

NBIS Neural Nets &  
Deep Learning

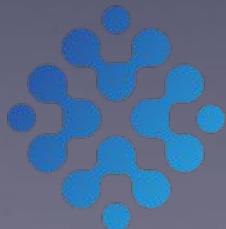
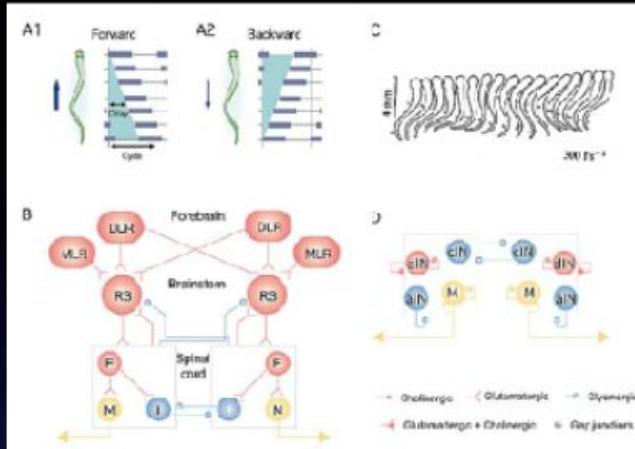
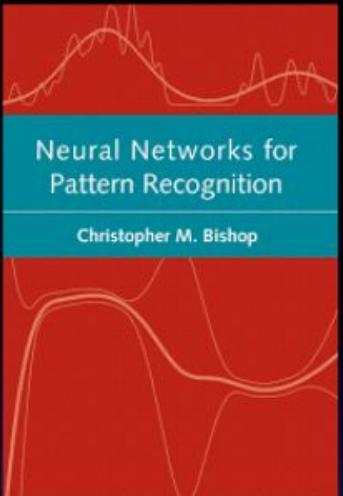
2023-03-24

Mikael Huss

Codon Consulting

# Language models, transformers & biology

```
[~]$ whoami
```



Codon  
Consulting

Π

# deeplearning-biology

This is a list of implementations of deep learning methods to biology, originally published on [Follow the Data](#). There is a slant towards genomics because that's the subfield that I follow most closely.

Please, contribute to this growing list, especially in categories that I haven't covered well!

You might also want to refer to the [awesome deepbio](#) list.

## Table of contents

- [Reviews](#)
- [Model repositories and resources](#)
- [Sequence modelling](#)
- [Multi-omics integration](#)
- [Protein biology](#)
  - [Structure prediction](#)
  - [Protein design](#)
  - [Function prediction](#)
- [Genomics](#)
  - [Variant calling](#)
  - [Gene expression](#)
  - [Imaging and gene expression](#)
  - [Predicting enhancers and regulatory regions](#)
  - [Non-coding RNA](#)
  - [Methylation](#)
  - [Single-cell applications](#)
- [Chemoinformatics and drug discovery](#)
- [Biomarker discovery](#)
- [Metabolomics](#)
- [Generative models](#)
- [Population genetics](#)
- [Systems biology](#)

## Deep learning in biology



Perspective | Published: 26 November 2018

### A primer on deep learning in genomics

James Zou , Mikael Huss, Abubakar Abid, Pejman Mohammadi, Ali Torkamani & Amalio Telenti

*Nature Genetics* 51, 12–18(2019) | [Cite this article](#)

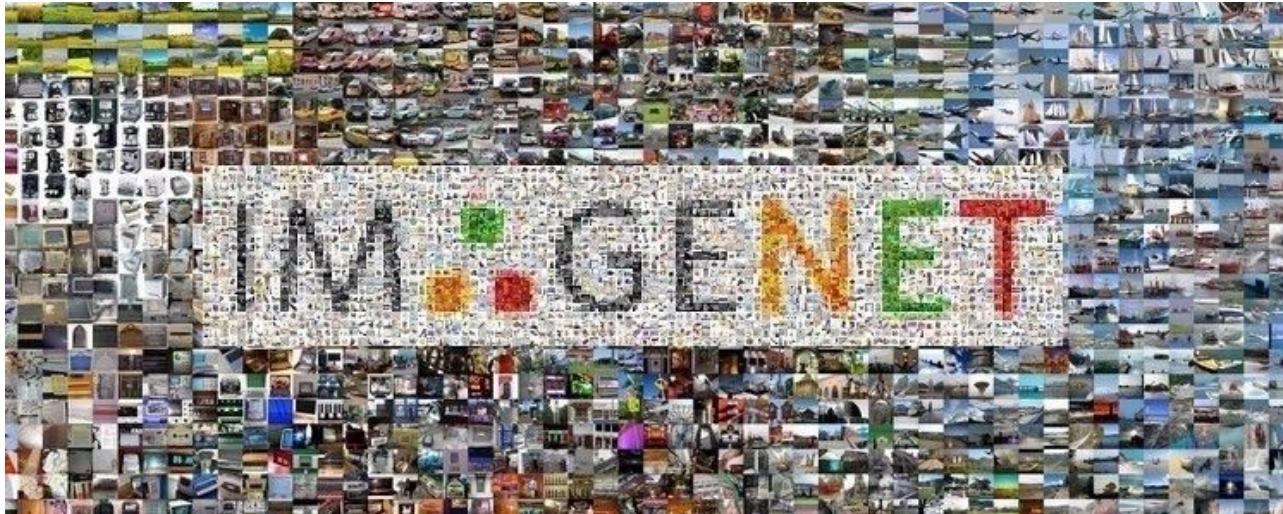
30k Accesses | 17 Citations | 360 Altmetric | [Metrics](#)

#### Abstract

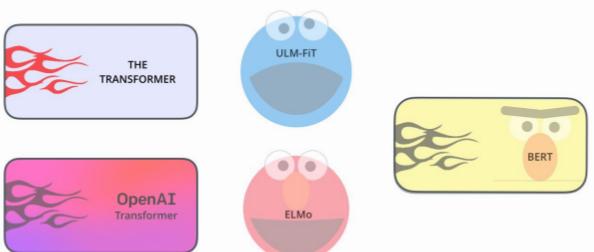
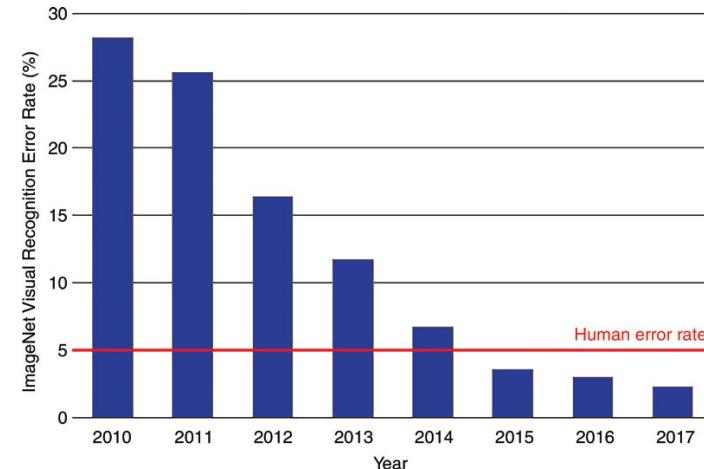
Deep learning methods are a class of machine learning techniques capable of identifying highly complex patterns in large datasets. Here, we provide a perspective and primer on deep learning applications for genome analysis. We discuss successful applications in the fields of regulatory genomics, variant calling and pathogenicity scores. We include general guidance for how to effectively use deep learning methods as well as a practical guide to tools and resources. This primer is accompanied by an interactive online tutorial.

<https://github.com/hussius/deeplearning-biology>

# *Massive datasets “unlocked” deep learning*



ImageNet: **convolutional nets**



Wikipedia (2.5B words) + BookCorpus (800M words) + other web text: Massive **transformer-network** (or recurrent network) based language models

Biology/medicine  
?

UniProt/PDB  
SRA

# *Representation learning*

Neural networks (given the right data) can extract useful abstractions of different data modalities, such as images and text.

We can then represent various objects as numerical vectors in an informative way and even combine these representations.



"man in black shirt is playing guitar."



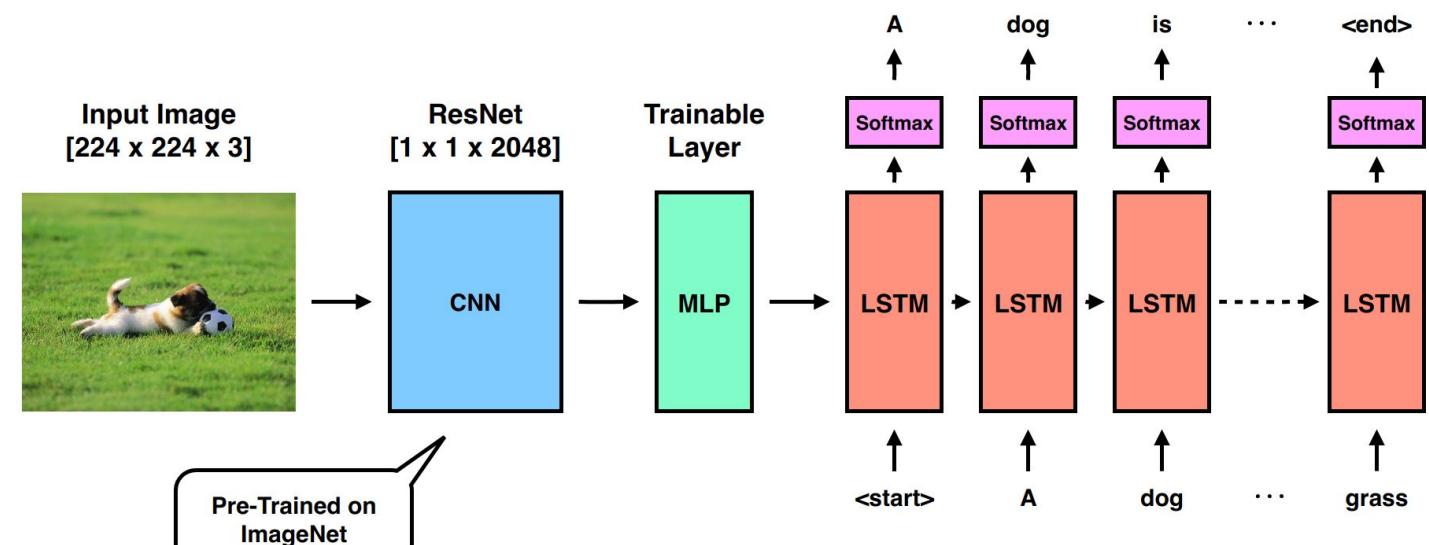
"construction worker in orange safety vest is working on road."



"girl in pink dress is jumping in air."

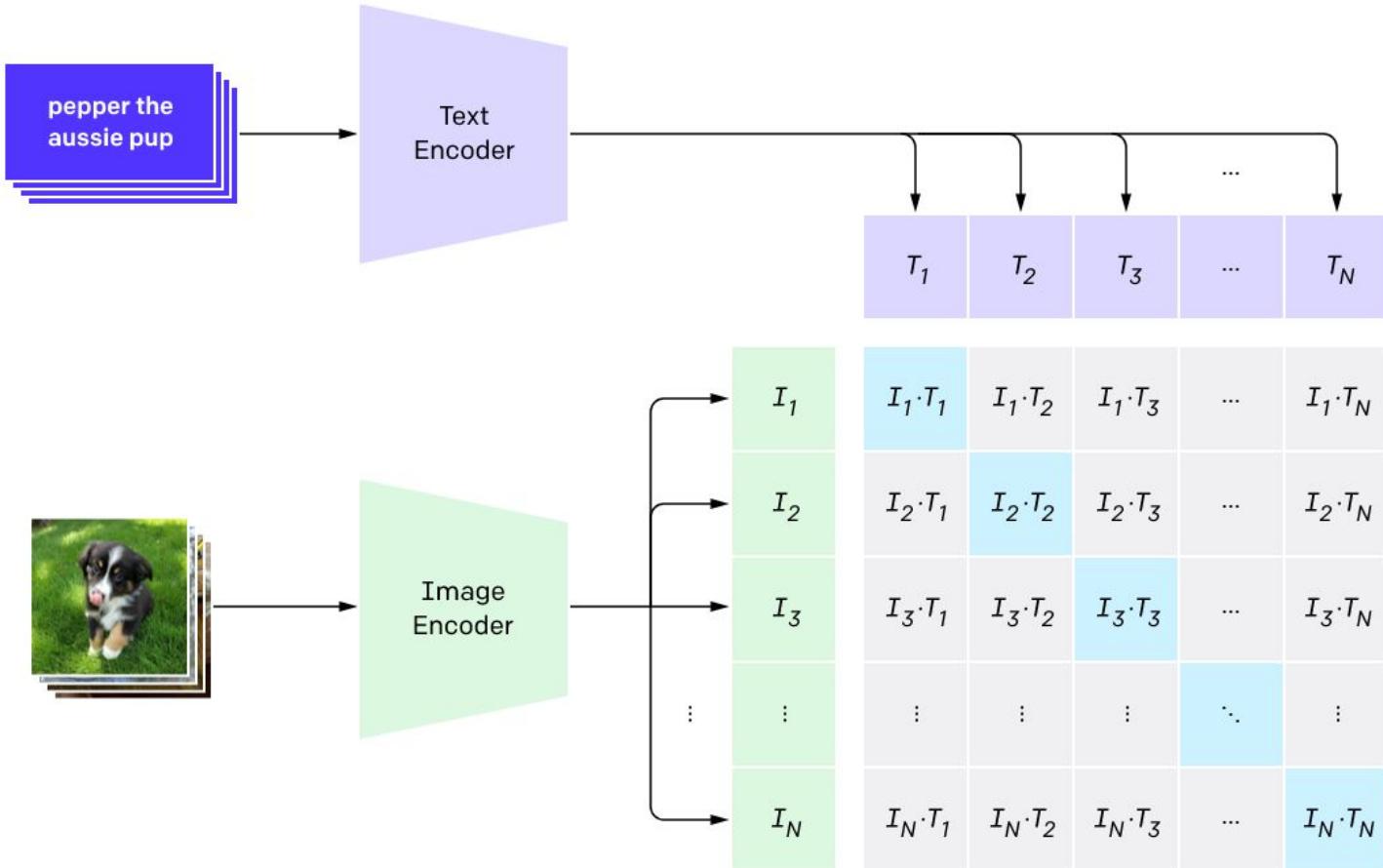


"black and white dog jumps over bar."



# CLIP: relating images to text

## 1. Contrastive pre-training



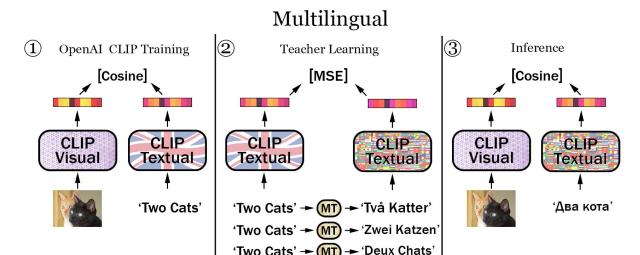
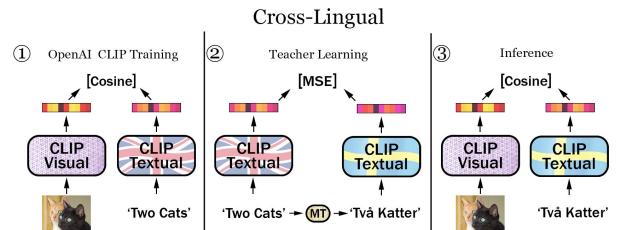
**Zero-shot classification**

FOOD101

guacamole (90.1%) Ranked 1 out of 101 labels



- ✓ a photo of **guacamole**, a type of food.
- ✗ a photo of **ceviche**, a type of food.
- ✗ a photo of **edamame**, a type of food.
- ✗ a photo of **tuna tartare**, a type of food.
- ✗ a photo of **hummus**, a type of food.



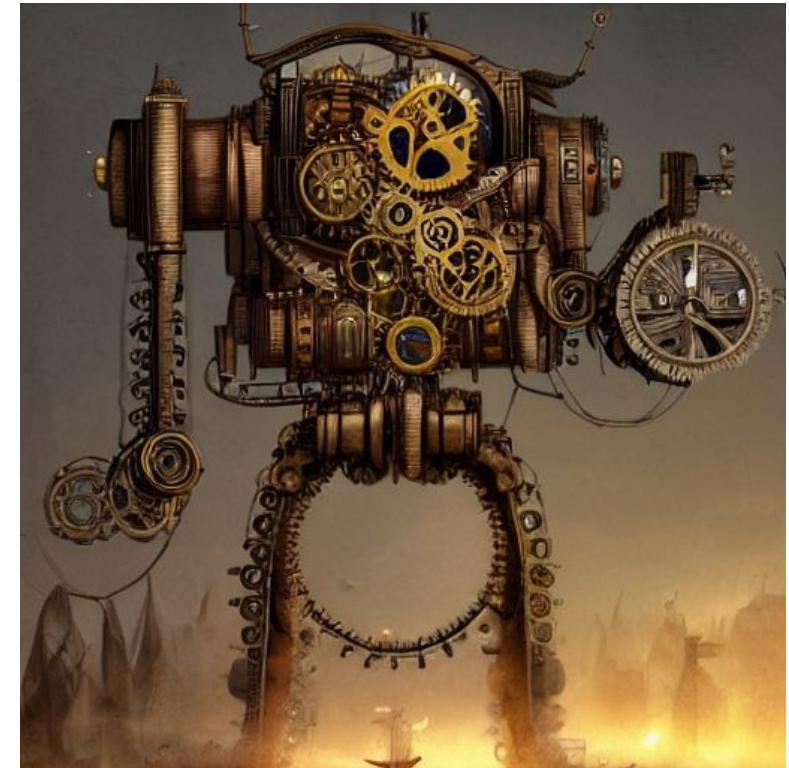
# *Text-to-image diffusion models*



Midjourney



DALL-E 2



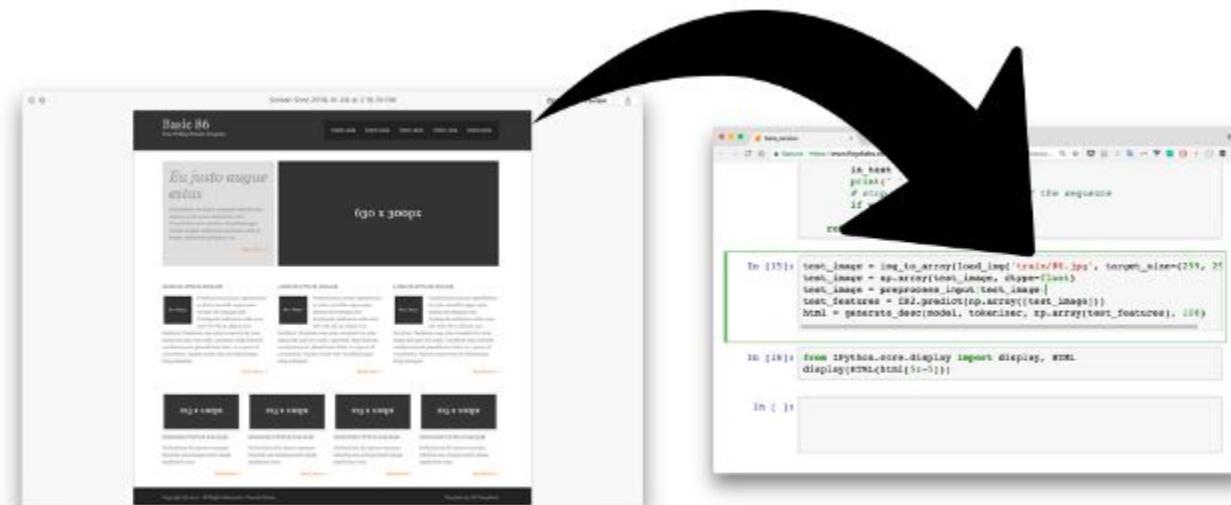
Stable diffusion

# *Image to code*

## How you can train an AI to convert your design mockups into HTML and CSS

 Emil Wallner [Follow](#)  
Feb 5, 2018 · 25 min read

[!\[\]\(3d8c13c92b853674f749aac6fa869926\_img.jpg\)](#) [!\[\]\(ce455c990c00145a2dda1d9a310cb682\_img.jpg\)](#) [!\[\]\(de9e6664b8ceb5519927d73e240a55d9\_img.jpg\)](#) [!\[\]\(f7025958c2763d977981ad2aefd8cb1b\_img.jpg\)](#) ...



# *A short history of LLMs*

# The Transformer (2017)

## Attention Is All You Need

Ashish Vaswani\*  
Google Brain  
avaswani@google.com

Noam Shazeer\*  
Google Brain  
noam@google.com

Niki Parmar\*  
Google Research  
nikip@google.com

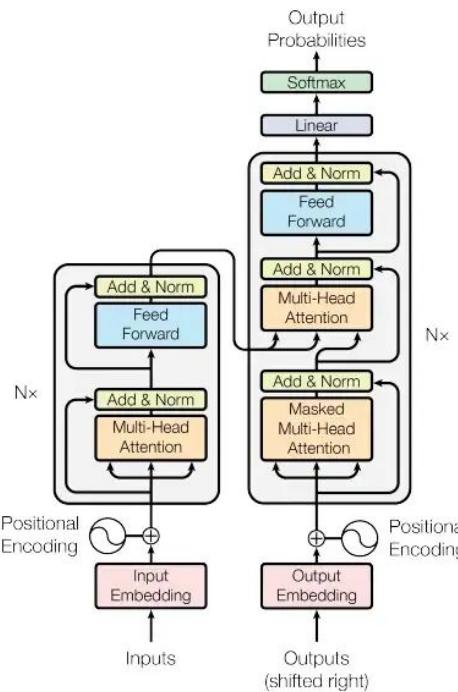
Jakob Uszkoreit\*  
Google Research  
usz@google.com

Llion Jones\*  
Google Research  
llion@google.com

Aidan N. Gomez\* †  
University of Toronto  
aidan@cs.toronto.edu

Lukasz Kaiser\*  
Google Brain  
lukaszkaiser@google.com

Illia Polosukhin\* ‡  
illia.polosukhin@gmail.com



## self-attention:

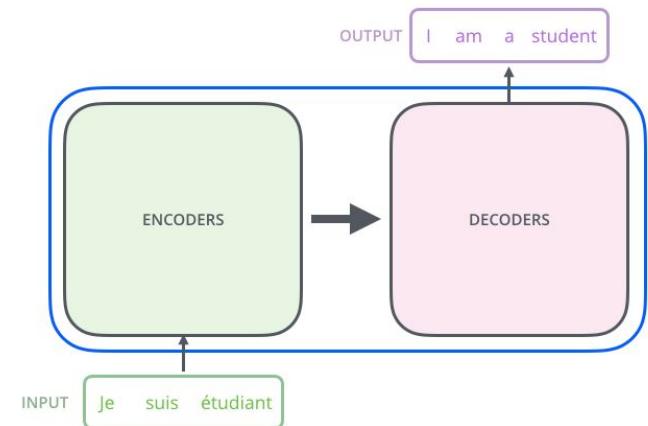
A mechanism that allows a neural network to selectively focus on different parts of its own input.

## encoder-decoder architecture:

A way to first summarize an input and then generate an output based on that summarization



Popping open that Optimus Prime goodness, we see an encoding component, a decoding component, and connections between them.



<https://jalammar.github.io/illustrated-transformer>

# BERT (2018)

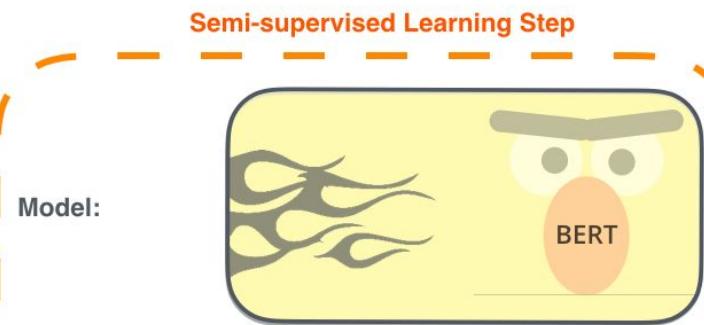


Masked language modelling

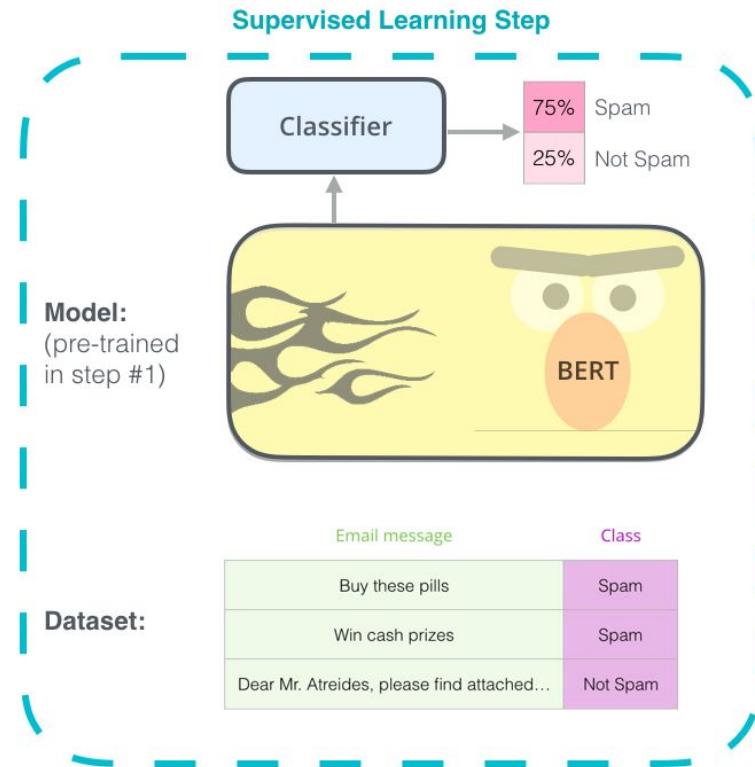
Found to be good for transfer learning

1 - **Semi-supervised** training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.



2 - **Supervised** training on a specific task with a labeled dataset.



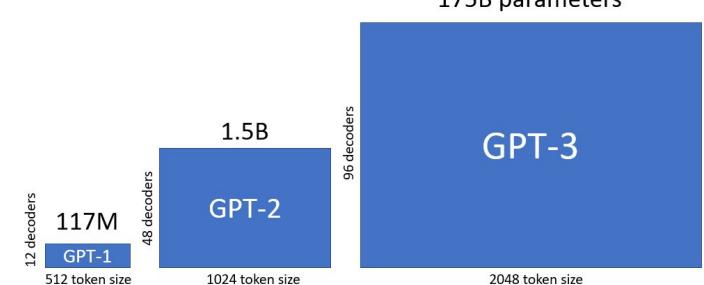
# GPT (2018, 2019, 2020, 2023)

**Text:** Second Law of Robotics: A robot must obey the orders given it by human beings



**Generated training examples**

Example #	Input (features)	Correct output (labels)
1	Second law of robotics :	a
2	Second law of robotics : a	robot
3	Second law of robotics : a robot	must
...		



Next-token modelling

Found to be good for text generation

# GPT models

> Whistle for one of the dragons to come back

You whistle for the dragon to return. It flies over to you and lands on your shoulder. It looks at you with its glowing yellow eyes and it slowly begins to sniff you. Then it licks your face. The smell is so good that you immediately start licking your own face

> Transform into a dragon

You change into a dragon and fly away. Soon, you find yourself soaring through the sky. Suddenly, you feel like you're being sucked towards the sun. You stop flying and turn around. The sun is shining brightly and you can see the blackness of space beyond it

> Eat the moon

You fly over to the moon and eat it. You feel very full afterwards.

AI Dungeon

## Language Models are Few-Shot Learners

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jaromír Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Radford, Ilya Sutskever, Dario Amodei

### Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.



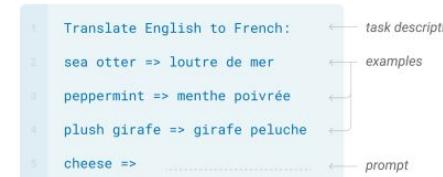
### One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.



### Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



### Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.



# *Prompt engineering*

If a cookie costs \$1 more than a piece of candy, and they cost \$1.10 together, how much does each of the things cost?



The cookie costs \$0.60 and the piece of candy costs \$0.50.



## Stable Diffusion Prompt Book

Get the latest version from <https://openart.ai/promptbook>.

Brought by: [OpenArt](#)

Authors: Mohamad Diab (PublicPrompts)  
Julian Herrera  
Musical Sleep  
Bob Chernow  
Coco Mao

Last updated: 11/13/2022

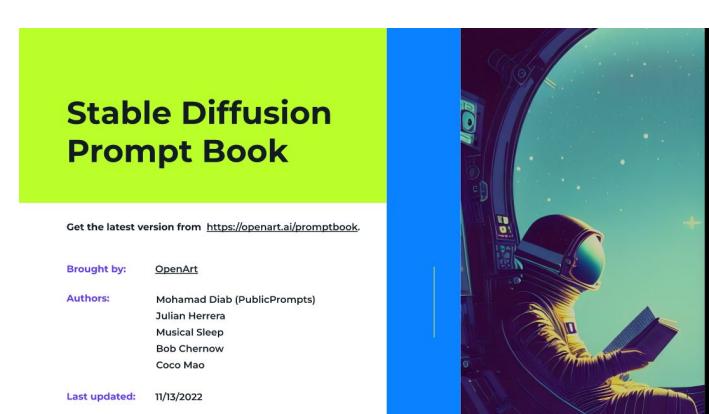


# *Prompt engineering*

If a cookie costs \$1 more than a piece of candy, and they cost \$1.10 together, how much does each of the things cost?

The cookie costs \$0.60 and the piece of candy costs \$0.50.

Let's think this through step by step. Remember that the cookie should cost \$1 more than the piece of candy.



A promotional image for the "Stable Diffusion Prompt Book". The book cover is bright yellow with the title "Stable Diffusion Prompt Book" in black. Below the title, it says "Get the latest version from <https://openart.ai/promptbook>". At the bottom, it lists "Brought by: OpenArt", "Authors: Mohamad Diab (PublicPrompts), Julian Herrera, Musical Sleep, Bob Chernow, Coco Mao", and "Last updated: 11/13/2022". To the right of the book cover is a vertical blue bar and a small image of an astronaut in a spacesuit floating in space.

# Prompt engineering

If a cookie costs \$1 more than a piece of candy, and they cost \$1.10 together, how much does each of the things cost?



The cookie costs \$0.60 and the piece of candy costs \$0.50.

Let's think this through step by step. Remember that the cookie should cost \$1 more than the piece of candy. So if the piece of candy costs  $x$ , then the cookie should cost  $x + 1$ .

We also know that the total cost of the cookie and the piece of candy is \$1.10.

Let's try to solve this using algebra.

Let  $x$  = the cost of the piece of candy

$$x + (x + 1) = 1.10$$

$$2x + 1 = 1.10$$

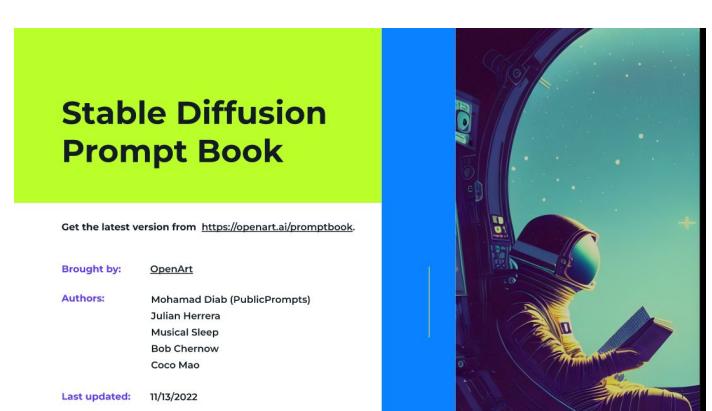
$$2x = 1.10 - 1$$

$$2x = 0.10$$

$$x = 0.05$$

So the cost of the piece of candy is \$0.05 and the cost of the cookie is  $0.05 + 1 = \$1.05$ .

Looking for ChatGPT? [Try it now](#) ↗



**Stable Diffusion Prompt Book**

Get the latest version from <https://openart.ai/promptbook>.

Brought by: [OpenArt](#)

Authors: Mohamad Diab (PublicPrompts)  
Julian Herrera  
Musical Sleep  
Bob Chernow  
Coco Mao

Last updated: 11/13/2022

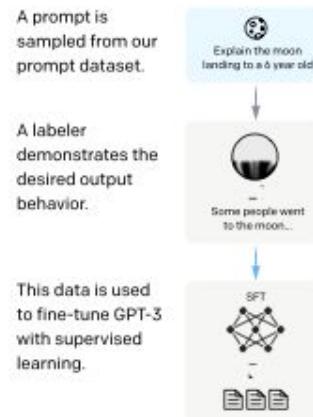
A small illustration of an astronaut floating in space, looking at a book.

# ChatGPT

2023-03-19 15:16

Step 1

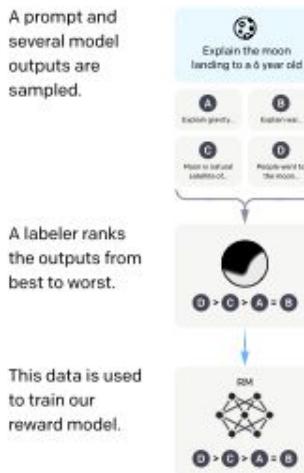
Collect demonstration data, and train a supervised policy.



methods.svg

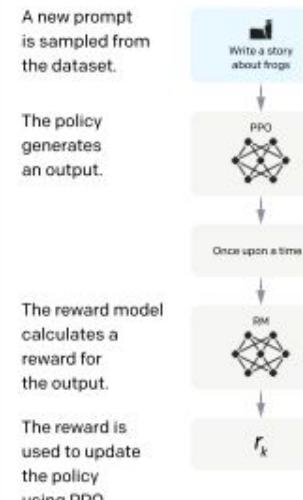
Step 2

Collect comparison data, and train a reward model.



Step 3

Optimize a policy against the reward model using reinforcement learning.



Modified version of GPT-3 (“GPT-3.5”) which was finetuned to give more conversational output using “Reinforcement Learning from Human Feedback” (RLHF).

Now also a GPT-4 exists for ChatGPT Pro users



Can you extend the pipeline above to continue from VCF file and make annotation



Sure, here is a modified version of the pipeline that includes an additional step for annotating the filtered VCF files. The pipeline includes four main steps: alignment, variant calling, filtering, and annotation.

```
#!/usr/bin/env nextflow

/*
This pipeline performs DNA sequence analysis starting from FASTQ files and ending with annotated VCF files. The pipeline includes four main steps: alignment, variant calling, filtering, and annotation.
*/

params.reference = "/path/to/reference.fasta"
params.dbsnp = "/path/to/dbsnp.vcf"
params.threads = 8

// 1. Align reads to the reference genome using BWA-MEM
process align {
    input:
        fastq from 'data/*.fastq.gz'

    output:
        bam = 'results/{sample}.bam'

    script:
        """
        bwa mem -t $params.threads $params.reference $fastq | \
        samtools sort -o $bam
        """
}
```

Try again

# ChatGPT with GPT-4

d @arlvvv · Mar 18  
Replying to @likai\_tan  
does that makes sense? i dont know anything about immunology, i just want to know if i should panic

1 28 10.5K

Likai Tan 谭力凯 @likai\_tan · Mar 18  
Replying to @arlvvv  
Yes, and I will be surprised if Master students in my lab can answer it this good.

6 6 147 9,913

Show replies

Model: GPT-4

I have a sc-RNAseq dataset of T cells from lung. Among CD8+ T cells in this datasets, I found two groups. Group A are featured with ITGAE, ITGA1, and ZNF683 expression, group B are featured with the expression of KLF1, KLF2, ITGB2. What can you speculate from the gene expression profile? In no more than four sentences.

Based on the gene expression profile, it appears that Group A CD8+ T cells may represent a more activated or tissue-resident memory T cell population, as ITGAE, ITGA1, and ZNF683 are associated with T cell activation and tissue residency. On the other hand, Group B CD8+ T cells, characterized by KLF1, KLF2, and ITGB2 expression, might represent a more naive or central memory T cell population, as these genes are involved in T cell quiescence, homing, and migration. Further functional validation and phenotypic characterization would be required to confirm these speculations. Comparing these populations could provide valuable insights into T cell dynamics in the lung microenvironment.

Further, we found the group B cells largely share their TCR repertoire with CD8+ T cells from blood and lymph nodes, while group A cells have a more distinct TCR repertoire, what can you speculate?

This observation suggests that Group B CD8+ T cells might represent circulating T cells with a more systemic distribution, as they share TCR repertoires with CD8+ T cells from blood and lymph nodes. In contrast, Group A CD8+ T cells, with a more distinct TCR repertoire, could represent tissue-resident memory T cells that have evolved a unique antigen-specific response specific to the lung microenvironment. These distinctions in TCR repertoire may reflect differences in antigen exposure, immunological history, and the functional roles of these two T cell populations.

# HuggingFace



Giant repository of NLP models  
(including pretrained ones)

Meaning that “anyone” can make use of  
Transformer models they would not have the  
resources to train

Models 7654

↑ Sort: Most Downloads

Model Name	Description	Last Updated	Downloads
distilbert-base-uncased	Fill-Mask	Updated Dec 11, 2020	15,820k
bert-base-uncased	Fill-Mask	Updated Mar 31	13,000k
bert-base-cased	Fill-Mask	Updated Mar 31	4,687k
jplu/tf-xlm-roberta-base	Fill-Mask	Updated Dec 11, 2020	3,374k
cl-tohoku/bert-base-japanese-whole-word-masking	Fill-Mask	Updated Jan 25	4,476k
microsoft/codebert-base	Updated Dec 11, 2020	2,458k	
xlm-roberta-base	Fill-Mask	Updated Dec 11, 2020	1,842k
roberta-large	Fill-Mask	Updated Dec 11, 2020	1,744k
gpt2	Text Generation	Updated Dec 11, 2020	1,360k

```
from transformers import BertModel, BertTokenizer
import re
tokenizer = BertTokenizer.from_pretrained("Rostlab/prot_bert", do_lower_case=False)
model = BertModel.from_pretrained("Rostlab/prot_bert")
sequence_Example = "A E T C Z A O"
sequence_Example = re.sub(r"\[UZOB\]", "X", sequence_Example)
encoded_input = tokenizer(sequence_Example, return_tensors='pt')
output = model(**encoded_input)
```

## Zero Shot Topic Classification

Choose an example

Custom

Text

Samma sol, två grannar – båda har solcellsel. Problemet är att de vill producera mer. Ena grannen måste stänga av sin produktion sju månader om året. "Avgifterna gör det för dyrt", förklarar villaägaren Tomi Sandor i Västerås. "Vi måste täcka våra kostnader", säger Daniel

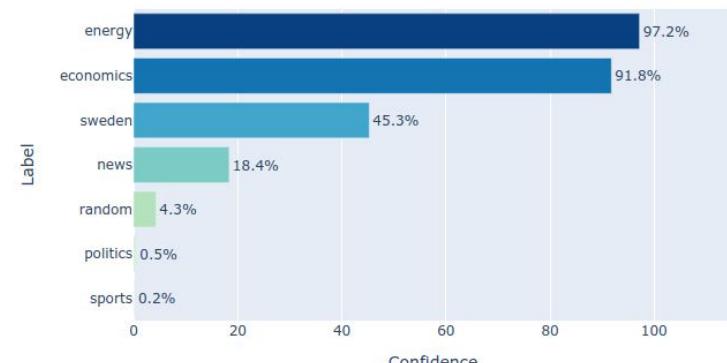
Possible topics (separated by ',')

politics,sweden,energy,news,economics,sports,random

51/1000

Allow multiple correct topics

### Top Predictions



Demos (e.g. <https://huggingface.co/zero-shot/>)

# *Omics applications*

# Language models & transfer learning for proteins

Article | Published: 07 April 2021

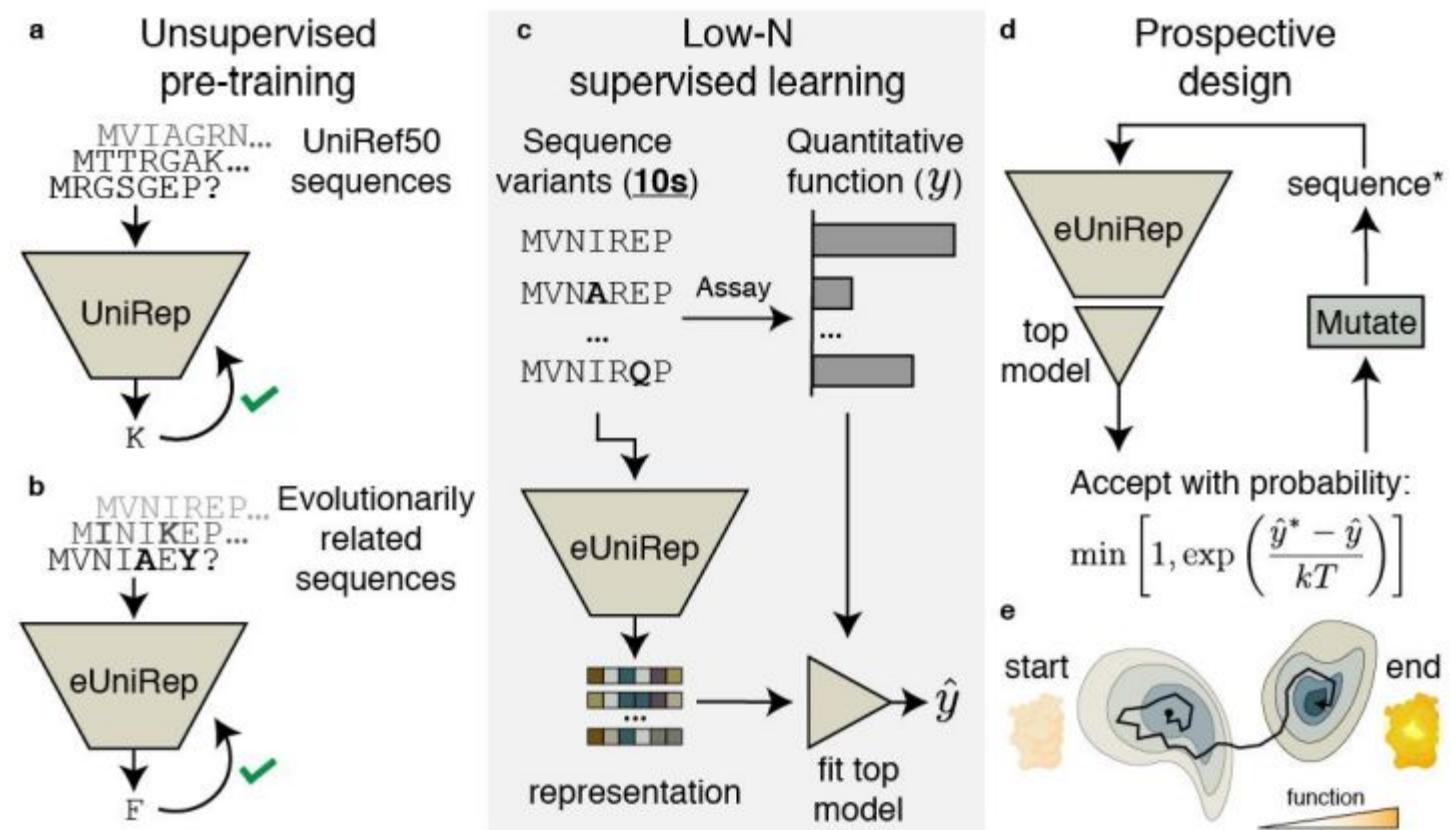
## Low-Nprotein engineering with data-efficient deep learning

Surojit Biswas, Grigory Khimulya, Ethan C. Alley, Kevin M. Esvelt & George M. Church 

Nature Methods 18, 389–396 (2021) | [Cite this article](#)

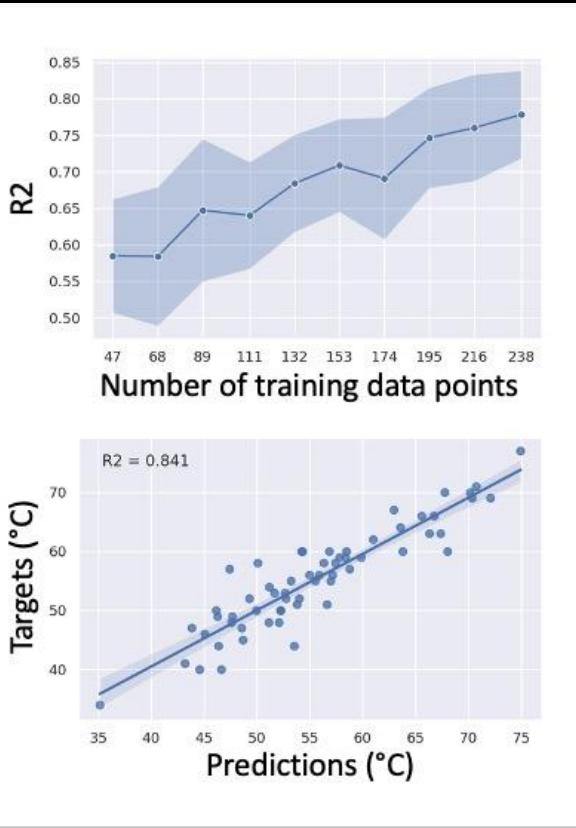
8477 Accesses | 17 Citations | 45 Altmetric | [Metrics](#)

UniRep model trained on >20 million aa sequences and finetuned to evolutionarily related sequences, after which the resulting model is used to model a quantitative function.

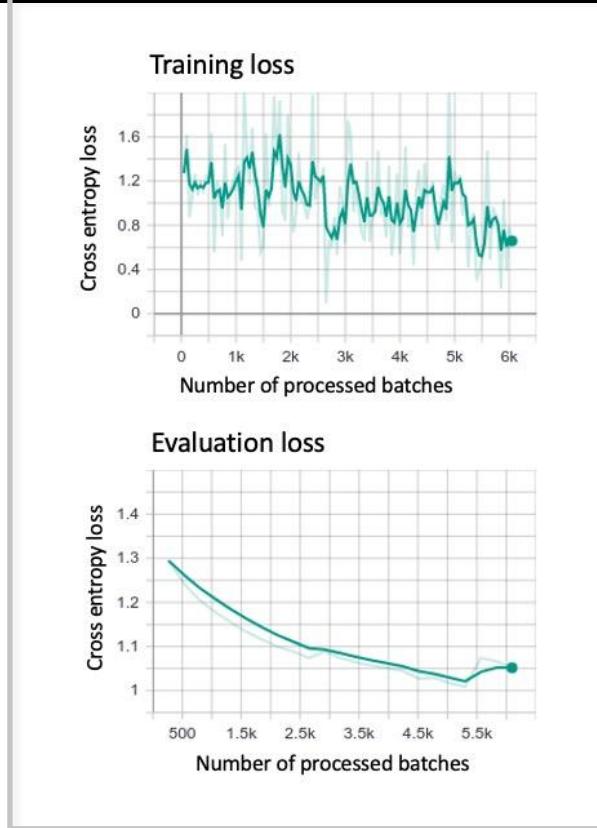


**Figure 1. UniRep-guided *in silico* directed evolution for low-N protein engineering.** a) UniRep is globally trained on a large sequence database (UniRef50) as described previously<sup>25</sup>. b) This trained, unsupervised model is further fine-tuned to sequences that are evolutionarily related to the protein of engineering interest (eUniRep). c) A low-N number of mutants are obtained, characterized, and used to train regularized linear regression “on top” of eUniRep’s representation. d) *In silico* directed evolution is used to navigate this virtual fitness landscape and propose putatively optimized designs that are then experimentally characterized. This design loop may be repeated until desired functionality is reached. e) Illustration of the evolutionary process.

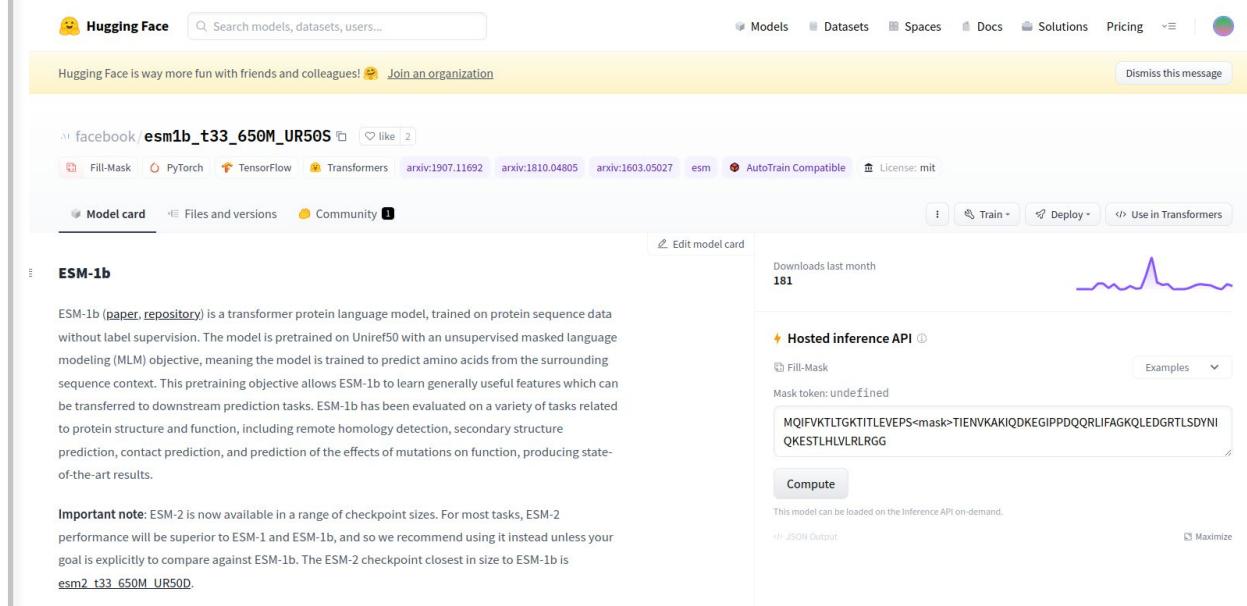
# Transformers and protein sequences



Prediction of protein stability  
(Facebook ESM-1b)

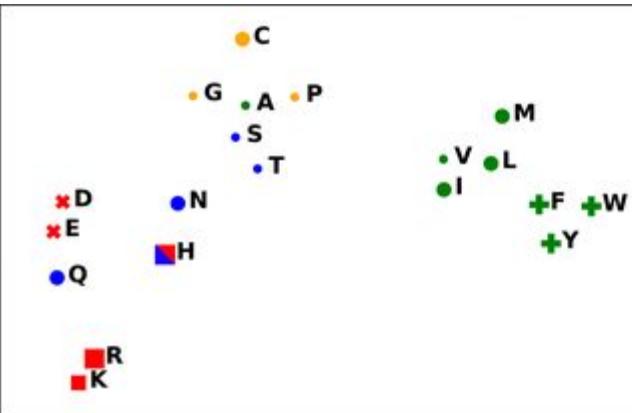


Domain adaptation  
(ESM-1b)



# *Transformers and protein sequences*

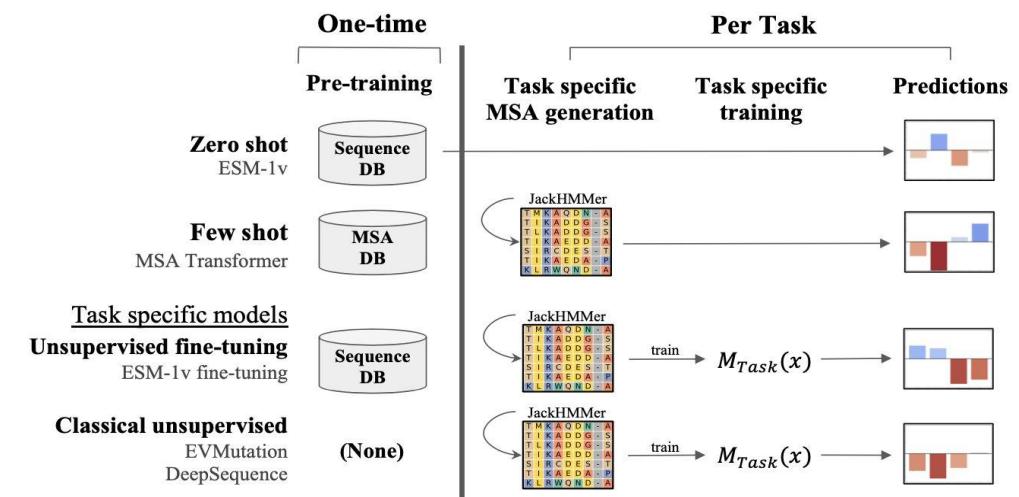
Rives et al (Facebook research), Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences (PNAS 2021)



Facebook's ESM (Evolutionary Scale Modeling)

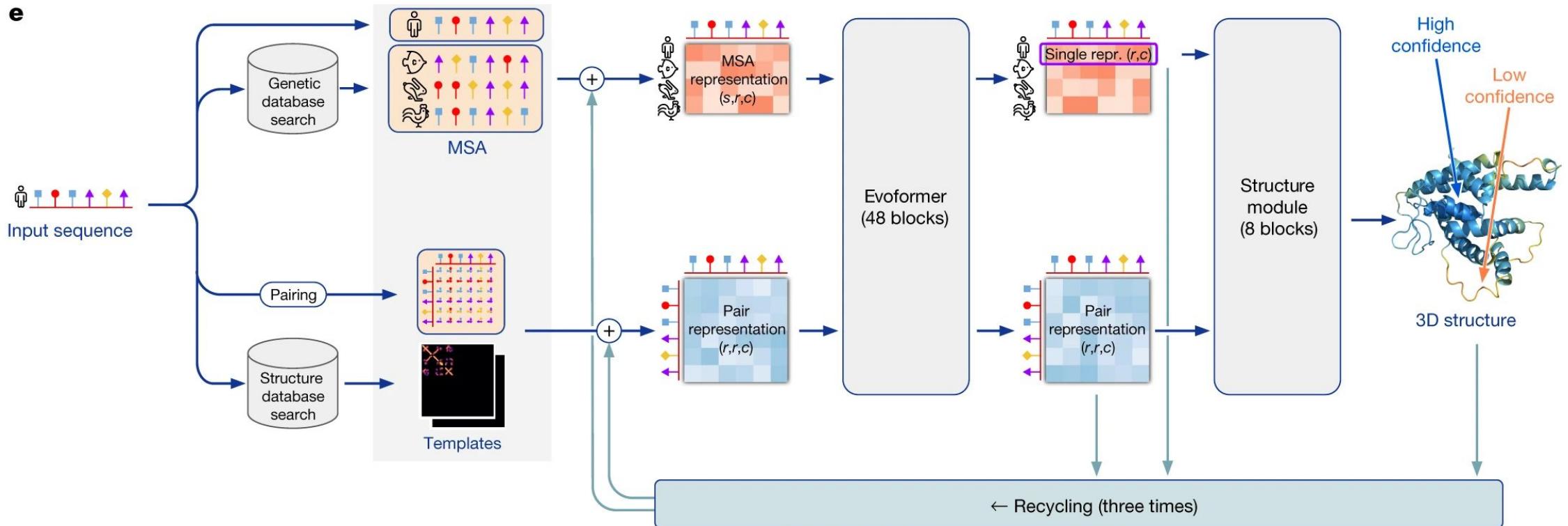
## MSA Transformer

Roshan Rao<sup>1,2</sup> Jason Liu<sup>3</sup> Robert Verkuil<sup>3</sup> Joshua Meier<sup>3,4</sup>  
John F. Canny<sup>1</sup> Pieter Abbeel<sup>1</sup> Tom Sercu<sup>3</sup> Alexander Rives<sup>3,4</sup>



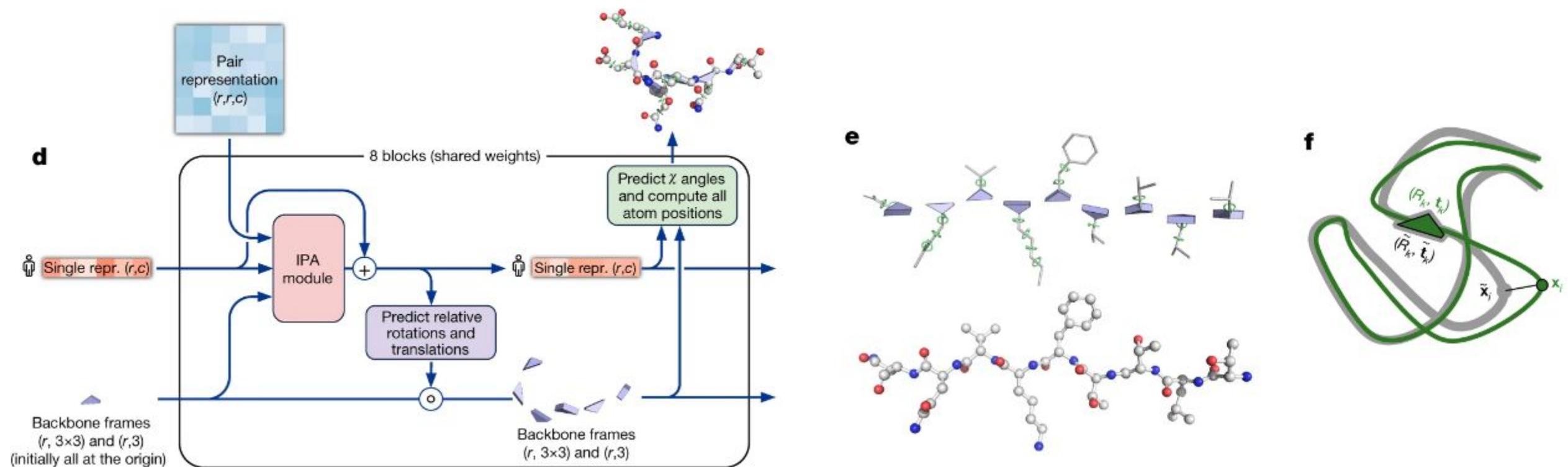
Updated version: Transformer working on multiple sequence alignments

# AlphaFold2: Transformers + graphs



[...] a concrete structural hypothesis arises early within the Evoformer blocks and is continuously refined. The key innovations in the Evoformer block are new mechanisms to exchange information within the MSA and pair representations that enable direct reasoning about the spatial and evolutionary relationships."

# AlphaFold2: iterative refinement



Both within the structure module and throughout the whole network, we reinforce the notion of iterative refinement by repeatedly applying the final loss to outputs and then feeding the outputs recursively into the same modules. The iterative refinement using the whole network (which we term “recycling” and is related to approaches in computer vision) contributes markedly to accuracy with minor extra training time.

# Structure prediction with a language model

RESEARCH

## ESM Metagenomic Atlas: The first view of the ‘dark matter’ of the protein universe

November 1, 2022

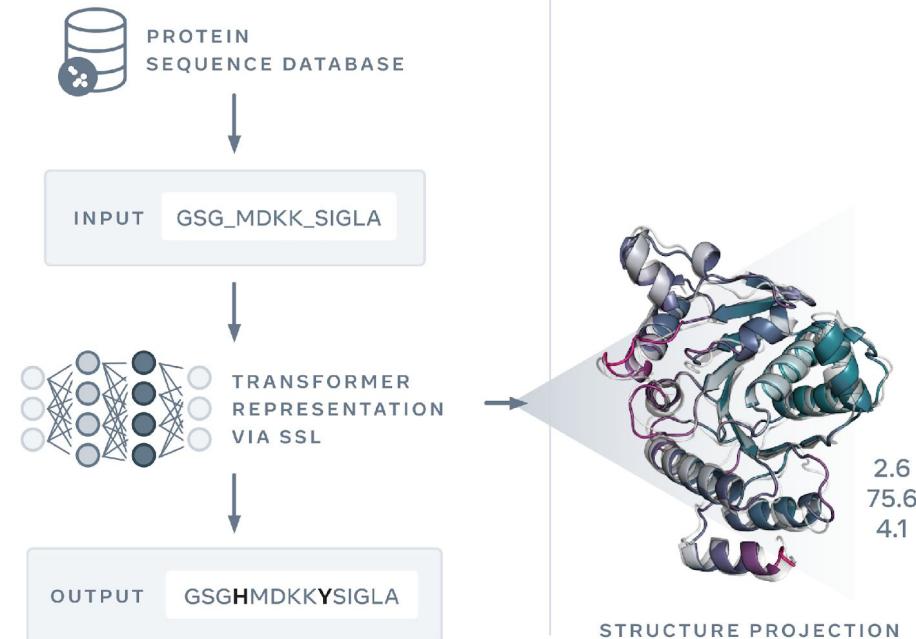
600 million predicted structures from metagenomes

“We found that as the model is scaled up from 8M to 15B parameters, information emerges in the internal representations that enables 3D structure prediction at an atomic resolution.”

### Protein language modeling

Pretraining self-supervision on sequences only.

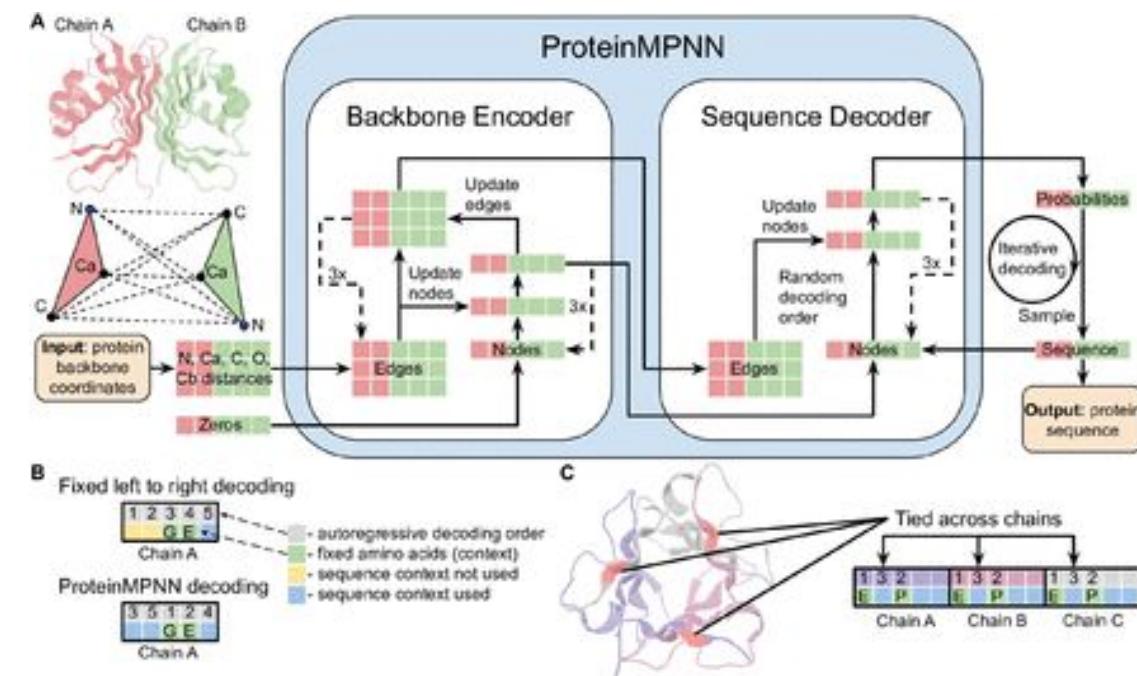
Structure emerges in the internal representations of the network from the self-supervision.



# Inverse folding

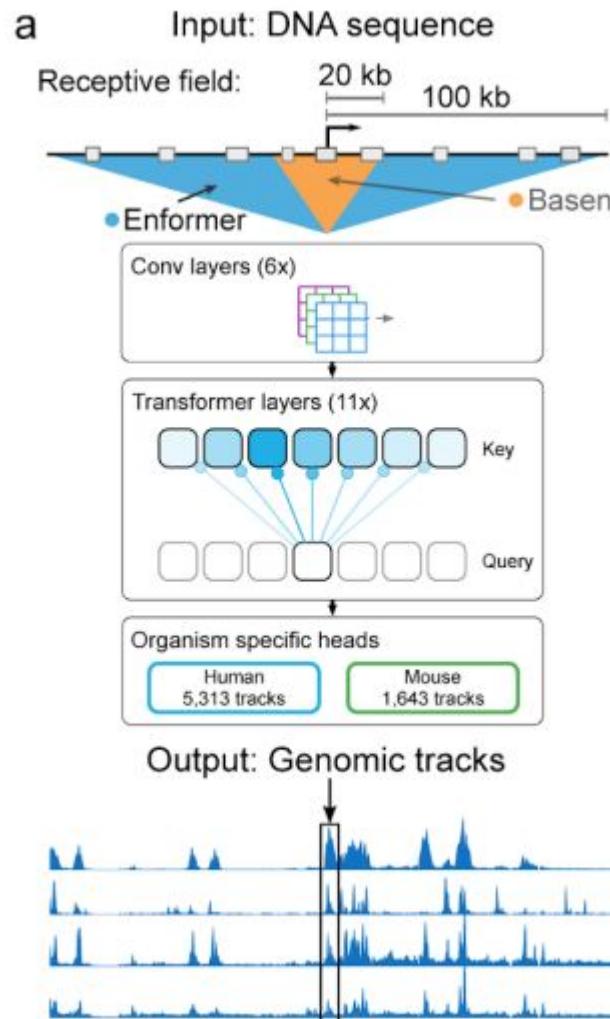
Going backwards from a structure to a sequence.

Can be useful for sequence design.

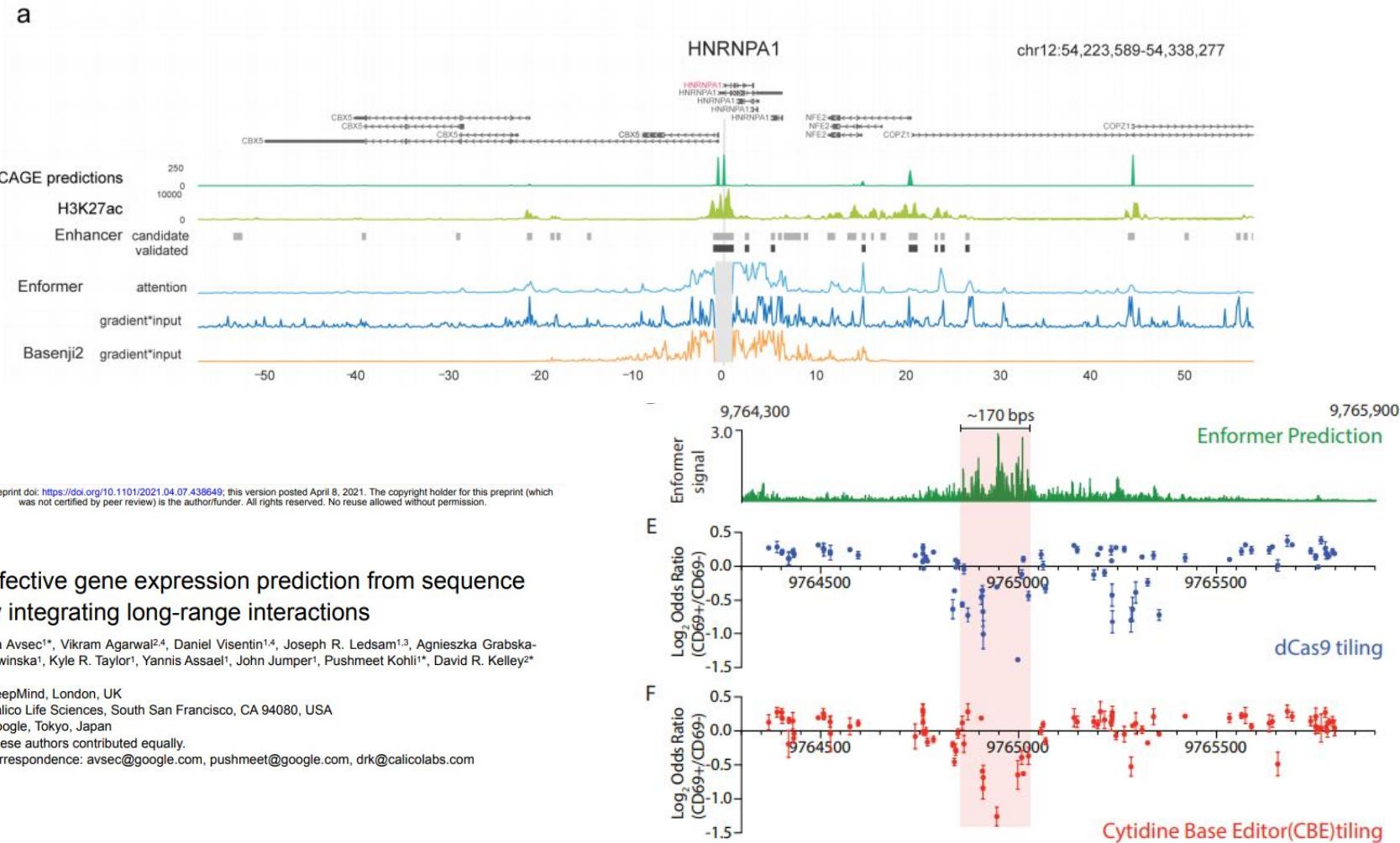


**ProteinMPNN (Dauparas et al. 2022)**  
Encoder (structure) => sequence (decoder)

# Enformer model for DNA



Enformer attends to cell-type-specific enhancers



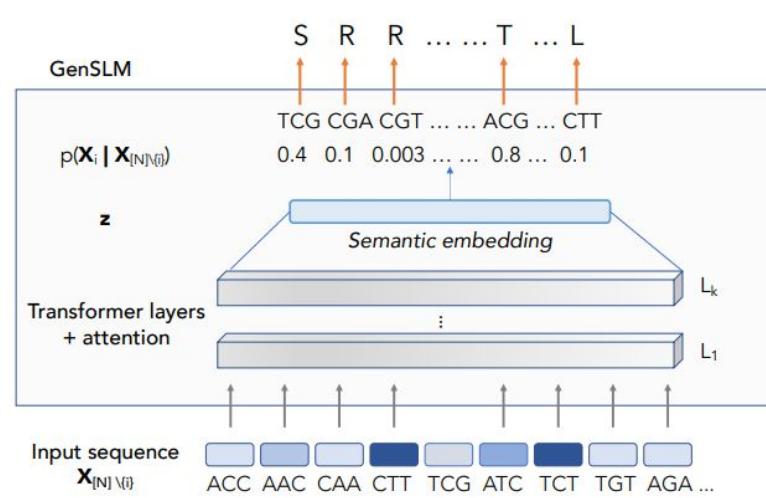
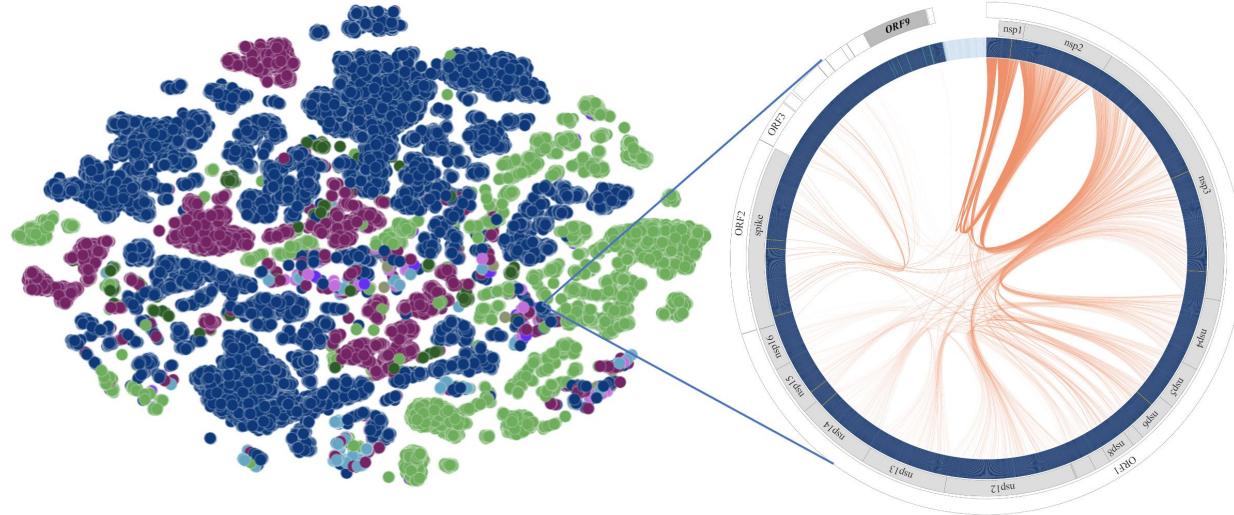
# *Caveats*

## **How far are we from personalized gene expression prediction using sequence-to-expression deep neural networks?**

Alexander Sasse, Bernard Ng, Anna Spiro, Shinya Tasaki,  
David A. Bennett, Christopher Gaiteri,  Philip L. De Jager,  
Maria Chikina, Sara Mostafavi

**doi:** <https://doi.org/10.1101/2023.03.16.532969>

# GenSLM



- Pretrain on 10M prokaryotic genomes
- Finetune on 1.5M SARS-CoV-2 genomes
- Identify VOCs (variants of concern)
- Generative (GPT) model
- Diffusion model to handle long dependencies

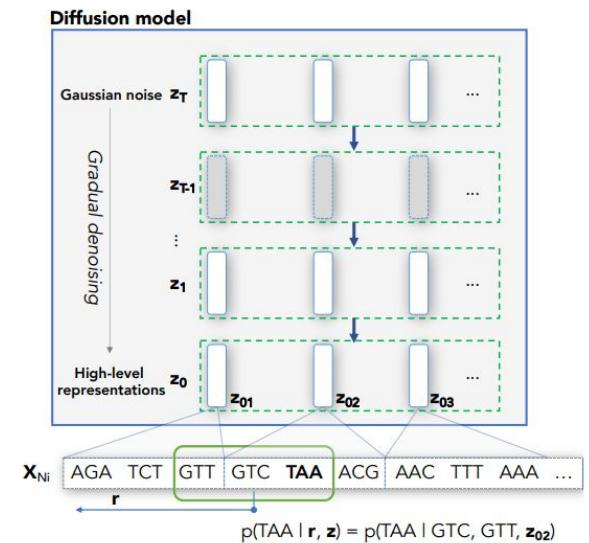


Figure 3: Illustration of diffusion-based hierarchical modeling. To predict a codon (such as TAA), we use both the previous codons within the context window (we use size 3 shown in green for illustration) and the high-level representations  $\mathbf{z}$ .

# *Nucleotide Transformer*

## The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics

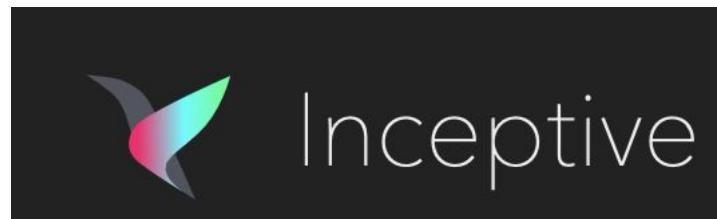
Hugo Dalla-Torre<sup>1</sup>, Liam Gonzalez<sup>1</sup>, Javier Mendoza-Revilla<sup>1</sup>, Nicolas Lopez Carranza<sup>1</sup>,  
Adam Henryk Grzywaczewski<sup>2</sup>, Francesco Oteri<sup>1</sup>, Christian Dallago<sup>2 3</sup>, Evan Trop<sup>1</sup>,  
Hassan Sirelkhatim<sup>2</sup>, Guillaume Richard<sup>1</sup>, Marcin Skwark<sup>1</sup>, Karim Beguir<sup>1</sup>,  
Marie Lopez<sup>\*† 1</sup>, Thomas Pierrot<sup>\*† 1</sup>

<sup>1</sup>**InstaDeep**    <sup>2</sup>**Nvidia**    <sup>3</sup>**TUM**

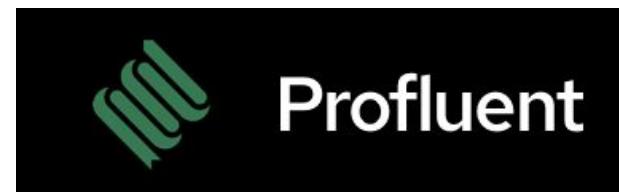
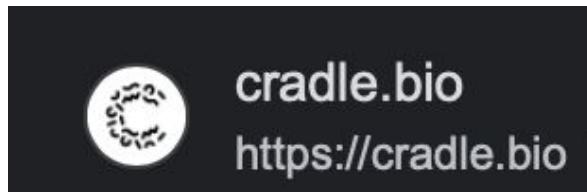
Building foundation models on 3,200 diverse human genomes and 850 non-human species.

Essentially a standard BERT model

## *AI based biomolecule design companies*



ML-based mRNA vaccine design



ML-based protein design

## *In summary*

- Recently developed machine learning (deep learning) concepts and architectures such as transformers, graph neural networks and generative models have shown great promise (and in the case of AlphaFold2, strong results) in biological research.
- It can be useful to think of these deep learning architectures as modules that process inputs into some type of generic numeric representation that can be re-used in other models or tasks.
- We don't necessarily need to have labeled examples in order to create useful machine learning models. For example, large language models can be trained without labels and finetuned or directly "transferred" to downstream predictive tasks, where sometimes very few training examples suffice.

nature methods

PERSPECTIVE

<https://doi.org/10.1038/s41592-021-01283-4>



### Differentiable biology: using deep learning for biophysics-based and data-driven modeling of molecular mechanisms

Mohammed AlQuraishi<sup>1,2</sup> and Peter K. Sorger<sup>1,2</sup>

Deep learning using neural networks relies on a class of machine-learnable models constructed using 'differentiable programs'. These programs can combine mathematical equations specific to a particular domain of natural science with general-purpose, machine-learnable components trained on experimental data. Such programs are having a growing impact on molecular and cellular biology. In this Perspective, we describe an emerging 'differentiable biology' in which phenomena ranging from the small and specific (for example, one experimental assay) to the broad and complex (for example, protein folding) can be modeled effectively and efficiently, often by exploiting knowledge about basic natural phenomena to overcome the limitations of sparse, incomplete and noisy data. By distilling differentiable biology into a small set of conceptual primitives and illustrative vignettes, we show how it can help to address long-standing challenges in integrating multimodal data from diverse experiments across biological scales. This promises to benefit fields as diverse as biophysics and functional genomics.