LLMs outside Natural Language Processing Applications

Luiz Felipe Vecchietti

23. 06. 2025

MAX PLANCK INSTITUTE
FOR SECURITY AND PRIVACY

Max Planck Institute for Security and Privacy

Luiz Felipe Vecchietti
Postdoctoral Researcher
Data Science for Humanity Group
Max Planck Institute for Security and Privacy (MPI-SP)
Bochum, Germany

About Me

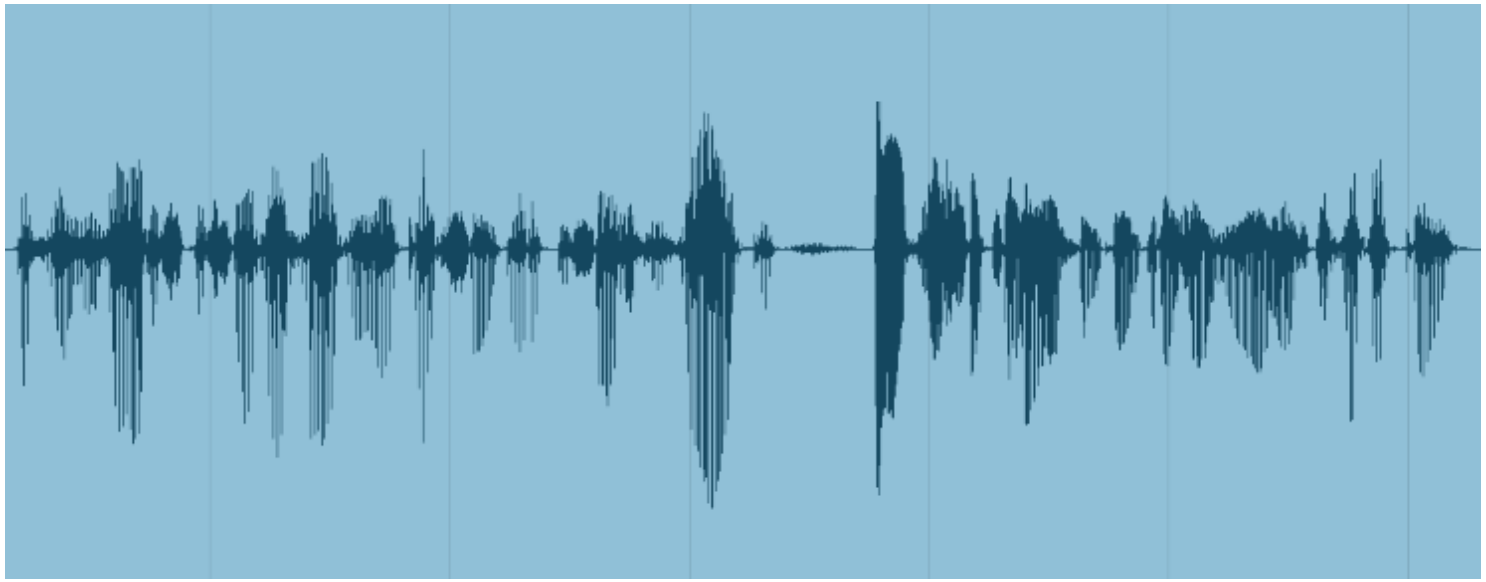
- PhD in Robotics from the Korea Advanced Institute of Science and Technology.
- At MPI-SP since last year: working on AI theory, AI alignment of LLMs and Robots, and AI applications such as Computational Biology and Demographic Research.

- **Alternative title for this lecture:** thinking about applications outside the Natural Language Processing domain.

NLP Applications

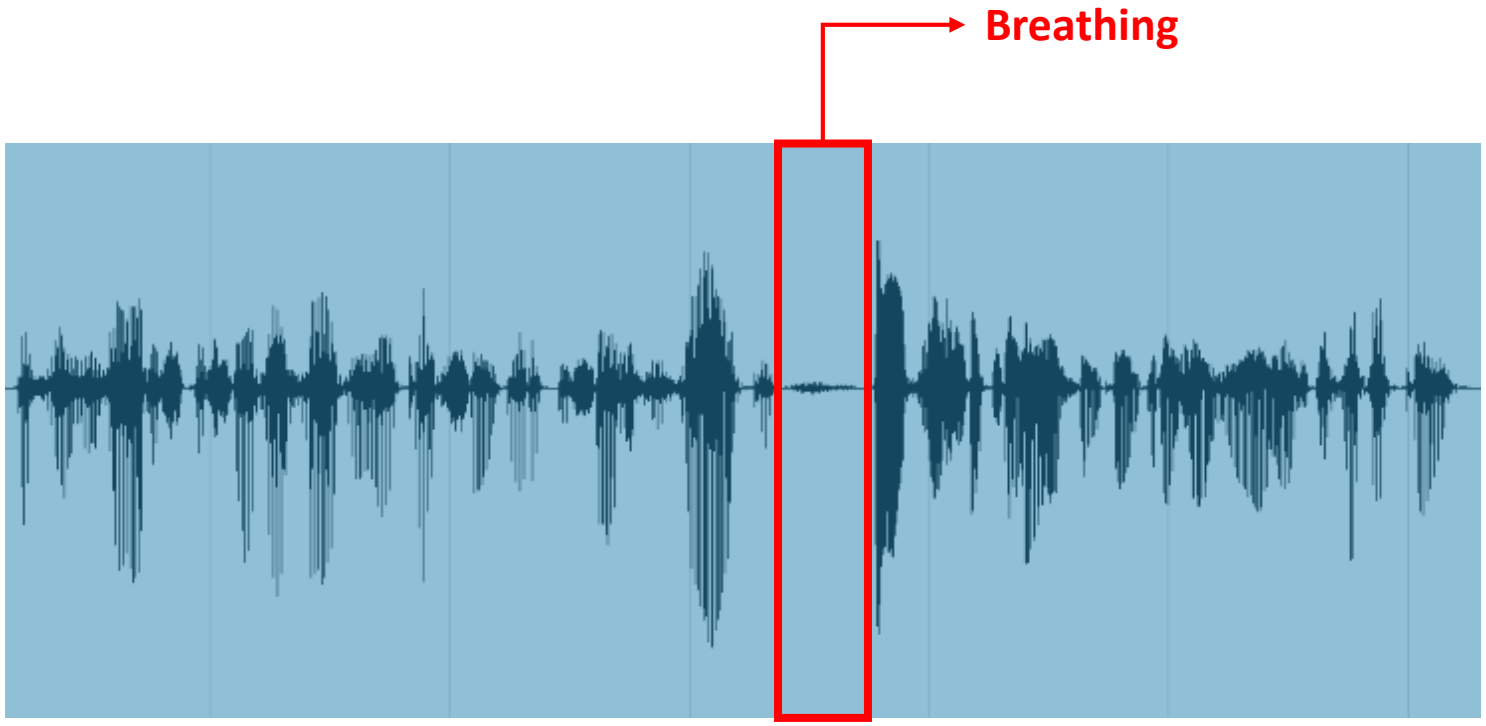
Part 1: LLMs Outside NLP Applications

2014



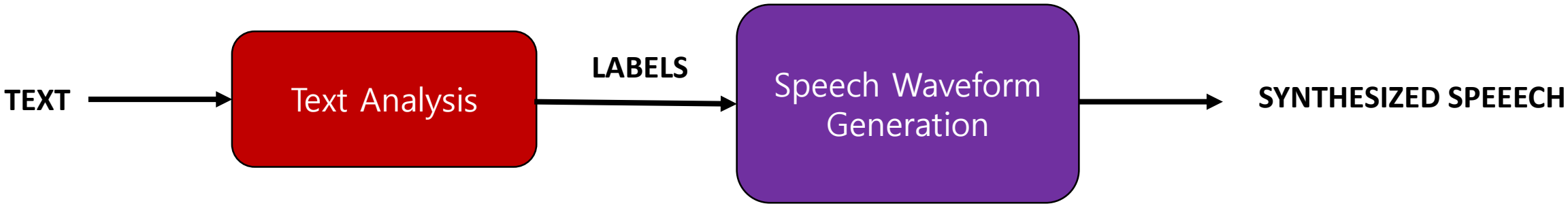
Transcription: Busco o que meu coração, minha alma e minha inteligência querem.

2014



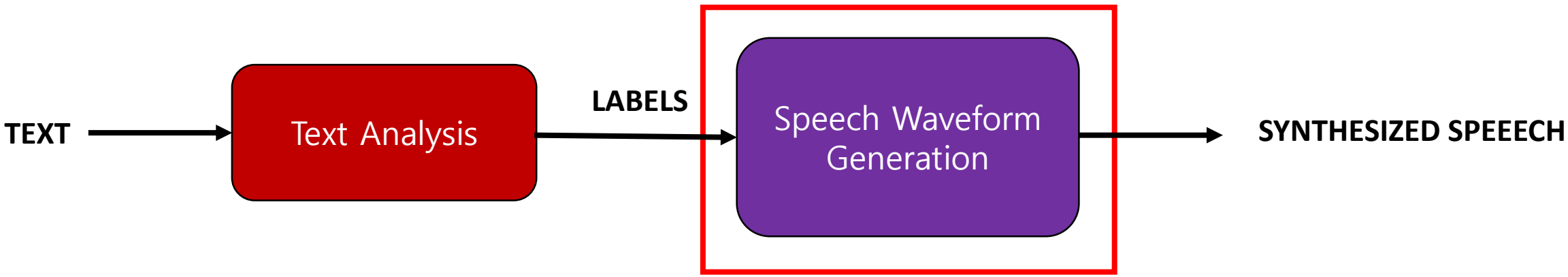
Transcription: Busco o que meu coração, minha alma e minha inteligência querem.

2015-2017



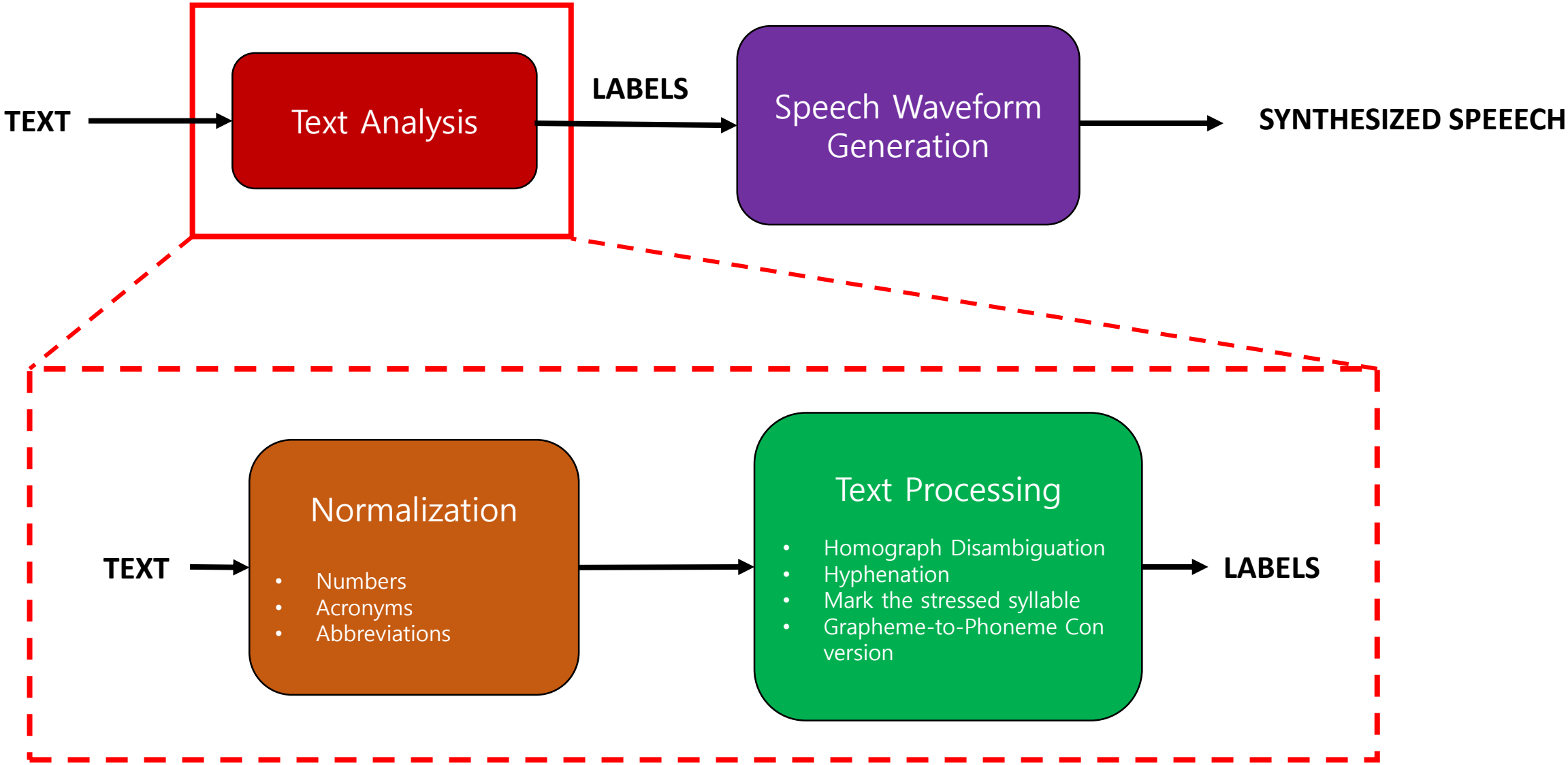
Basic architecture of a Text-to-Speech (TTS) system.

2015-2017



Basic architecture of a Text-to-Speech (TTS) system.

2015-2017


TEXT
Text Analysis
LABELS
Speech Waveform Generation
SYNTHESIZED SPEEECH
TEXT
Normalization
Numbers
Acronyms
Abbreviations
Text Processing
Homograph Disambiguation
Hyphenation
Mark the stressed syllable
Grapheme-to-Phoneme Con version
LABELS

Rule-based Grapheme-to-Phoneme (G2P) Conversion

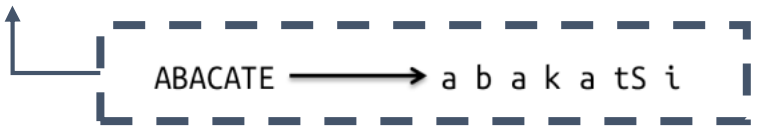
Table 3.3: Rules for the grapheme <a, b, c, d>.

Rule	Algorithm for the Grapheme <a>	Phone	Example
1	...<an><**Pont**>...	[a˜]	Iv<u>an</u>, Itapo<u>an</u>
2	...<am><**Pont**>...	[a˜w˜]	and<u>am</u>, cresç<u>am</u>
3	...<a (m,n)><**C**-h>...	[a˜]	<u>an</u>tena, <u>am</u>pola
4	...<a(**V_ton**)><m,n>...	[a˜]	c<u>a</u>ma, b<u>a</u>nho
5	...<â (m,n)><**C**-h>...	[a˜]	l<u>â</u>mpada, c<u>â</u>ntico
6	...<ão>...	[a˜w˜]	avi<u>ão</u>
7	...<ã,â>...	[a˜]	amanh<u>ã</u>, c<u>â</u>mara
8	...<á,à>...	[a]	Ant<u>á</u>rtica, <u>à</u>quela
9	...<a>...	[a]	<u>a</u>r<u>a</u>cnofobia

Rule	Algorithm for the Grapheme <b>	Phone	Example
1	...<b>...	[b]	<u>b</u>acate

Rule	Algorithm for the Grapheme <c>	Phone	Example
1	...<c><e,i>...	[s]	a<u>c</u>eitar, ja<u>c</u>into
2	...<ç>...	[s]	almo<u>ç</u>o
3	...<c h>...	[S]	a<u>ch</u>o
4	...<c>...	[k]	<u>c</u>laro

Rule	Algorithm for the Grapheme <d>	Phone	Example
1	...<d><i,[i]>...	[dZ]	<u>d</u>ia, tar<u>d</u>e
2	...<d><**C**-r,l>...	[dZ]	a<u>d</u>vogado
3	...<d><**Pont**>...	[dZ]	Raid
4	...<d>...	[d]	<u>d</u>ote



Luiz Felipe Vecchietti. “Comparison between rule-based and data-driven natural language processing algorithms for Brazilian Portuguese speech synthesis ." (2017).

Sequence-to-Sequence Models

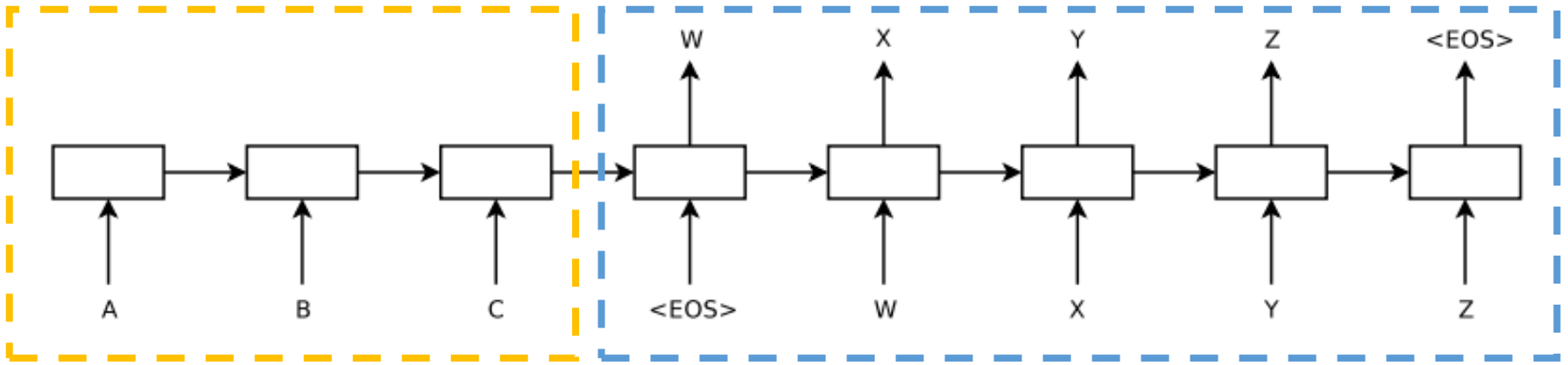
2014

Sequence to Sequence Learning with Neural Networks

Ilya Sutskever
Google
ilyasu@google.com

Oriol Vinyals
Google
vinyals@google.com

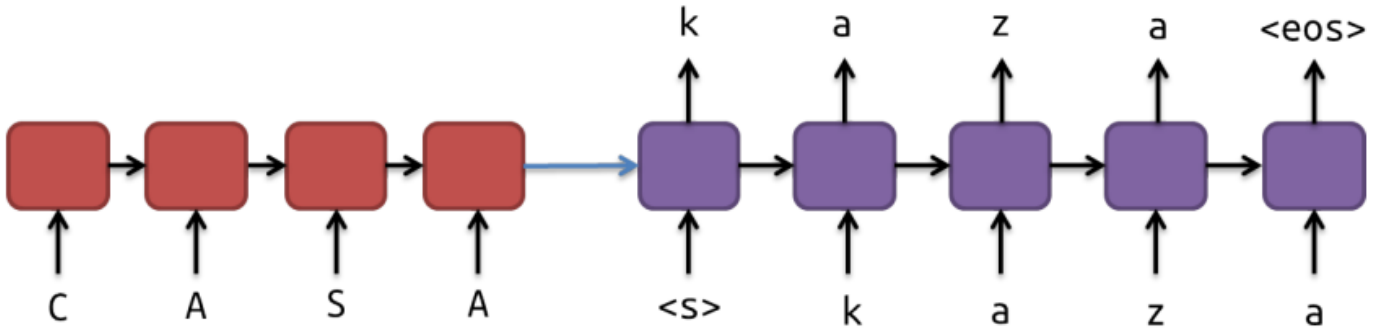
Quoc V. Le
Google
qvl@google.com



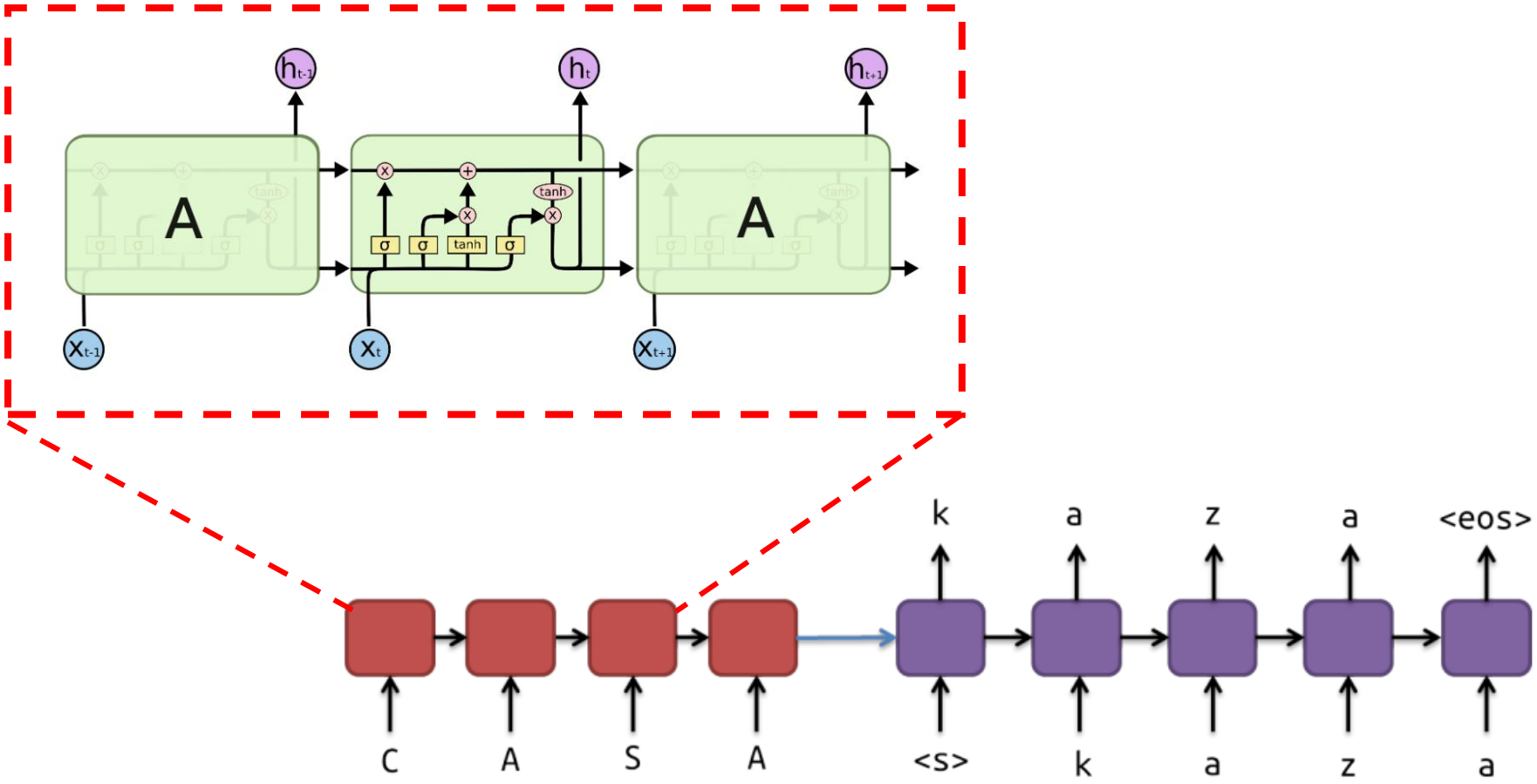
Encoder

Decoder

G2P Conversion modeled as Sequence-to-Sequence

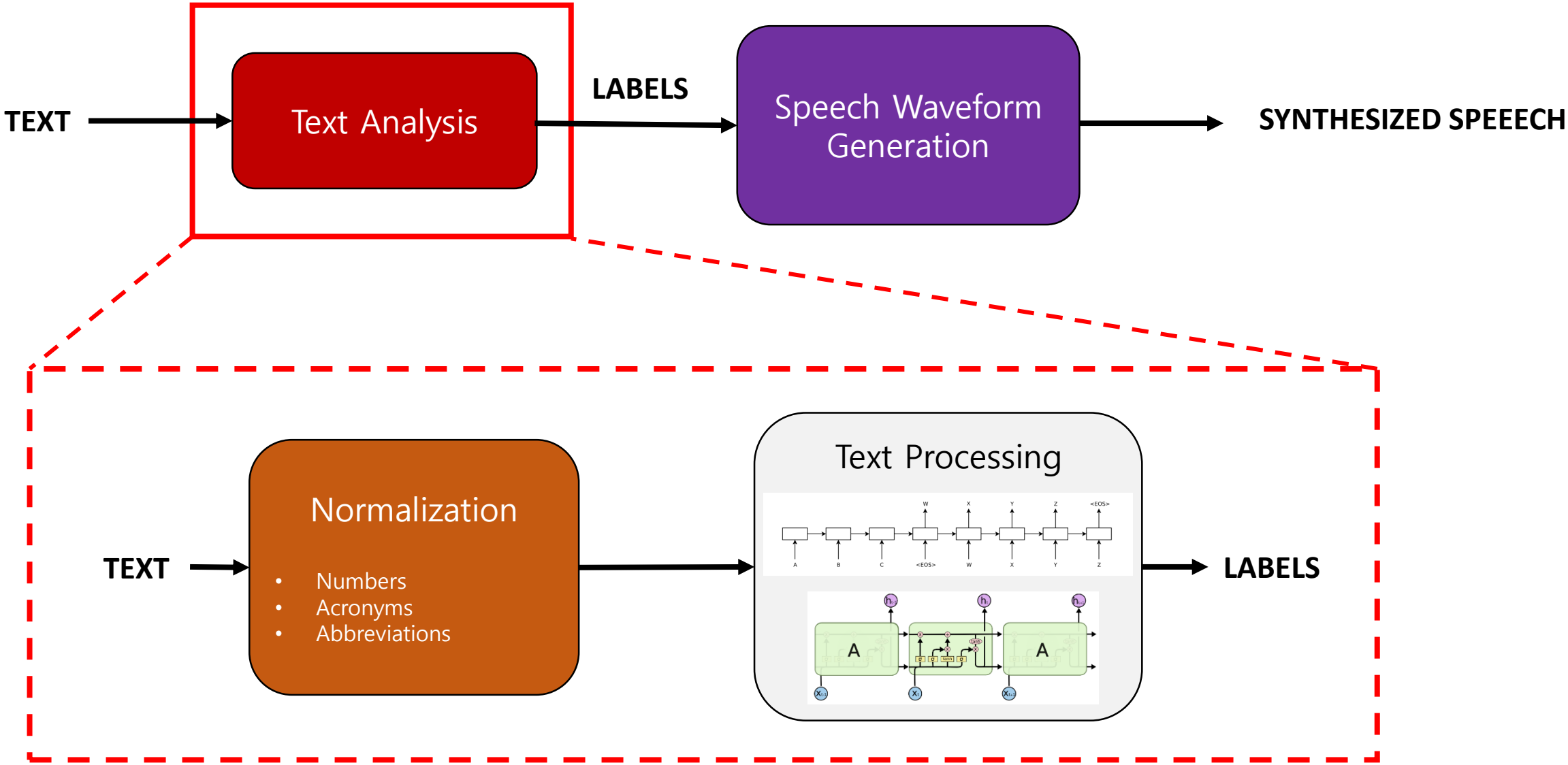


G2P Conversion modeled as Sequence-to-Sequence



https://colah.github.io/posts/2015-08-Understanding-LSTMs/

2015-2017



Luiz Felipe Vecchietti. “Comparison between rule-based and data-driven natural language processing algorithms for Brazilian Portuguese speech synthesis ." (2017).

Key Takeaways

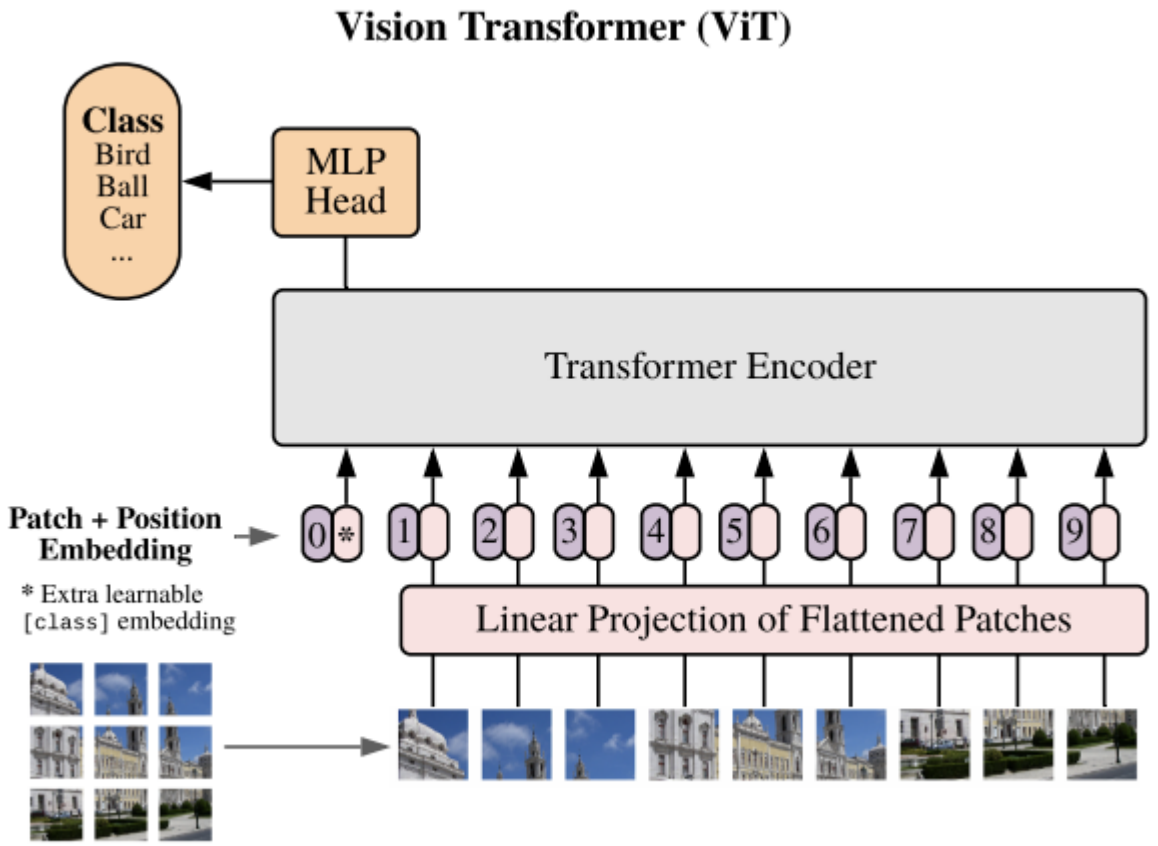
- Algorithms like LSTMs and Transformers in encoder-only or encoder-decoder fram eworks are very flexible and can be applied to a variety of different problems.
- Importance of flexibility, generalizability, and computational efficiency when develo ping new algorithms.
- If you can tokenize your data and model it as a sequence, you can use these pow erful architectures for generation, making predictions in downstream tasks, and lea rn meaningful context-based representations.

Popular Applications in Other Domains: Generating Code

Copilot

```
// start a server with the echo library
// take the address as a parameter
// add the usual middlewares and an inline health check endpoint
func Start(addr string) error {
    e := echo.New()
    e.Use(middleware.Logger())
    e.Use(middleware.Recover())
    e.GET("/health", func(c echo.Context) error {
        return c.String(http.StatusOK, "OK")
    })
    return e.Start(addr)
}
```

Popular Applications in Other Domains: Vision Transformers



Applications in Other Domains: Demographic Research

Article | Published: 18 December 2023

Using sequences of life-events to predict human lives

Germans Savcisens, Tina Eliassi-Rad, Lars Kai Hansen, Laust Hvas Mortensen, Lau Lilleholt, Anna Rogers, Ingo Zettler & Sune Lehmann

Nature Computational Science **4**, 43–56 (2024) | Cite this article

Applications in Other Domains: Demographic Research

- Dataset

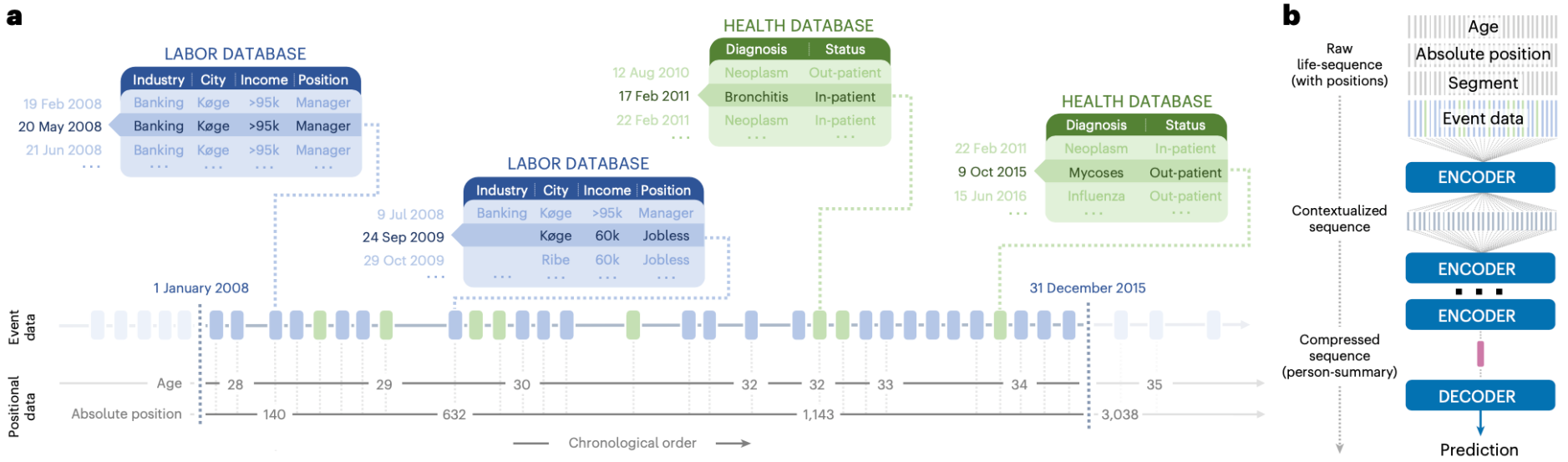
Dataset

We worked with the Labour Market Account (AMRUN)[24] and National Patient Registry (LPR) datasets[6,27]. Within the Labour Market Account dataset are event data for every resident of Denmark. For Danish residents who have been in contact with secondary healthcare services, primarily hospitals, the events are recorded in the National Patient Registry. We limited ourselves to data recorded in the period from 2008 until the end of 2015. The datasets were pseudonymized before our work by de-identifying addresses, Central Person Register numbers (CPRs) and names. The data are stored within Statistics Denmark, and all access/use of data is logged.

The total number of residents in the filtered dataset was 3,252,086 (1,630,082 men and 1,622,004 women). For our research, we chose people who (1) were alive and lived in Denmark on 31 December 2015, (2) had at least 12 records in the labor data during 2015 (corresponds to 12 incomes over one year, for example salary, pension and so on; we did not set requirements on the health-set, as not every resident had any records in the health dataset), (3) had consistent sex and birthday attributes over the whole residency period, (4) were between 25 and 70 years old on 31 December 2015.

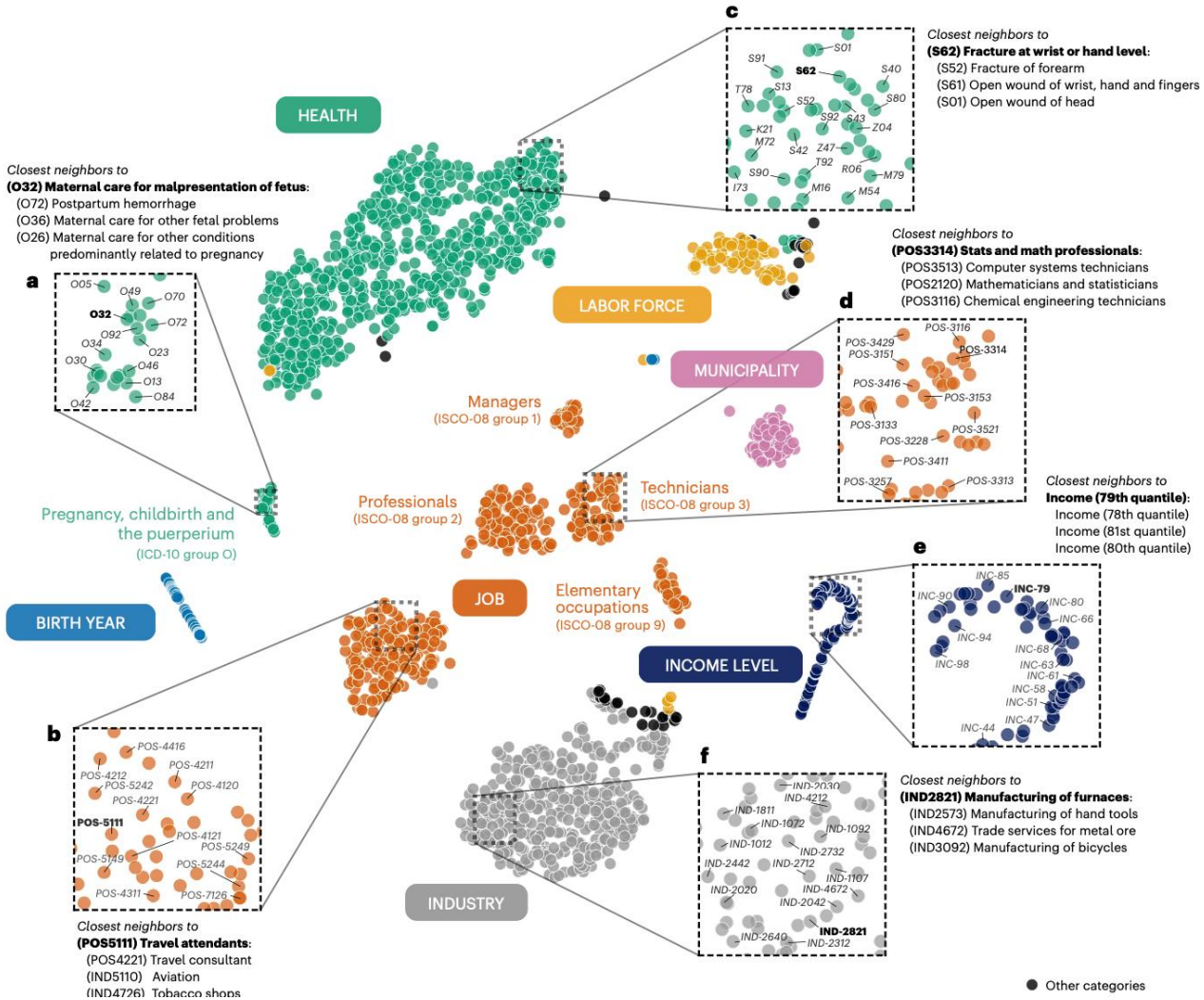
Applications in Other Domains: Demographic Research

- life2vec model



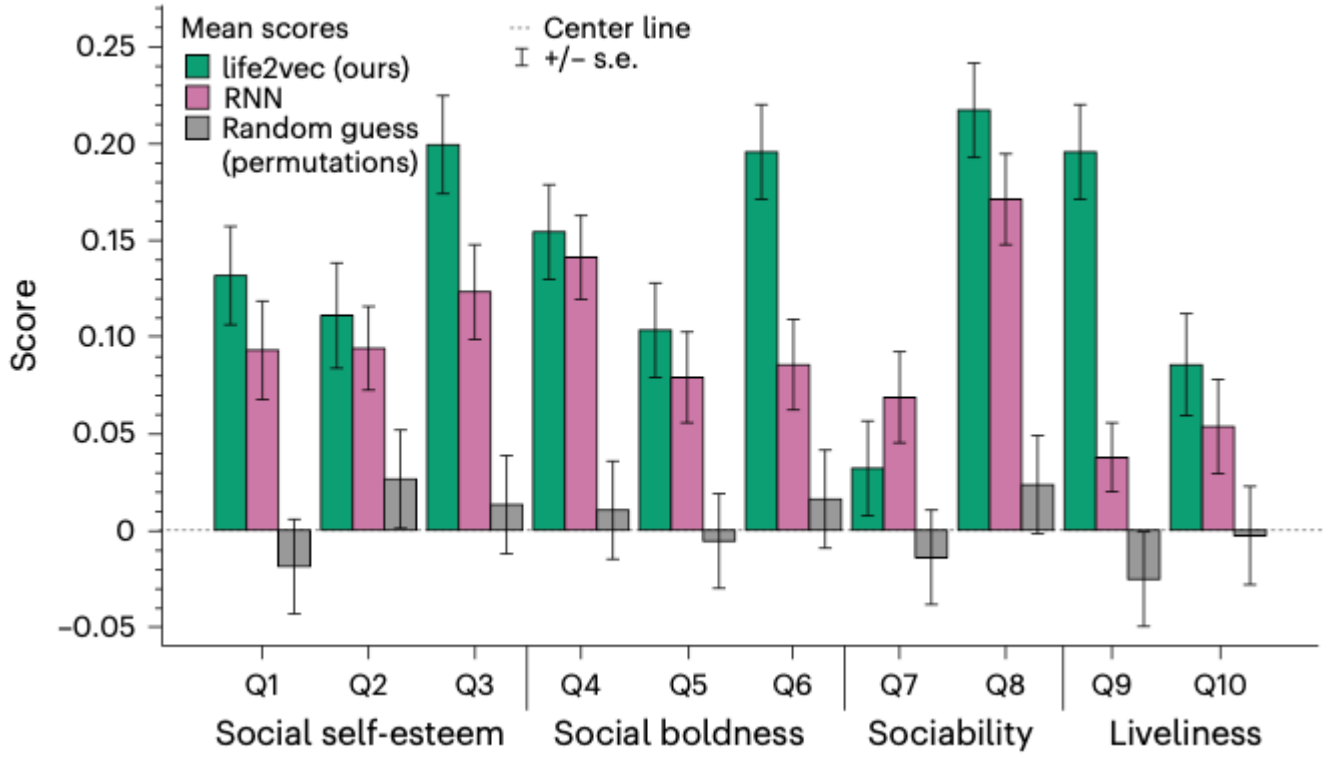
Applications in Other Domains: Demographic Research

- The life2vec concept space



Applications in Other Domains: Demographic Research

- Predicting personality nuances



Personality nuances prediction task. The personality nuances prediction task is an ordinal classification task where labels correspond to the five levels of agreement with a particular item/statement. We predict the response to ten different items corresponding to the extraversion facet (Fig. 5):

1. I feel that I am an unpopular person,
2. I feel reasonably satisfied with myself overall,
3. I sometimes feel that I am a worthless person,
4. When I'm in a group of people, I'm often the one who speaks on behalf of the group,
5. In social situations, I'm usually the one who makes the first move,
6. I rarely express my opinions in group meetings,
7. The first thing that I always do in a new place is to make friends,
8. I prefer jobs that involve active social interaction to those that involve working alone,
9. Most people are more upbeat and dynamic than I generally am,
10. On most days, I feel cheerful and optimistic.

Questions 1–3 correspond to social self-esteem, 4–6 to social boldness (feeling comfortable in diverse social settings), 7–8 to sociability, or enjoyment of social interactions, and, finally, 9–10 evaluates liveliness (which includes enthusiasm and overall energy)[72].

Applications in Other Domains: Music Generation

Jukebox: A Generative Model for Music

Prafulla Dhariwal [*1] **Heewoo Jun** [*1] **Christine Payne** [*1] **Jong Wook Kim** [1] **Alec Radford** [1] **Ilya Sutskever** [1]

Applications in Other Domains: Music Generation

- Abstract

Abstract

We introduce Jukebox, a model that generates music with singing in the raw audio domain. We tackle the long context of raw audio using a multi-scale VQ-VAE to compress it to discrete codes, and modeling those using autoregressive Transformers. We show that the combined model at scale can generate high-fidelity and diverse songs with coherence up to multiple minutes. We can condition on artist and genre to steer the musical and vocal style, and on unaligned lyrics to make the singing more controllable. We are releasing thousands of non cherry-picked samples, along with model weights and code.

Applications in Other Domains: Music Generation

- Learning the Codebook/Tokens

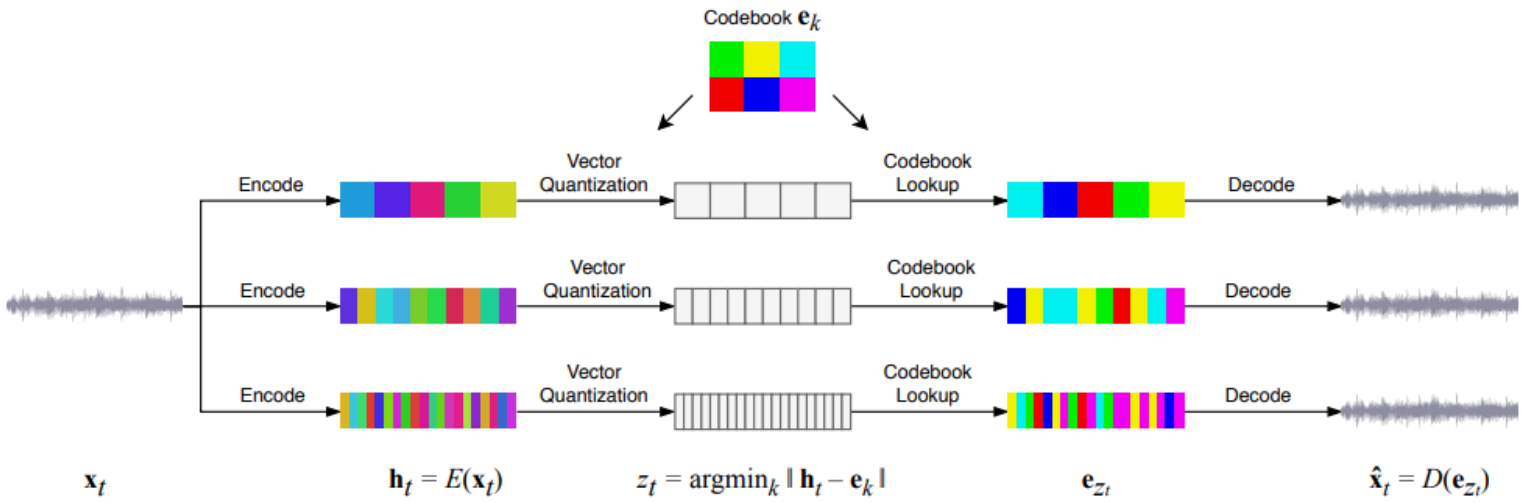
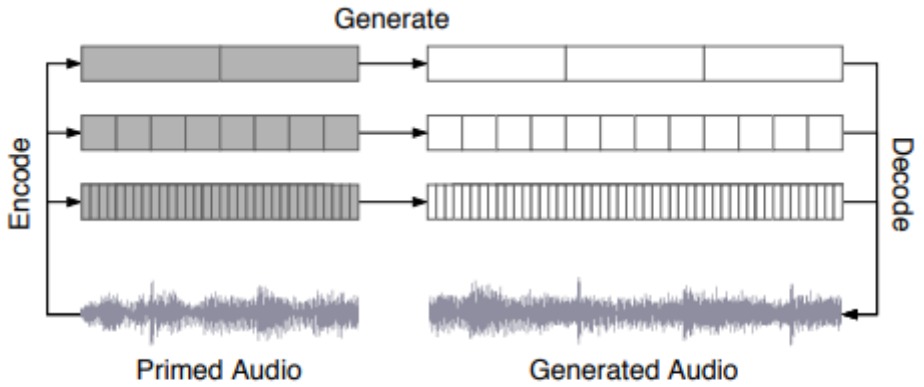


Figure 1. We first train three separate VQ-VAE models with different temporal resolutions. At each level, the input audio is segmented and encoded into latent vectors $\mathbf{h}_t$, which are then quantized to the closest codebook vectors $\mathbf{e}_{z_t}$. The code z_t is a discrete representation of the audio that we later train our prior on. The decoder takes the sequence of codebook vectors and reconstructs the audio. The top level learns the highest degree of abstraction, since it is encoding longer audio per token while keeping the codebook size the same. Audio can be reconstructed using the codes at any one of the abstraction levels, where the least abstract bottom-level codes result in the highest-quality audio, as shown in Figure 4. For the detailed structure of each component, see Figure 7.

Applications in Other Domains: Music Generation

- Music Generation



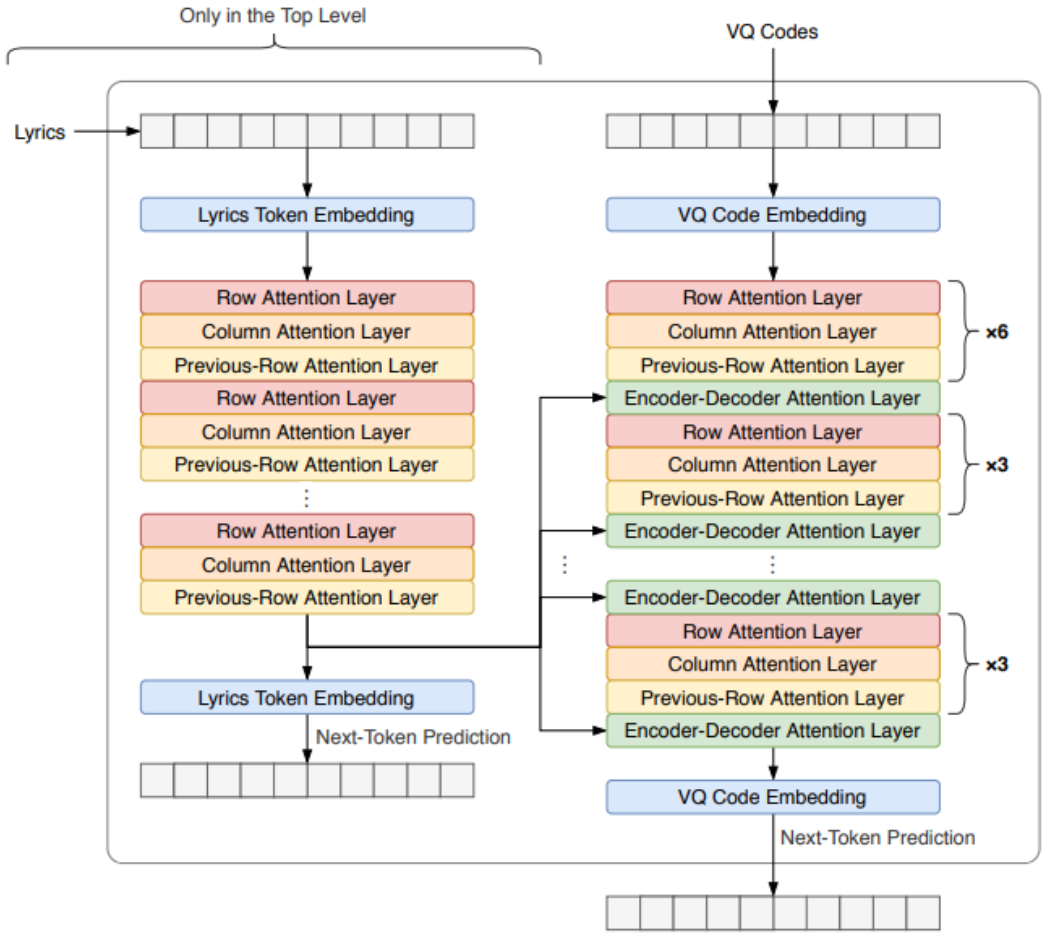
(c) **Primed sampling**: The model can generate continuations of an existing audio signal by converting it into the VQ-VAE codes and sampling the subsequent codes in each level.

Applications in Other Domains: Music Generation

- Optimizations in the Transformer architecture

A. Scalable Transformer

We make the Sparse Transformer (Child et al., 2019) more scalable and easier to implement by a few small changes. We implement a simpler attention pattern that has the same performance without needing custom kernels to implement. We simplify the initialization by using the same initalization scale in the whole model without rescaling the weights based on fan-in and depth, and we optimize the memory footprint with fully half-precision training, i.e. storing the model weights, gradients and the optimizer states in half precision and performing computations in half precision as well. To cope with the narrower dynamic range of the fp16 format, we use dynamic scaling of the gradient and Adam optimizer states.



Summary of Part 1: LLMs Outside NLP Applications

- Concepts learned for LLM Engineering can be applied for different domains.
- Different applications offer different challenges from LLM Engineering perspectives that should be discussed with especialists of the domain of interest (importance of interdisciplinary research).
- Previous examples show creative ways to adapt recent architectures, e.g., changing positional embeddings.

Exercise 1: LLMs Outside NLP Applications

- 30 minutes: search and think about possible applications outside the NLP domain.
- 15 minutes: sharing and discussion

- Main Task:

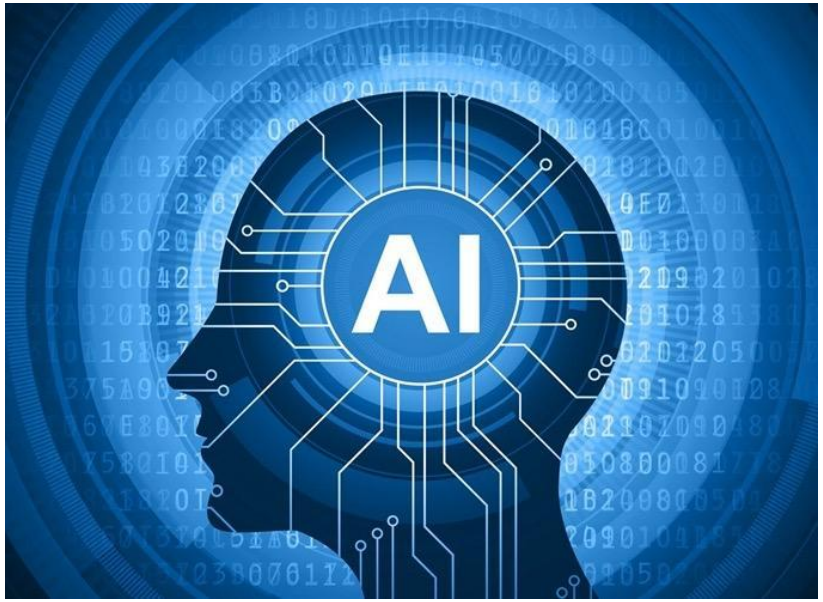
1. Search applications outside the NLP domain that apply Transformer architectures and have creative ways to tokenize their data.
2. Think about concepts for LLM Engineering methods that can be applied for these applications and if the same challenges that applies for the NLP domain also appl ies to other domains.
3. Feel free to come up with your own ideas of applications!

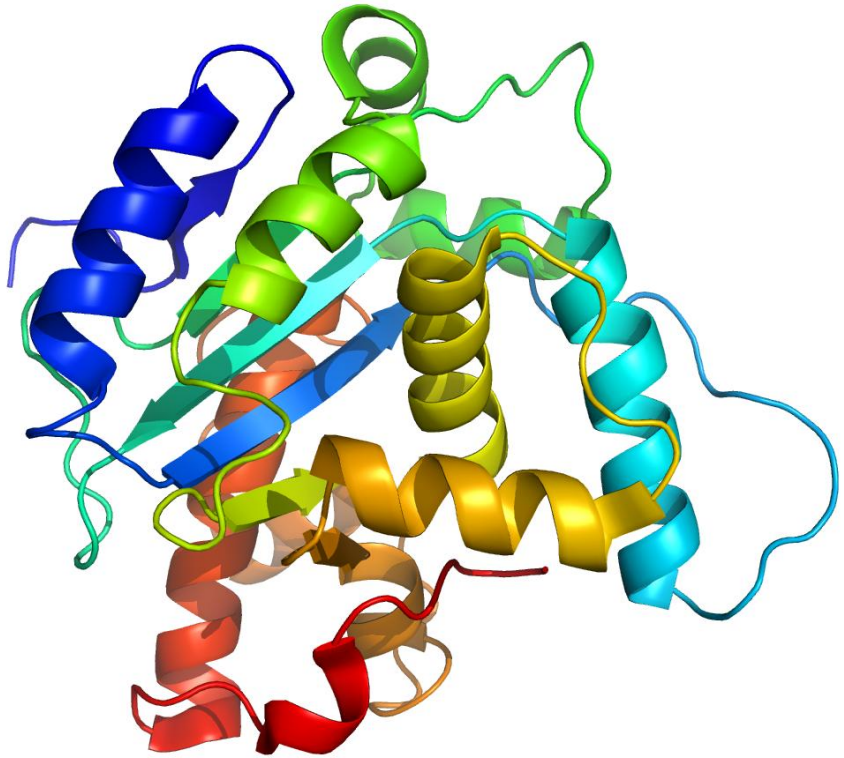
- During Sharing and Discussion: (if you feel comfortable)

1. Present the Application You Searched
2. How they tokenize the data and which architecture they use.

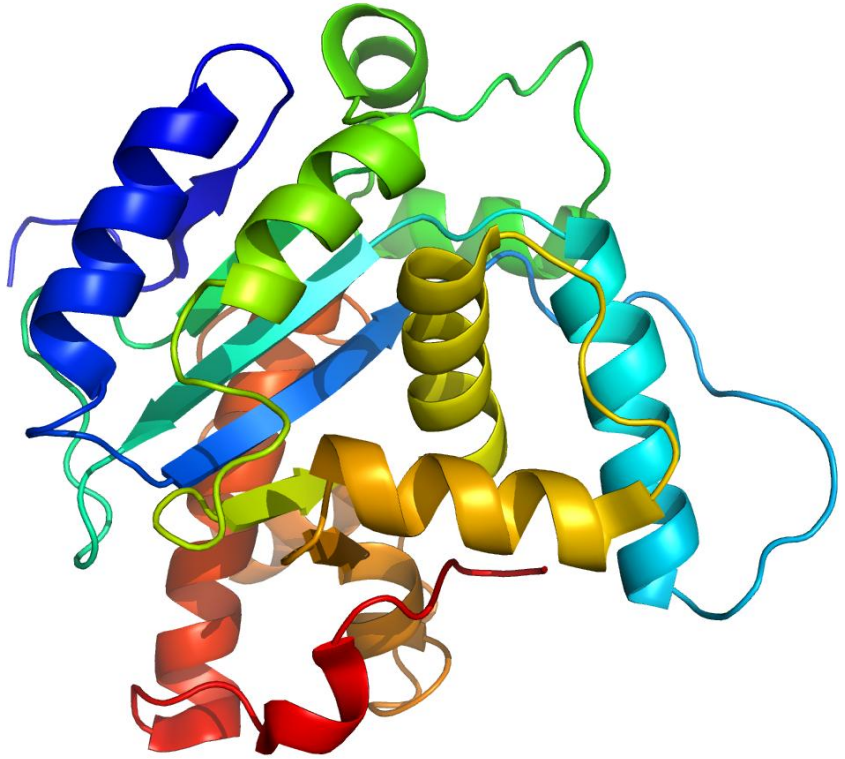
Part 2: A Case Study on Protein Language Models

2021-2025



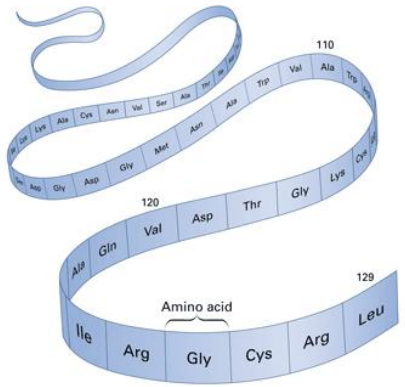


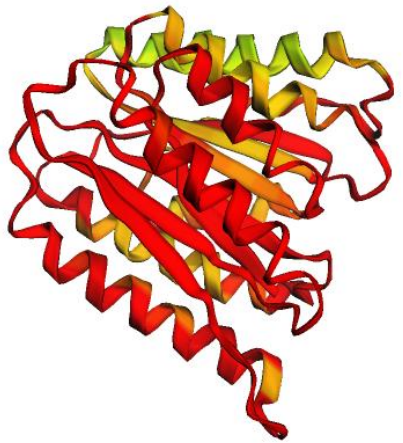
What do you know about proteins?



Protein

GSPFQKRLAKFNTNFNRCYGTCLKIAGCAL



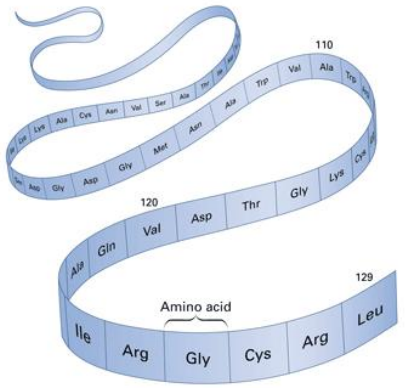


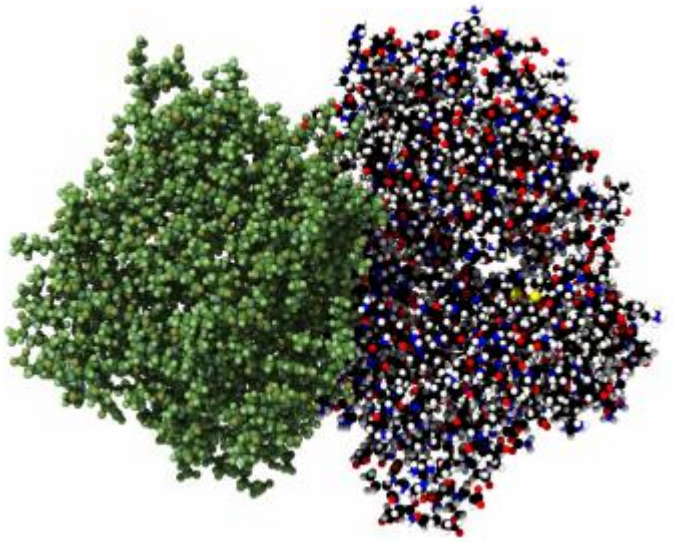
Proteins are large complex molecules made up of chains of amino acids

Proteins are responsible for essentially all biological processes. These processes depends on its unique 3 D structure.

Protein

GSPFQKRLAKFNTNFNRCYGTCLKIAGCAL



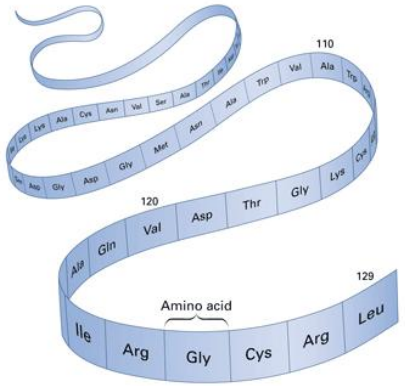


Proteins are large complex molecules made up of chains of amino acids

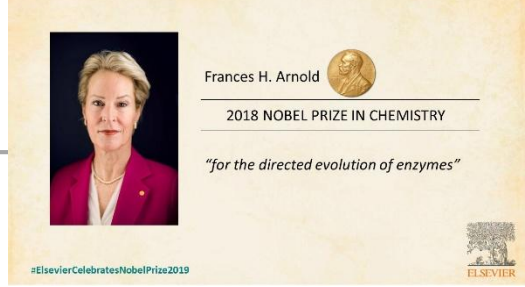
Proteins are responsible for essentially all biological processes. These processes depends on its unique 3 D structure.

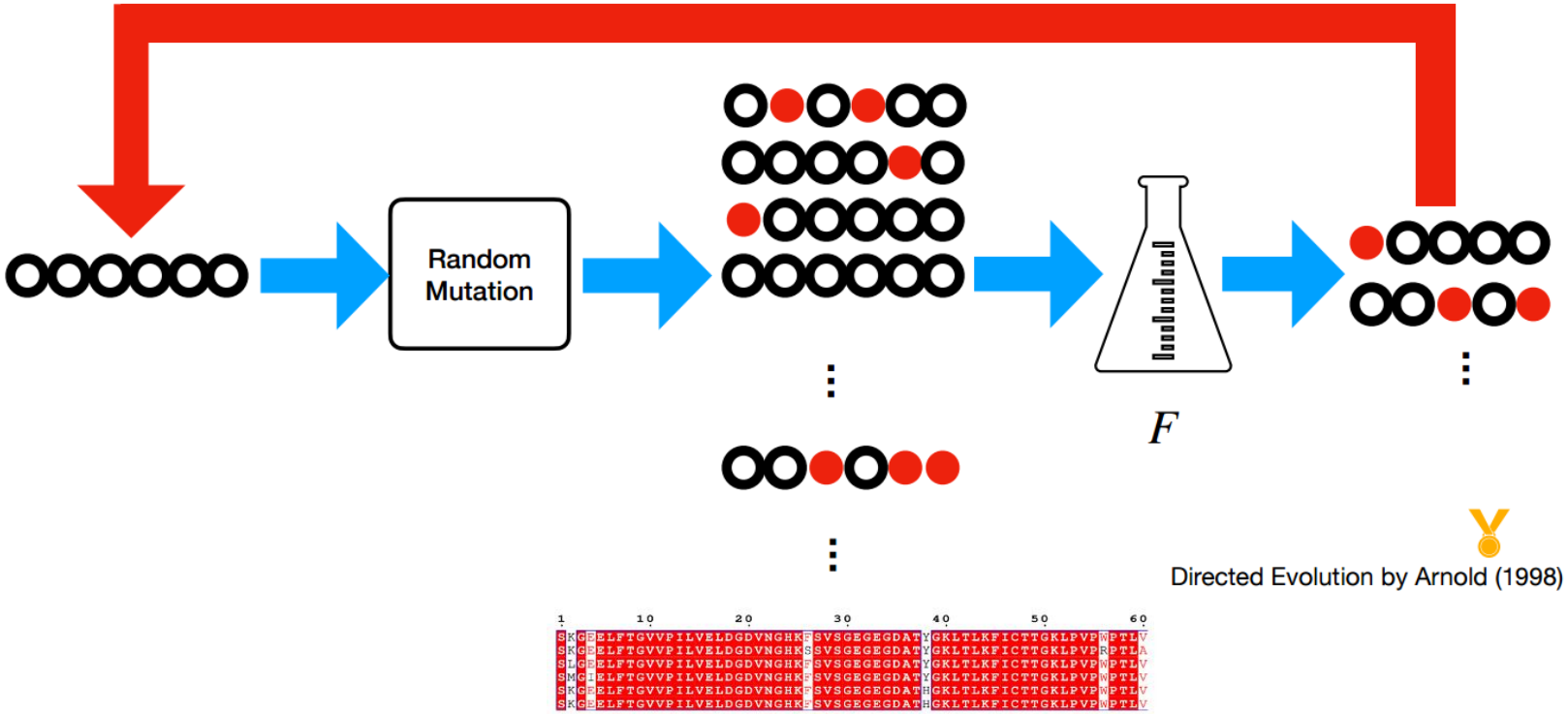
Protein Engineering: Directed Evolution

GSPFQKRLAKFNTNFNRCYGTCLKIAGCAL



Directed Evolution

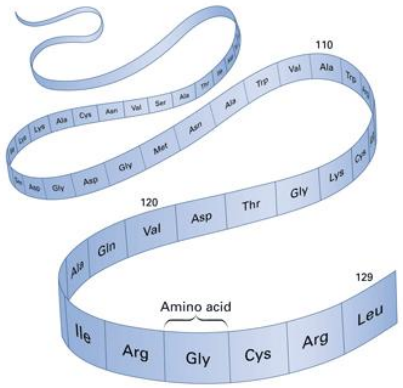




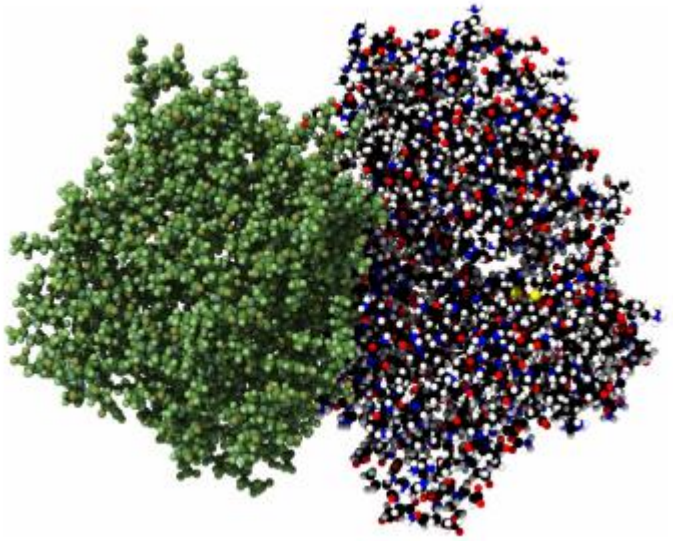
Directed Evolution

Protein Structure Prediction

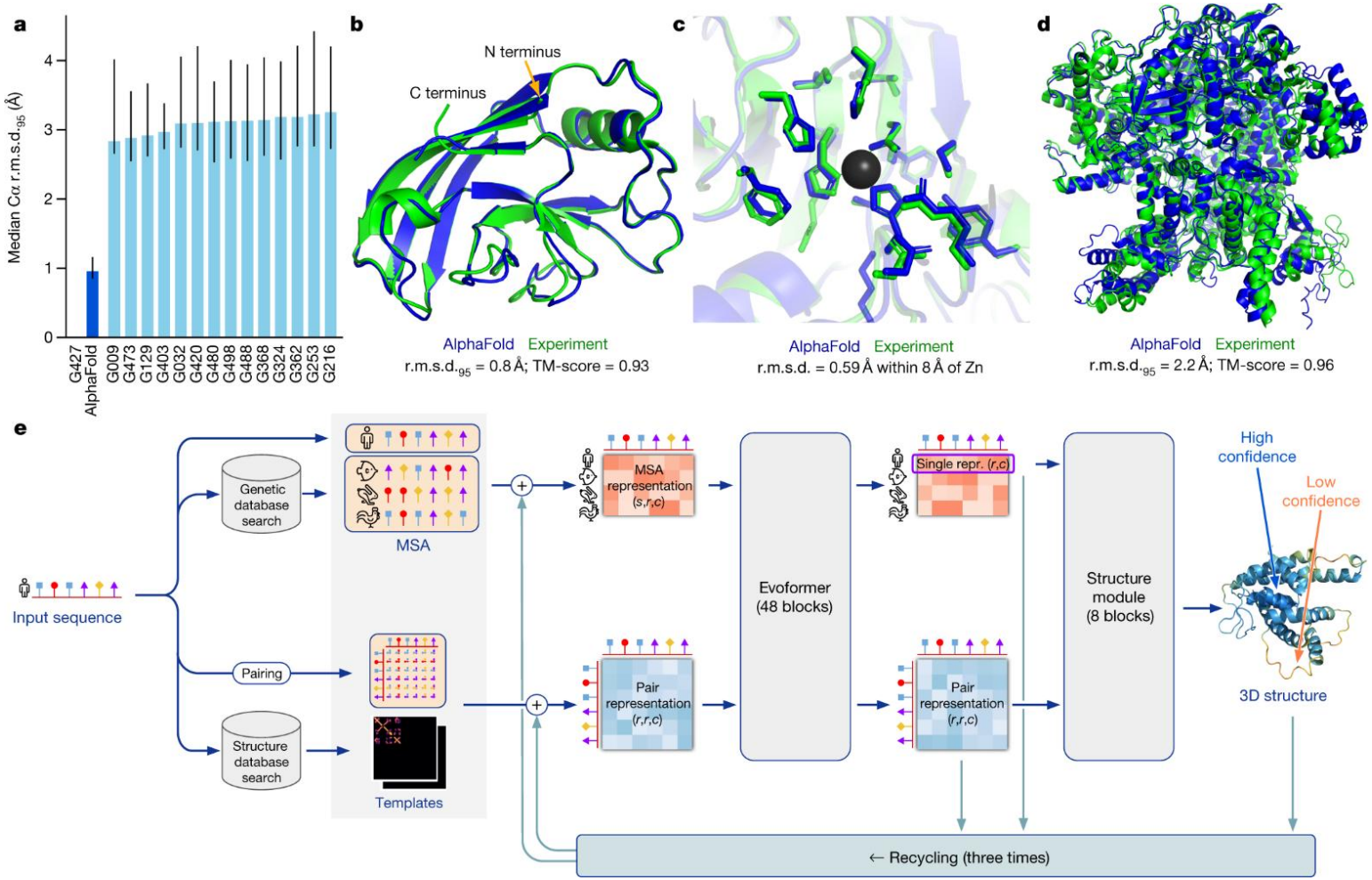
GSPFQKRLAKFNTNFNRCYGTCLKIAGCAL

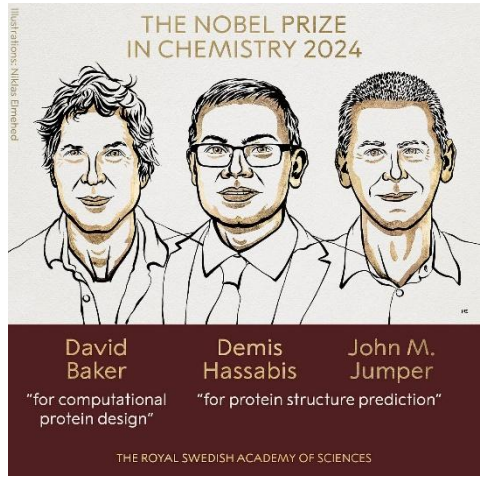


AI Model

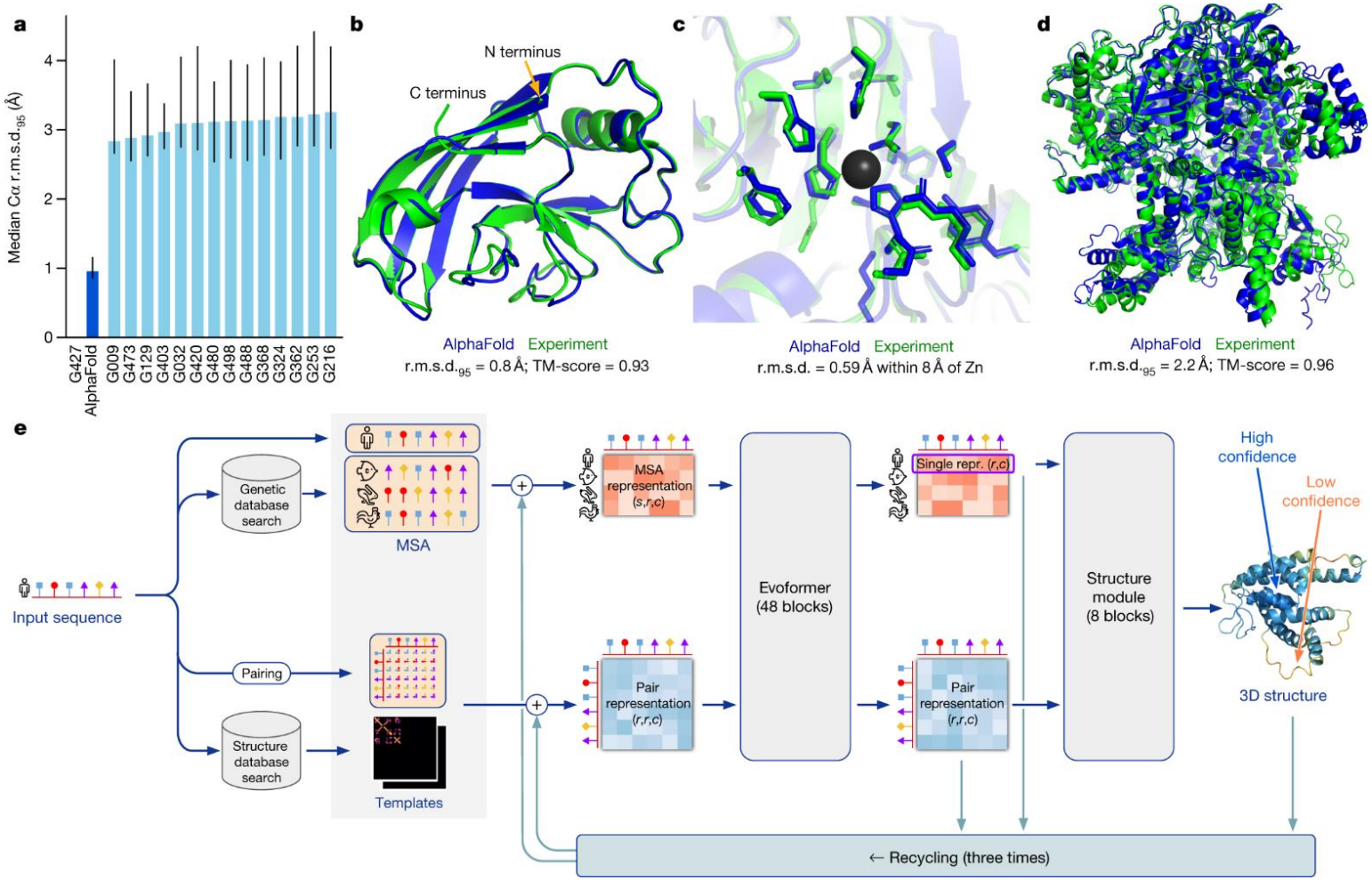


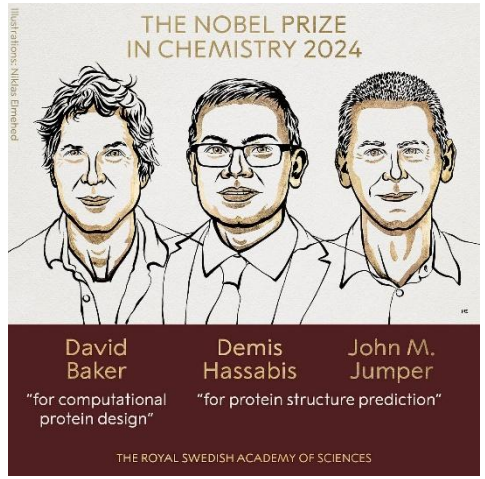
Protein Structure Prediction: AlphaFold2

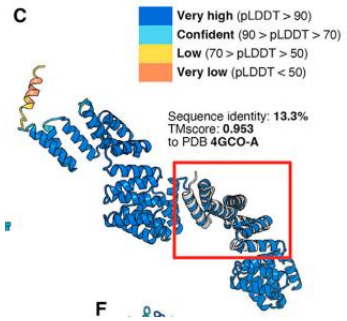




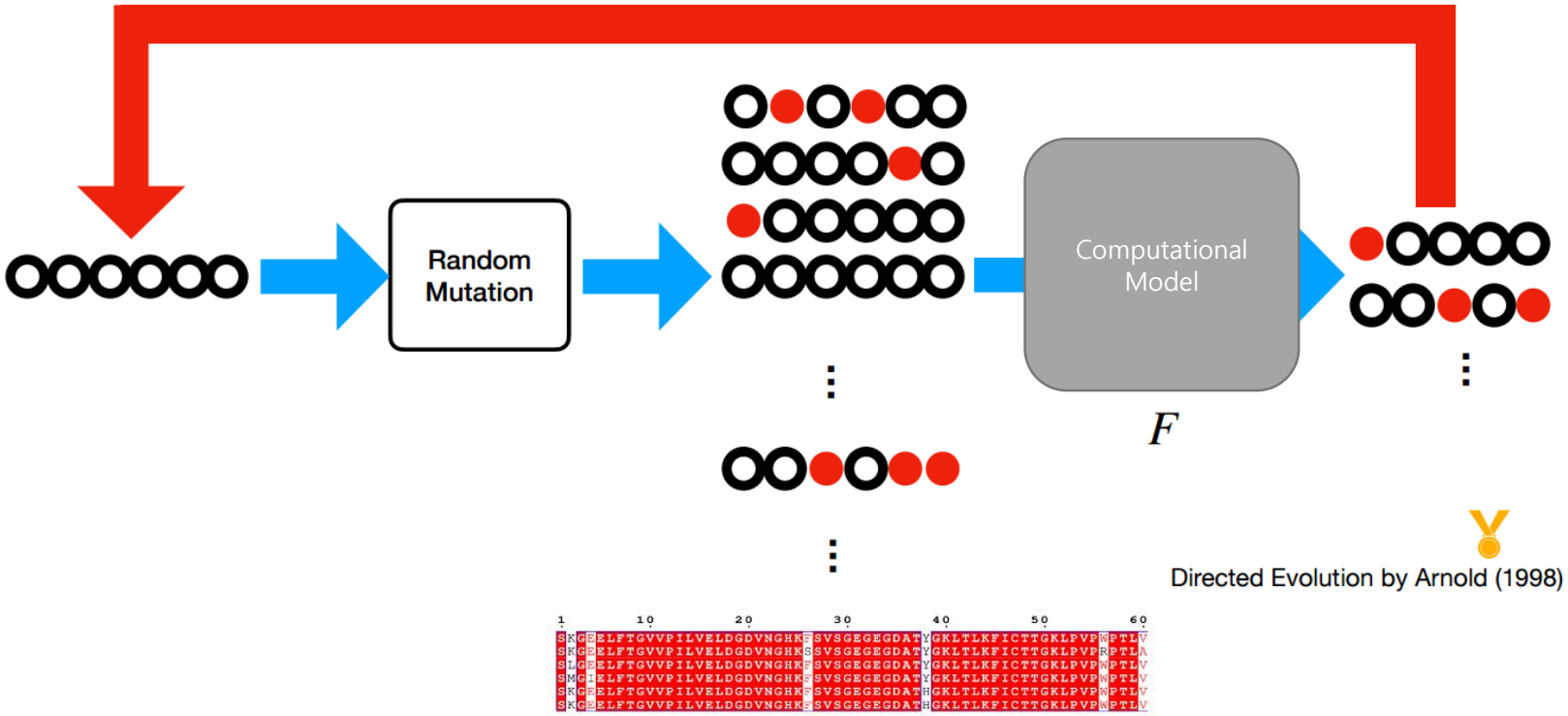
Protein Structure Prediction: AlphaFold2



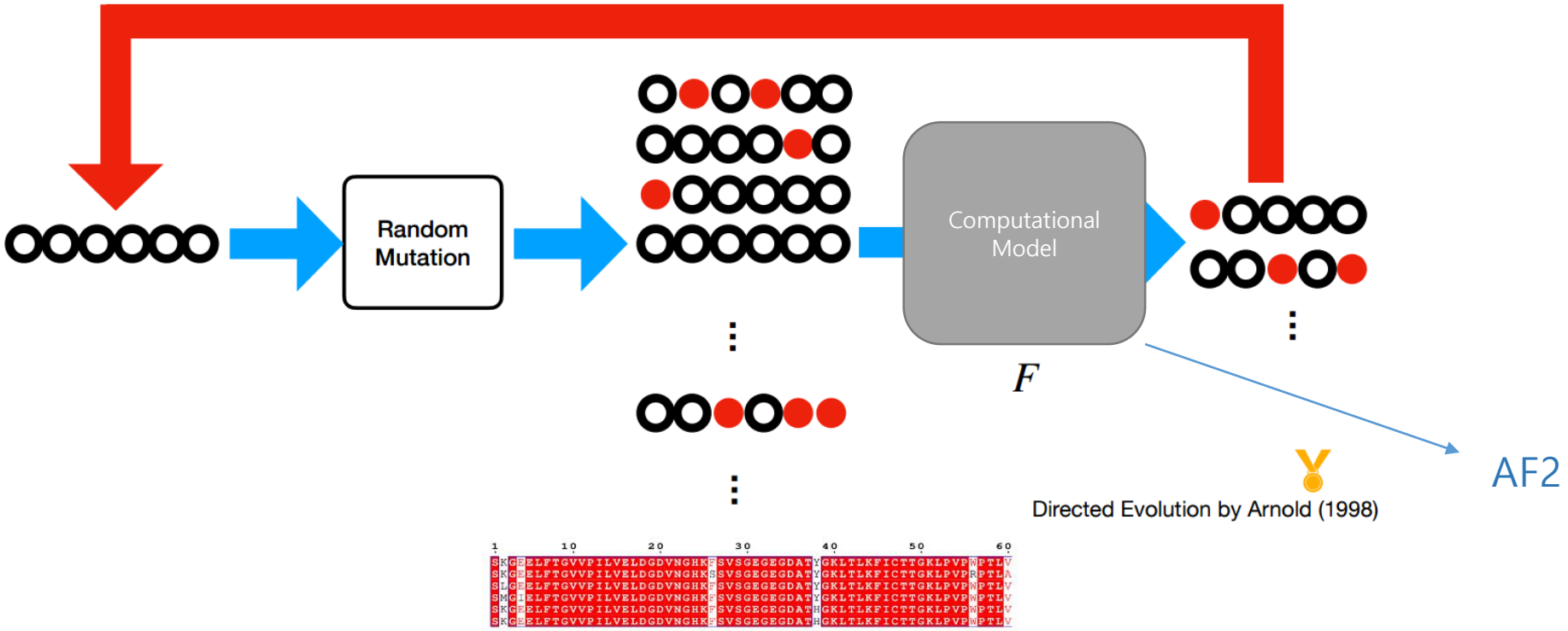




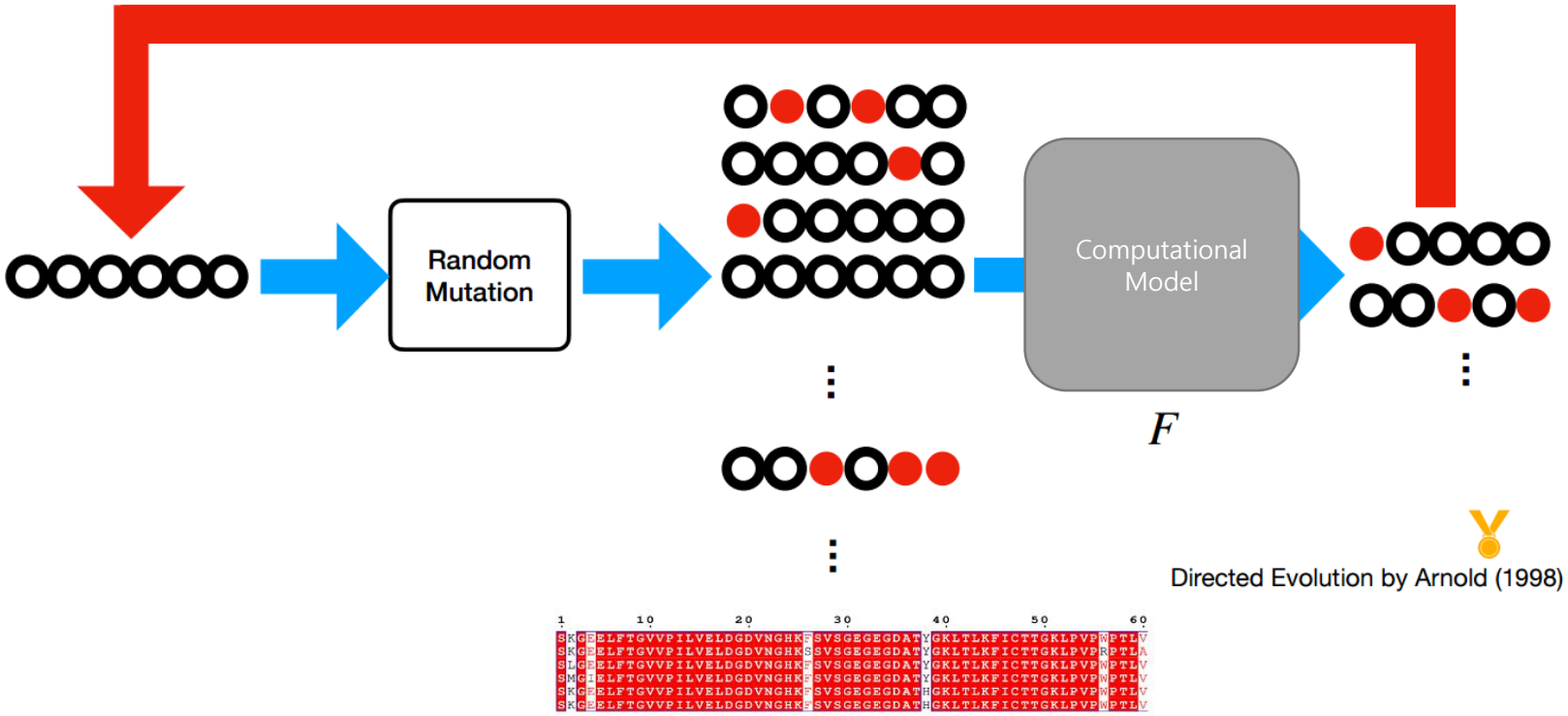
Directed Evolution



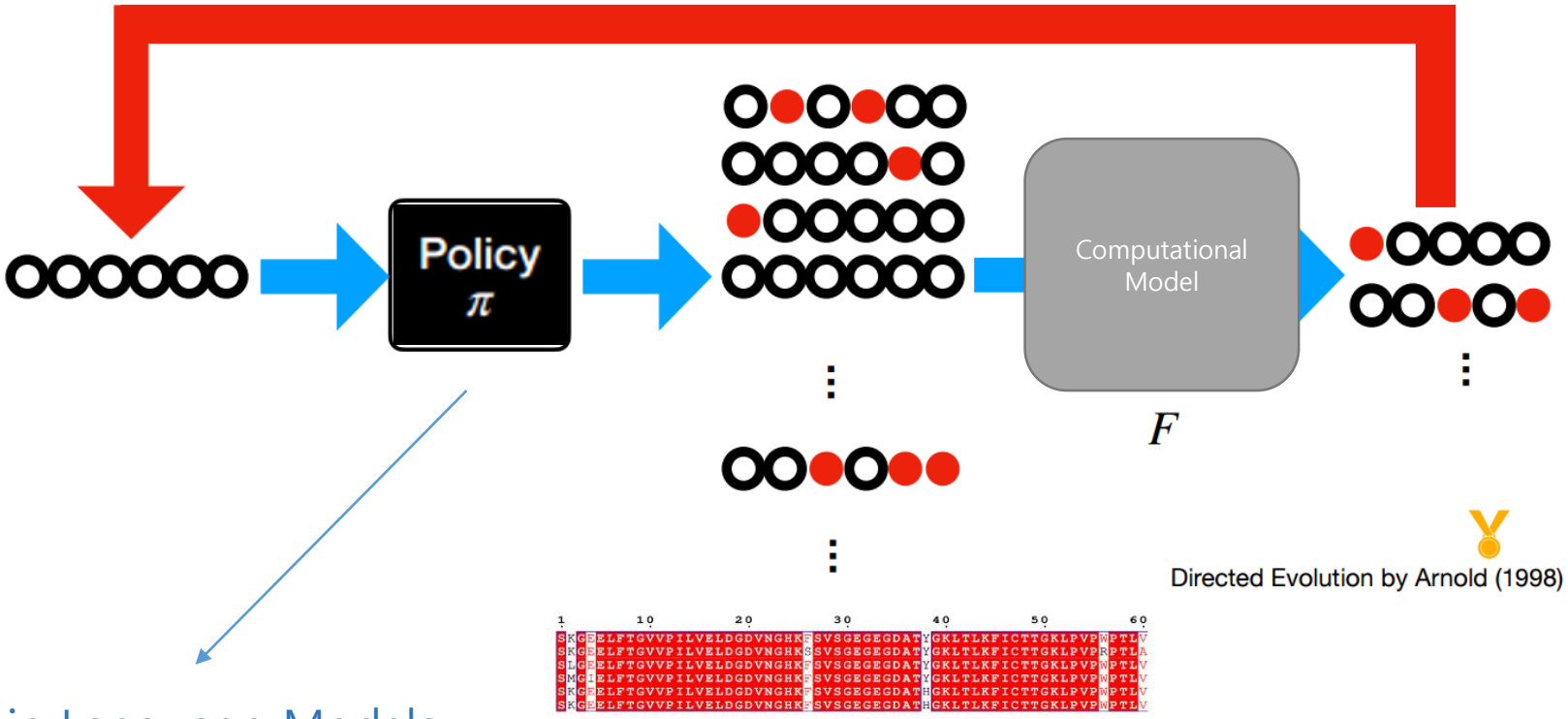
Directed Evolution



Is there a more effective way than to do random mutations?

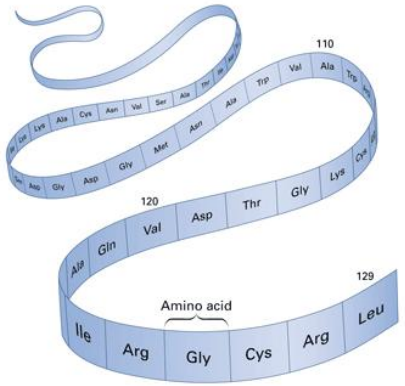


Protein Engineering



Protein Sequence

GSPFQKRLAKFNTNFNRCYGTCLKIAGCAL



Protein Sequence Data

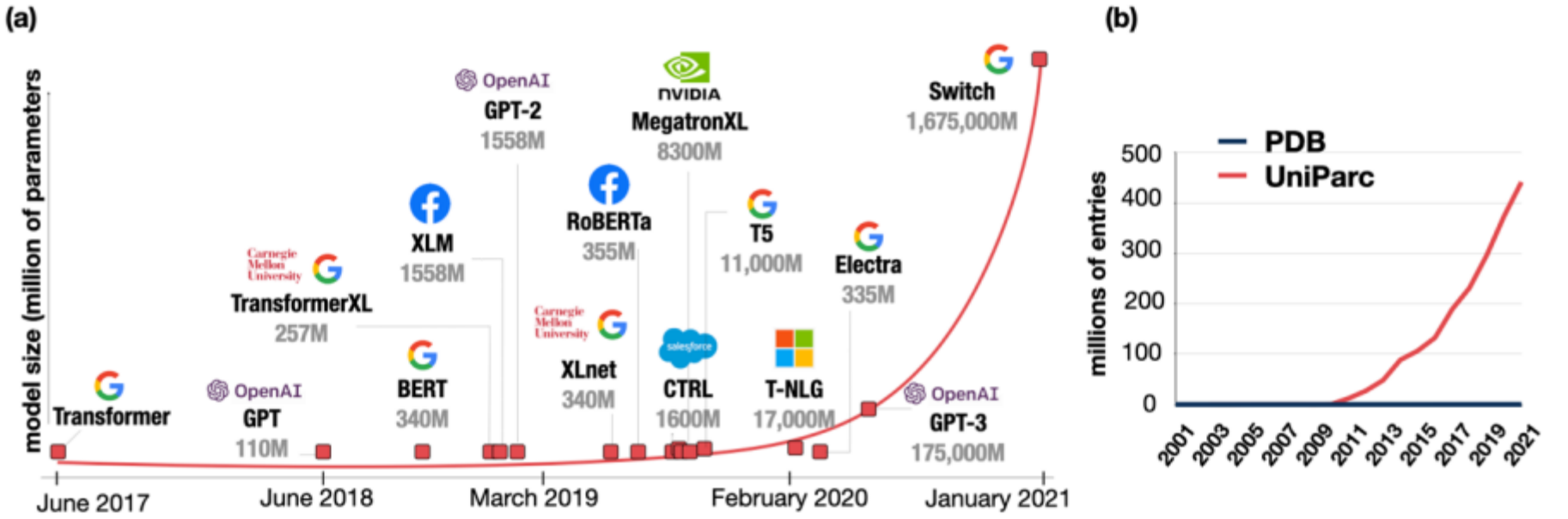
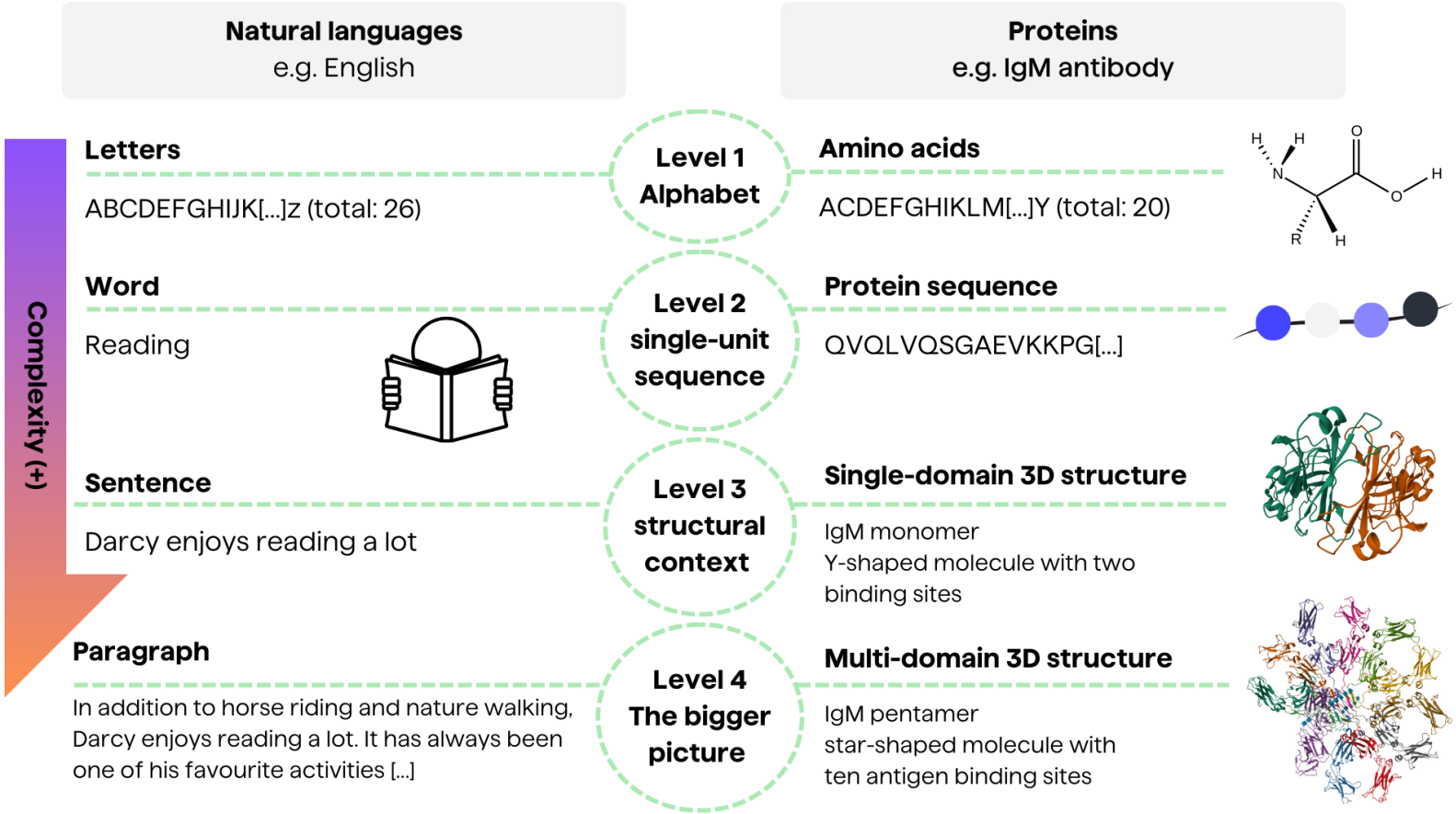


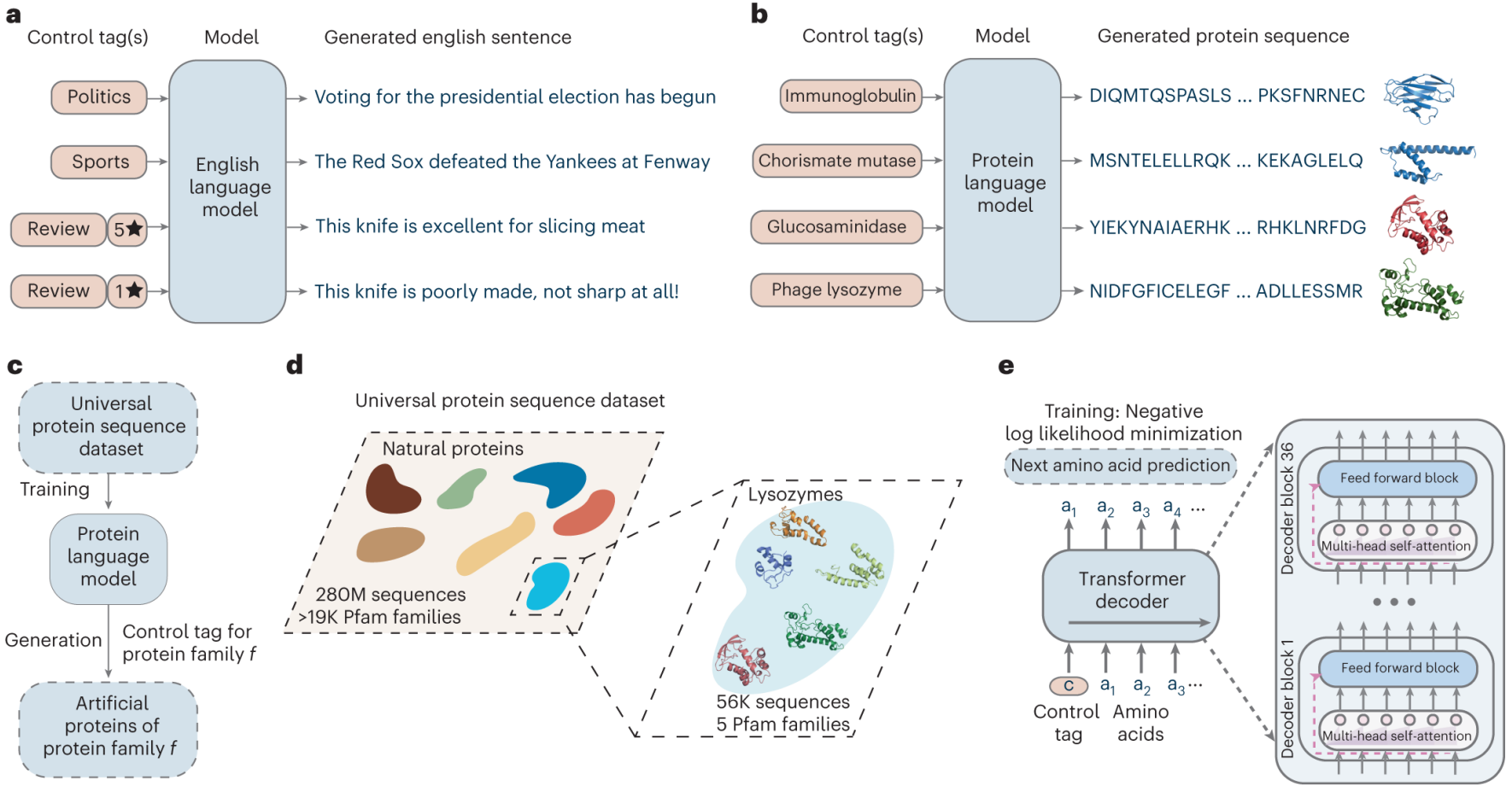
Figure 4: (a) Overview of most prominent Transformers released in the last years and their number of parameters. **(b)** Deposited entries in the protein databases PDB and UniParc.

https://www.deepmind.com/blog/alphafold-using-ai-for-scientific-discovery-2020

Protein Language Model and parallel with NLP



Protein Language Model and parallel with NLP



Madani, Ali, et al. "Large language models generate functional protein sequences across diverse families." *Nature Biotechnology* (2023): 1-8.

Key Takeaways

- The exponential increase in protein sequence data makes it a good domain for usi ng Transformer architectures and for the development of protein language models.
- Protein language models can be effective to select good candidate mutations that improve protein functions or to generate proteins with specific functions.
- The development of new architectures for protein language models and creative w ays to combine recent AI models to protein design pipeline are leading to breakthr oughs in this domain in the last 5 years.

Protein Language Models

- ProGen2

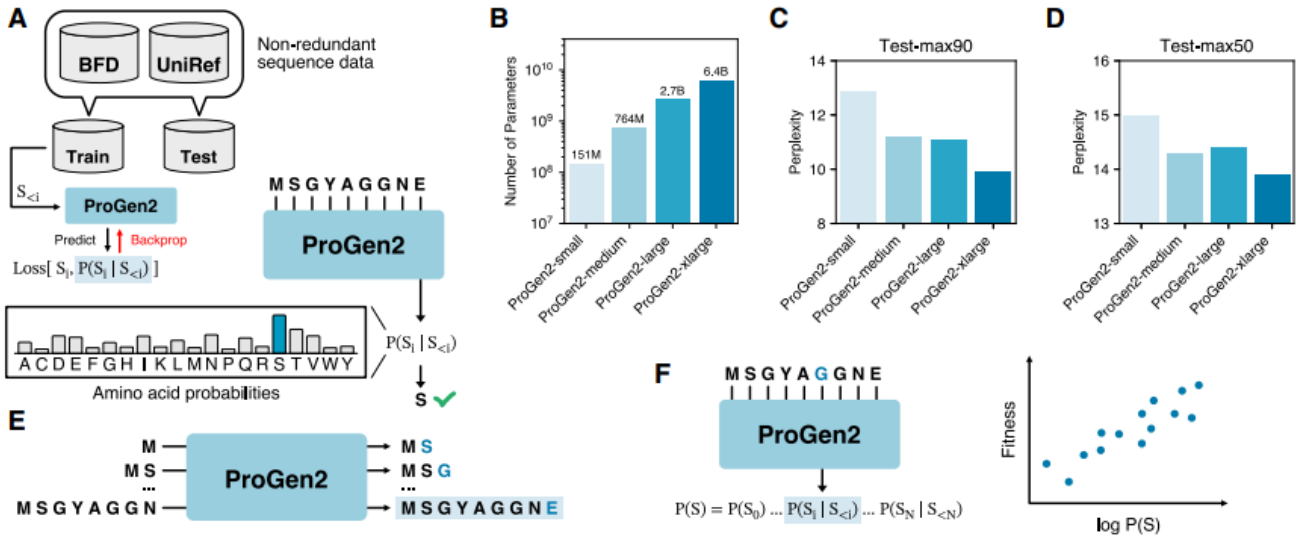
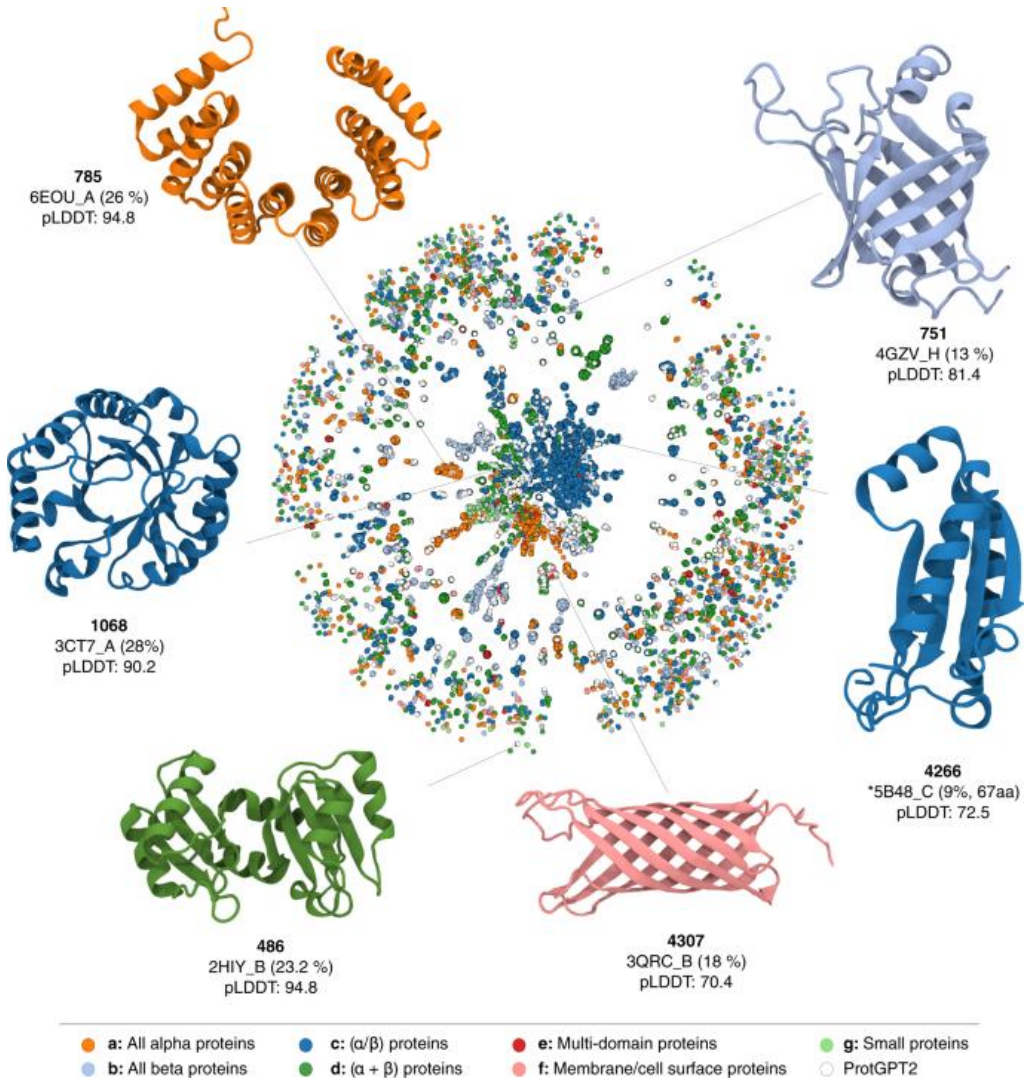


Figure 1. Overview of ProGen2 models for protein sequence generation and scoring

(A) Diagram of model pretraining scheme and autoregressive amino acid prediction.
(B) Number of parameters (log scale) for ProGen2 models.
(C) Perplexity (unitless) for sequences held out from pretraining dataset clustered at 90% sequence identity (Test-max90).
(D) Perplexity (unitless) for sequences held out from pretraining dataset clustered at 50% sequence identity (Test-max50).
(E) Diagram of sequence generation with an autoregressive language model.
(F) Diagram of sequence log likelihood calculation for protein fitness prediction.

Protein Language Models – Concept Space



Ferruz, Noelia, Steffen Schmidt, and Birte Höcker. "ProtGPT2 is a deep unsupervised language model for protein design." *Nature communications* 13.1 (2022): 4348.

Protein Language Models: ESM-2

Evolutionary-scale prediction of atomic-level protein structure with a language model

ZEMING LIN, HALIL AKIN, ROSHAN RAO, BRIAN HIE, ZHONGKAI ZHU, WENTING LU, NIKITA SMETANIN, ROBERT VERKUIL, ORI KABELI, [...], AND ALEXANDER RIVES +5 authors Authors Info & Affiliations

Ferruz, Noelia, Steffen Schmidt, and Birte Höcker. "ProtGPT2 is a deep unsupervised language model for protein design." *Nature communications* 13.1 (2022): 4348.

Training Protein Language Models: ESM-2

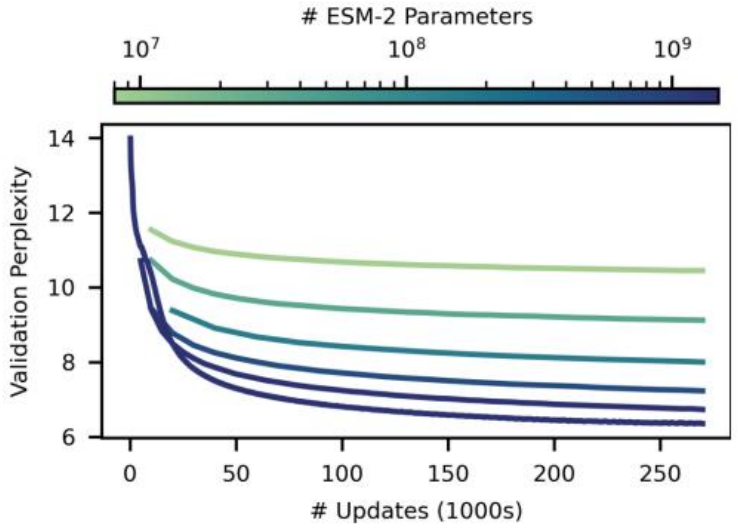


Figure S1. ESM-2 masked language modeling training curves. Training curves for ESM-2 models from 8M (highest curve, light) to 15B parameters (lowest curve, dark). Models are trained to 270K updates. Validation perplexity is measured on a 0.5% random-split holdout of UniRef50. After 270K updates the 8M parameter model has a perplexity of 10.45, and the 15B model reaches a perplexity of 6.37.

Training Protein Language Models: ESM-2

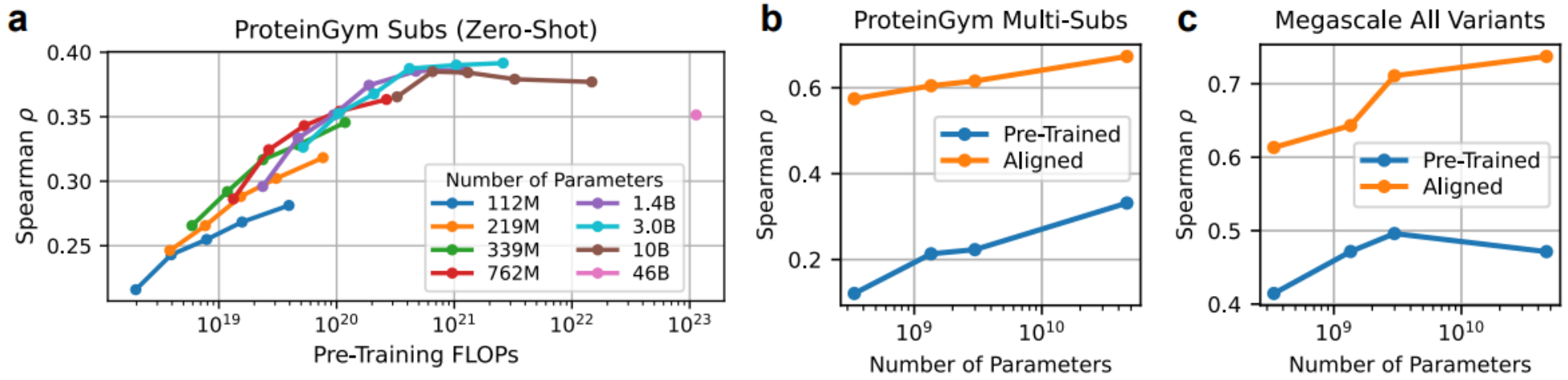
Model	# Params	# Updates	Validation Perplexity	LR P@L	LR P@L/5	CASP14	CAMEO
ESM-2	8M	270K	10.45	0.16	0.28	0.37	0.48
	35M	270K	9.12	0.29	0.49	0.41	0.56
	150M	270K	8.00	0.42	0.68	0.47	0.63
	650M	270K	7.23	0.50	0.77	0.51	0.68
	3B	270K	6.73	0.53	0.80	0.51	0.71
	8M	500K	10.33	0.17	0.29	0.37	0.48
	35M	500K	8.95	0.30	0.51	0.41	0.56
	150M	500K	7.75	0.44	0.70	0.49	0.65
	650M	500K	6.95	0.52	0.79	0.51	0.70
	3B	500K	6.49	**0.54**	0.81	0.52	**0.72**
	15B	270K	**6.37**	**0.54**	**0.82**	**0.55**	**0.72**
ESM-1b	650M	—	—	0.41	0.66	0.42	0.64
Prot-T5-XL (UR50) (21)	3B	—	—	0.48	0.72	0.50	0.69
Prot-T5-XL (BFD) (21)	3B	—	—	0.36	0.58	0.46	0.63
CARP (24)	640M	—	—	—	—	0.42	0.59

Scaling Protein Language Models

Scaling unlocks broader generation and deeper functional understanding of proteins

Aadyot Bhatnagar, Sarthak Jain, Joel Beazer, Samuel C. Curran, Alexander M. Hoffnagle, Kyle Ching, Michael Martyn, Stephen Nayfach, Jeffrey A. Ruffolo, Ali Madani

doi: https://doi.org/10.1101/2025.04.15.649055



Limits of scaling also for Protein LLMs

Key Takeaways

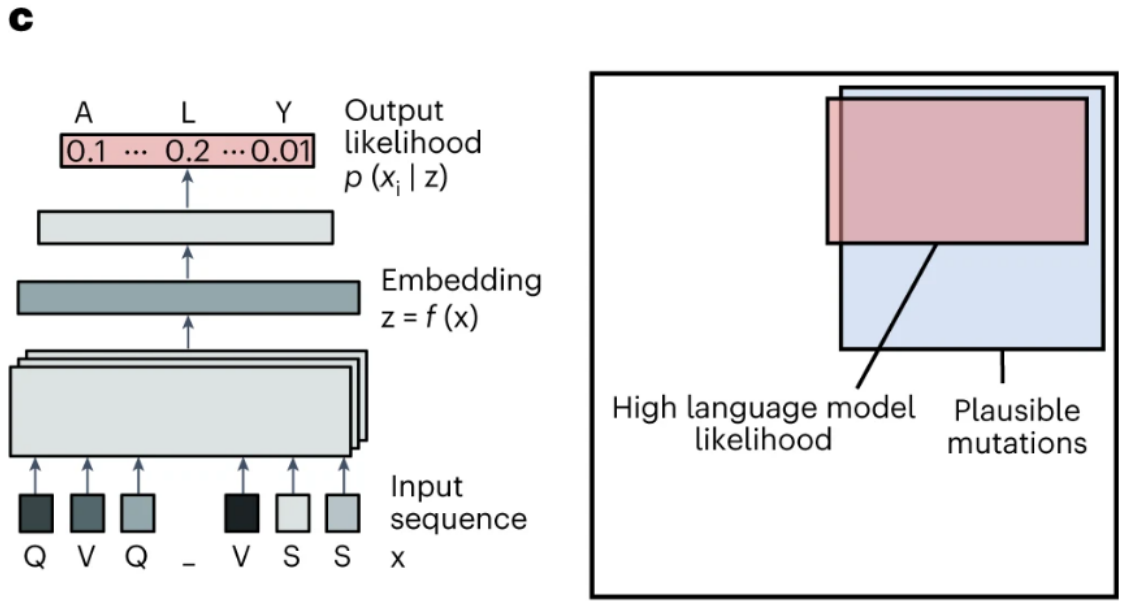
- Collection of protein language models have been trained for protein research latel y.
- Important part of the development of protein LMs is understanding the scaling of these methods and applying LLM engineering concepts.
- It seems that we approaching the limits of parameter scaling for these methods.

Applications of pLMs

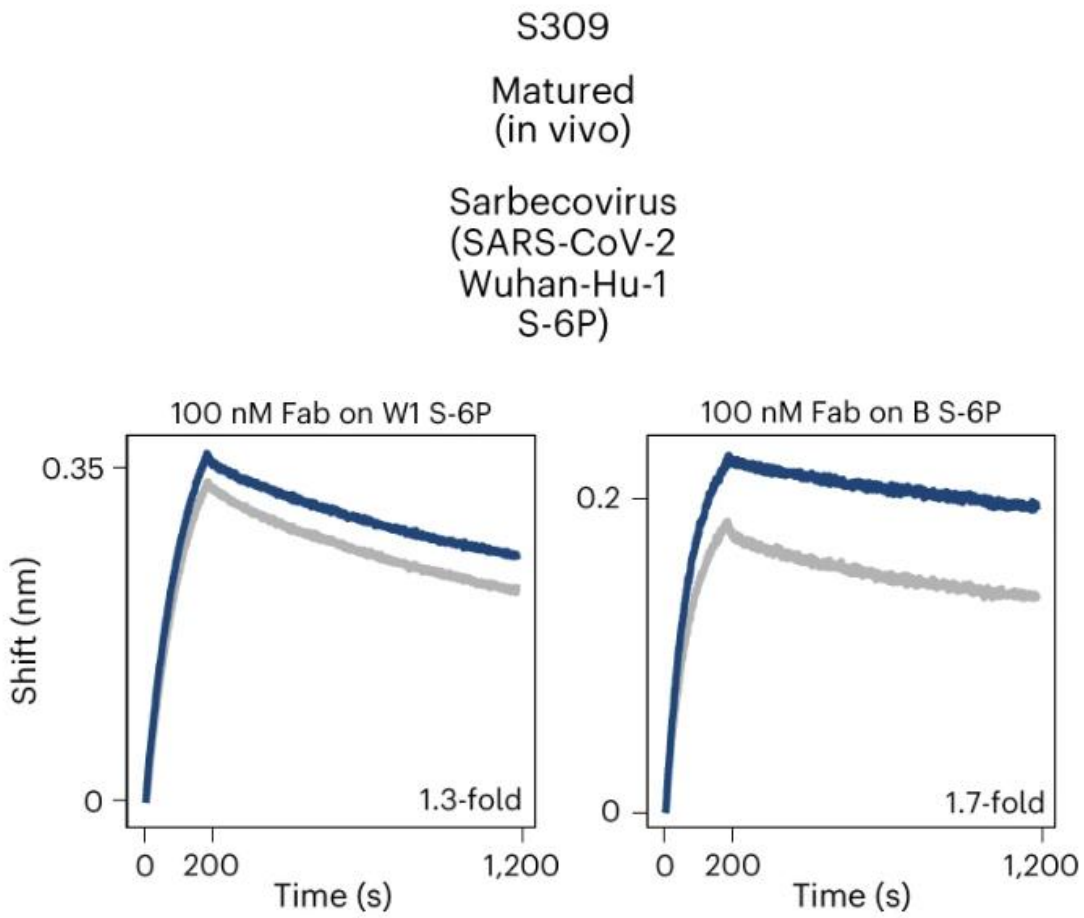
Article | Open access | Published: 24 April 2023

Efficient evolution of human antibodies from general protein language models

Brian L. Hie, Varun R. Shanker, Duo Xu, Theodora U. J. Bruun, Payton A. Weidenbacher, Shaogeng Tang, Wesley Wu, John E. Pak & Peter S. Kim



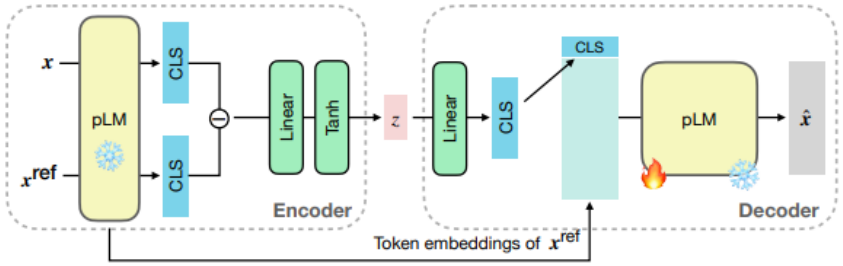
Experiment: use high language model likelihood to sample evolutionarily plausible mutations, measure fitness



Applications of pLMs

Robust Optimization in Protein Fitness Landscapes Using Reinforcement Learning in Latent Space

Minji Lee [*1] Luiz Felipe Vecchietti [*2] Hyunkyu Jung [1 2] Hyunjoo Ro [3] Meeyoung Cha [1 2] Ho Min Kim [4 3]



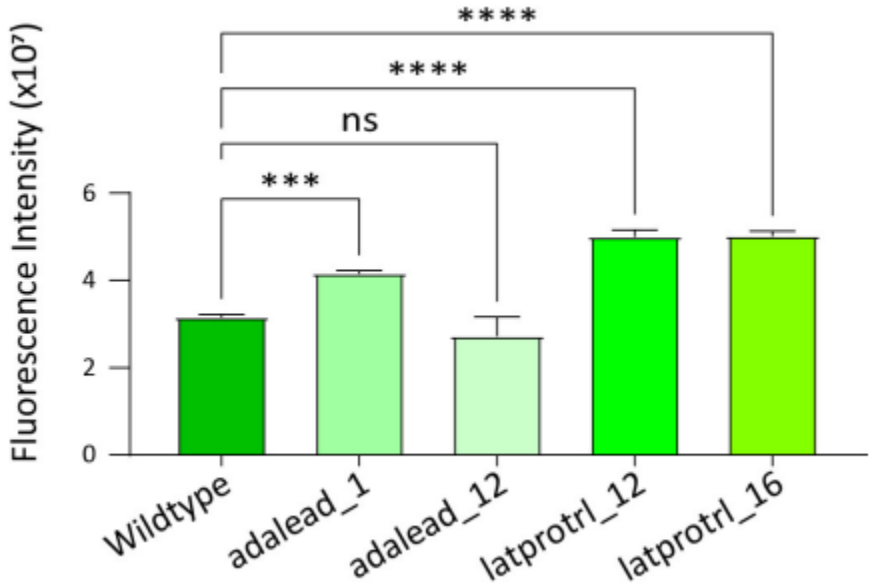
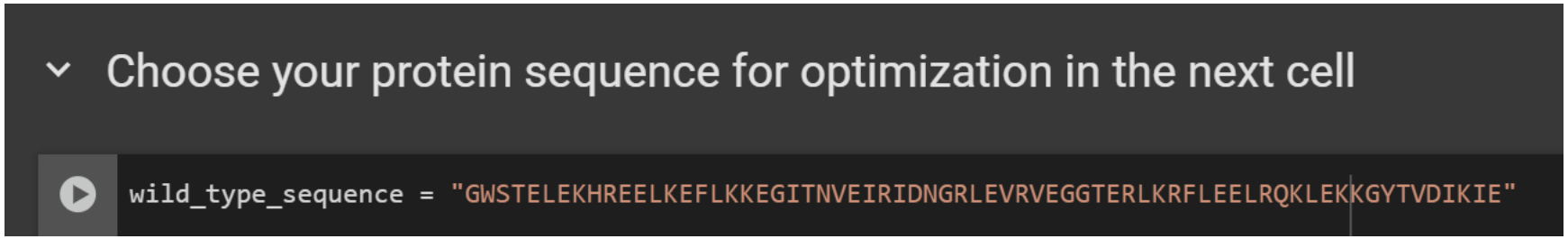


Figure 2. **Variant Encoder-Decoder Architecture**. Given an input sequence, the encoder calculates a representation that is used by the decoder to reconstruct the original sequence. The term CLS represents the embeddings for the classification token in ESM-2.

An optimization example…



An optimization example…

Run **ESMFold** Prediction (~1min)

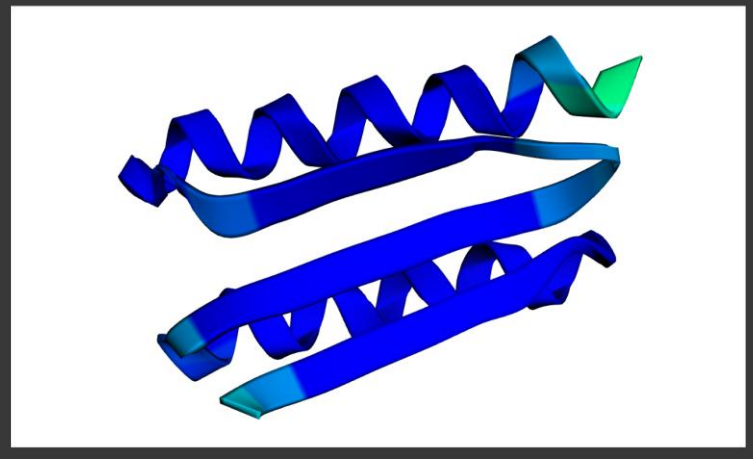
jobname: wild_type_protein

Show code

```
length 68
ptm: 0.817 plddt: 90.521
CPU times: user 11.9 s, sys: 8.58 s, total: 20.5 s
Wall time: 1min 1s
```

Display Wild Type Protein

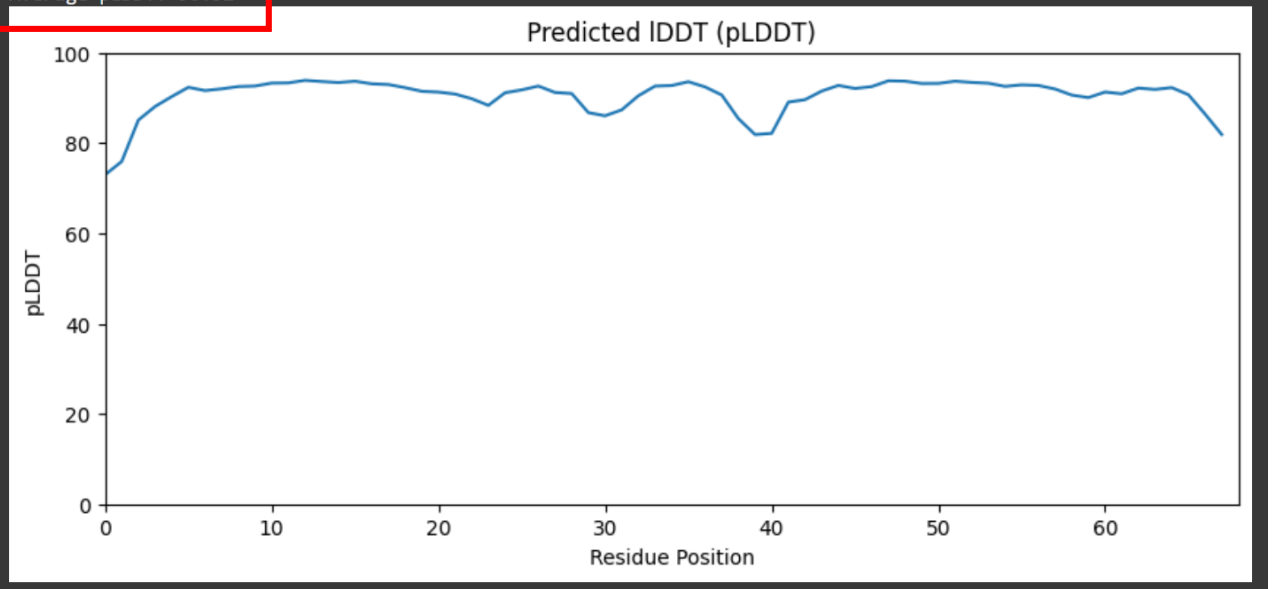
Show code



Plot Confidence Metrics for the Wild Type

Show code

```
Average pLDDT: 90.52
```



An optimization example…

```
Original sequence: GWSTELEKHREELKEFLKKEGITNVEIRIDNGRLEVRVEGGTERLKRFLEELRQKLEKKGYTVDIKIE
Masked sequence (10% random): GWSTELEKHREELKEFLKKEGITNVEIRI**GRLEVRVEGGTERLKRFLE*LR**L*KKGYTVDIKIE
Masked indices (ESM token indices): [30, 31, 51, 54, 55, 57]

--- Assign the generated sequence to a string ---
Generated: GWSTELEKHREELKEFLKKEGITNVEIRIARGRLEVRVEGGTERLKRFLEKLRRDLEKKGYTVDIKIE
```

An optimization example…

Run **ESMFold** Prediction for Generated Sequence (~1min)

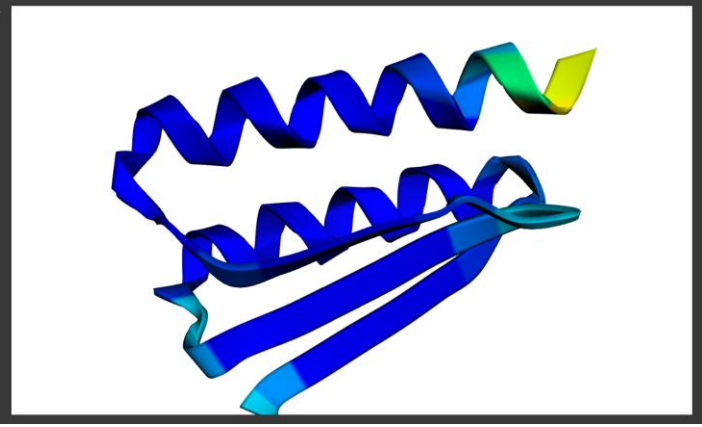
jobname: generated_protein

Show code

```
length 68
ptm: 0.807 plddt: 89.265
CPU times: user 1.83 s, sys: 17 ms, total: 1.84 s
Wall time: 1.82 s
```

Display Generated Protein

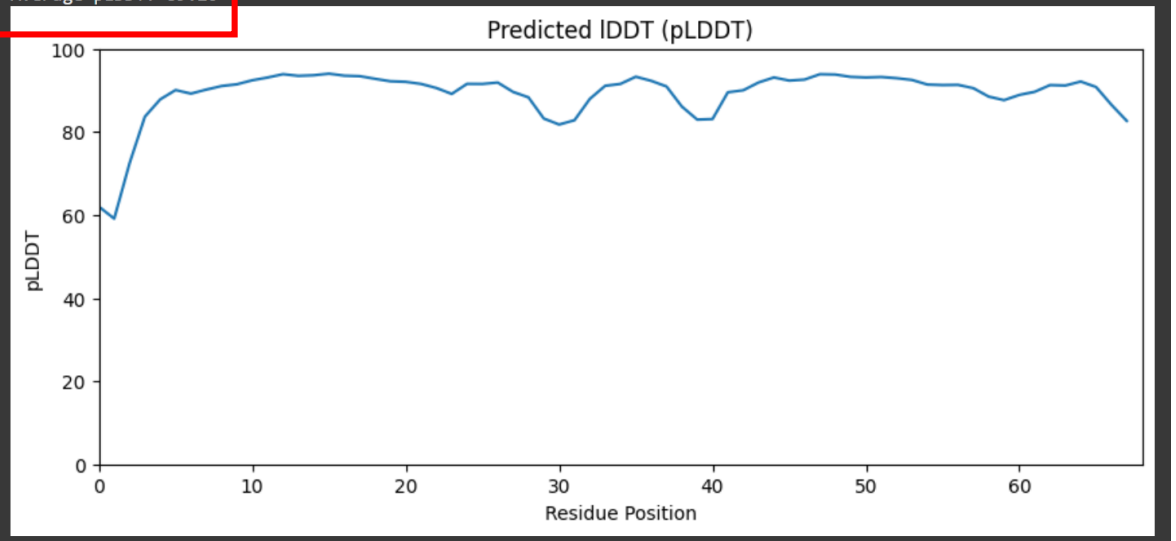
Show code



Plot Confidence Metrics for the Generated Sequence

Show code

```
Average pLDDT: 89.26
```



Summary of Part 2: Case Study on Protein Language Models

- Protein Research is a domain in which many of the concepts for LLM engineering i s being applied and one of the domains with the highest potential in AI for Scienc e.
- Protein LMs bring different challenges. For example, scaling laws for NLP might no t hold for proteins. Data preprocessing and assumptions about redundancy are als o challenging.
- Many protein LMs are currently being used in protein engineering pipelines to opt imize proteins that are being used as vaccines, therapeutics or in domains such as agriculture.

Exercise 2: Protein Engineering with ESM-2 and ESMFold

- 02_exercise.ipynb
- 45 minutes: try to optimize a protein using ESM-2 and ESMFold.

- Main Task:

1. Run ESMFold to predict the structure and the confidence of ESMFold for a protein sequence of your choice.
2. Mask part of the sequence, generate mutants using ESM-2 or random mutations, and run ESMFold again trying to find mutants with higher confidence.

- If you find a higher confidence mutant:

1. Feel free to share with colleagues or to discuss your results!

Thank you