Artificial Intelligence Application's in Computational Biology

Eamon McAndrew

Institut de Pharmacologie Moléculaire et Cellulaire, Sophia Antipolis, France

Mcandrew@ipmc.cnrs.fr

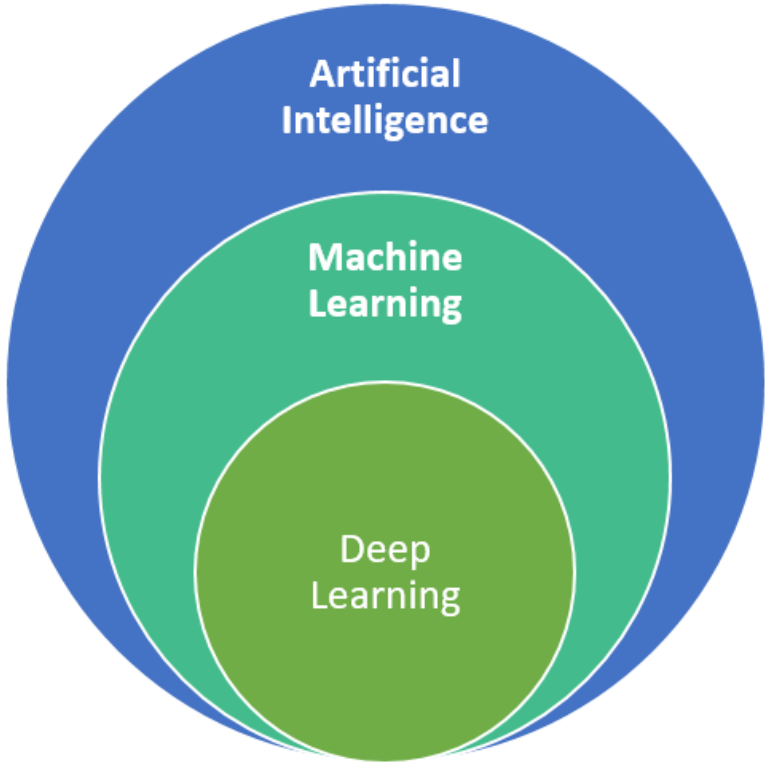
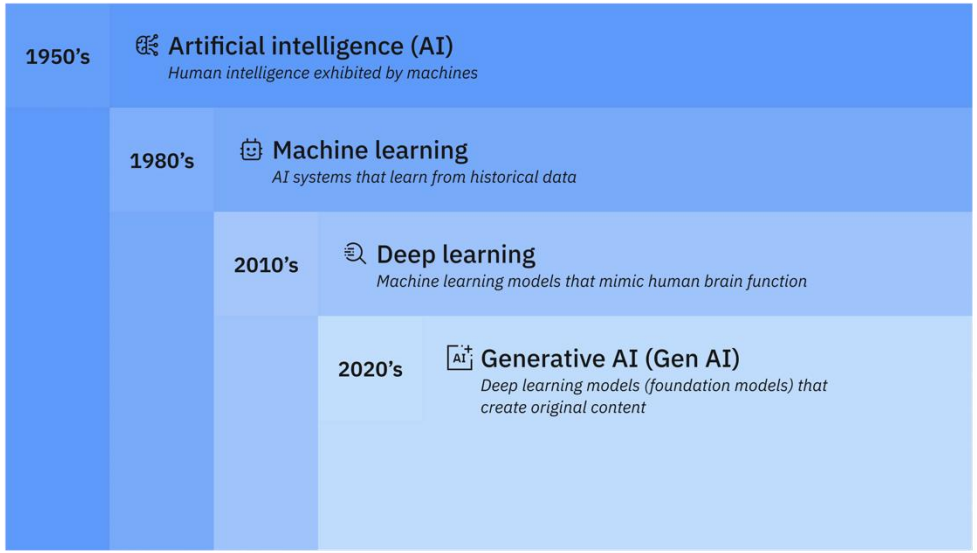
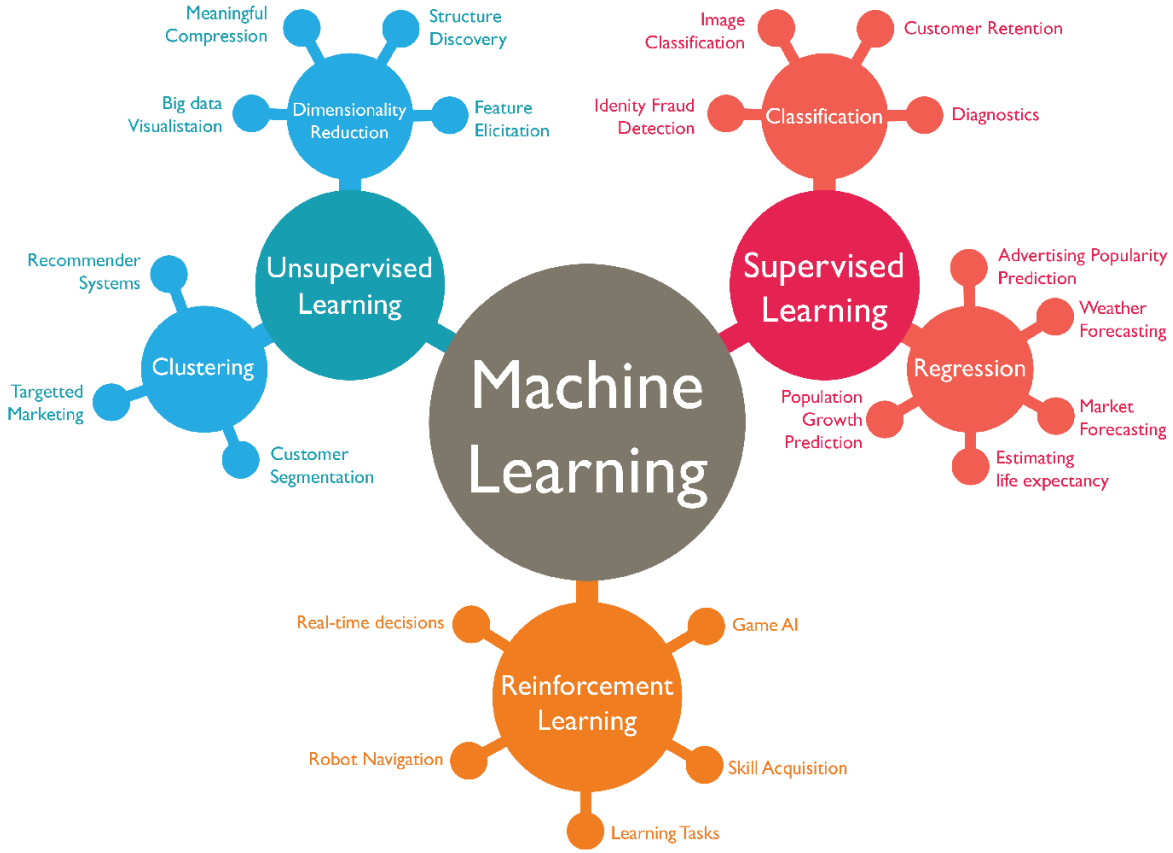
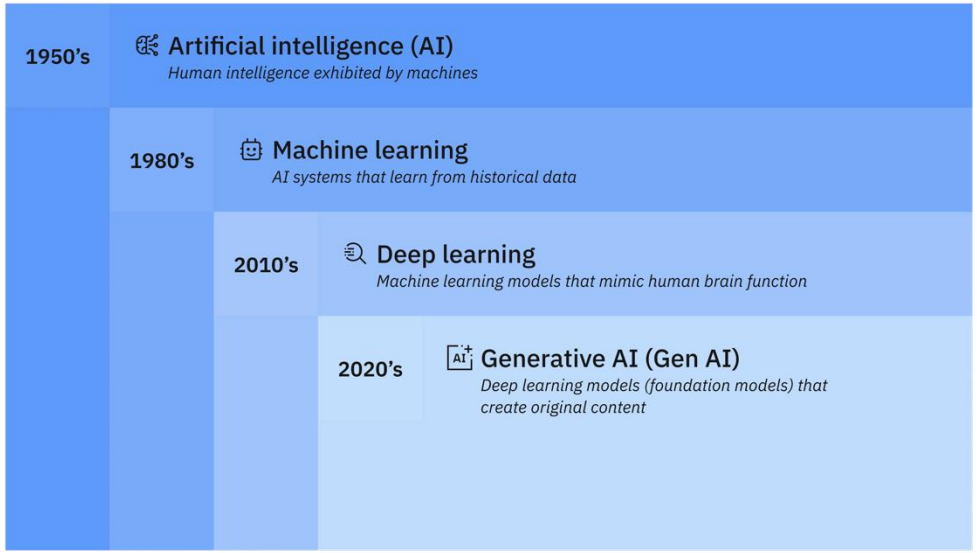


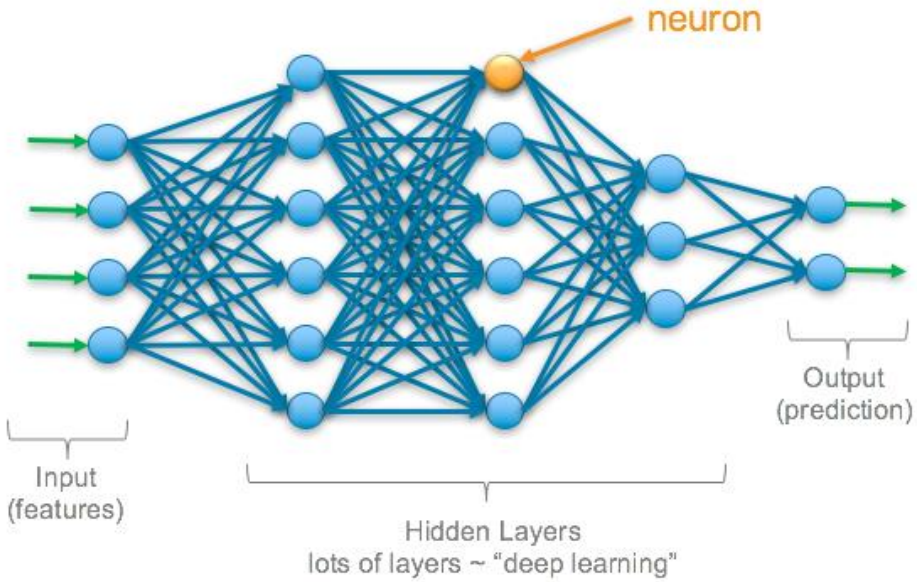
Figure 1: artificial intelligence, machine leaning and deep learning Source: Nadia BERCHANE (M2 IESCI, 2018)

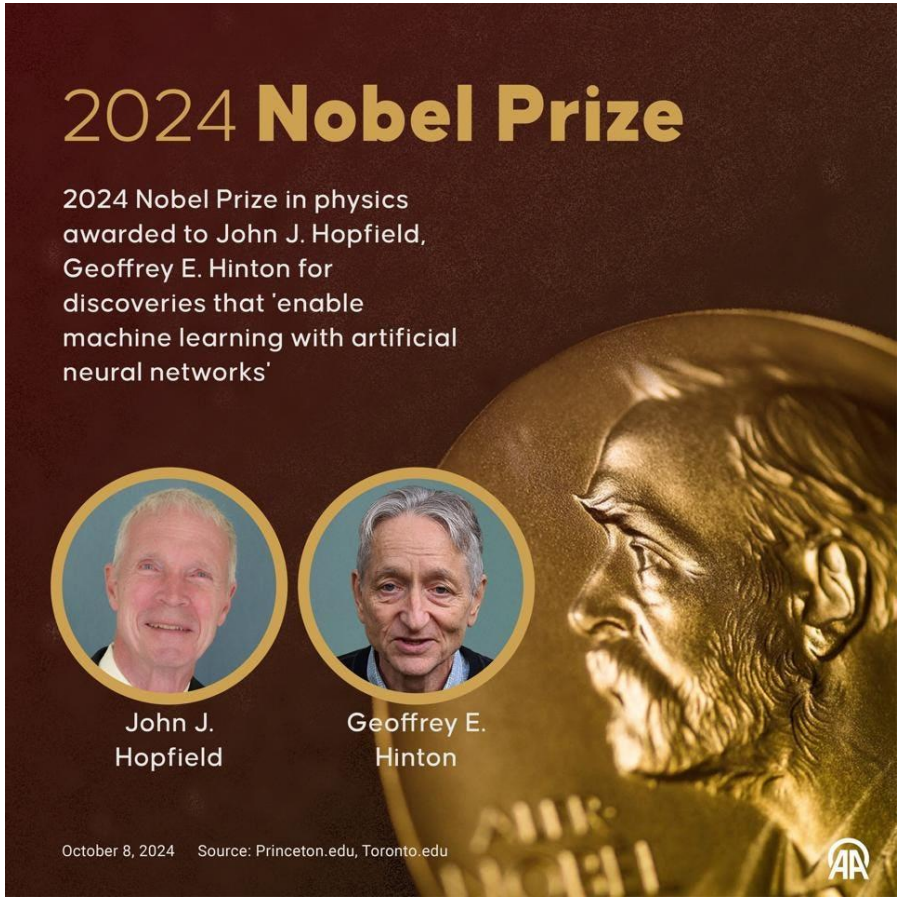
Funded by the European Union






1950's
Artificial intelligence (AI)
Human intelligence exhibited by machines
1980's
Machine learning
AI systems that learn from historical data
2010's
Deep learning
Machine learning models that mimic human brain function
2020's
Generative AI (Gen AI)
Deep learning models (foundation models) that create original content


neuron
Input (features)
Hidden Layers
lots of layers ~ "deep learning"
Output (prediction)


2024 Nobel Prize
2024 Nobel Prize in physics awarded to John J. Hopfield, Geoffrey E. Hinton for discoveries that 'enable machine learning with artificial neural networks'
John J. Hopfield
Geoffrey E. Hinton
October 8, 2024
Source: Princeton.edu, Toronto.edu

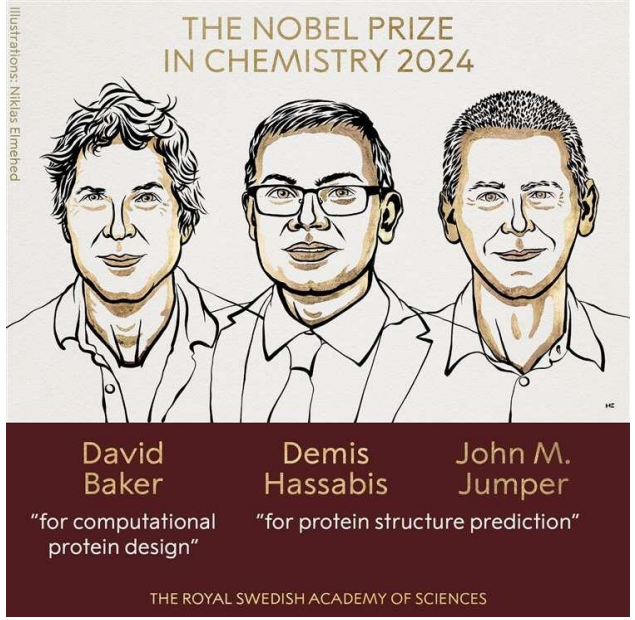
LongTREC
ipmc
cnrs
UNIVERSITÉ CÔTE D'AZUR
Funded by the European Union

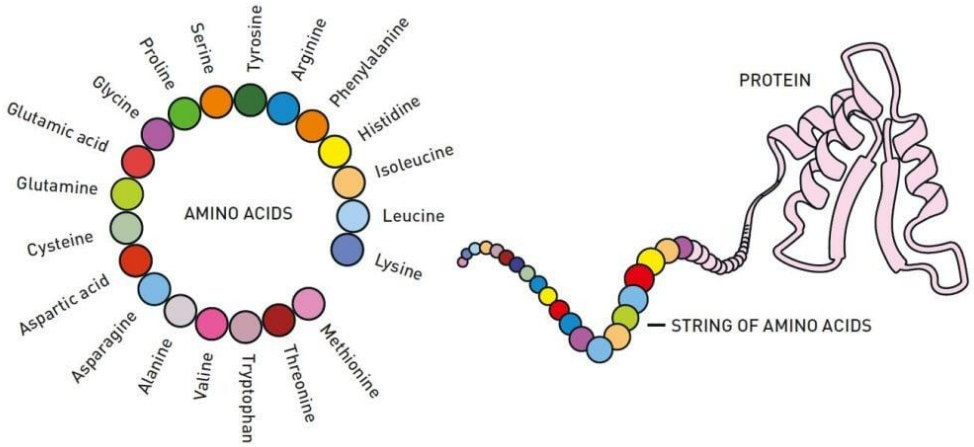
1950's — **Artificial intelligence (AI)**
Human intelligence exhibited by machines

1980's — **Machine learning**
AI systems that learn from historical data

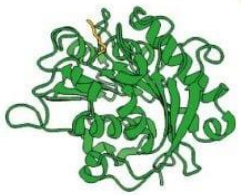
2010's — **Deep learning**
Machine learning models that mimic human brain function

2020's — **Generative AI (Gen AI)**
Deep learning models (foundation models) that create original content





2022: Part of a huge molecular structure in the human body. More than a thousand proteins form a pore through the membrane surrounding the cell nucleus.



2022: Natural enzymes that can decompose plastic. The aim is to design proteins that can be used to recycle plastic.

2023: A bacterial enzyme that causes antibiotic resistance. The structure is important for discovering ways of preventing antibiotic resistance.

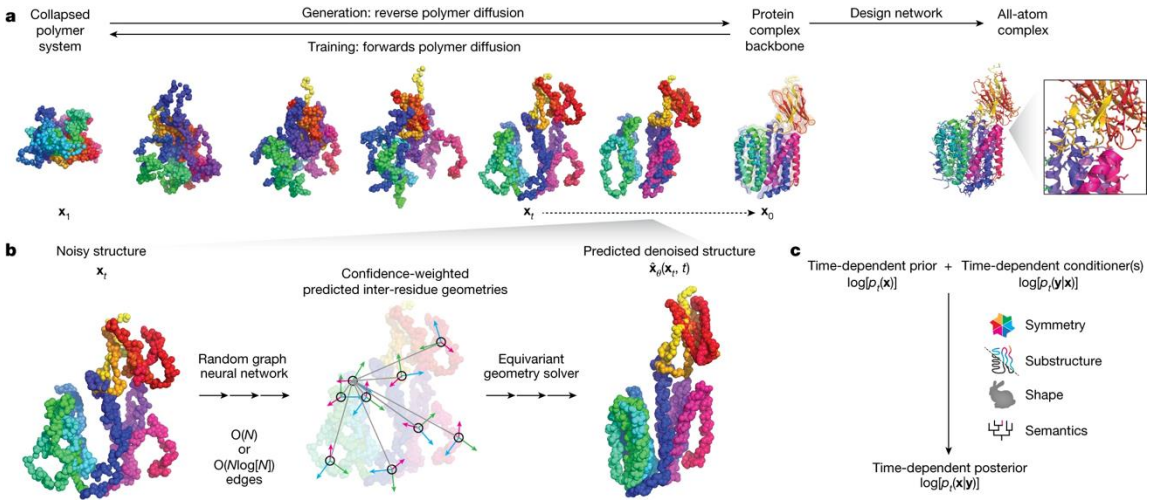
LongTREC · ipmc · cnrs · UNIVERSITÉ CÔTE D'AZUR · Funded by the European Union

1950's — **Artificial intelligence (AI)**
Human intelligence exhibited by machines

1980's — **Machine learning**
AI systems that learn from historical data

2010's — **Deep learning**
Machine learning models that mimic human brain function

2020's — **Generative AI (Gen AI)**
Deep learning models (foundation models) that create original content



ChatGPT

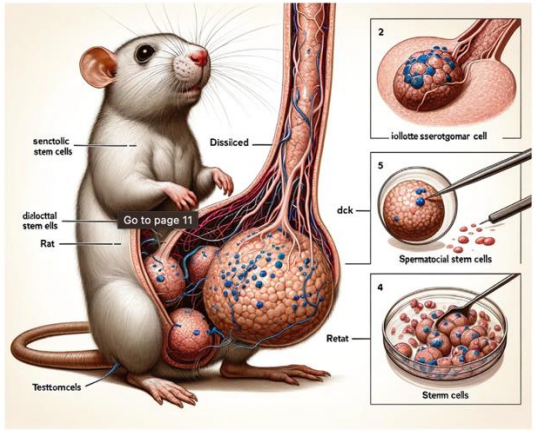
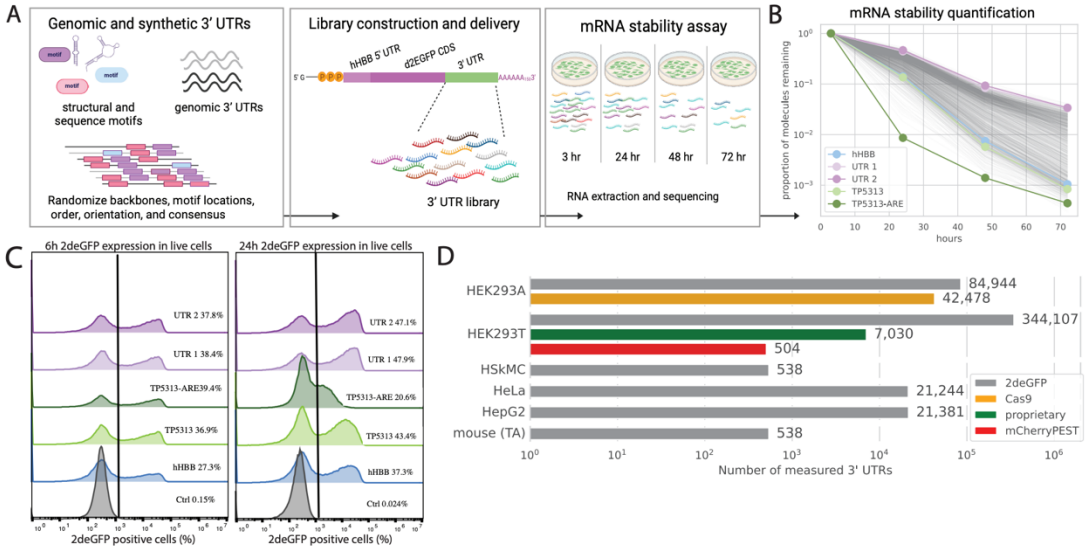


FIGURE 1
Spermatogonial stem cells, isolated, purified and cultured from rat testes.



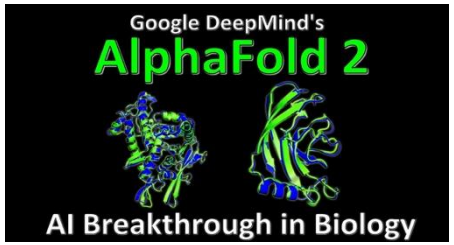
(Ingraham et al., 2023)
(Morrow et al., 2024)

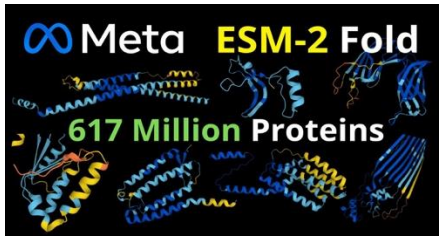
LongTREC ipmc cnrs UNIVERSITÉ CÔTE D'AZUR Funded by the European Union

Images

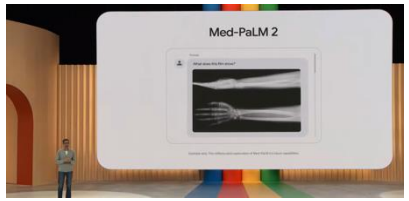
Text / Language

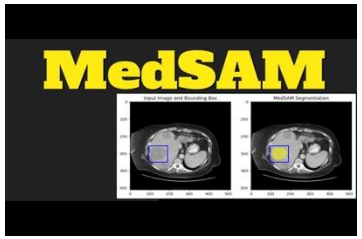
Protein Structure Prediction



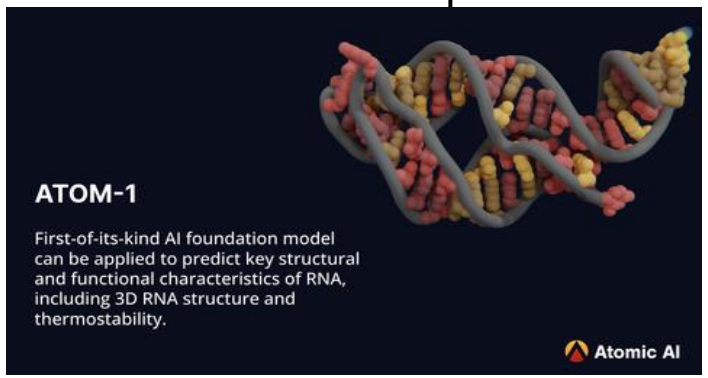


Clinical Applications



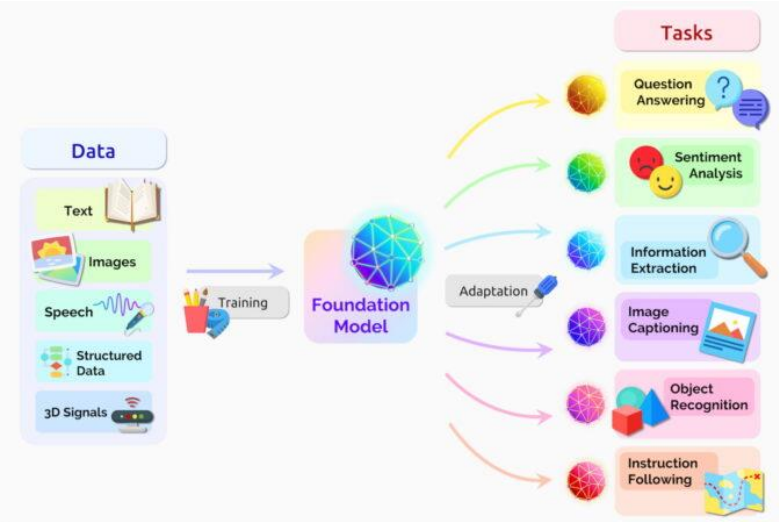


RNA structure prediction

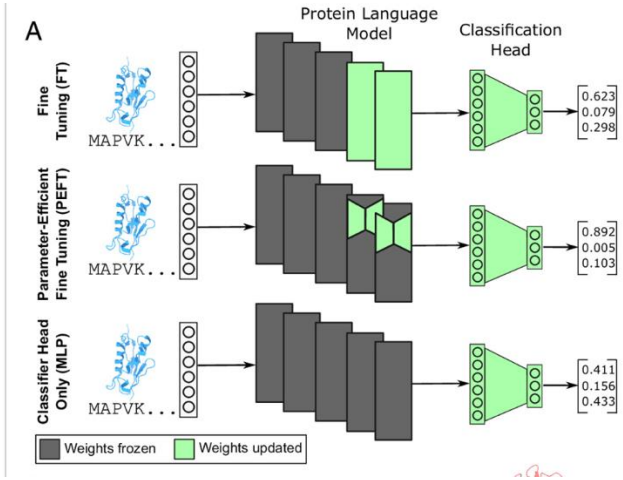


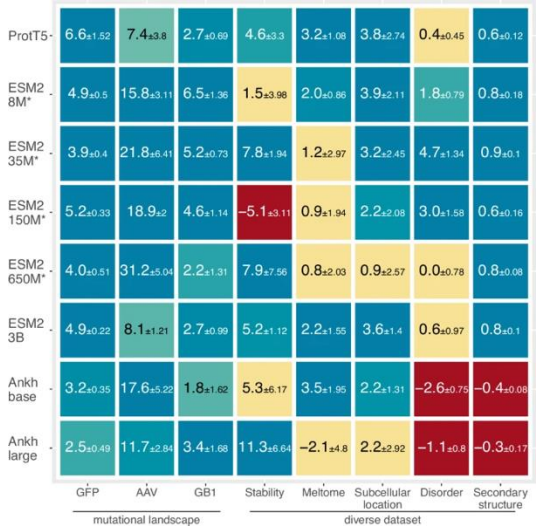
Video and ????

Foundation Models



Fine-tuning



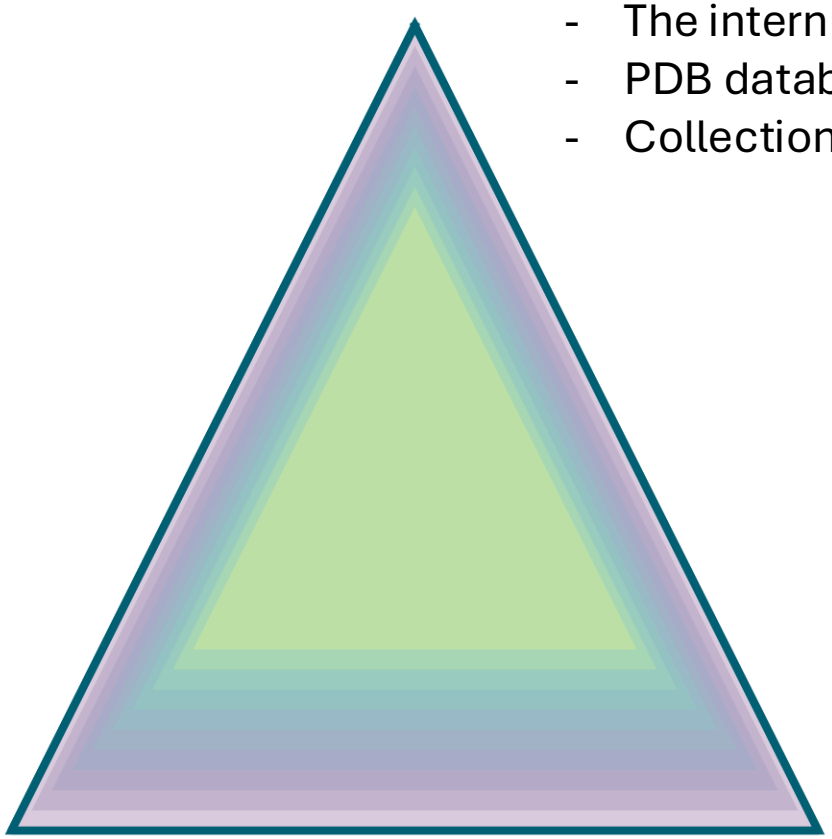


(Schmirler et al., 2024)

A model Needs -

DATA

- The internet
- PDB database
- Collection of images



COMPUTE

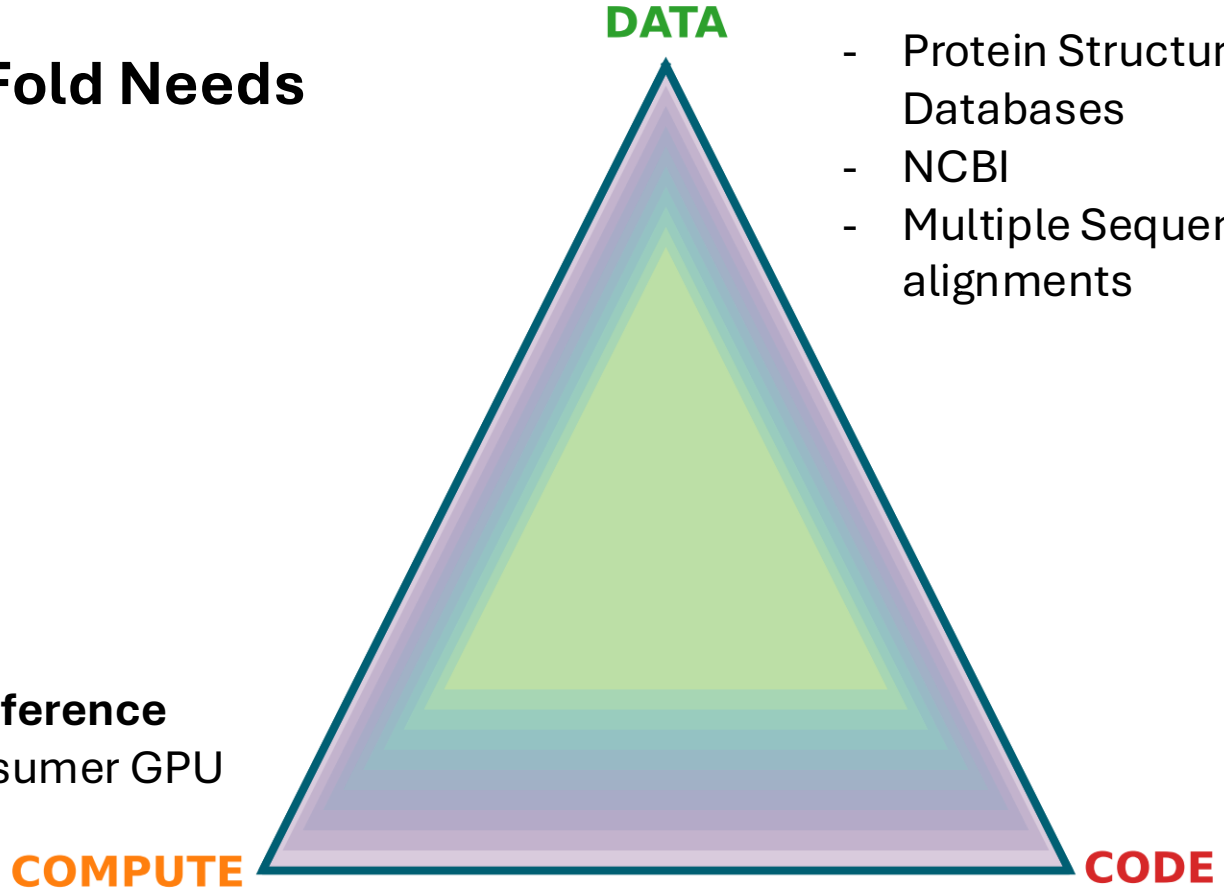
- GPU
- TPU
- Nuclear power stations?
- Inference VS Compute

CODE

- Pytorch / Tensorflow / Jax
- Architecture
- Training schemes
- Transformers
- CNNs
- GNNs
- Diffusion

All three scale with model complexity and performance

Alpha Fold Needs



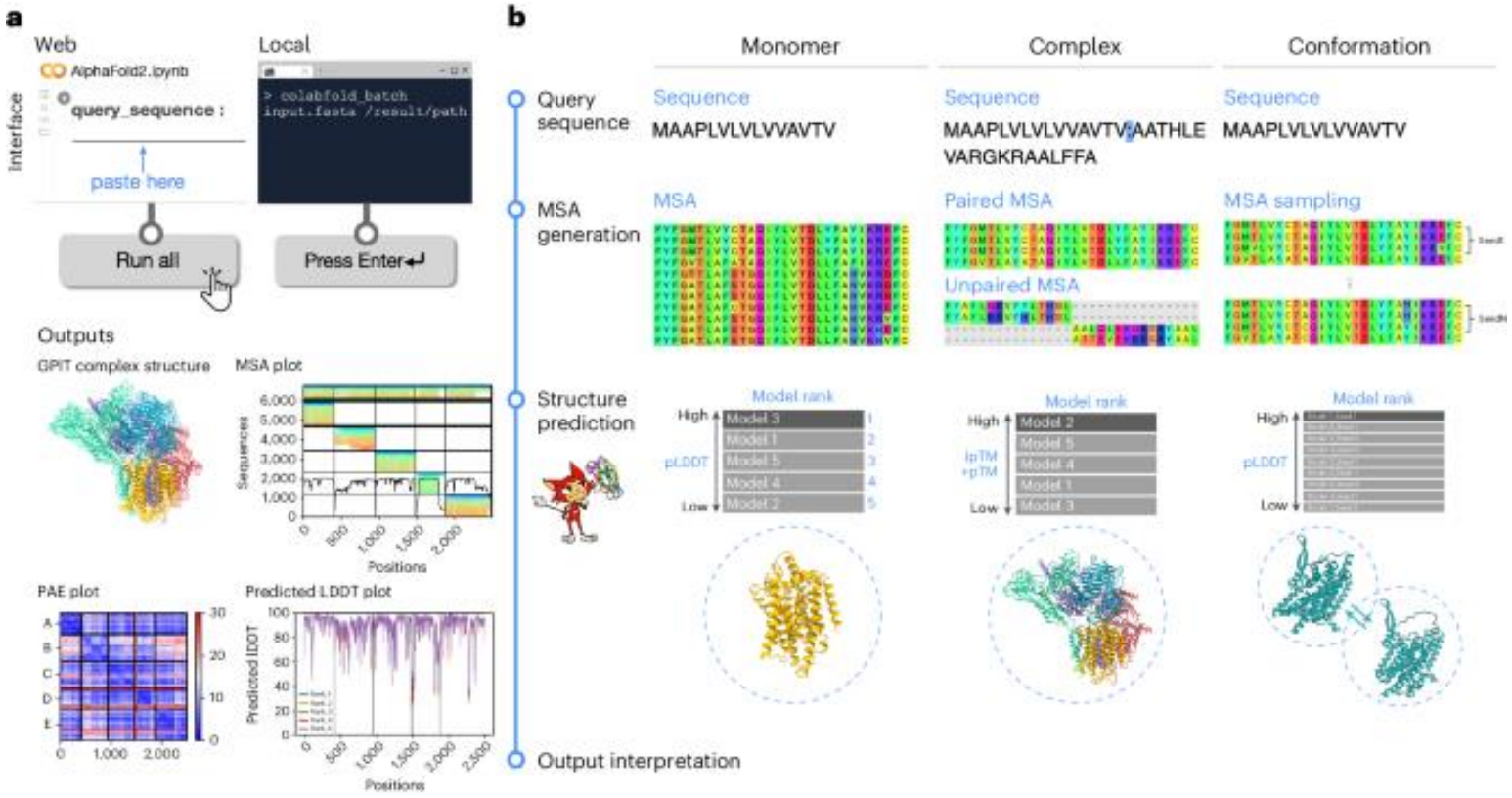
- Protein Structure Databases
- NCBI
- Multiple Sequence alignments

Training
Many TPUv3 cores.

Inference
Consumer GPU

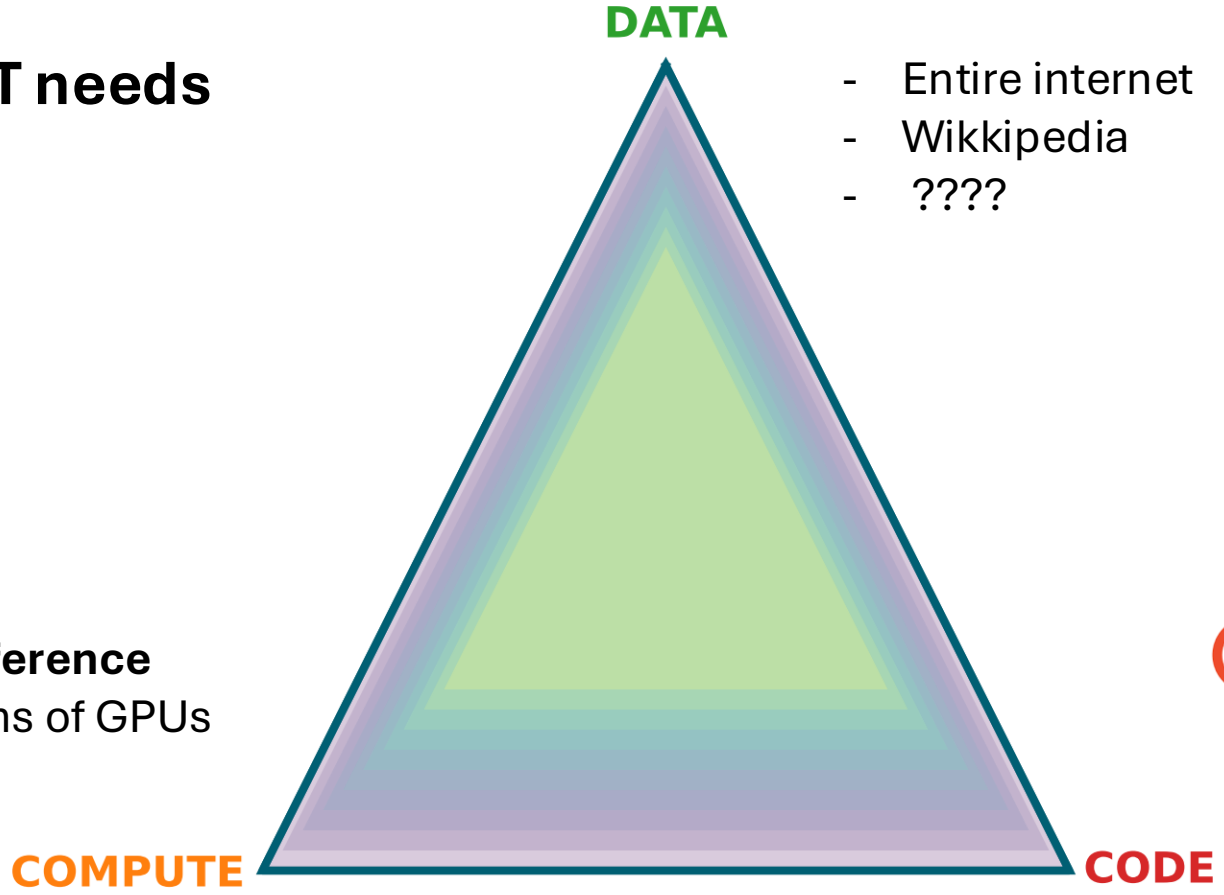
- Pytorch / Tensorflow / Jax
- Transformer

Run for free or pay as you need for larger tasks to run the whole AlphaFold workflow in a Jupyter notebook hosted entirely online.



(Kim et al., 2024)

Chat GPT needs



- Entire internet
- Wikkipedia
- ????

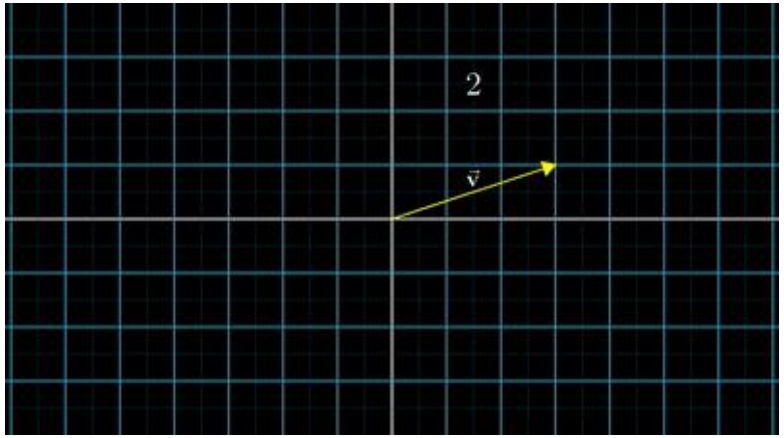
Training
Hundreds of thousands of GPUs

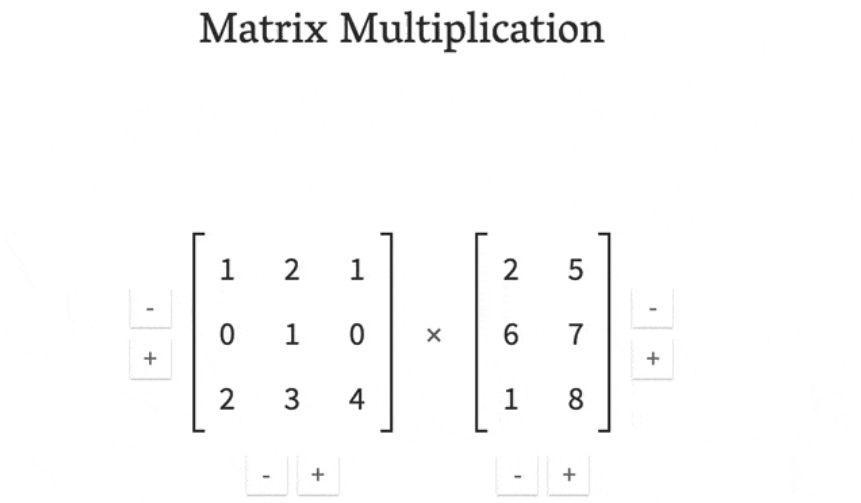
Inference
Dozens of GPUs

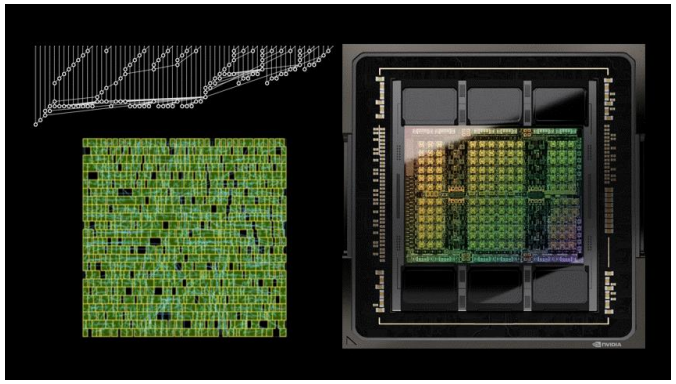
PyTorch

TensorFlow

- Pytorch / Tensorflow / Jax
- Transformers







CPU	GPU
Central Processing Unit	Graphics Processing Unit
Several cores	Many cores
Low latency	High throughput
Good for serial processing	Good for parallel processing
Can do a handful of operations at once	Can do thousands of operations at once

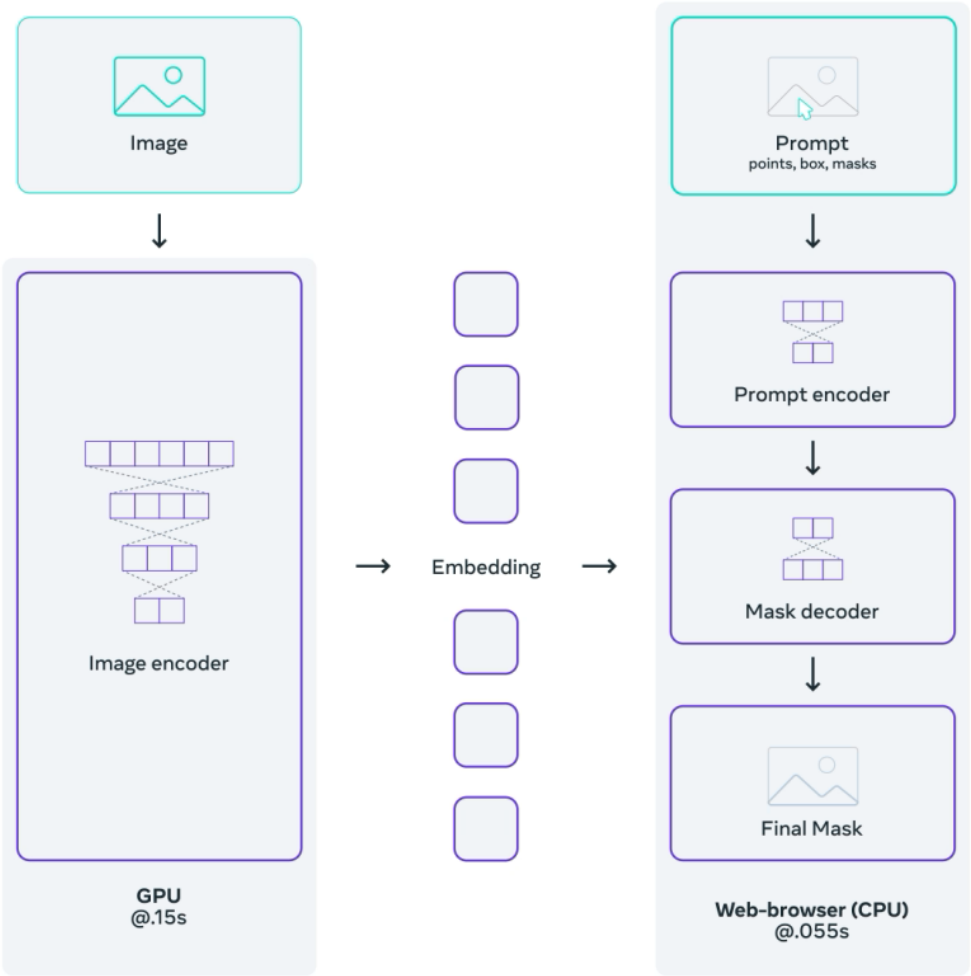
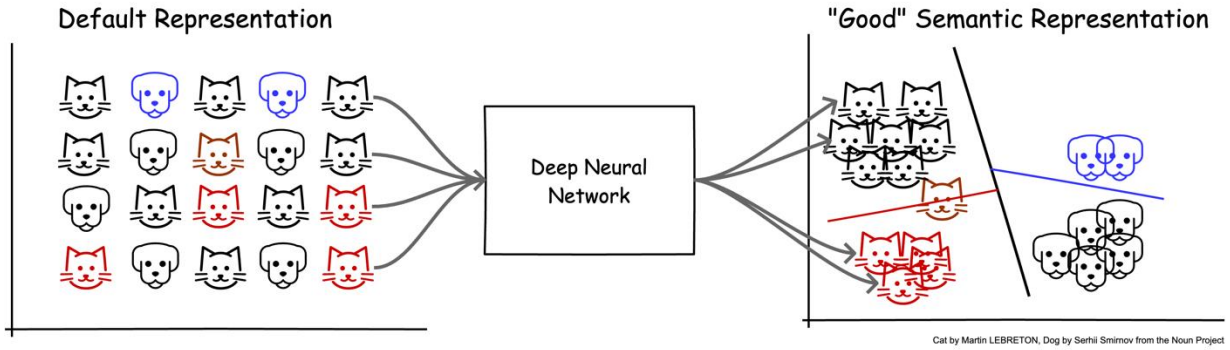
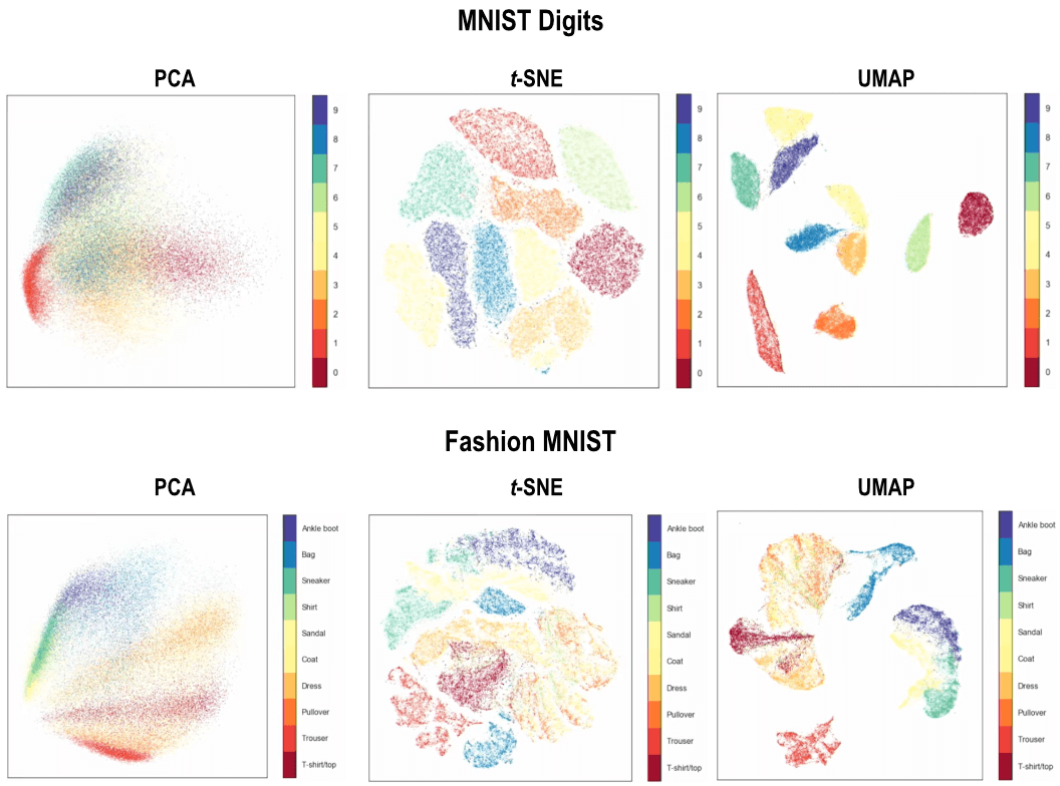
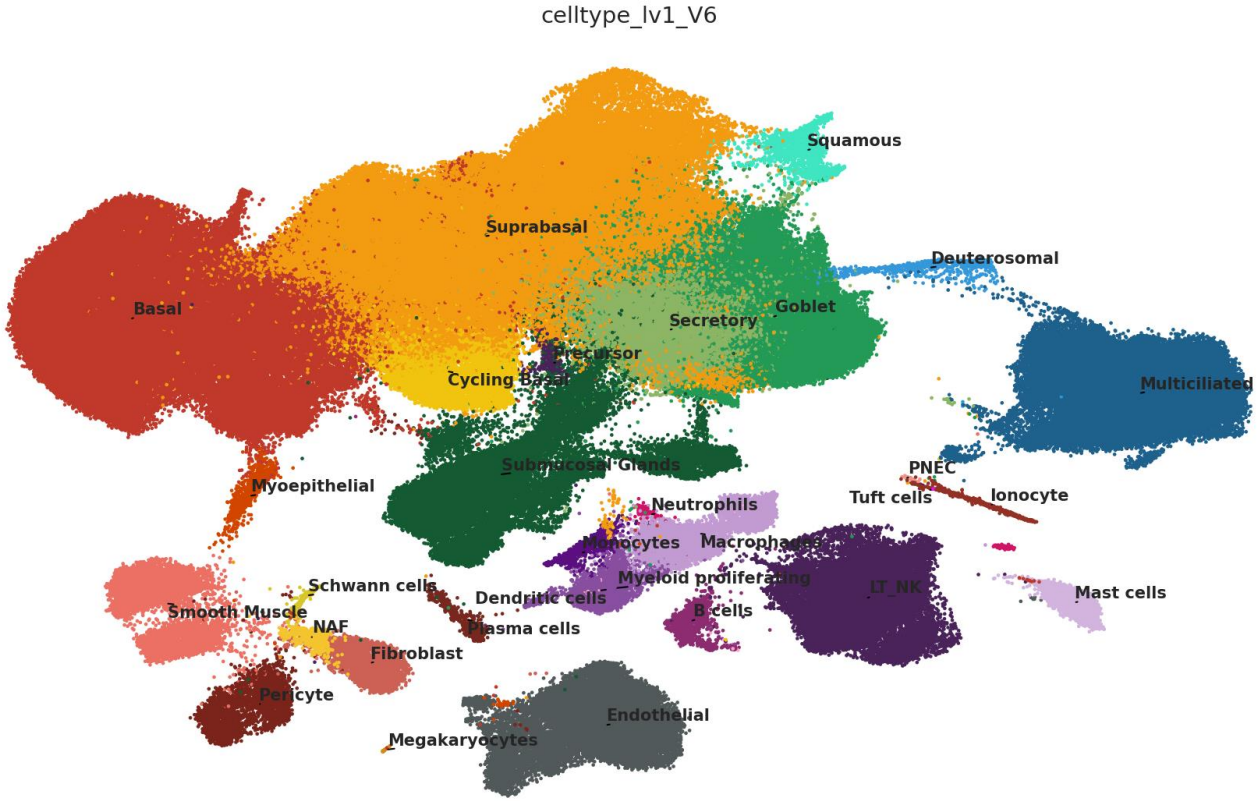


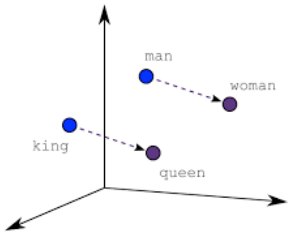
Image
Image encoder
GPU
@.15s
Embedding
Prompt
points, box, masks
Prompt encoder
Mask decoder
Final Mask
Web-browser (CPU)
@.055s


Default Representation
Deep Neural Network
"Good" Semantic Representation
Cat by Martin LEBRETON, Dog by Serhii Smirnov from the Noun Project

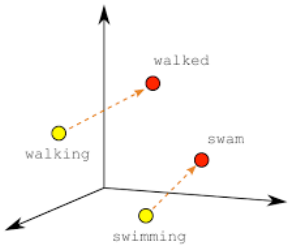
Funded by the European Union



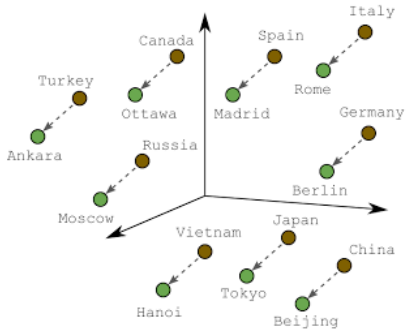




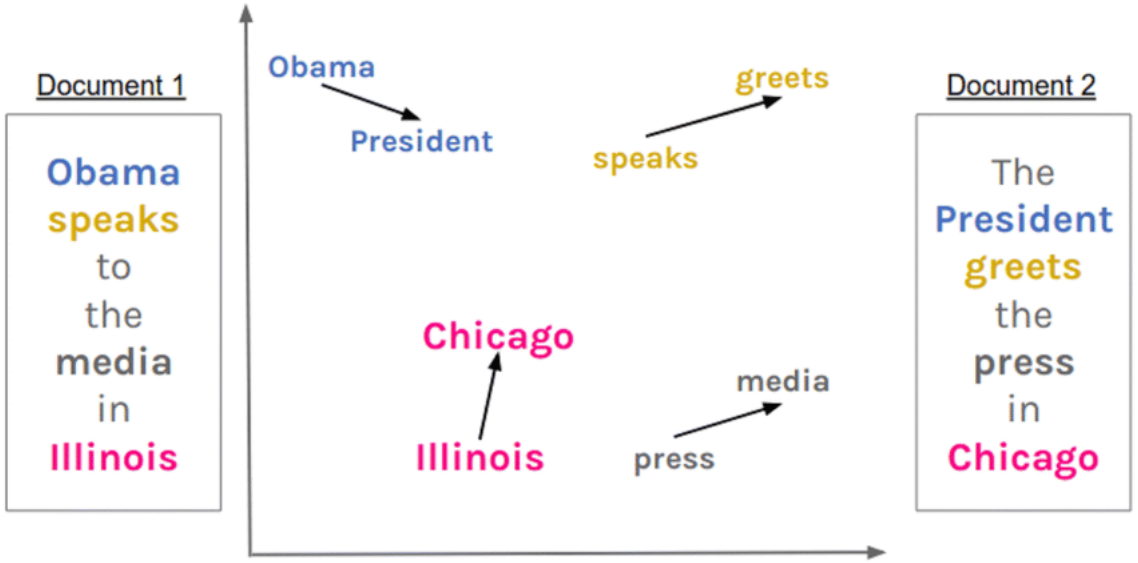
Male-Female



Verb Tense

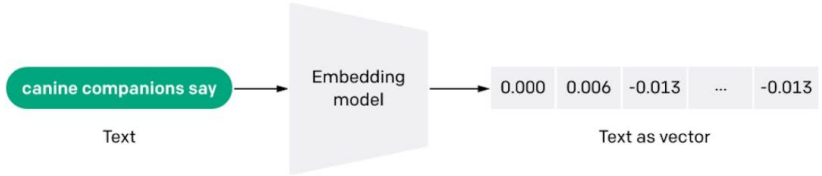


Country-Capital

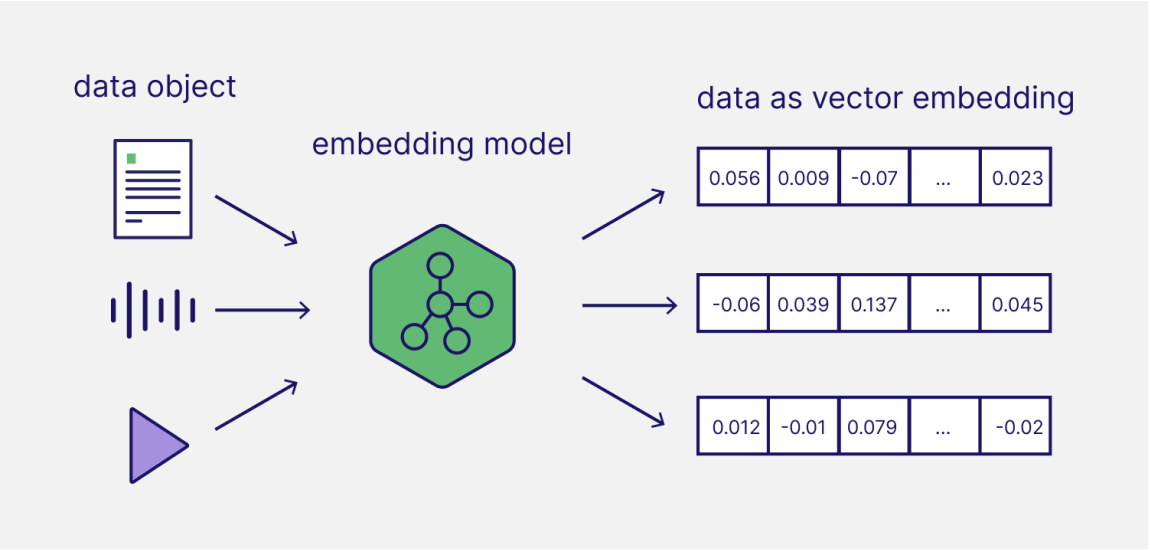


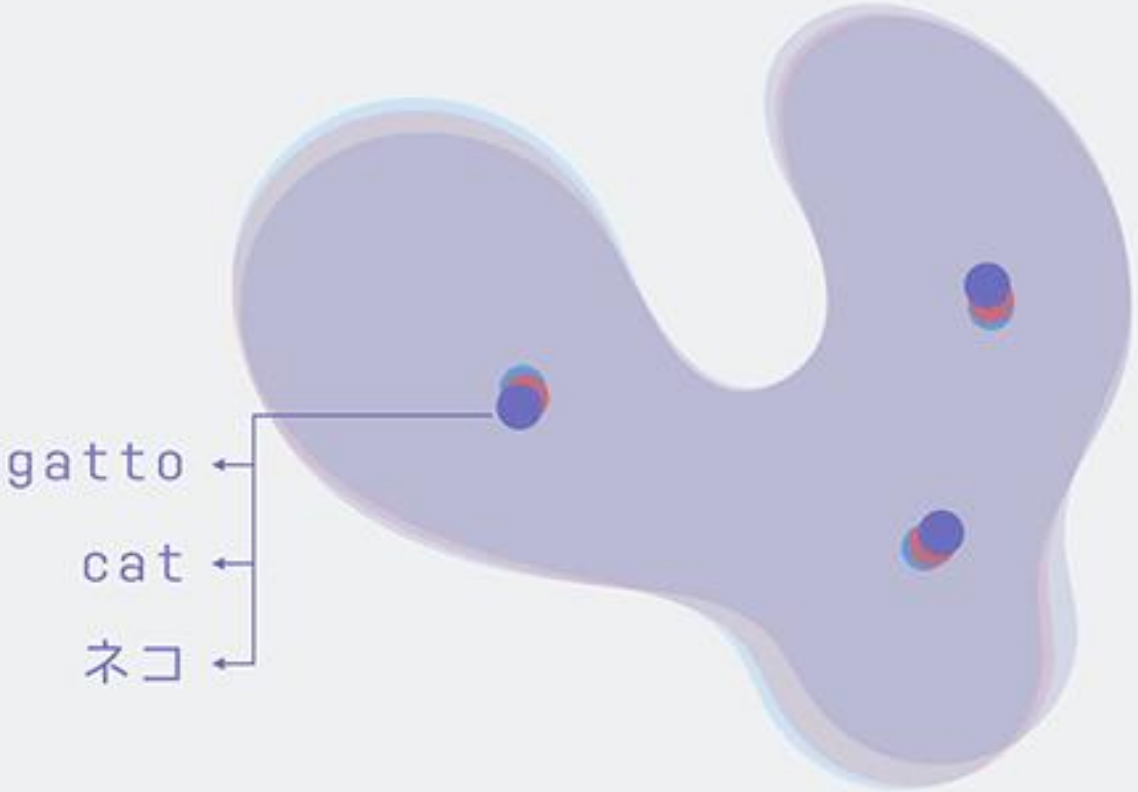
OpenAI Embedding

Embeddings Use Cases



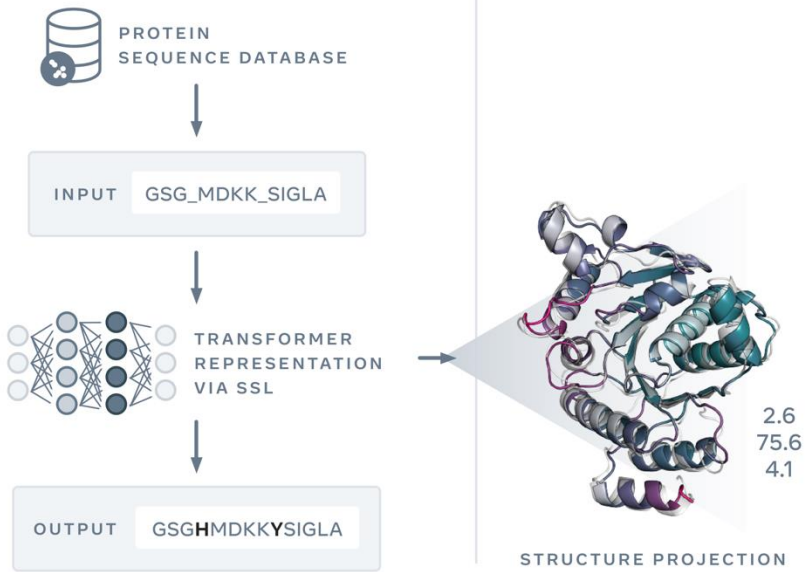
Text Similarity, Semantic Search, Clustering

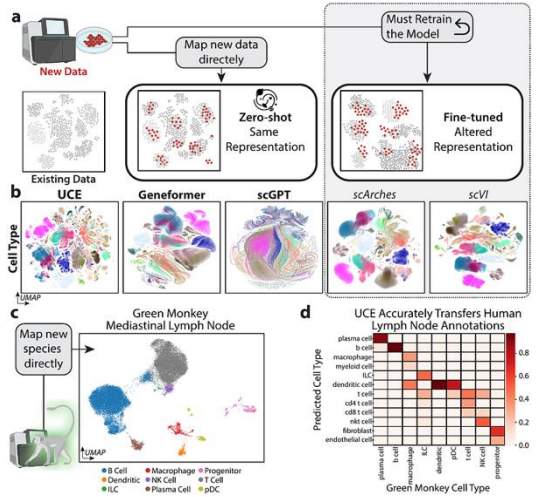


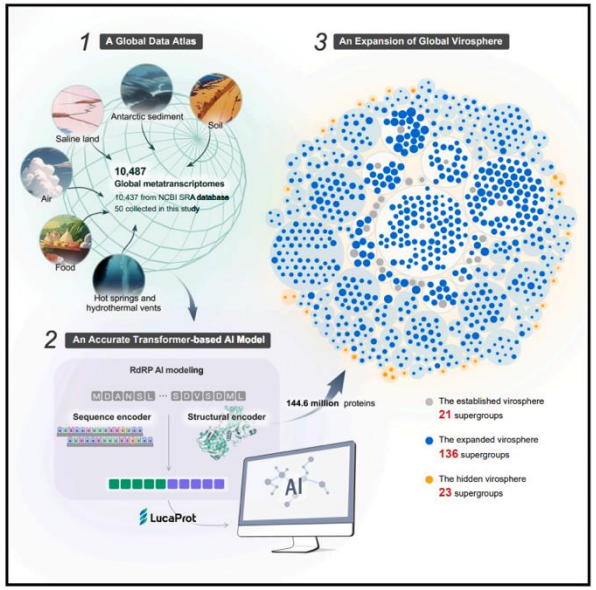

gatto
cat
ネコ

Protein language modeling

Pretraining self-supervision on sequences only. → Structure emerges in the internal representations of the network from the self-supervision.





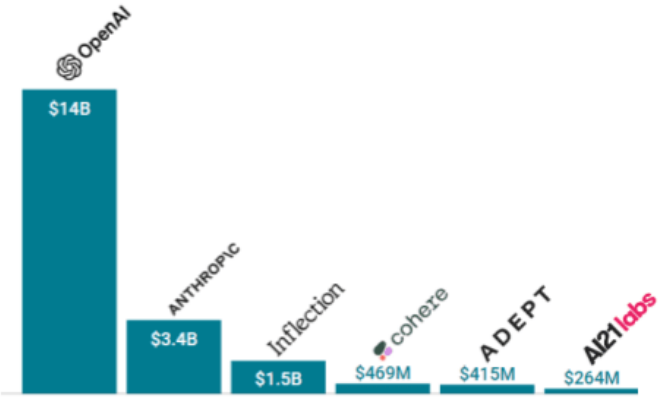


(Rosen et al., 2023)
(Hou et al., 2024)

The private market is split into open vs. closed

Disclosed equity funding to LLM developers (as of 10/27/2023)

Closed-source LLMs

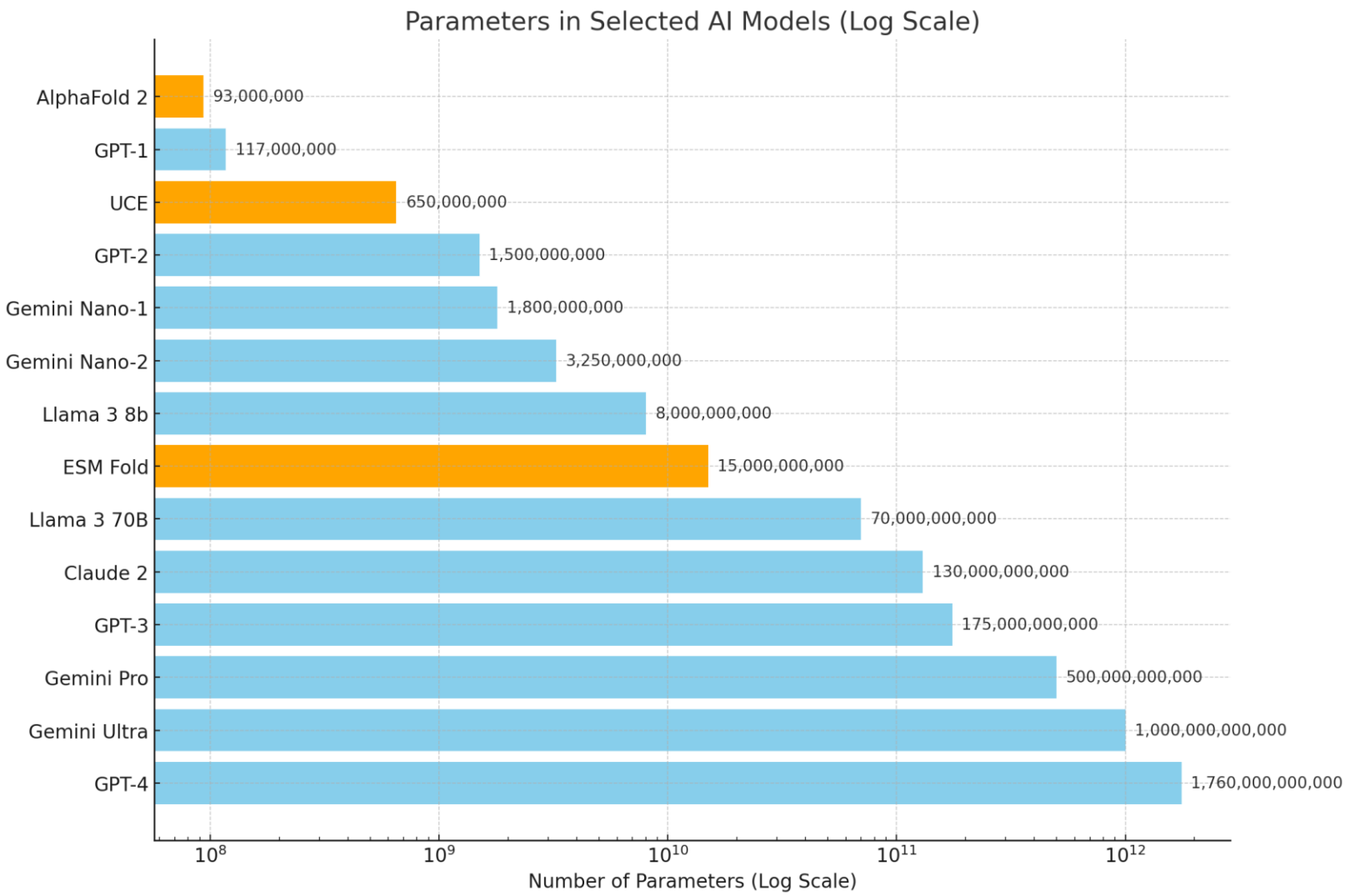


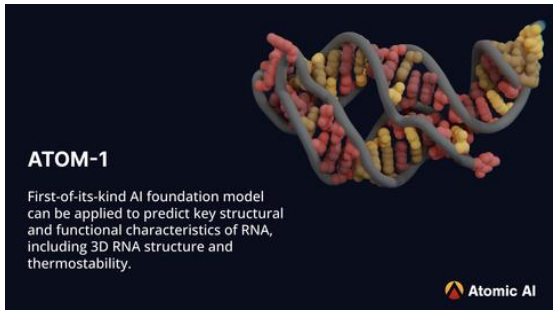
*Some developers may offer open-source versions of their models but keep their core models proprietary

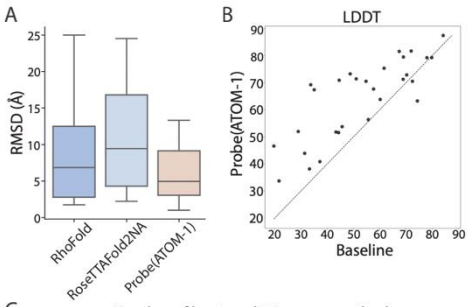
Open-source LLMs

*Excludes open-source developers that have not raised equity funding

CBINSIGHTS Source: CB Insights market report: Generative AI – large language model developers


Parameters in Selected AI Models (Log Scale)
AlphaFold 2
93,000,000
GPT-1
117,000,000
UCE
650,000,000
GPT-2
1,500,000,000
Gemini Nano-1
1,800,000,000
Gemini Nano-2
3,250,000,000
Llama 3 8b
8,000,000,000
ESM Fold
15,000,000,000
Llama 3 70B
70,000,000,000
Claude 2
130,000,000,000
GPT-3
175,000,000,000
Gemini Pro
500,000,000,000
Gemini Ultra
1,000,000,000,000
GPT-4
1,760,000,000,000
10^8
10^9
10^{10}
10^{11}
10^{12}
Number of Parameters (Log Scale)



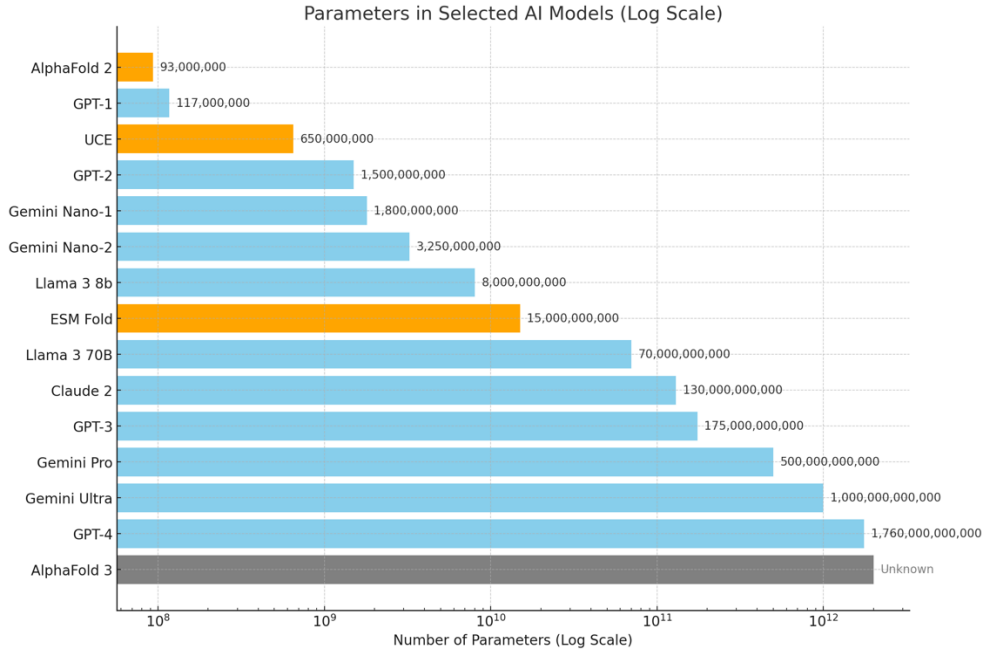


C

Number of best predictions per method

	Baseline	RhoFold	RoseTTAFold2NA	Probe(ATOM-1)
RMSD	1	6	3	19
LDDT	1	3	10	15

AI I for computational biology is beginning to look similar



(Boyd et al., 2023)
(Abramson et al., 2024)

LongTREC ipmc cnrs UNIVERSITÉ CÔTE D'AZUR Funded by the European Union

Article | Open access | Published: 08 May 2024

Accurate structure prediction of biomolecular interactions with AlphaFold 3

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, ... John M. Jumper + Show authors

AlphaFold Server BETA

But his enthusiasm was not universally shared. In contrast to the launch of AlphaFold 2 in 2021, *Nature's* publication of AlphaFold 3 lacked the open source code. That omission has sparked outcry from the research community, culminating in a protest letter signed by more than 1,000 scientists.

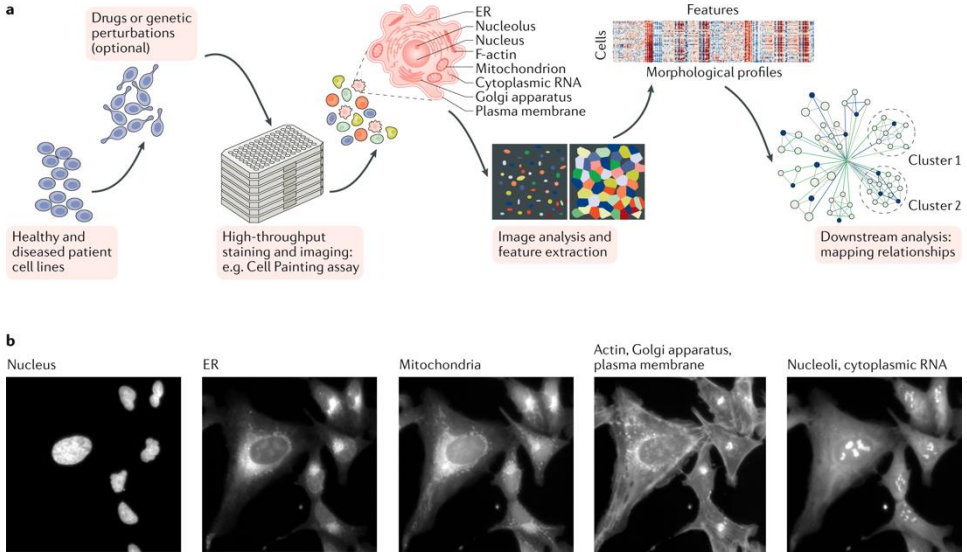
Why AlphaFold 3 needs to be open source

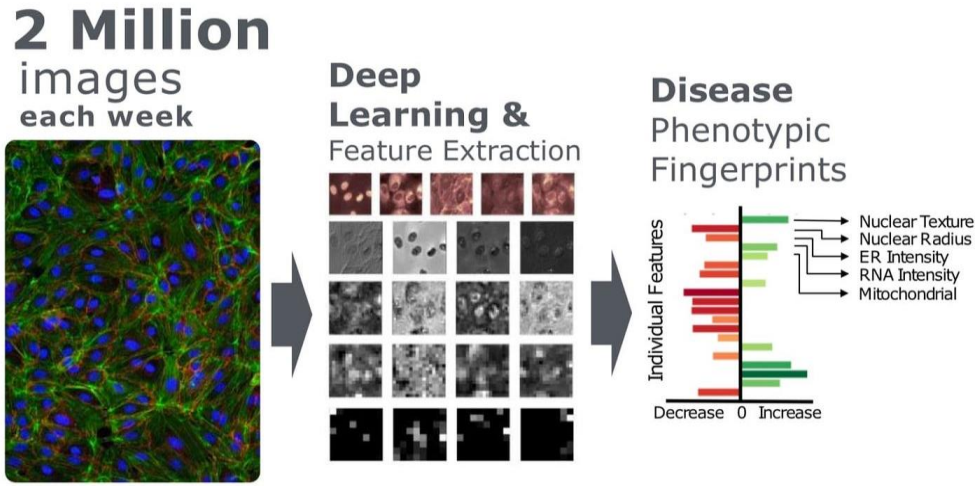
By Bryce Johnson
July 7, 2024

AlphaFold3 – why did *Nature* publish it without its code?

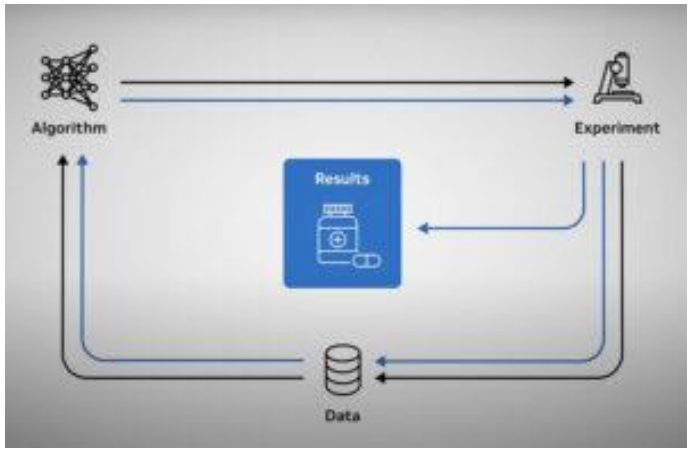
Criticism of our decision to publish AlphaFold3 raises important questions. We welcome readers' views.

Generate Large ML Centric Datasets





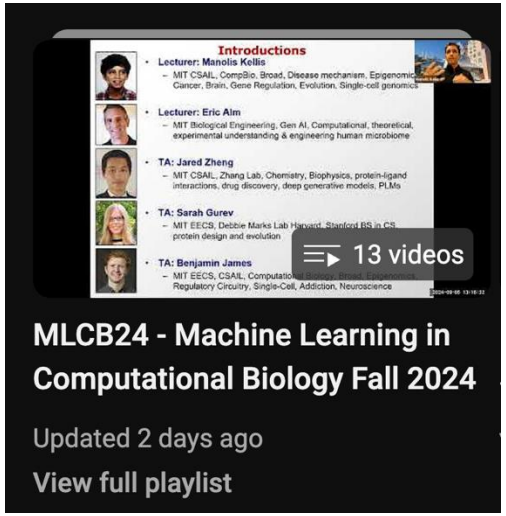
Lab In a Loop



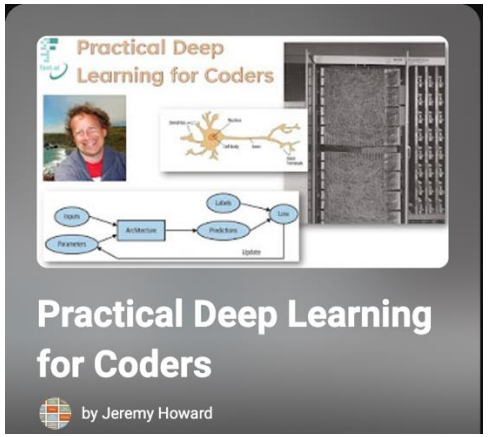
Iterate between experimental data generation and model training

(Celik et al., 2024)

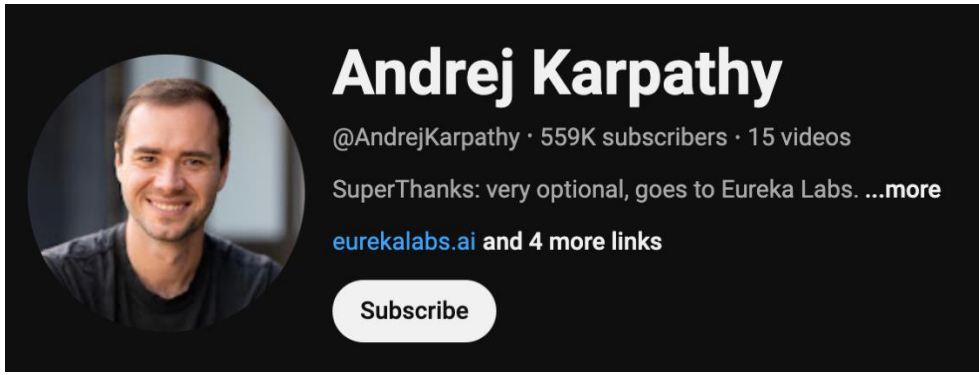
Thorough introduction centered on computational biology, course is ongoing and posted every week so its up to date.



Full course on deep learning from setup to training many popular architectures aimed at people with some coding and no deep learning experience



Some of the best explanations and teaching available – ranging in difficulty from public audiences to expert technical deep dives.



Extremely detailed explanations of the mathematical concepts underlying neural network, made mush more approachable through beautiful visualizations