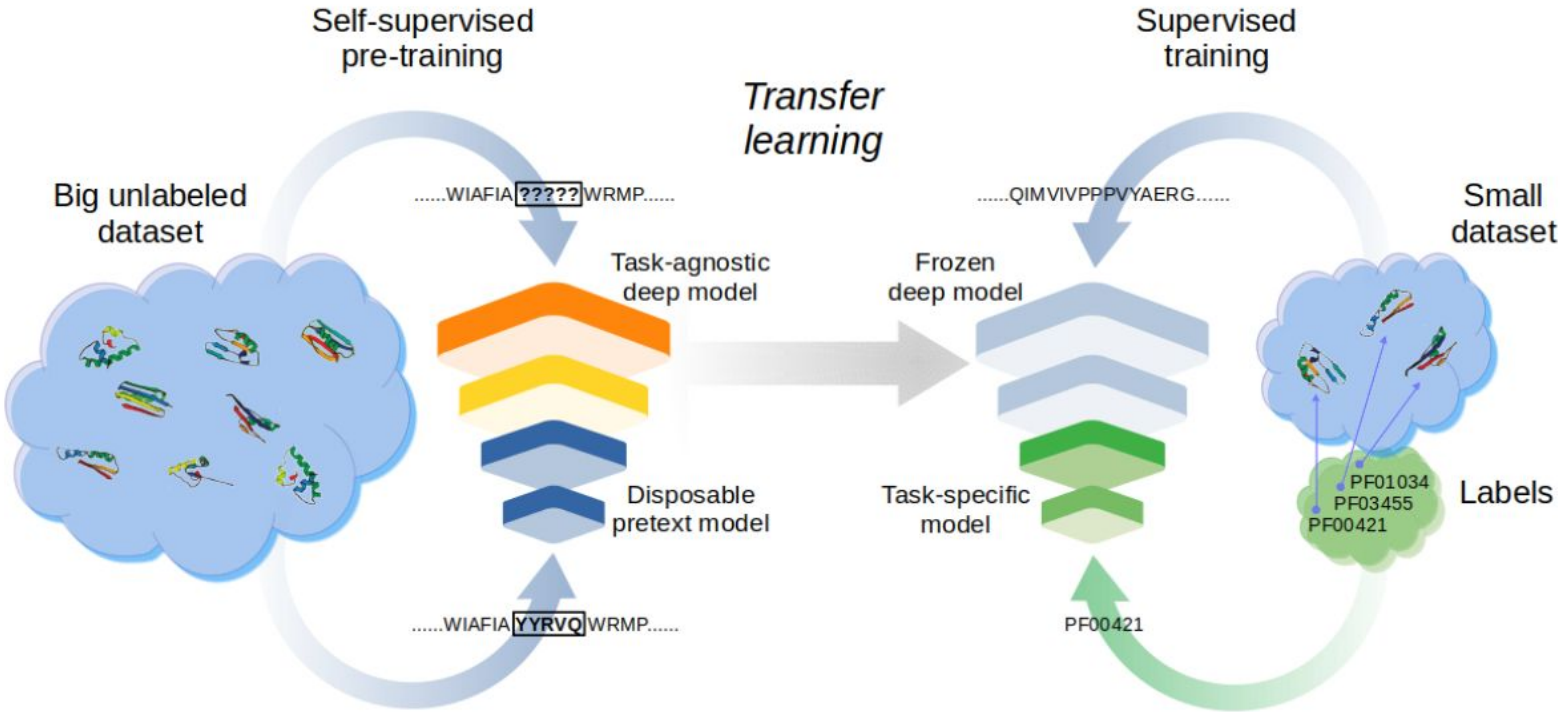


Prediciendo dominios con LLMs

Table 1: Pfam families classification without TL

No-TL	Error rate
HMM	18.10%
BLASTp	35.90%
ProtCNN	27.60%
ProtENN	**12.20%**

Prediciendo dominios con LLMs



Prediciendo dominios con LLMs

Table 1: Pfam families classification without TL

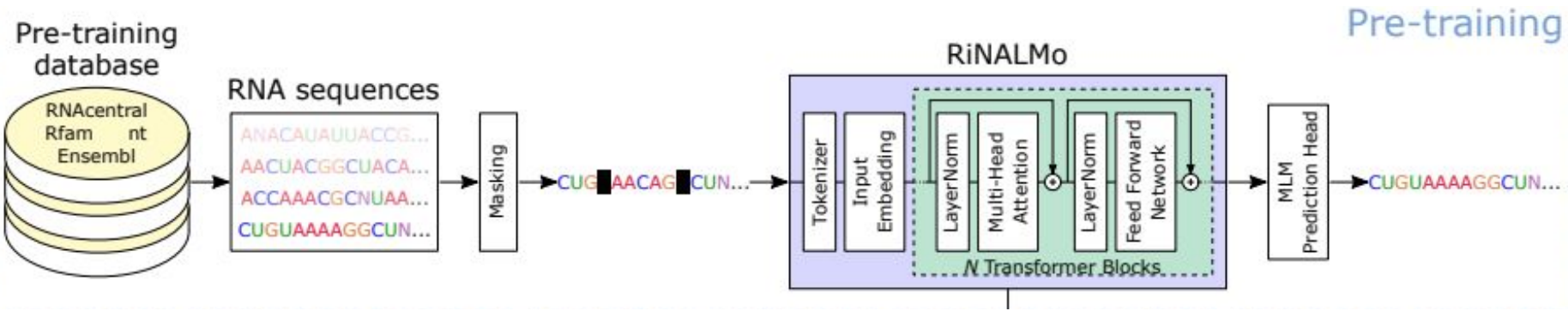
No-TL	Error rate
HMM	18.10%
BLASTp	35.90%
ProtCNN	27.60%
ProtENN	**12.20%**

Table 2: Pfam families classification with TL

ESM2	
KNN	15.55%
MLP	30.88%
CNN	17.48%
MLP-E	18.10%
CNN-E	**7.67%**

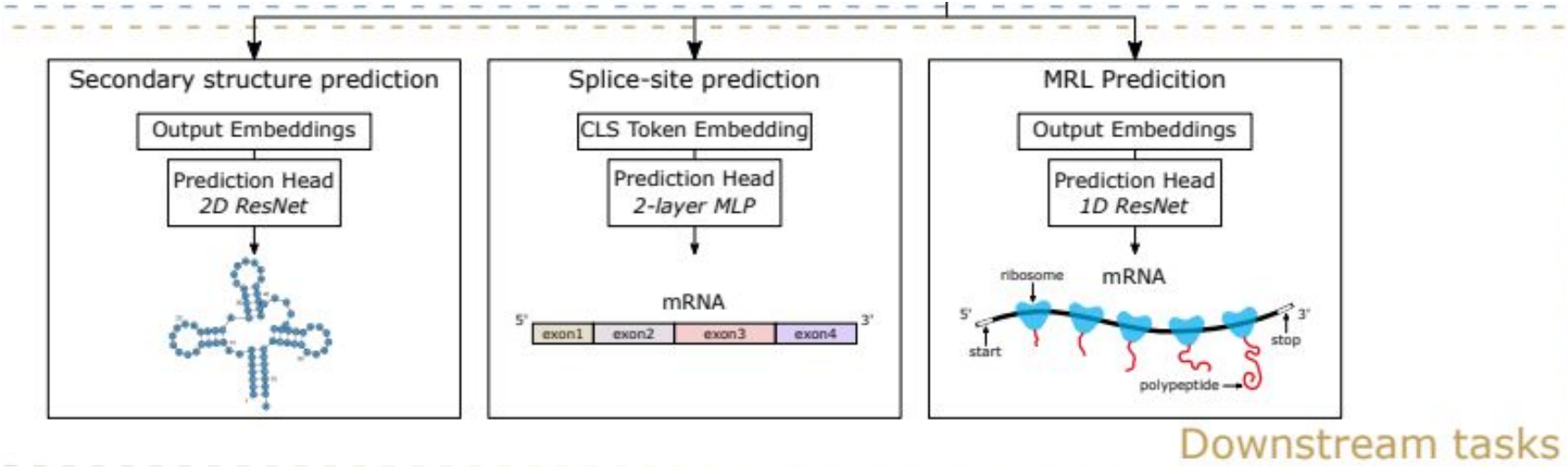
Evaluating large language models for annotating proteins,
Rosario Vitale, Leandro A Bugnon, Emilio Luis Fenoy, Diego H Milone, Georgina Stegmayer,
Briefings in Bioinformatics 2024

LLM para RNA



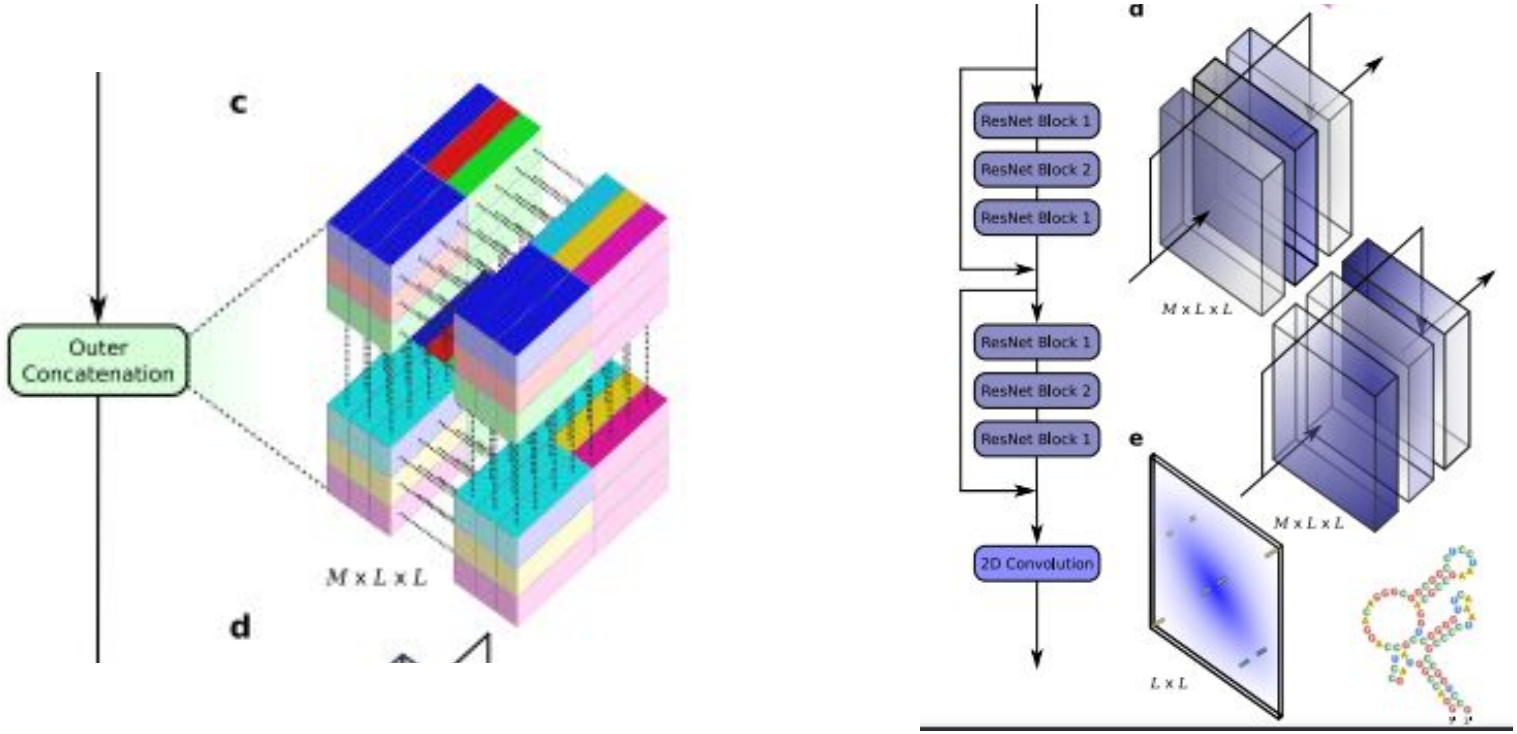
RiNALMo: General-Purpose RNA Language Models Can Generalize Well on Structure Prediction Tasks R. Penic´et al., arXiv 2024

LLM para RNA: “downstream tasks”



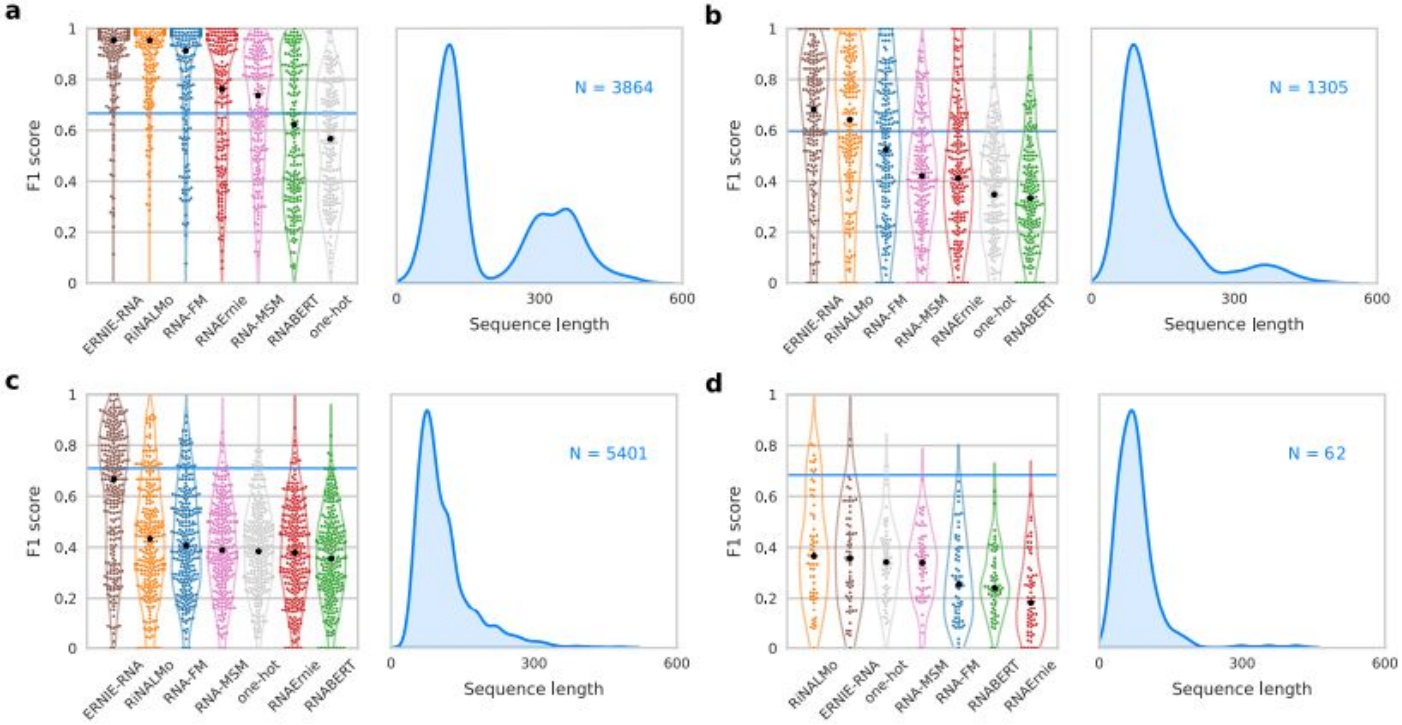
RiNALMo: General-Purpose RNA Language Models Can Generalize Well on Structure Prediction Tasks R. Penic´et al., arXiv 2024

Predicción de estructuras con LLMs



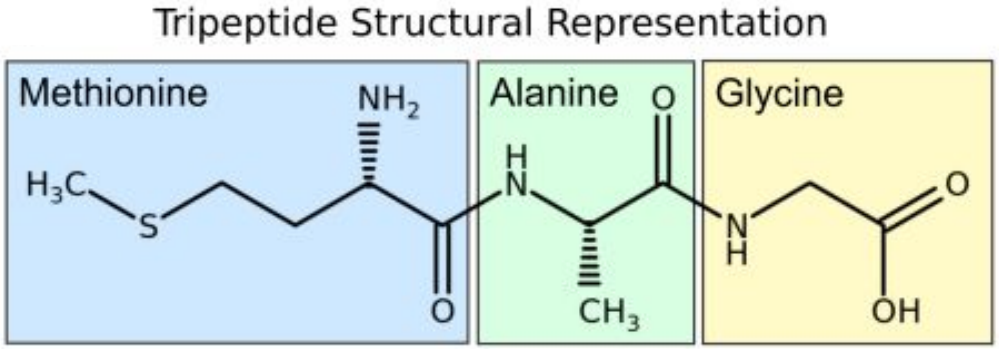
Comprehensive benchmarking of large language models for RNA secondary structure prediction, L.I. Zablocki, L.A. Bugnon, M. Gerard, L. Di Persia, G. Stegmayer, D.H. Milone, arXiv 2024

Predicción de estructuras con LLMs

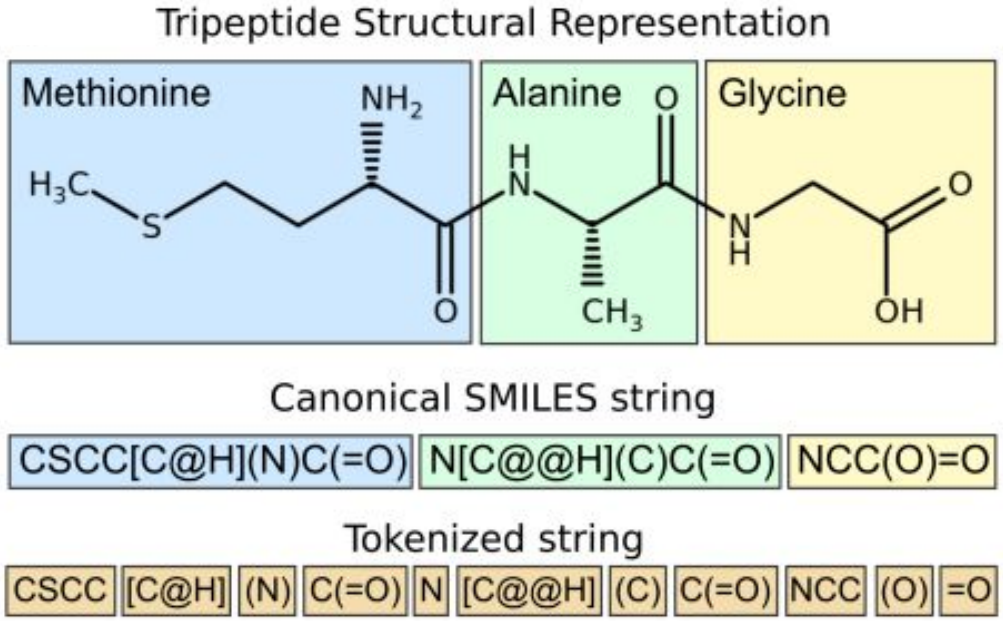


Comprehensive benchmarking of large language models for RNA secondary structure prediction, L.I. Zablocki, L.A. Bugnon, M. Gerard, L. Di Persia, G. Stegmayer, D.H. Milone, arXiv 2024

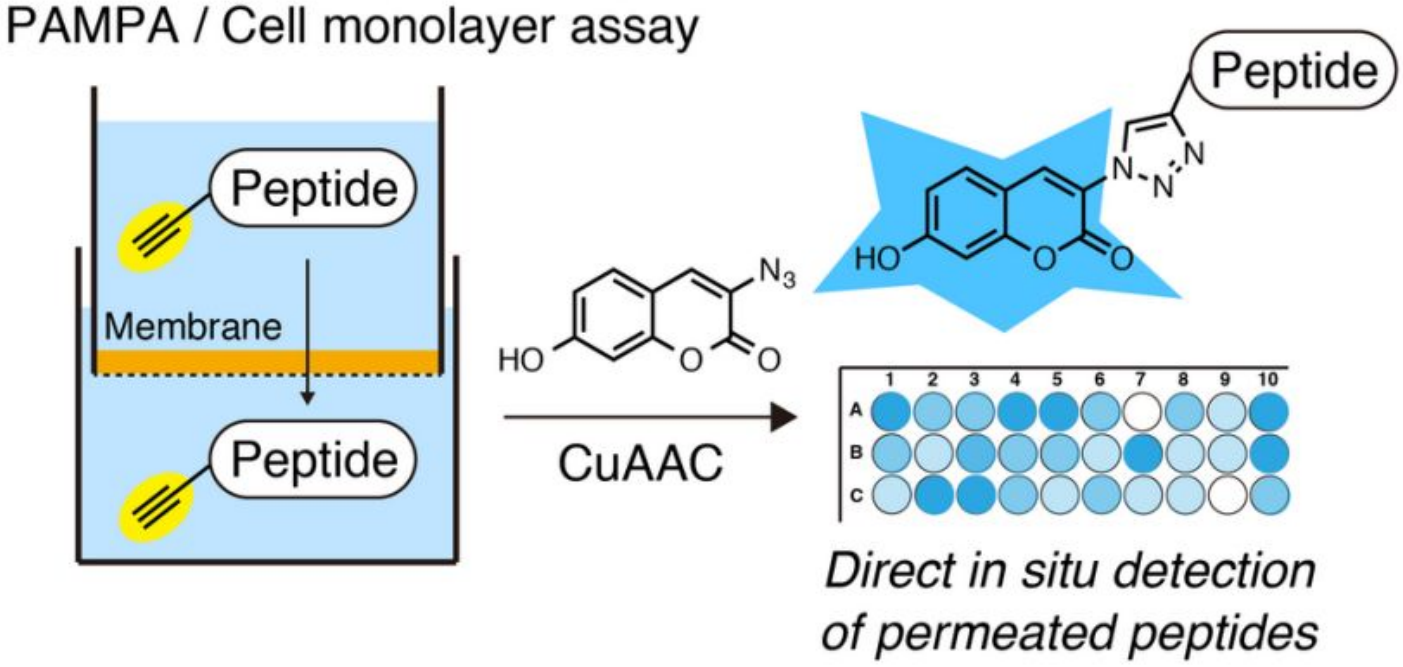
Modelando otras secuencias



Modelando otras secuencias

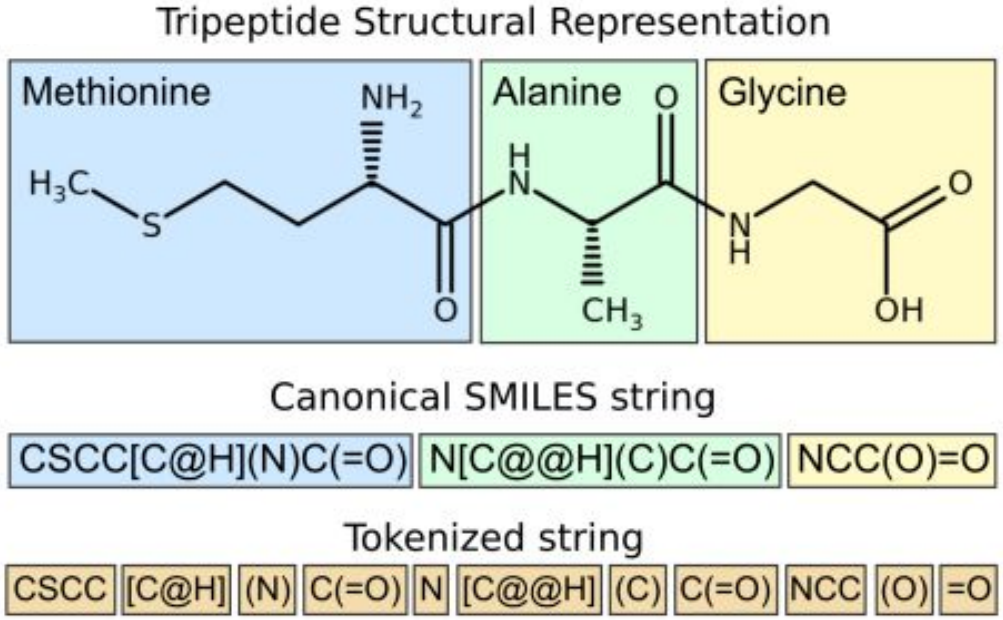


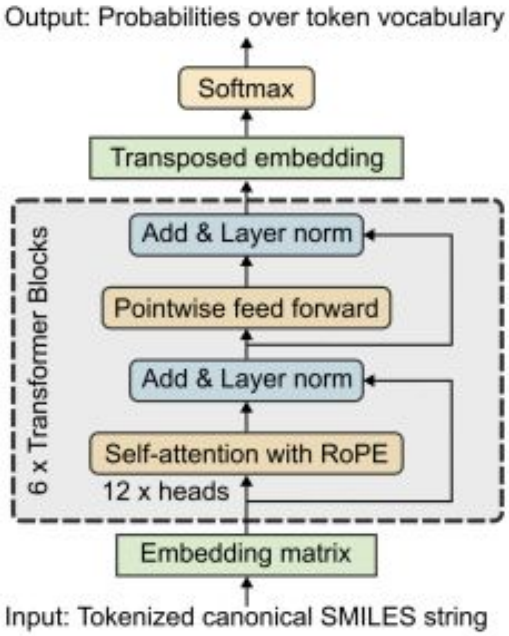
Modelando otras secuencias



Peptide-specific chemical language model successfully predicts membrane diffusion of cyclic peptides, A. Feller et al., bioRxiv 2024

Modelando otras secuencias

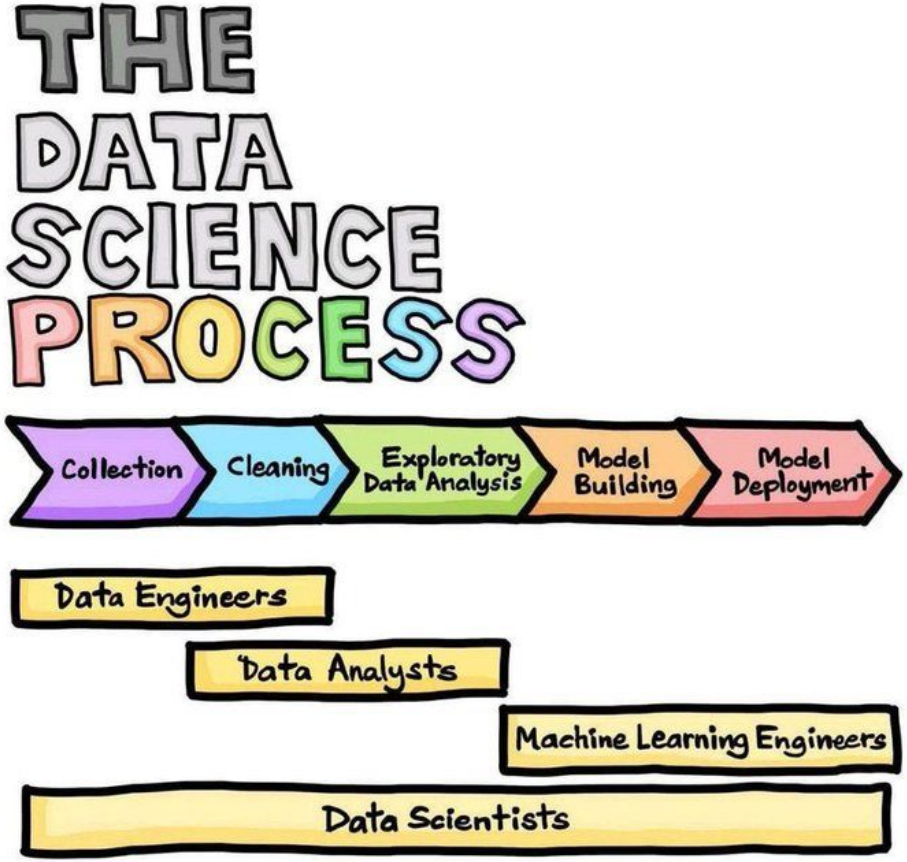




Peptide-specific chemical language model successfully predicts membrane diffusion of cyclic peptides, A. Feller et al., bioRxiv 2024

Herramientas: PyTorch & Google Colab

Representing SMILES


THE DATA SCIENCE PROCESS
Collection
Cleaning
Exploratory Data Analysis
Model Building
Model Deployment
Data Engineers
Data Analysts
Machine Learning Engineers
Data Scientists

Importancia de las particiones de datos

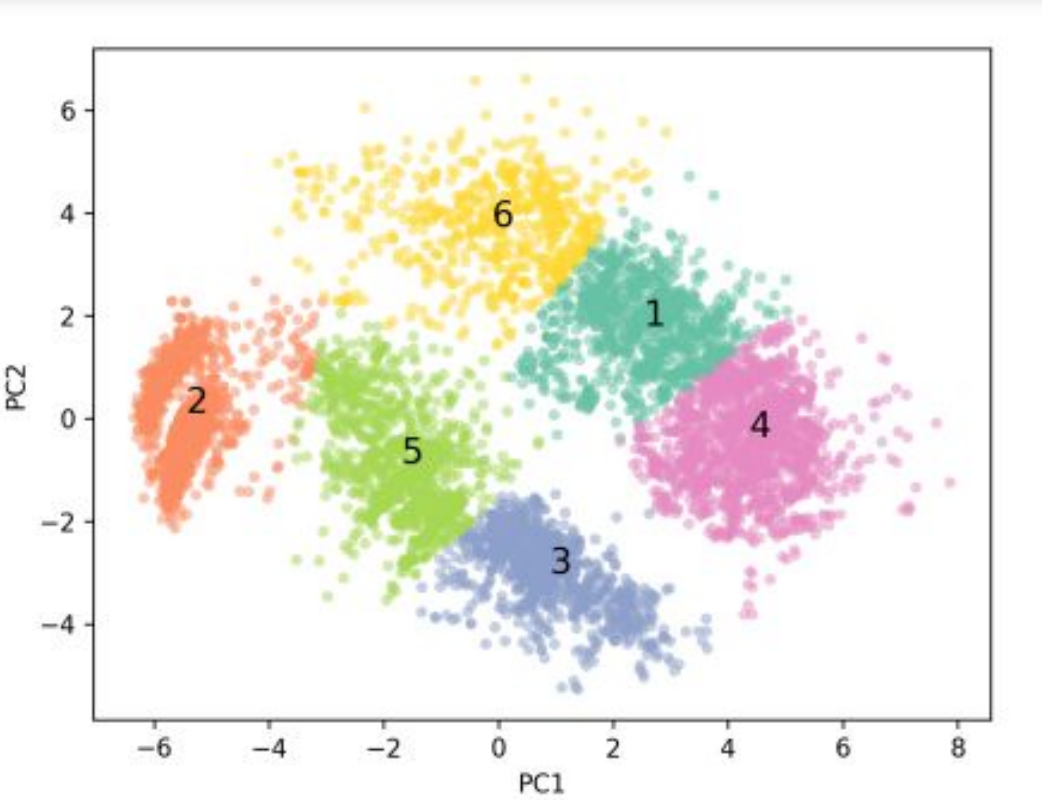
Los métodos de ML intentarán hacer una **inferencia** a partir de la muestra (los datos de entrenamiento).

La muestra tiene que ser representativa del problema.

Siempre debemos dejar parte de los datos para **test**

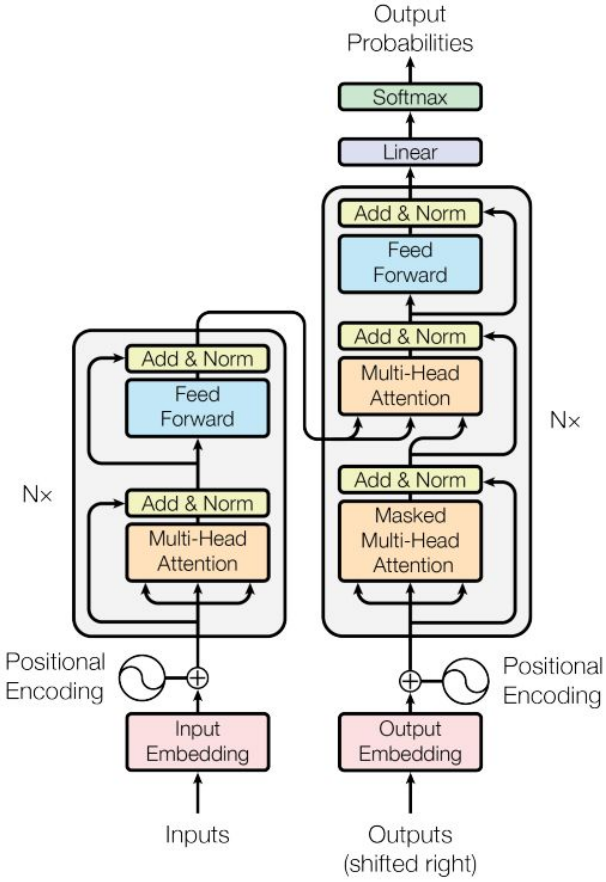
Y revisar que exista una distancia razonable entre test y tain…

Agrupando secuencias por similaridad



Tarea

Modelos generativos



Entrada: “Hola mundo!”
Salida: “Hello world!”

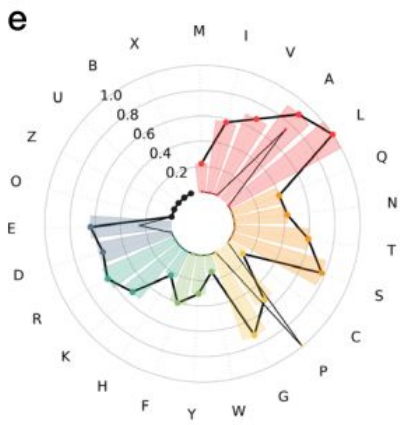
Inputs: [Hola, mundo, !]
Outputs: [<BOS>]
[<BOS>, Hello]
[<BOS>, Hello, world]
[<BOS>, Hello, world, !]

Attention Is All You Need y col., Viswani 2017

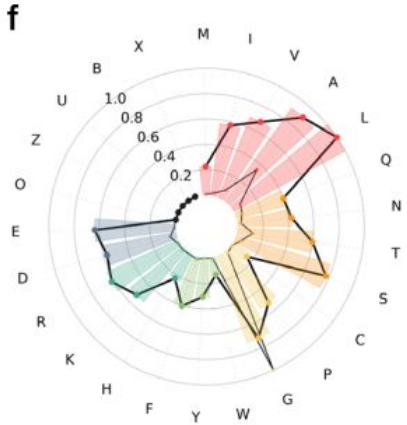
Modelos generativos

El objetivo es usar los modelos de lenguaje para **generar nuevas secuencias**

- Diseño de secuencias de-novo
- Generar mutaciones funcionales
- Condicionar la generación
  a propiedades de interés



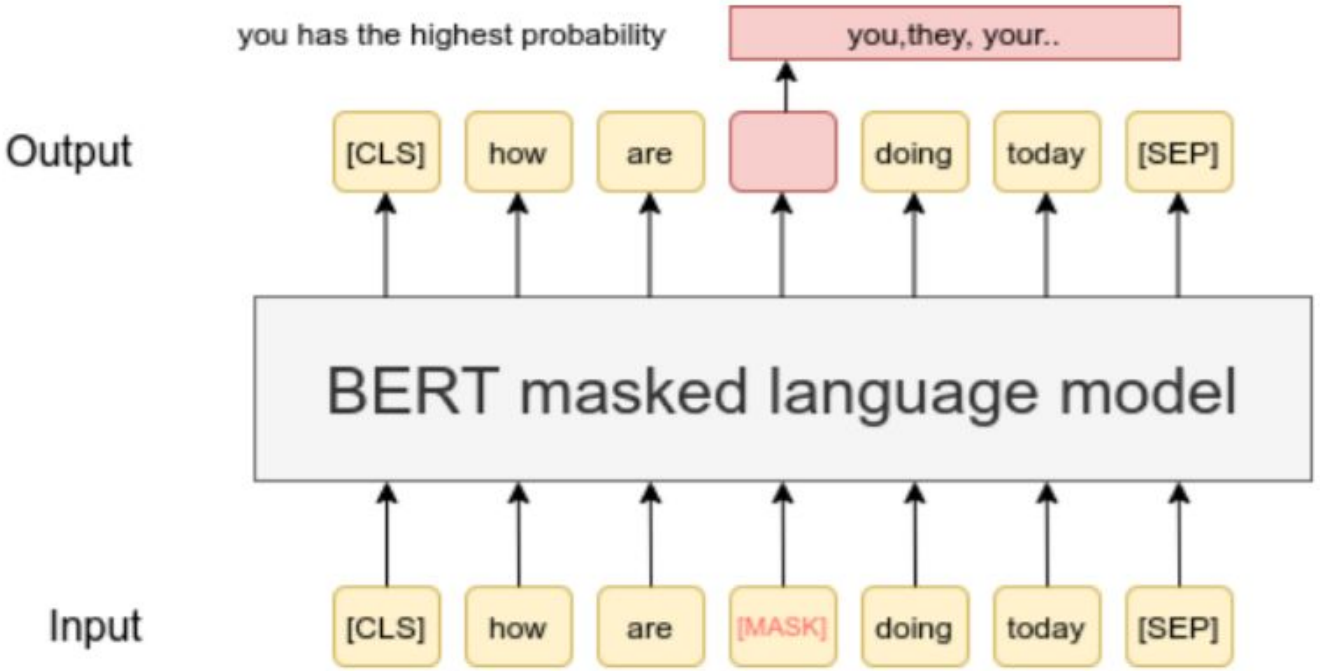
MSNDTPTHDPTPPAPAPAPEPAPAPAPEP
APAPAPEPAPAPAPEPAPAPAPEPAPAPA
PEPAPAPAPEPAPAPAPEPAPAPAPEPAP
APAPEPAPAPAPEPAPAPAPEPAPAPAPE
…



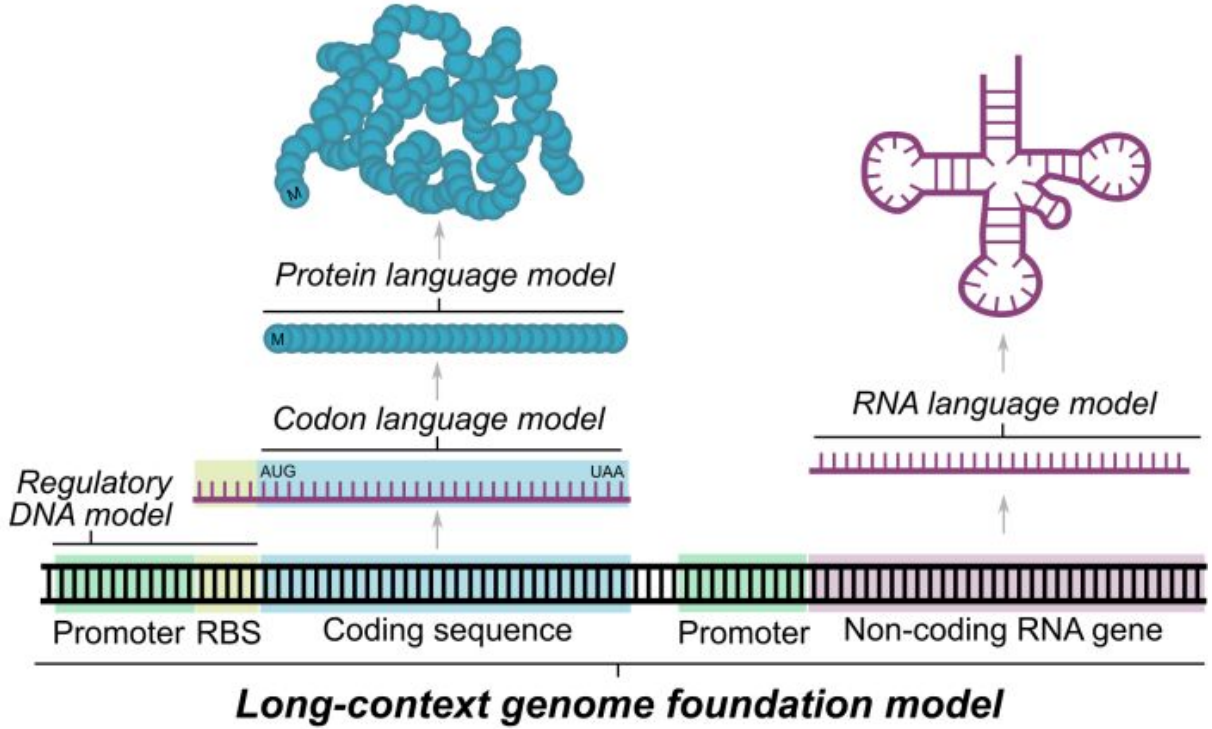
GAGGVGGAGGGTGGAGGRAELLFGAGGAG
GAGGAGTDGGPGATGGTGGHGGVGGDGGW
LAPGGAGGAGGQGGAGGAGSDGGALGGTG
GTGGTGGAGGAGGRGALLLGAGGQGGLGG
AGGQGGTGGAGG…

ProtGPT2 is a deep unsupervised language model for protein design
N. Ferruz et al., Nat comm 2022

Volviendo a la predicción por enmascaramiento

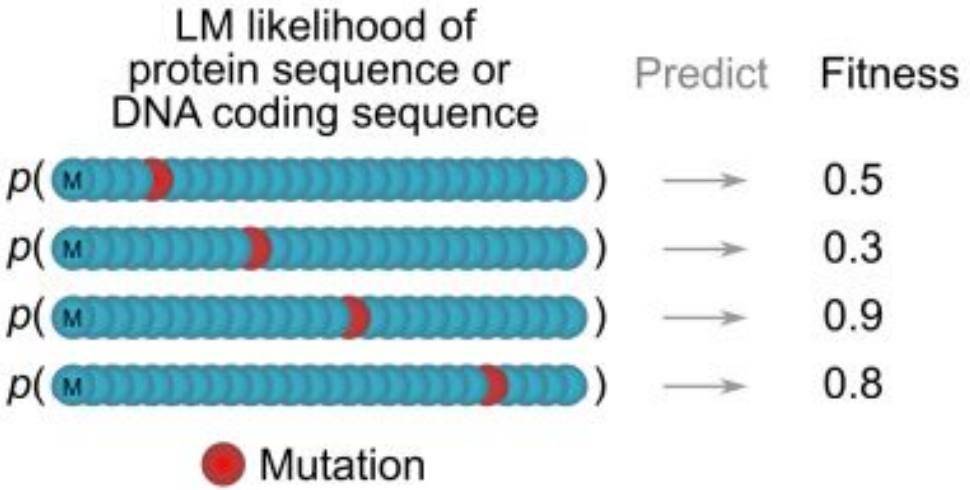


Evo: un modelo generativo para DNA



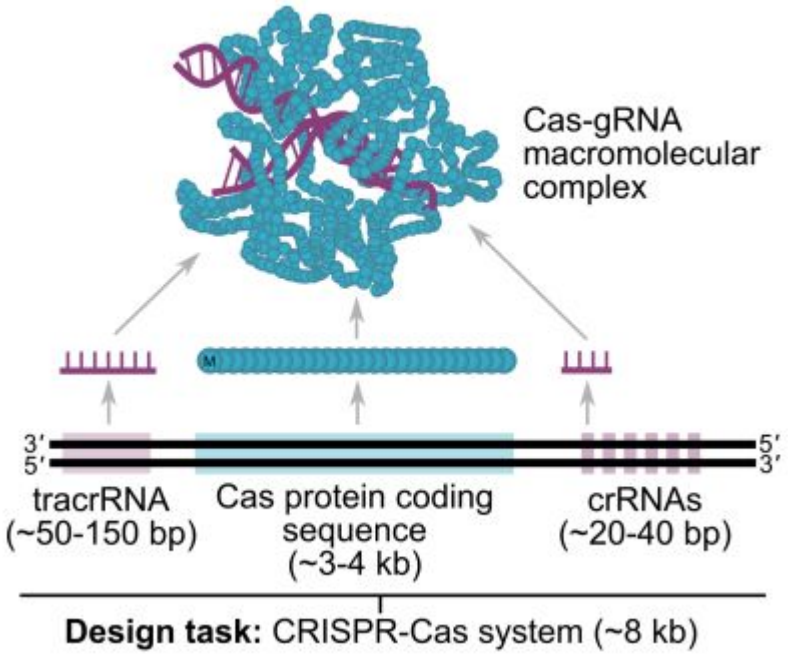
Sequence modeling and design from molecular to genome scale with Evo, E. Nguyen et al, bioRxiv 2024

Estimando propiedades a partir del LM



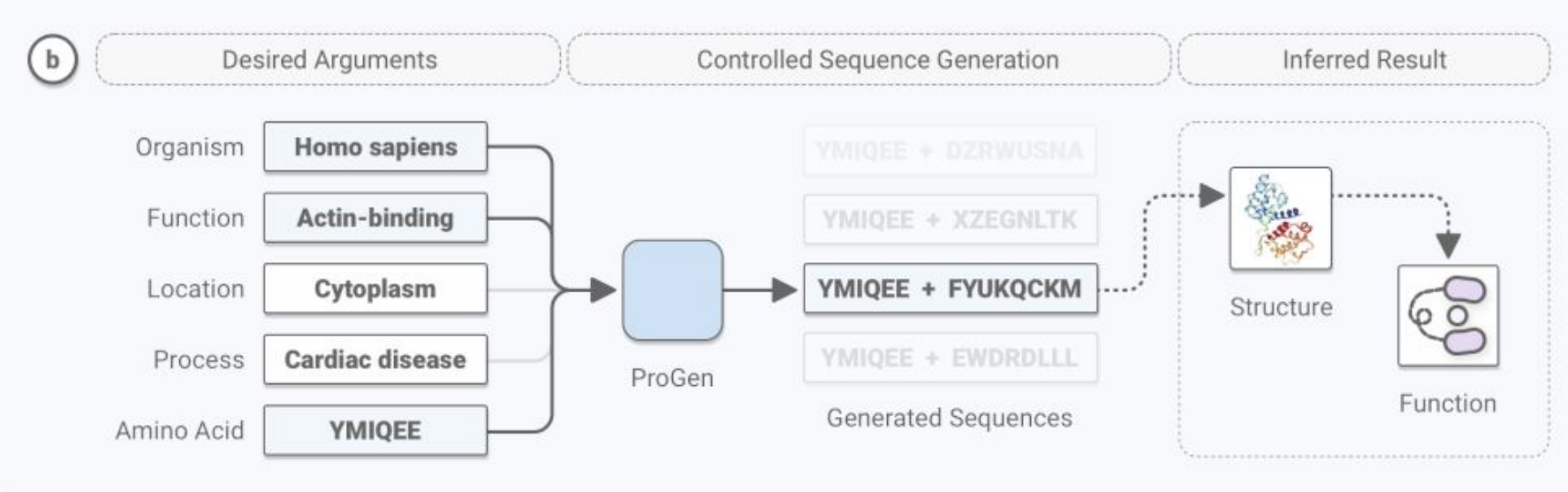
Sequence modeling and design from molecular to genome scale with Evo, E. Nguyen et al, bioRxiv 2024

Generando complejos funcionales



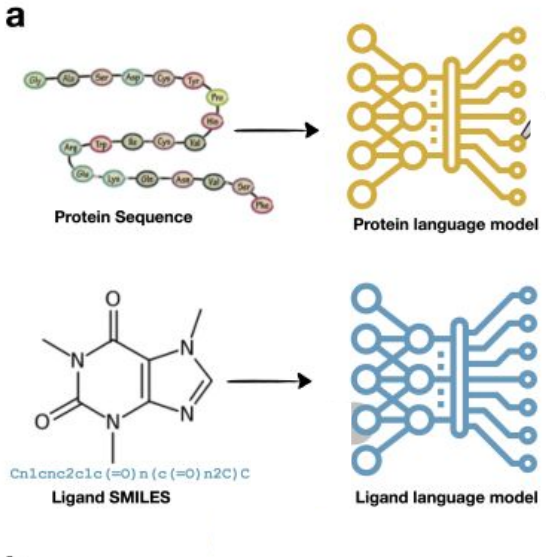
Sequence modeling and design from molecular to genome scale with Evo, E. Nguyen et al, bioRxiv 2024

Modelos condicionados



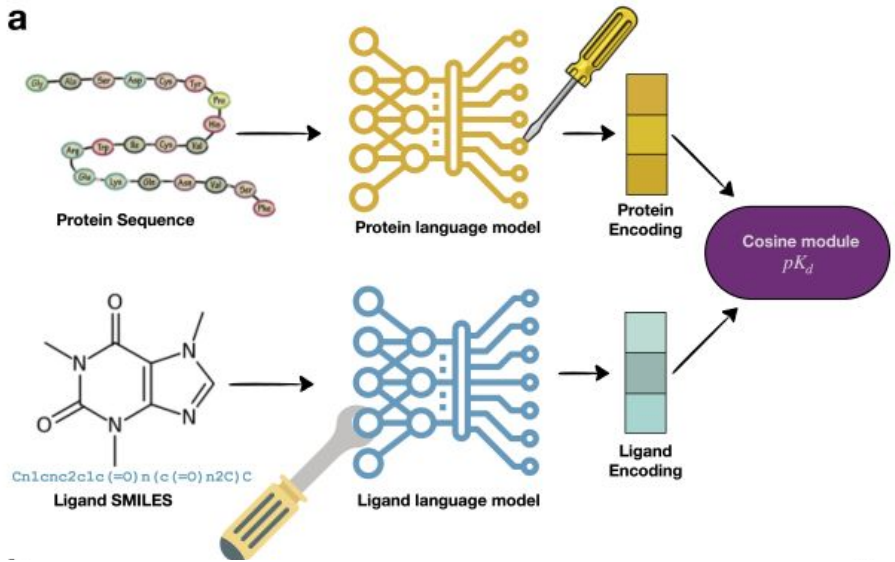
ProGen: Language Modeling for Protein Generation, arxiv 2023

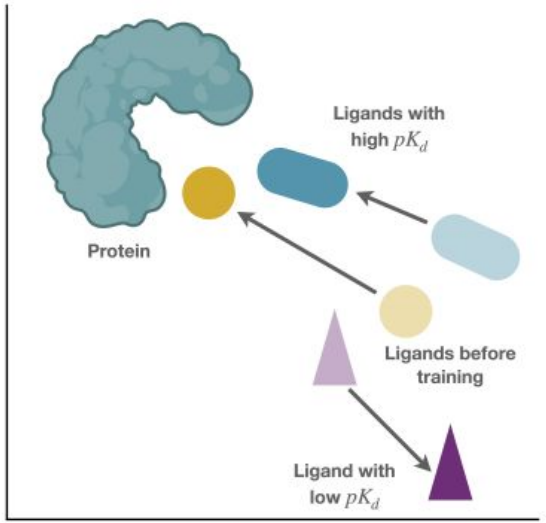
Combinando LLMs



Learning Binding Affinities via Fine-tuning of Protein and Ligand Language Models, R. Gorantla et al, biorXiv 2024

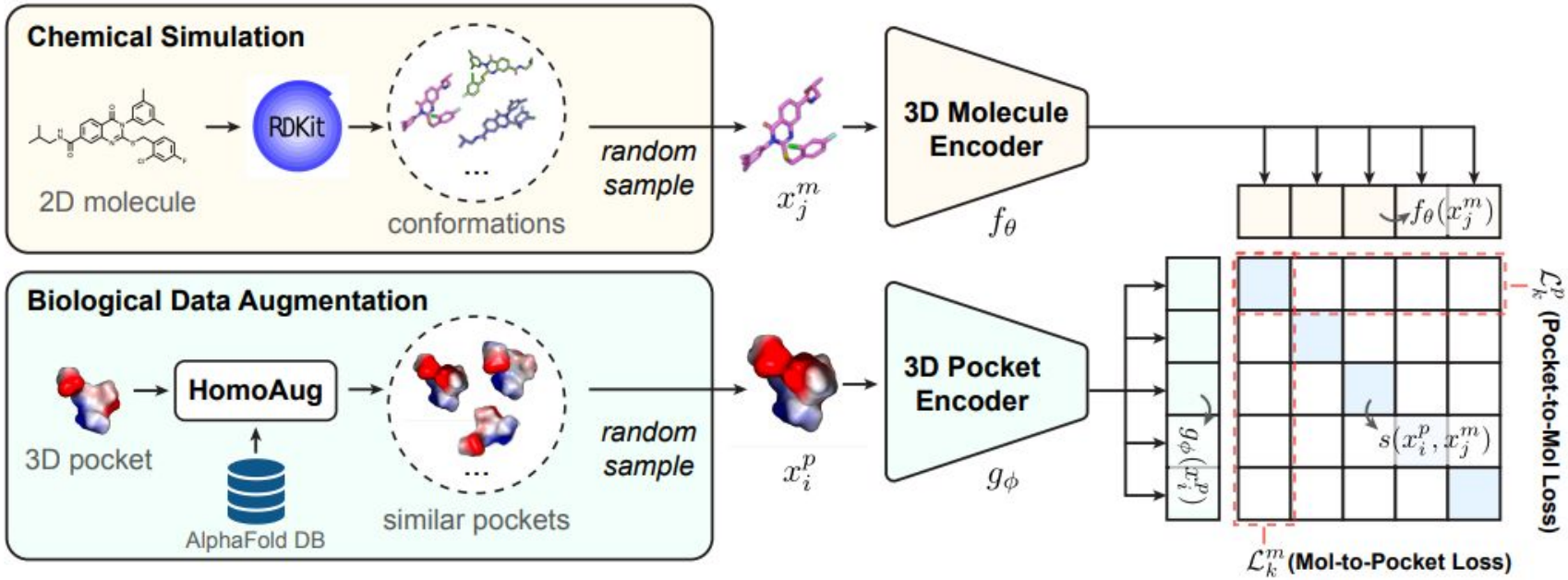
Combinando LLMs





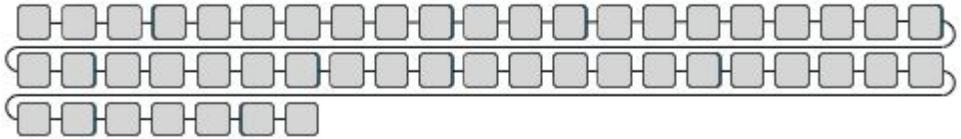
Learning Binding Affinities via Fine-tuning of Protein and Ligand Language Models, R. Gorantla et al, biorXiv 2024

Y no solo en secuencias



DrugCLIP: Contrastive Protein-Molecule Representation Learning for Virtual Screening Bowen, B. Gao 2023, arXiv

EvoDiff



Protein generation with evolutionary diffusion: sequence is all you need, S. Alamdari, biorXiv 2023

Taller SAJIB 9:
Modelos fundacionales para secuencias biológicas

Leandro Bugnon
lbugnon.github.io

UNL CONICET

s i n c (*i*)