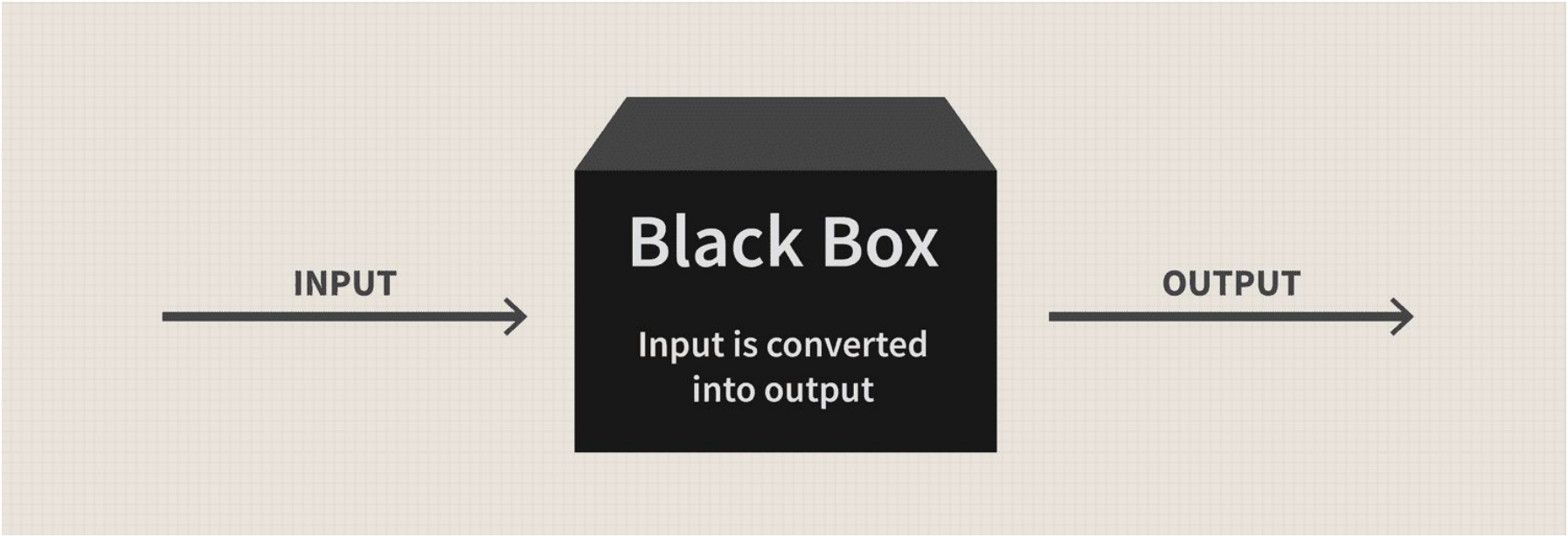
Teaching Language Models to Speak Biology

“If it’s machine learning, it’s probably written in Python. If it’s AI, it’s probably written in PowerPoint.”

-Somebody

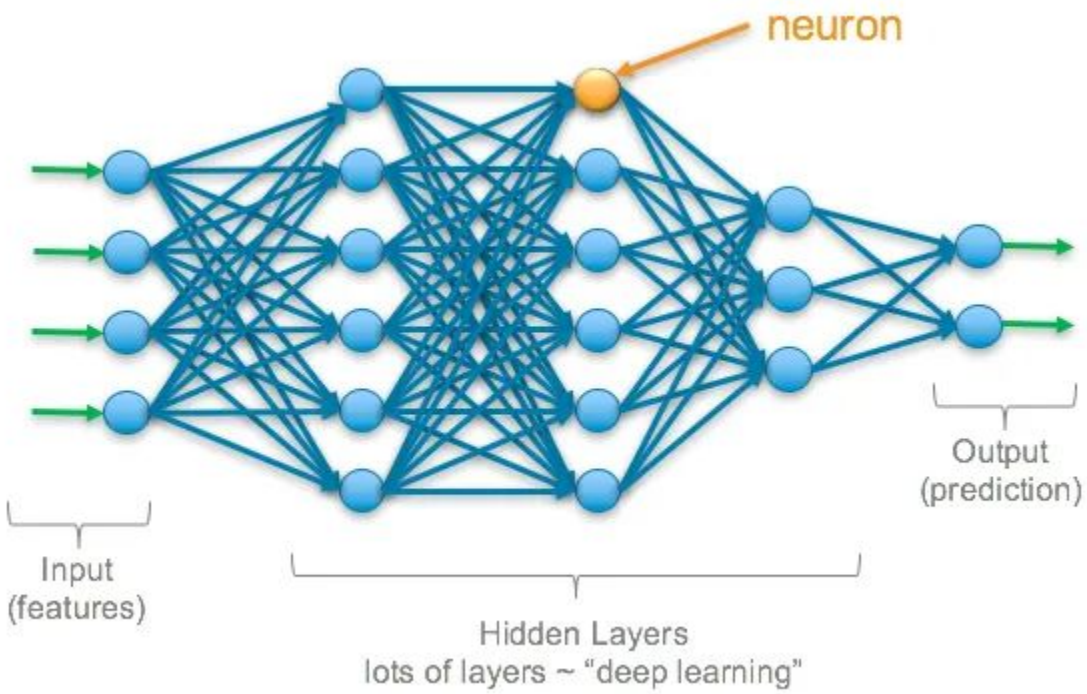
Roadmap

- Inside the black box: neural networks and software 2.0
- A two-neuron toy example
- Transformers: how do we get language in and out?
- The attention mechanism, the driver behind the transformer
- Applications, examples, and Google Colab

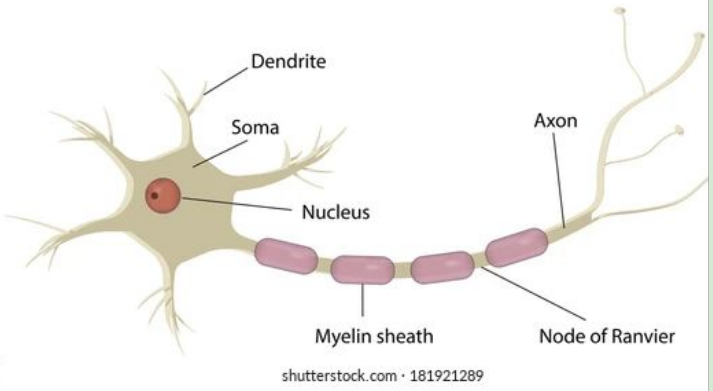

INPUT
Black Box
Input is converted into output
OUTPUT

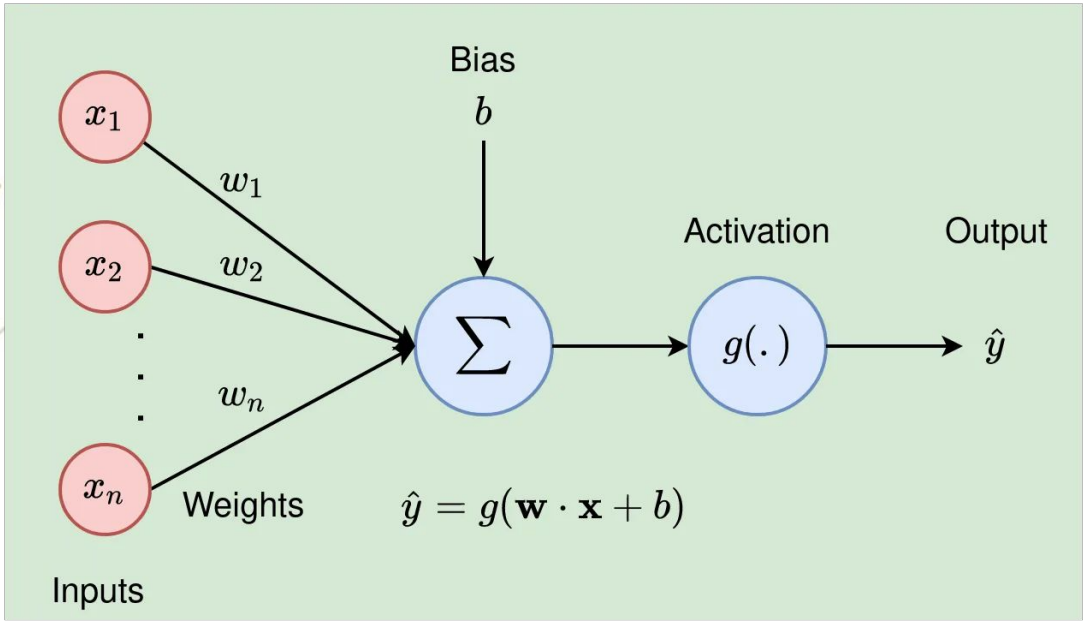
Machine learning derives logic from input data

A machine learning network is made up of neurons



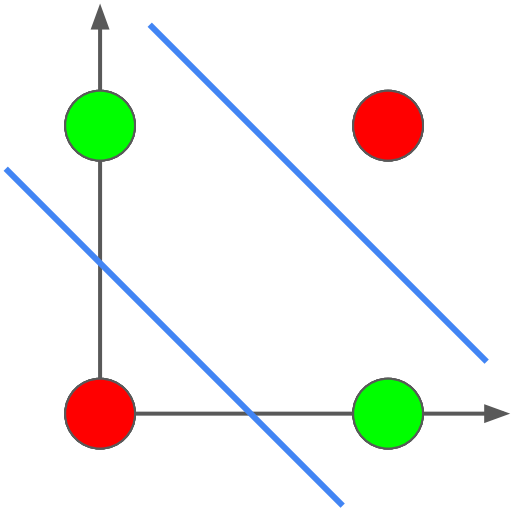
The neuron is the smallest unit of machine learning



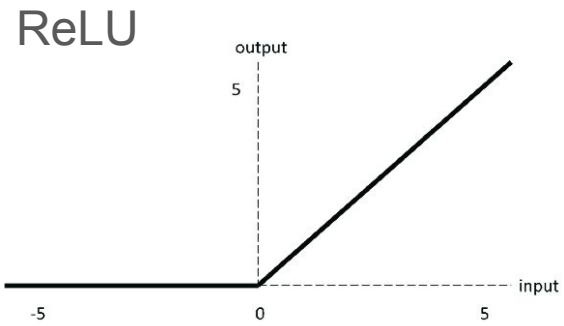


Why use neurons and ML?

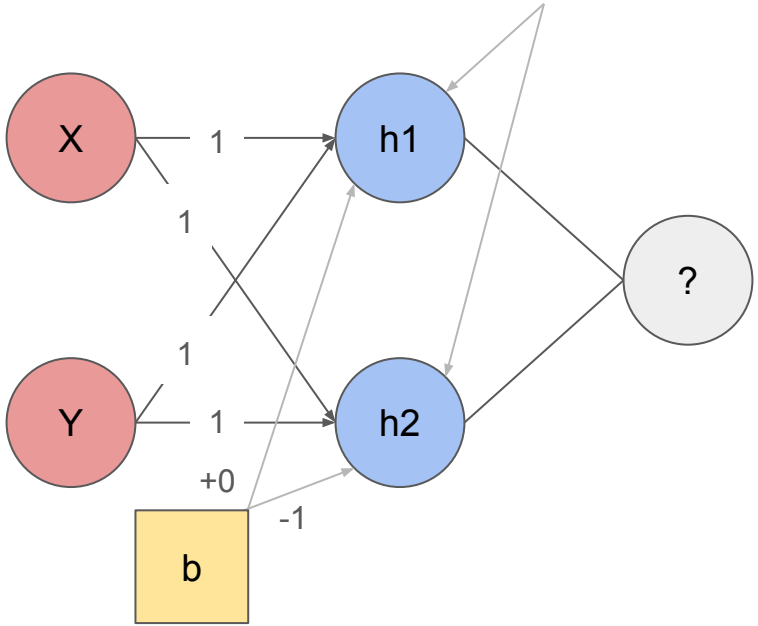
X	Y
0	0
0	1
1	0
1	1



Neural networks can restructure data

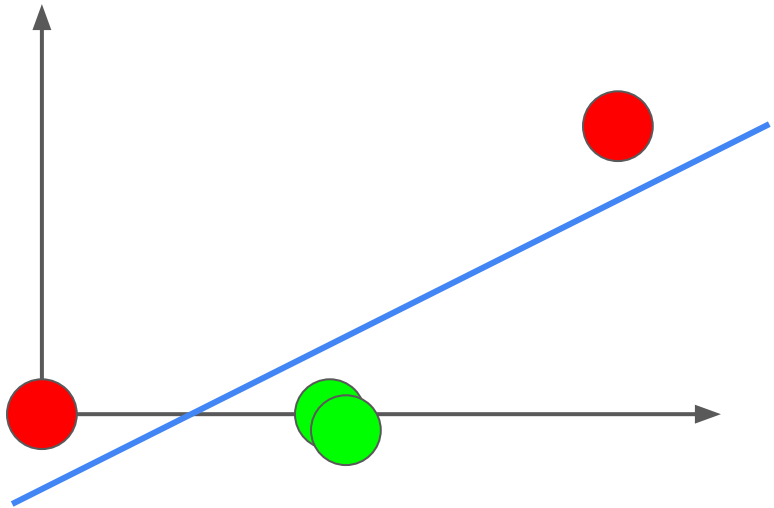


X	Y	h1	h2	XOR (X⊕Y)
0	0			0
0	1			1
1	0			1
1	1			0

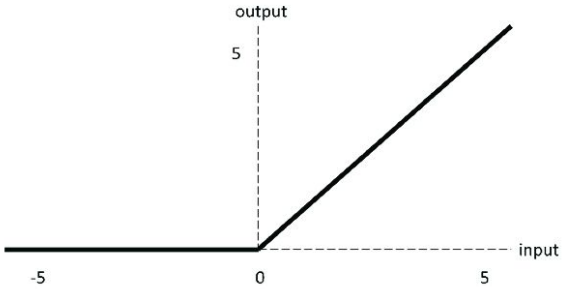


We can now linearly separate for the XOR function

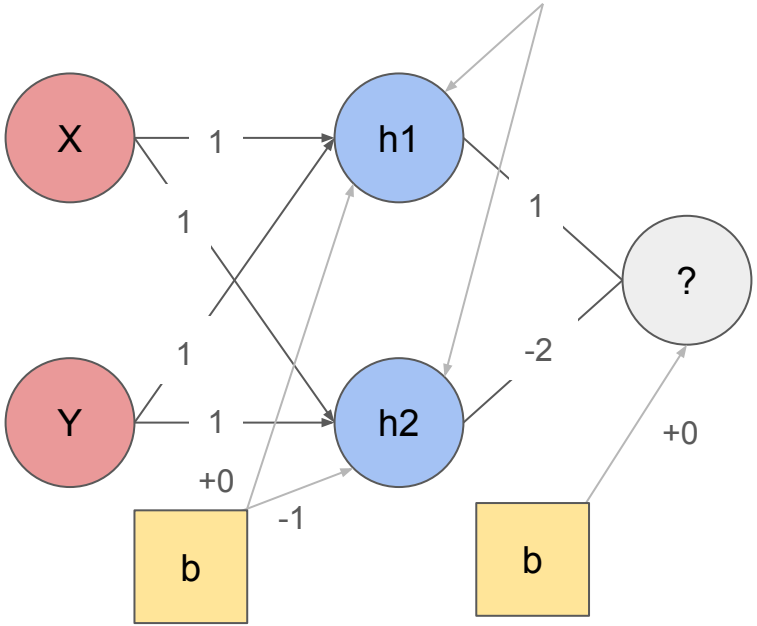
X	Y	h1	h2	XOR (X⊕Y)
0	0	0	(-1) 0	0
0	1	1	0	1
1	0	1	0	1
1	1	2	1	0



Final steps for computing the result

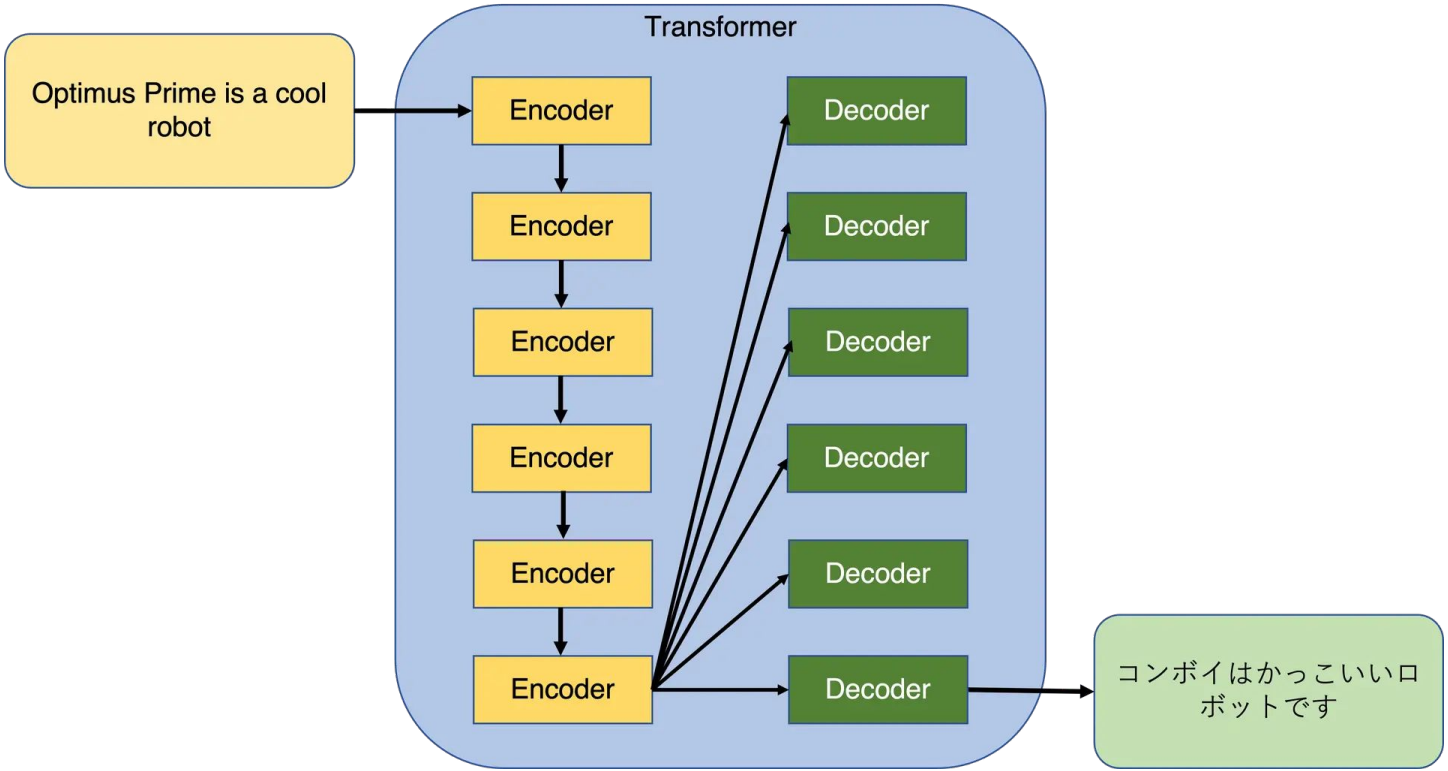


X	Y	h1	h2	XOR (X⊕Y)	Result
0	0	0	(-1) 0	0	
0	1	1	0	1	
1	0	1	0	1	
1	1	2	1	0	



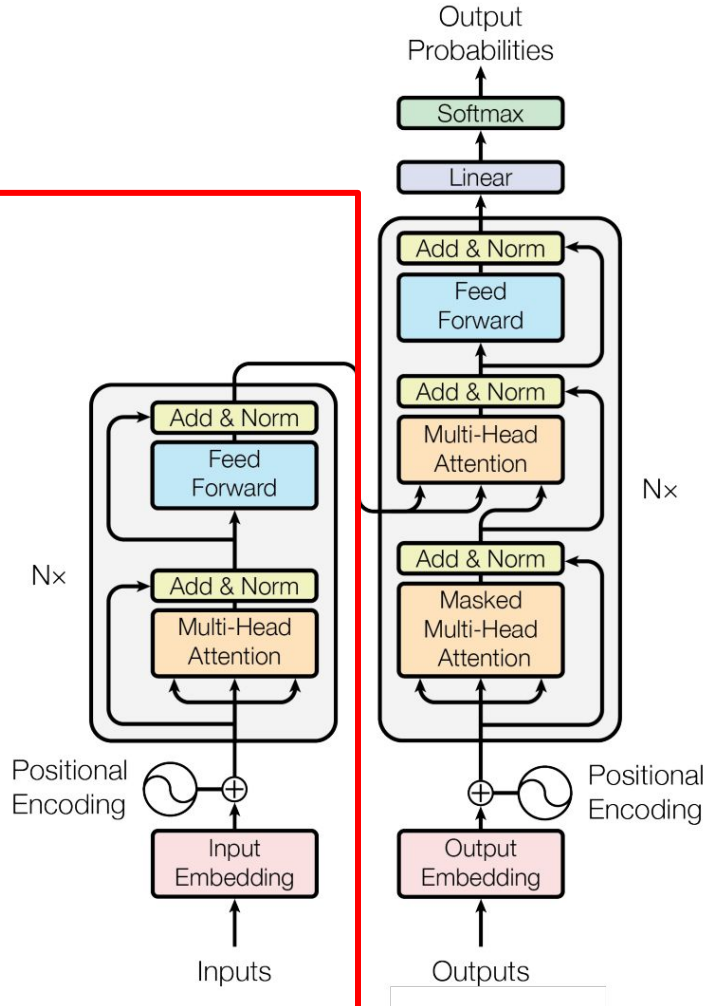
Enter the Transformer

The term transformer was coined because it transforms data



BERT

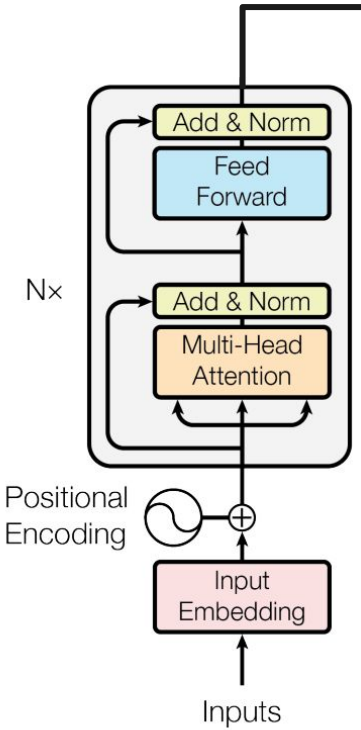
Encoder



GPT

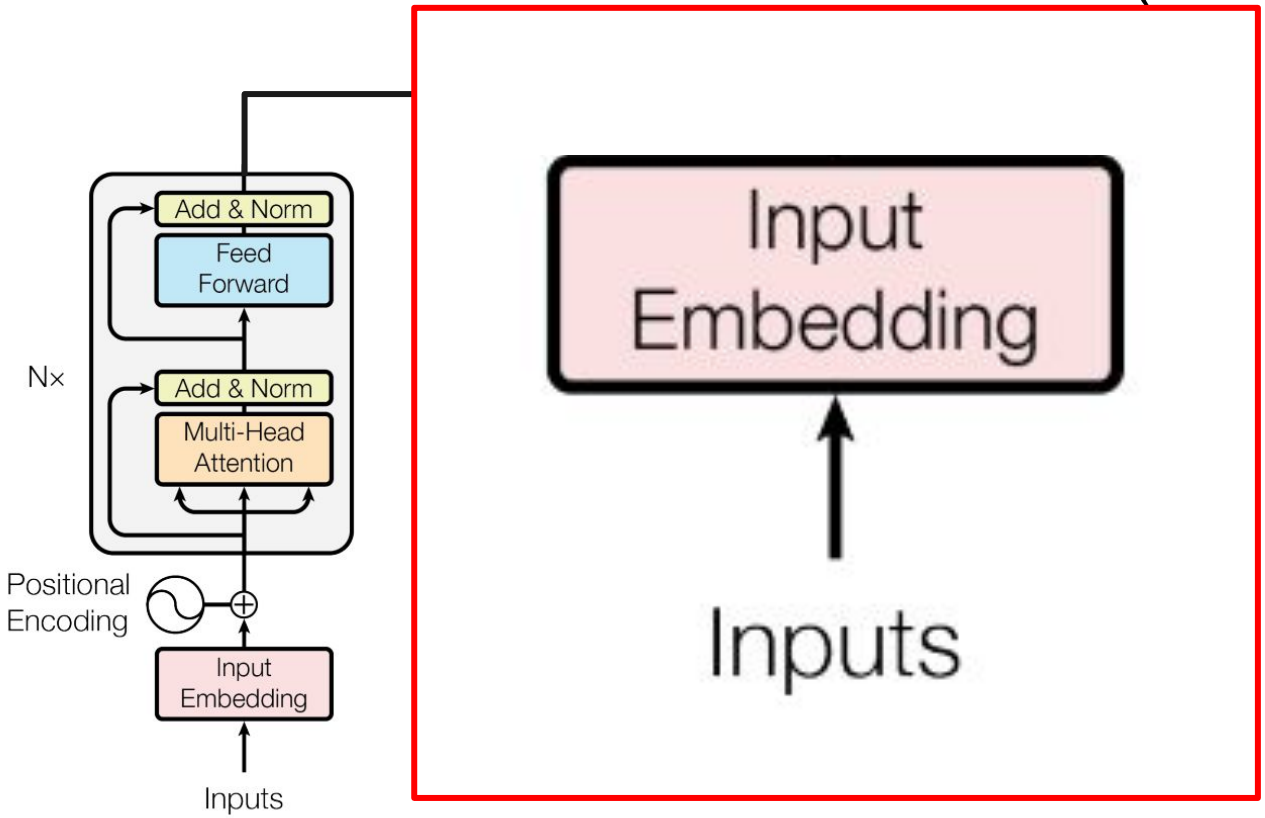
Decoder

The output of the encoder (BERT) is still functional!



[0.1150, -0.1438, 0.0555, 0.1442]

We can use the encoder for downstream tasks (transfer learning)

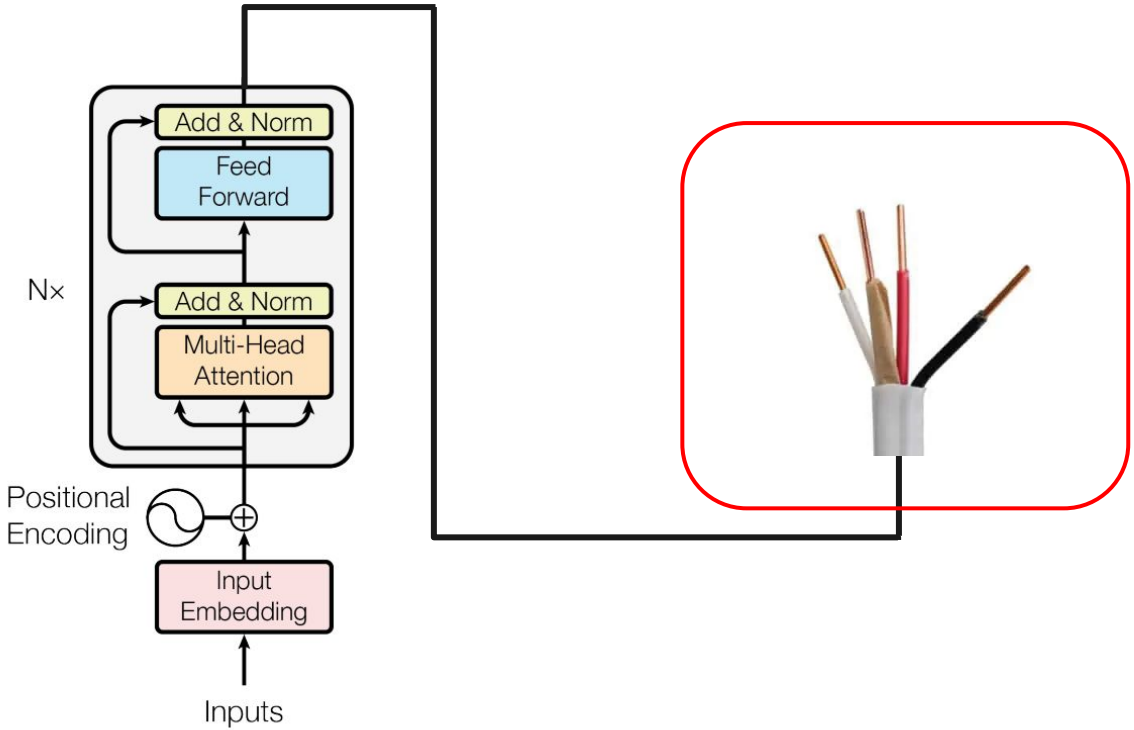


We get text into a model through tokenization

The token is converted into a vector of numbers

Token String		Token ID		Embedded Token Vector
this	->		->	

How do we get the outputs?

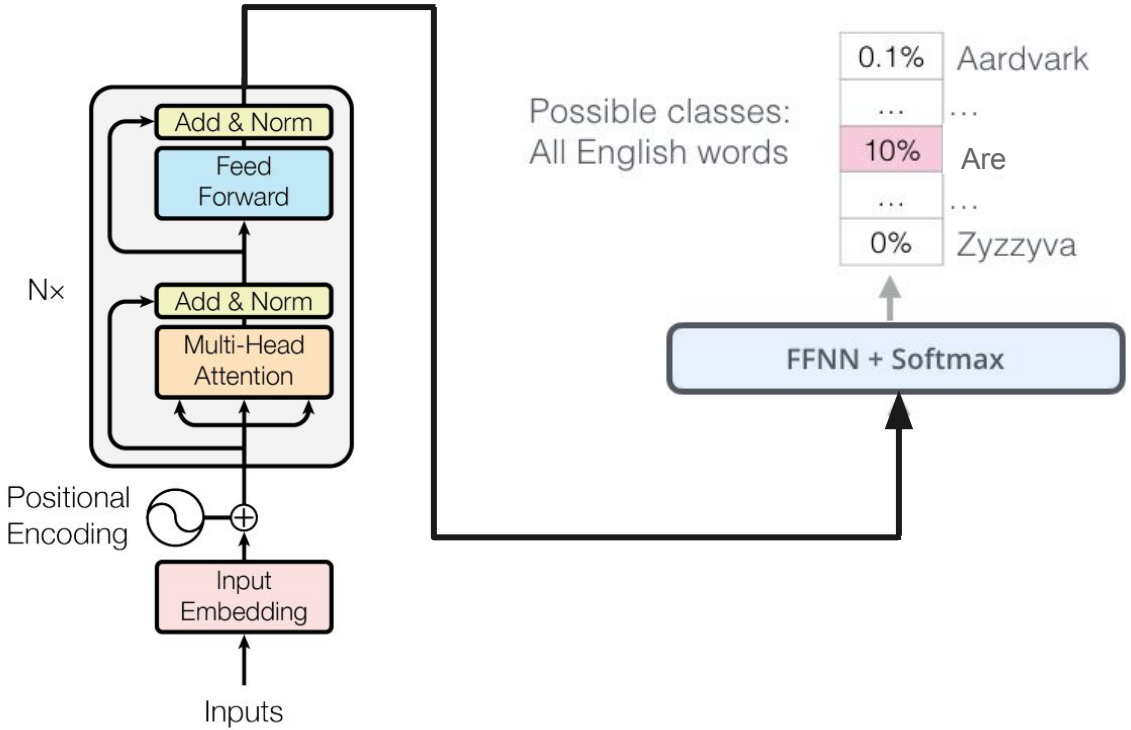


We train the model with an unlabeled corpus of data

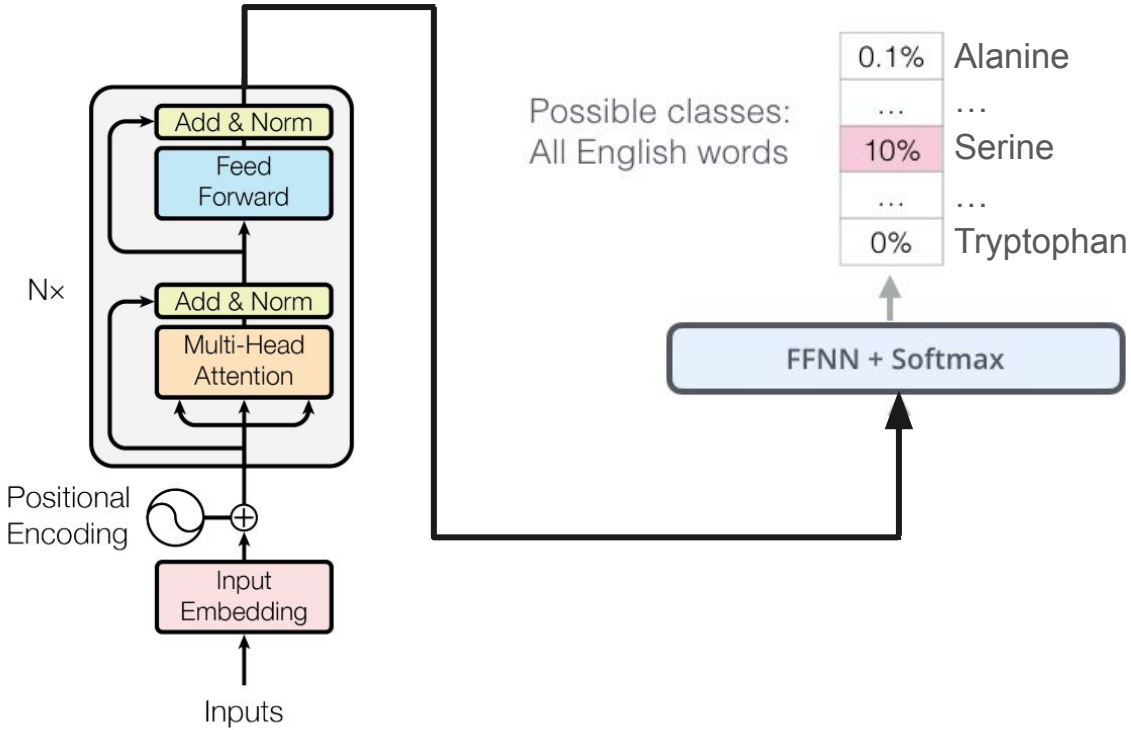
Input sentence:

“Hi how are you?”

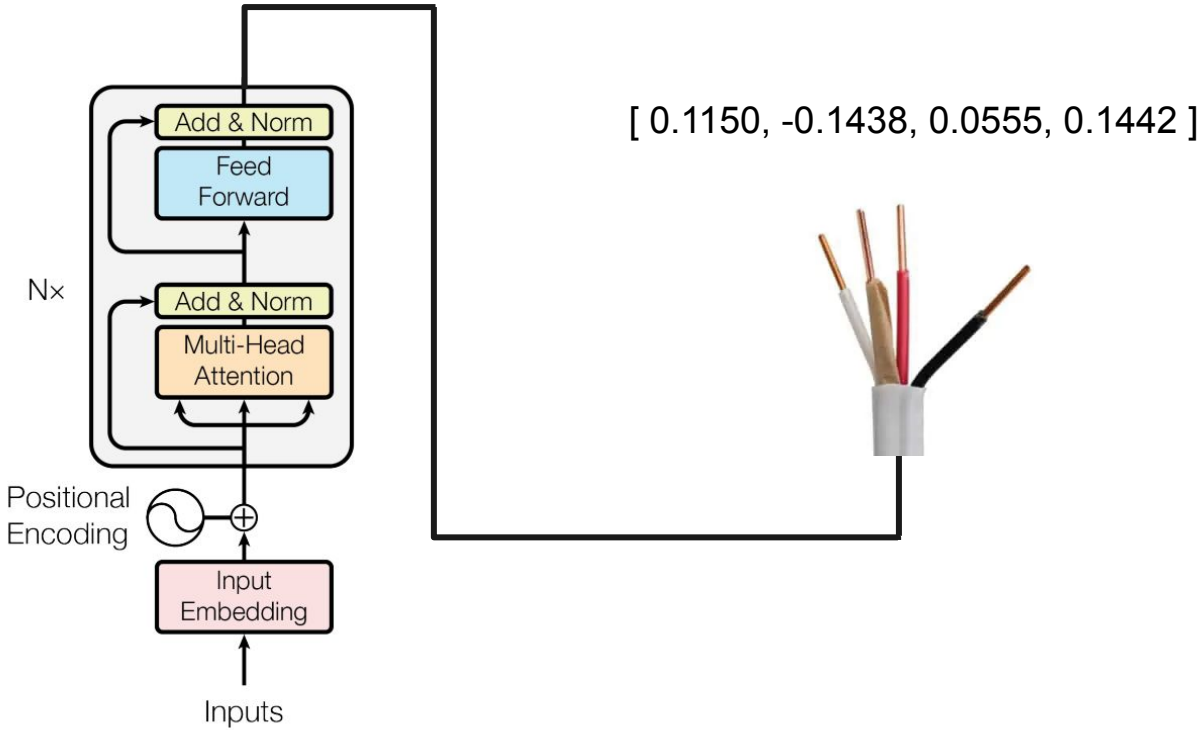
This training infuses the model with a general understanding of language



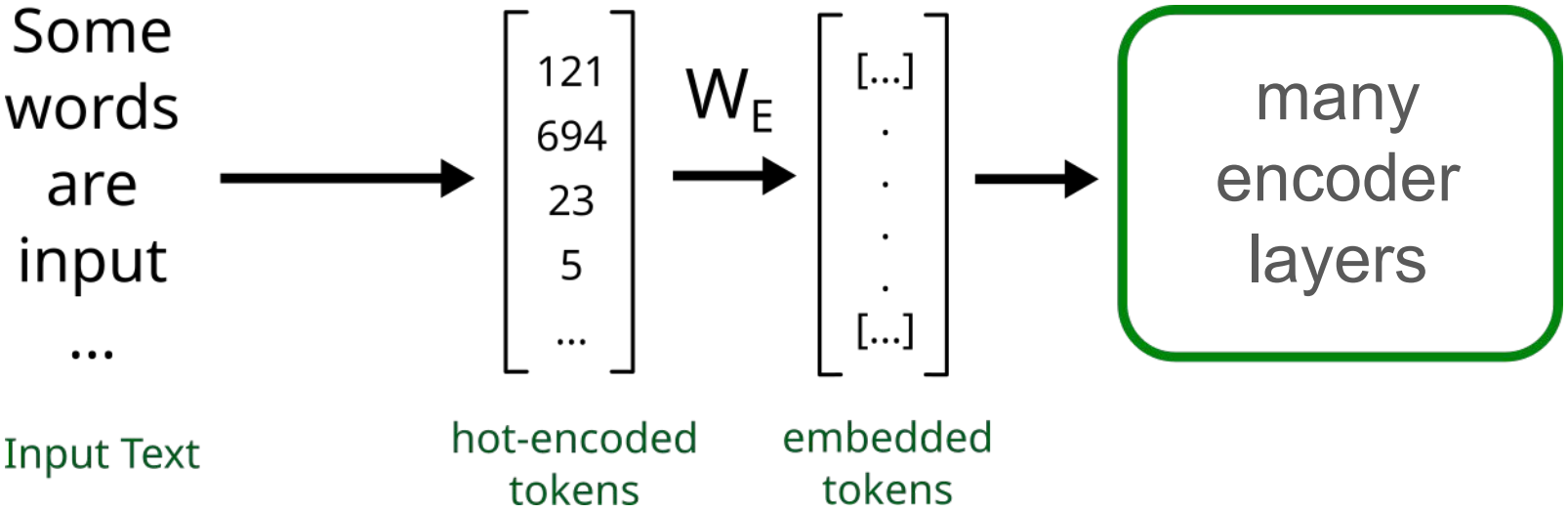
Or, in our case, proteins!



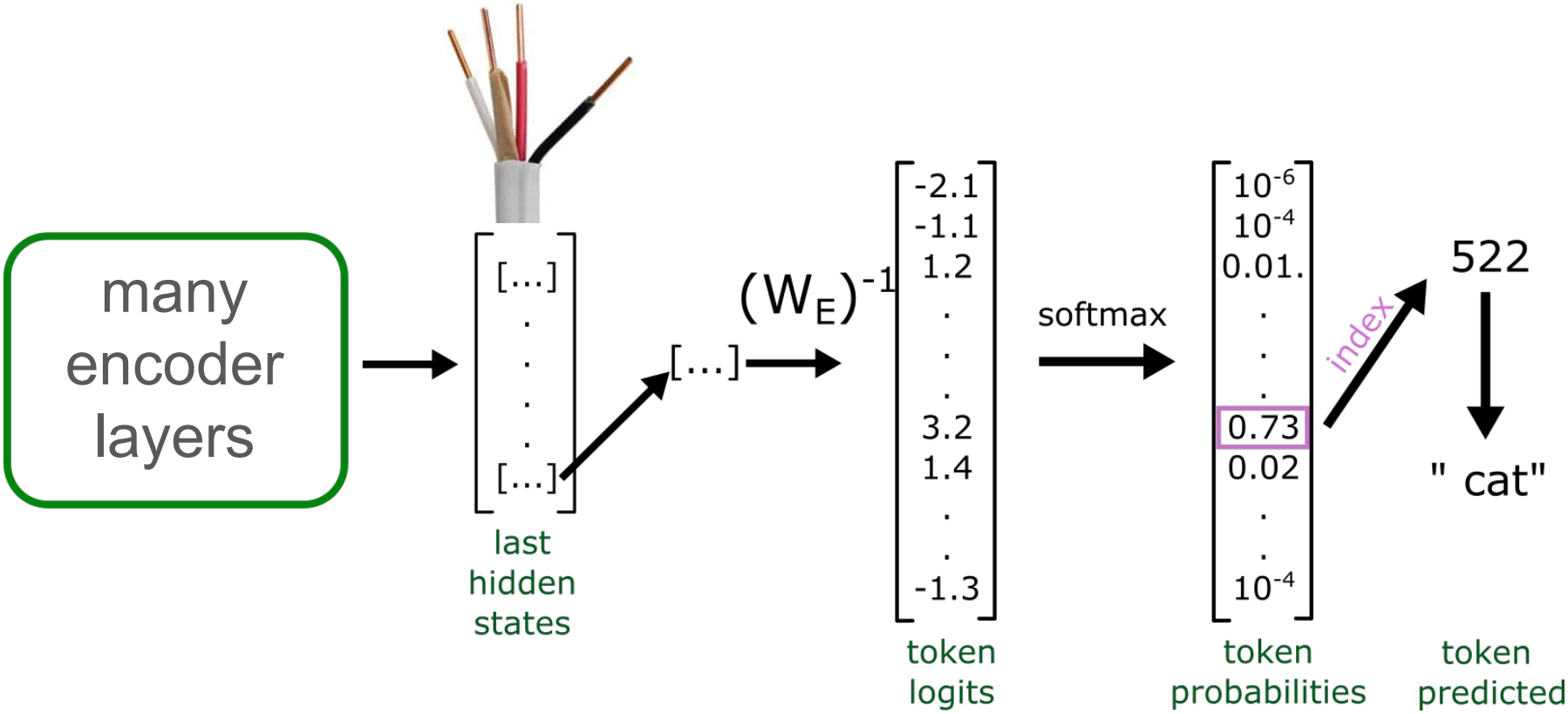
Once again: The output of the encoder is still functional!



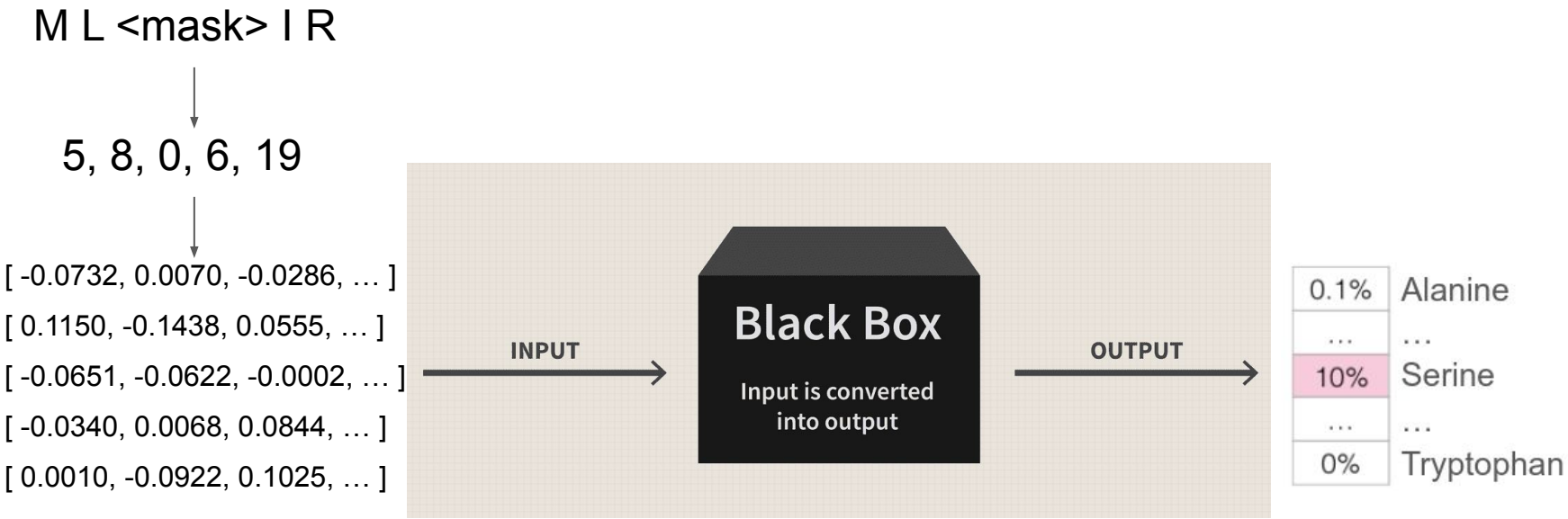
Overview of input



Overview of output



Let’s recap:


M L <mask> I R
5, 8, 0, 6, 19
[-0.0732, 0.0070, -0.0286, …]
[0.1150, -0.1438, 0.0555, …]
[-0.0651, -0.0622, -0.0002, …]
[-0.0340, 0.0068, 0.0844, …]
[0.0010, -0.0922, 0.1025, …]
INPUT
Black Box
Input is converted into output
OUTPUT
0.1% Alanine
… …
10% Serine
… …
0% Tryptophan

Transformers work by a mechanism called **attention**

Attention is **all you need**

A Vaswani, N Shazeer, N Parmar… - Advances in neural …, 2017 - proceedings.neurips.cc

… to attend to **all** positions in the decoder up to and including that position. **We need** to prevent

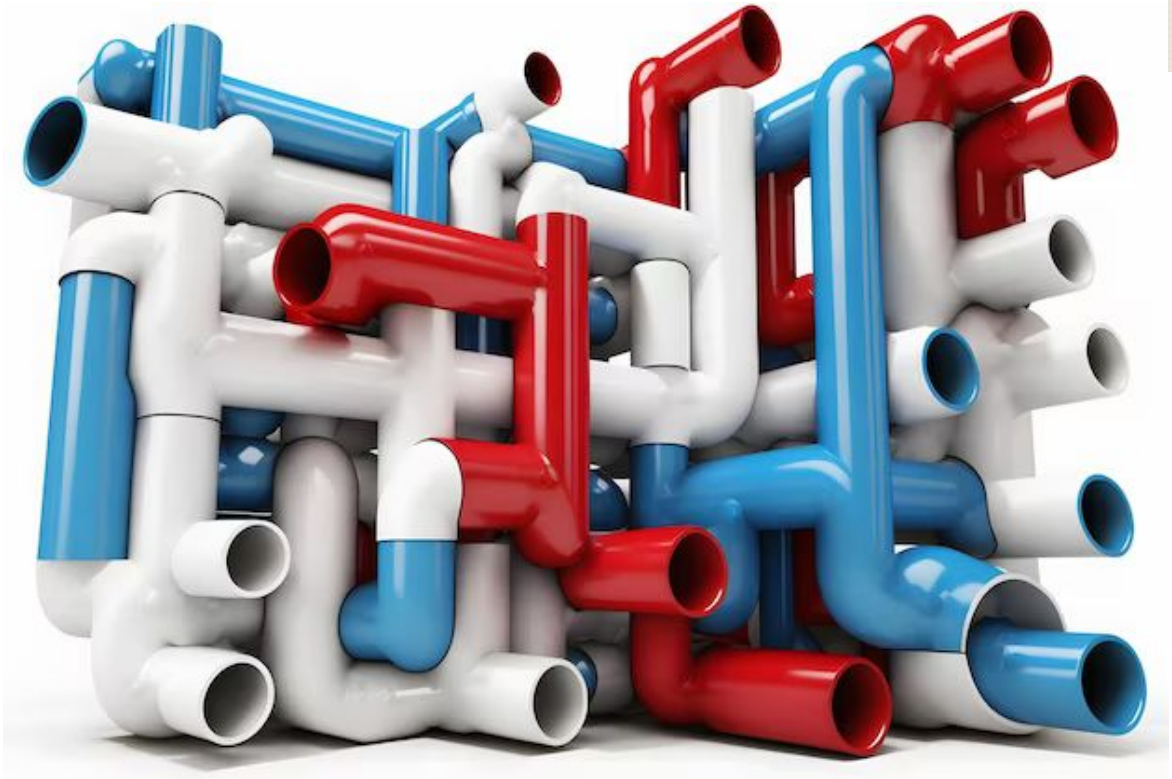
… **We** implement this inside of scaled dot-product **attention** by masking out (setting to −∞) …

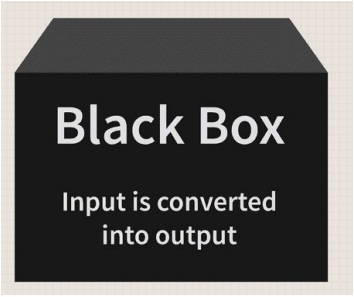
☆ Save 99 Cite Cited by 115250 Related articles All 87 versions »

These are the top cited papers (from a review in 2014)

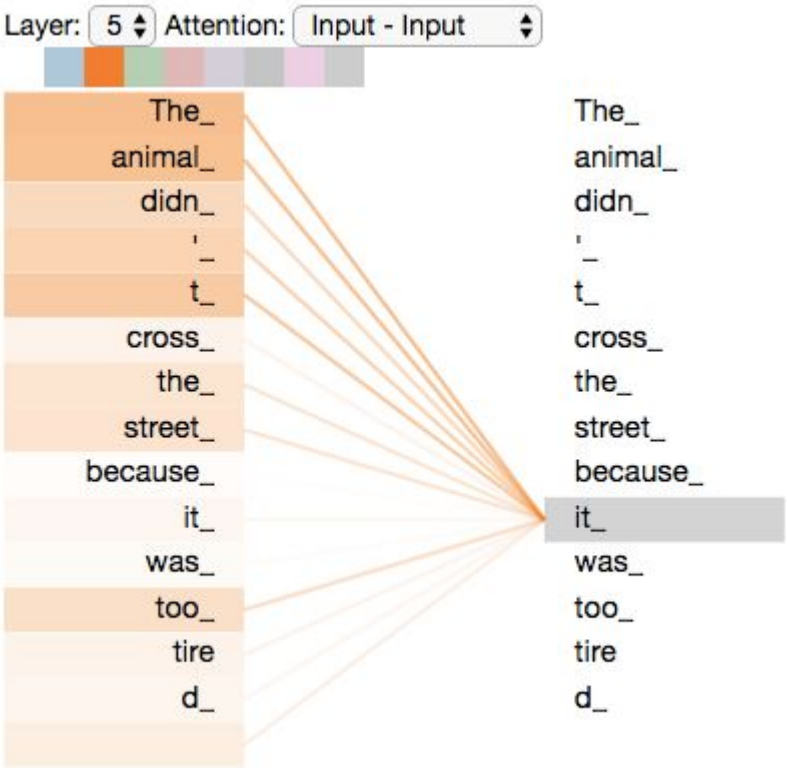
305,148	Protein measurement with the folin phenol reagent.	Biology lab technique	1951	*J. Biol. Chem.*
213,005	Cleavage of structural proteins during the assembly of the head of bacteriophage T4.	Biology lab technique	1970	*Nature*
155,530	A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding.	Biology lab technique	1976	*Anal. Biochem.*
65,335	DNA sequencing with chain-terminating inhibitors.	Biology lab technique	1977	*Proc. Natl Acad. Sci. USA*

How does attention work?

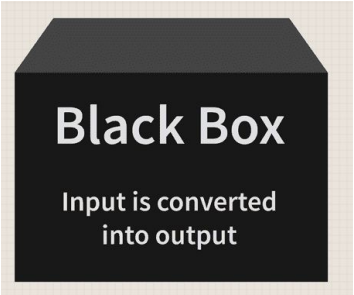


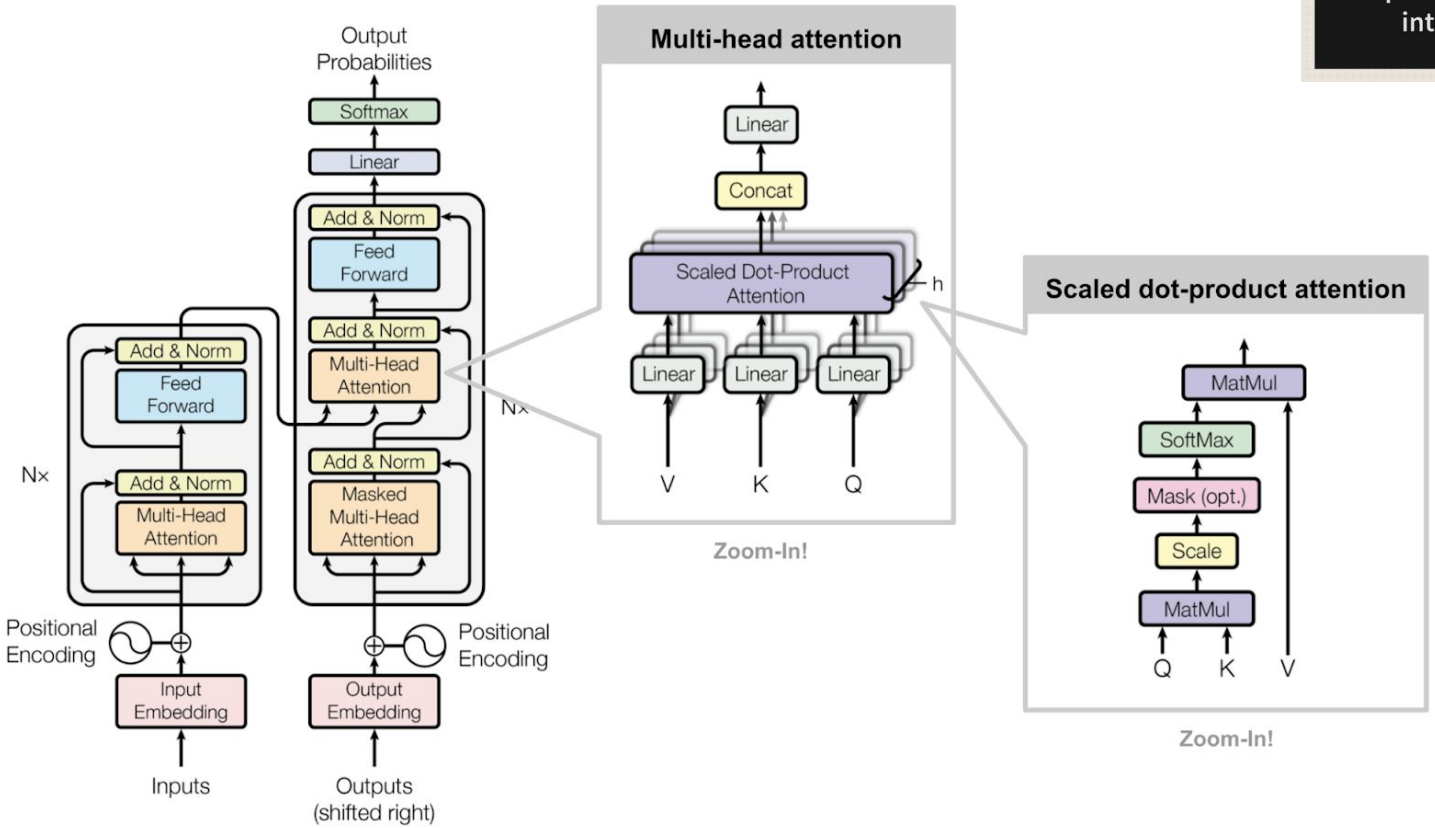


Because attention is numerical, we can visualize it!

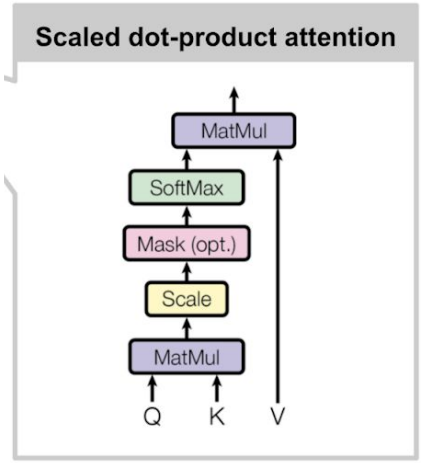


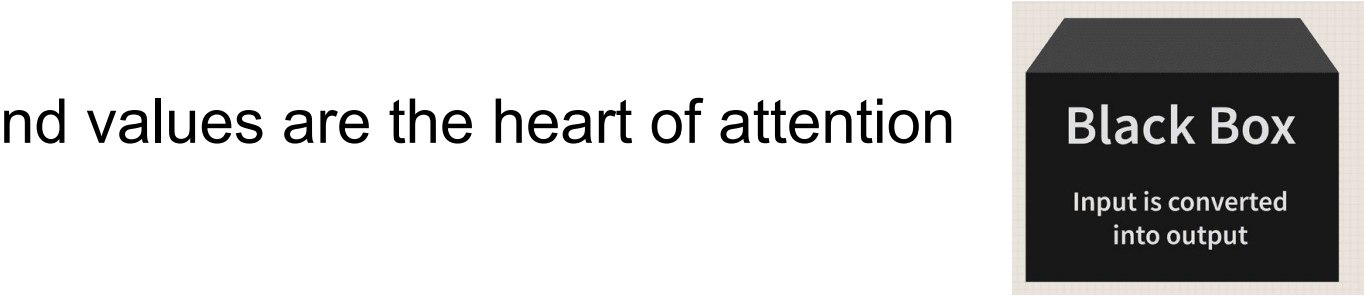
Dive into the model

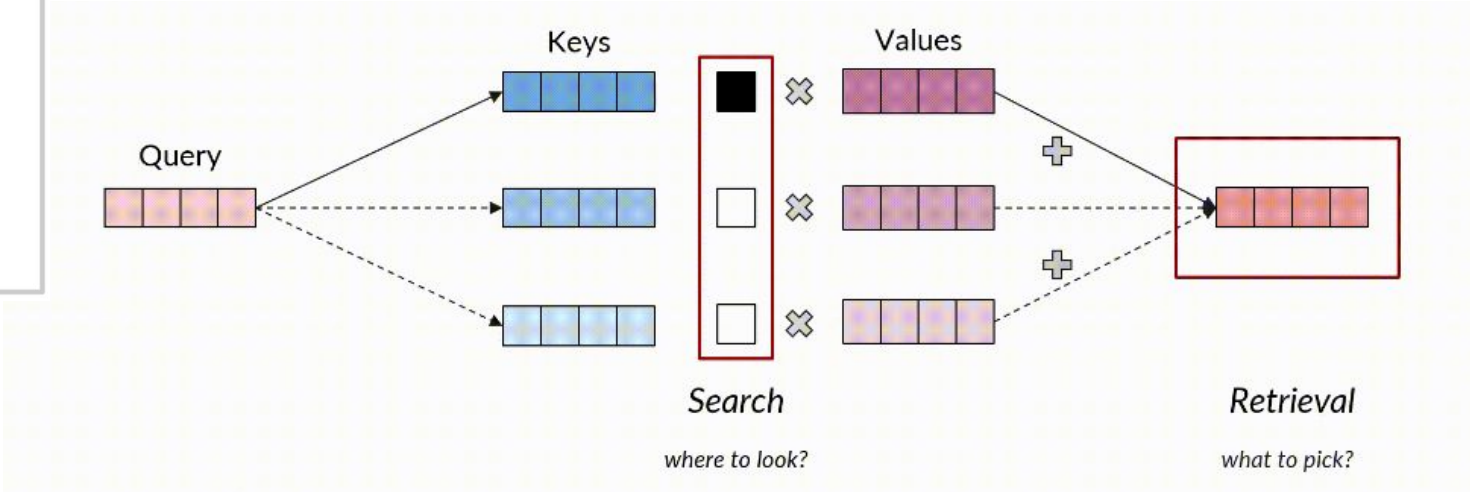




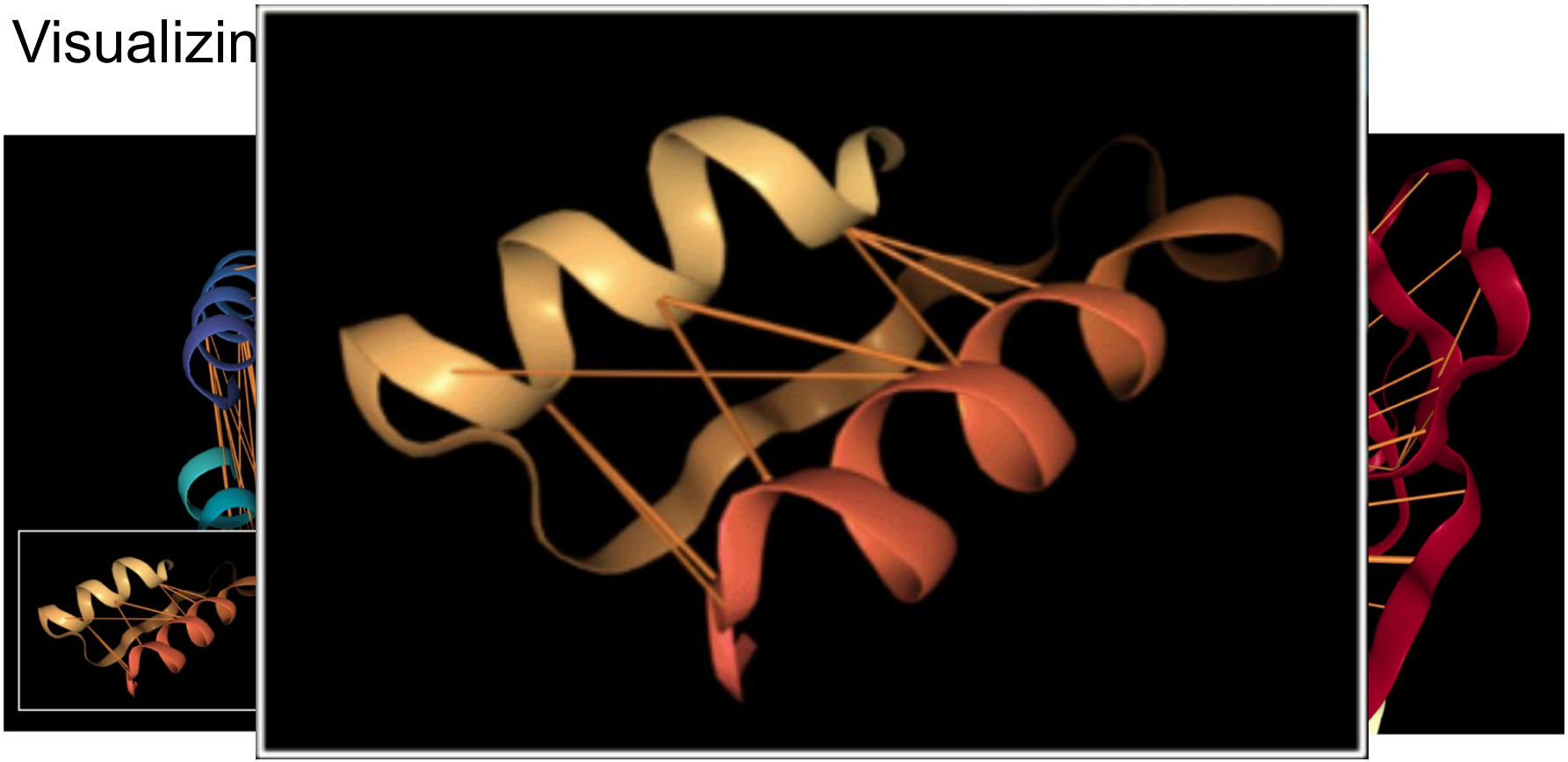
Queries, keys, and values are the heart of attention



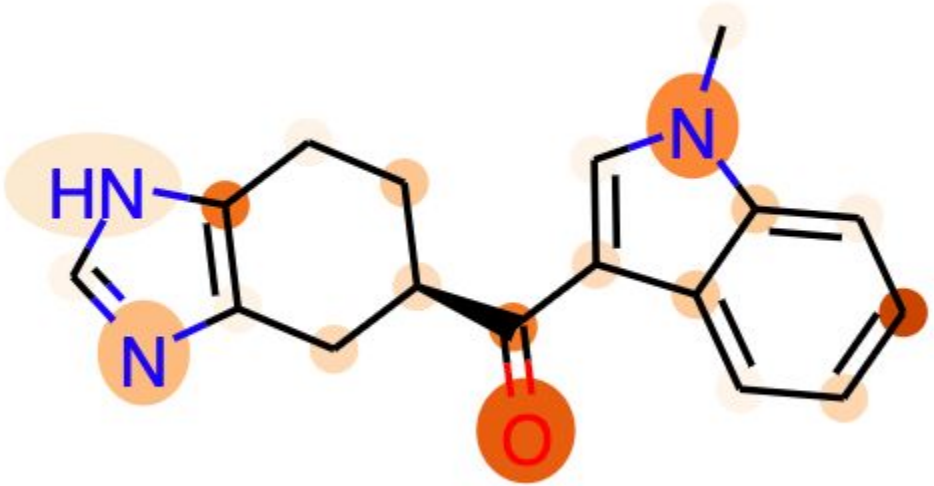




Visualizin

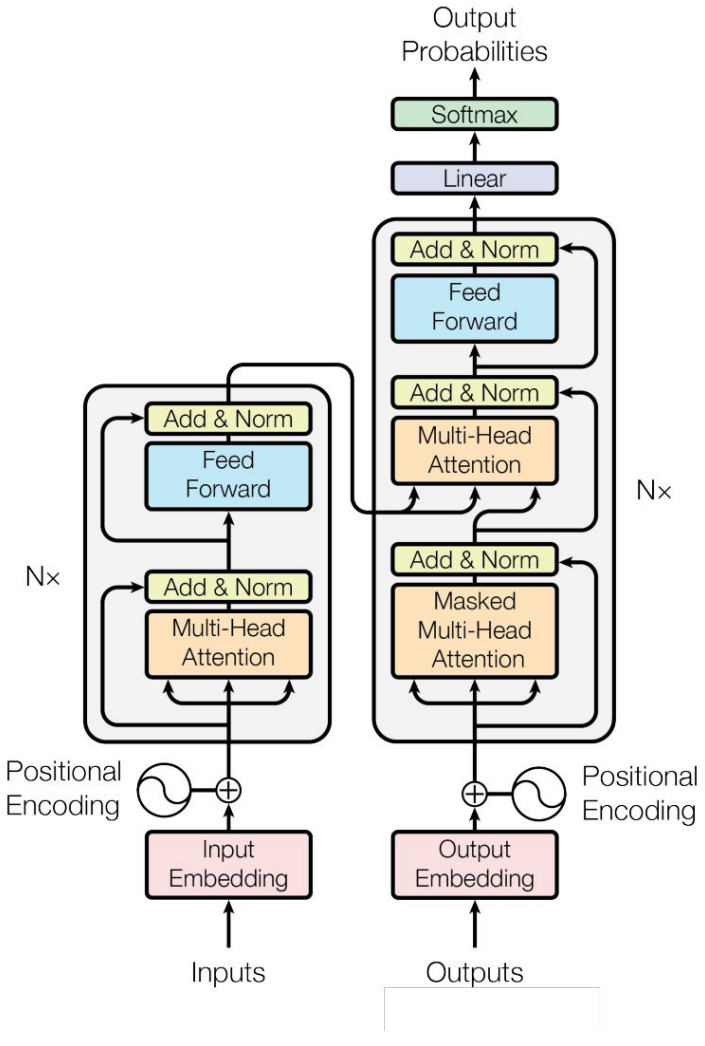


We can do the same thing for small molecules



BERT

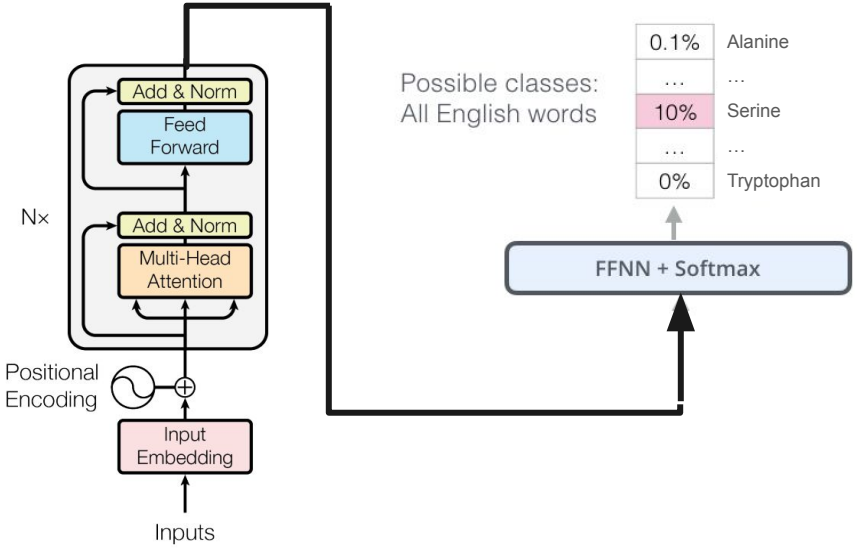
Encoder



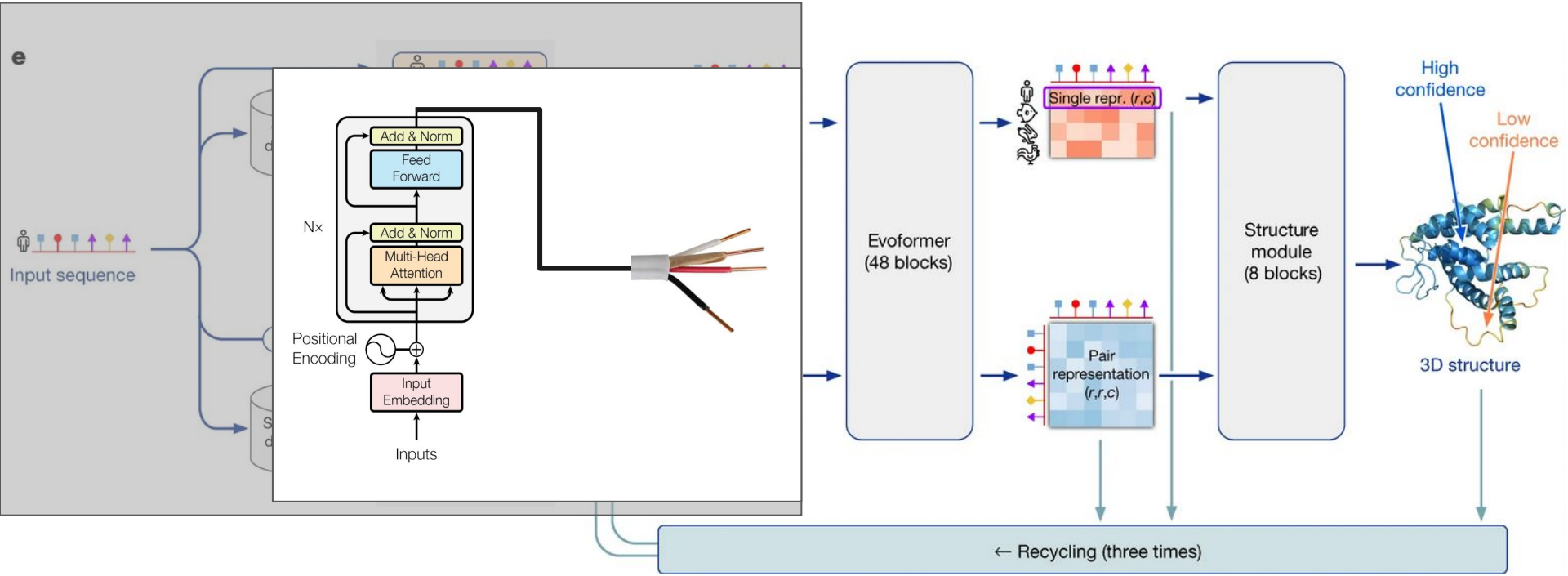
GPT

Decoder

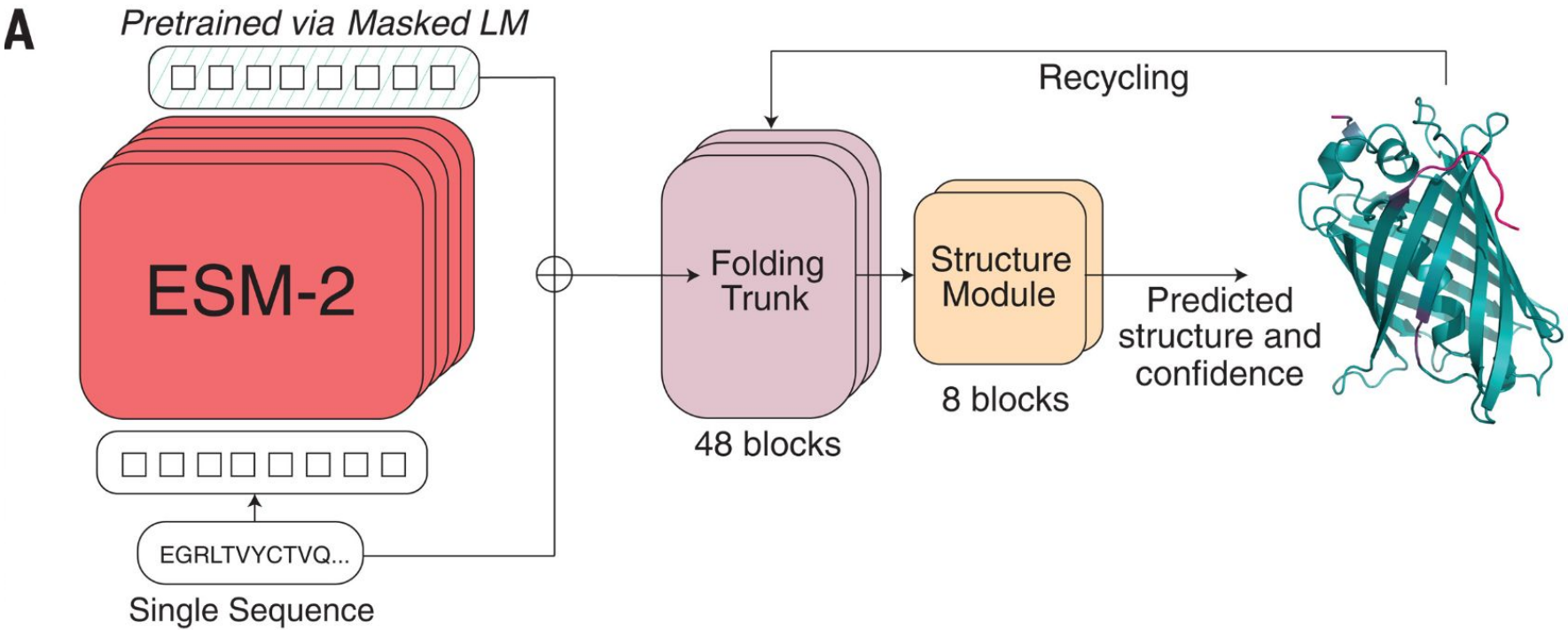
Meta trained a massive PLM called Evolutionary Scale Modeling (ESM)



Alphafold's Evoformer and Structure model

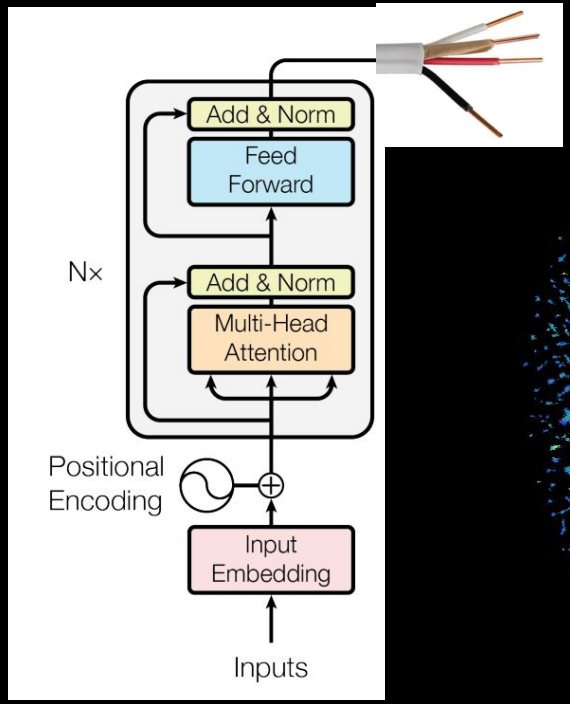


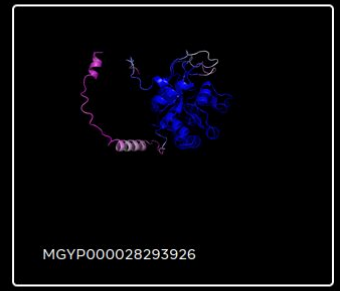
ESM-fold gets similar accuracy with a single sequence



ESM Metagenomic Atlas

Explore | Resources | About | Blog ↗



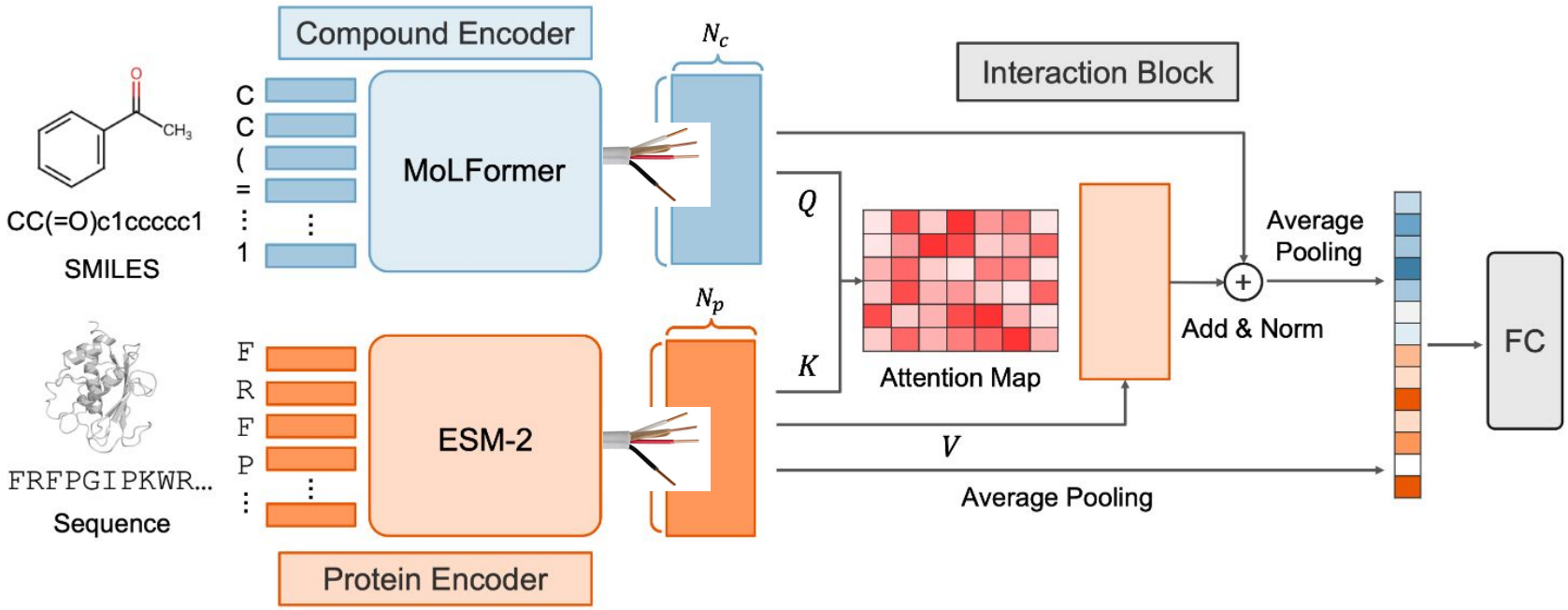


Drag to pan
Scroll to zoom in/out
Click to select a protein

Exploring 1 million out of 772M proteins
Unknown — Known

Meta AI

ChemGLaM: predicting protein-ligand interactions



Koyama, Takuto, et al. *bioRxiv* (2024): 2024-02.

ESM Atlas and ESM-fold API

https://esmatlas.com/

https://github.com/facebookresearch/esm

https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/ESMFold.ipynb

https://colab.research.google.com/drive/13DgnFzTUJWU8luyU0SKhSjlqjEqz1YO9?usp=sharing