PROTEIN EMBEDDING MODELS:

LEARNING THE LANGUAGE OF PROTEINS

GHENT UNIVERSITY

LABORATORY OF APPLIED BIOTECHNOLOGY

KERMIT

Dimi Boeckaerts

Overview of this workshop

VHDMCYIRCAHTSIRDSHDRDGG**INTRODUCTION**EYFFNKTTDMEFPWYAPTFTRLESTLPFLVAFWNP

D

V

W

PNFISGDQWAVCGHFWSPQPFFMI**HISTORY**FTLGMGEEKIDDG

M

L

P

DMW**STATE-OF-THE-ART**VWEHGPNFISGDQWYRSVHRGQDQHRHMIIY

P

H

IRCAHTSIRDSHDRDGGESIDNFTLPFLVAFWNPQTWLFKY**OUTLOOK**LHEIADCVYWRVLDIWH

W

N

RYASEYFFNKTTD**PRACTICE**TFTRLESGDQWAVCGHFWSPQPFFMIRIFHQWCFTLTNGTKFAFFVFFYVWEHGPN

The rapid progress in natural language processing

Natural language processing: the field of computer science that is concerned with automated text and language analysis.

Thought experiment generation

10 Thought experiments

Humor

500+ Openers for Tinder wri...

Deepfakes

AI Eminem

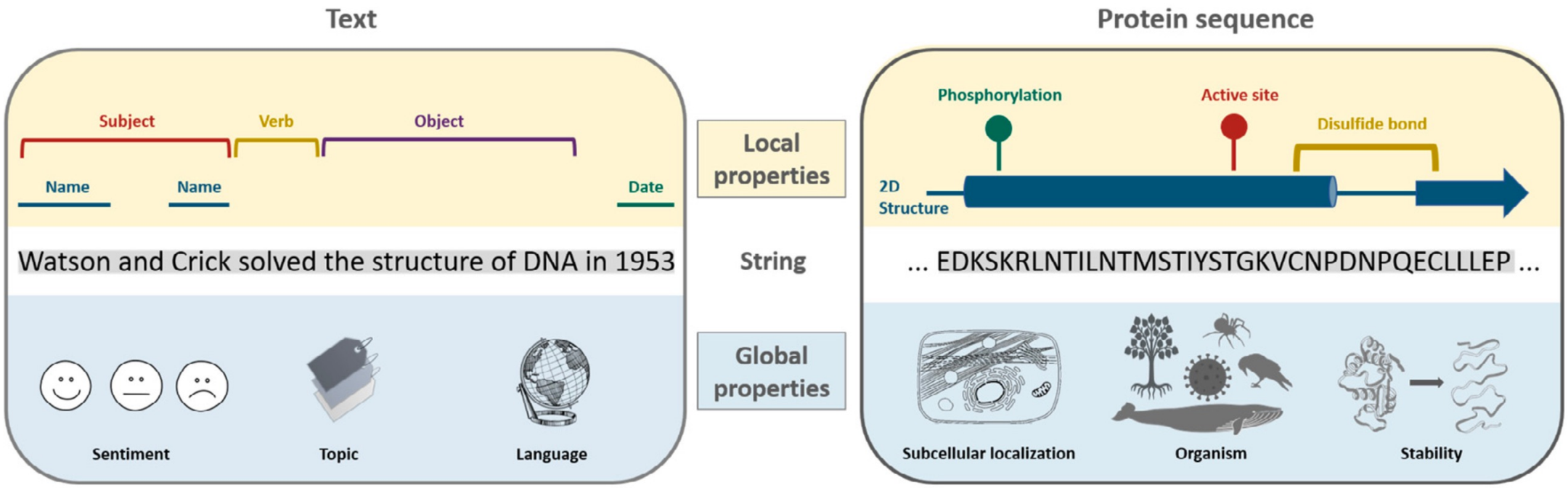
View details

Image Generation

DALL·E by OpenAI

OpenAI's GPT-3 examples

Proteins are similar to natural language…



Ofer et al. (2021). The language of proteins: NLP, machine learning & protein sequences.

… but not entirely similar

- We cannot ‘read’ proteins
- I love you vs. I loved you
- Protein 3D structure -> distant interactions
- Sequencing bias

Are proteins learnable?

The linguistic hypothesis

The space of naturally occurring proteins occupies a learnable manifold. This manifold emerges from evolutionary pressures that heavily encourage the reuse of components at many scales: from short motifs of secondary structure, to entire globular domains.

Mohammed AlQuraishi

Keep it simple: use a bag-of-words

	about	bird	heard	is	the	word	you
About the bird, the bird, bird bird bird	1	5	0	0	2	0	0
You heard about the bird	1	1	1	0	1	0	1
The bird is the word	0	1	0	1	2	1	0

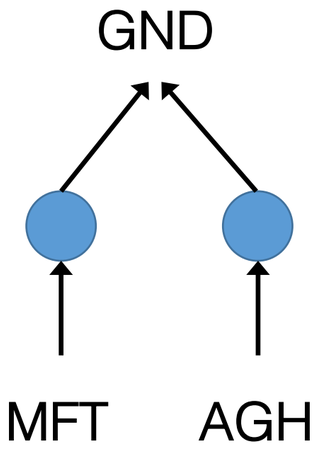
Keep it simple: use a bag-of-words

	DMC	MCY	YIR	HRG	SVH	QDQ	CAH
DMCYIRCAHTSIRD…	1	1	1	0	0	0	0
WYRSVHRGQDQH…	0	0	0	1	1	1	1

- Different k-mer lengths
- Overlapping vs. non-overlapping
- Binary vs. count vs. frequency

Embedding methods for protein language

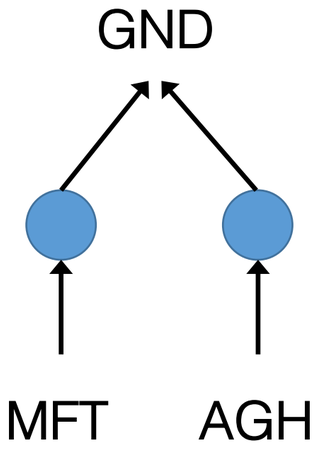
Word2Vec



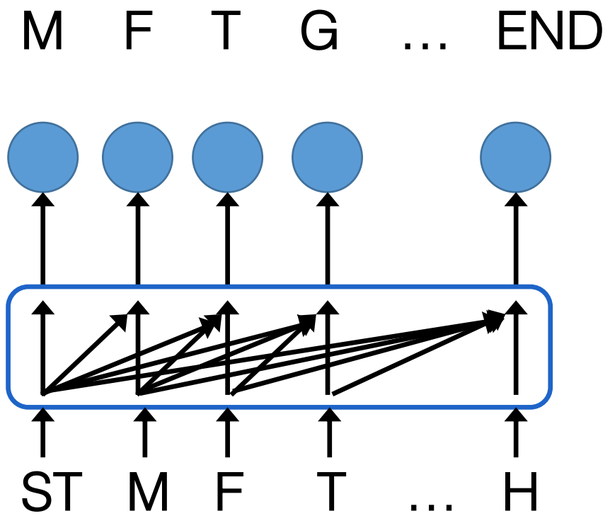
MFTGNDAGH

Embedding methods for protein language

Word2Vec



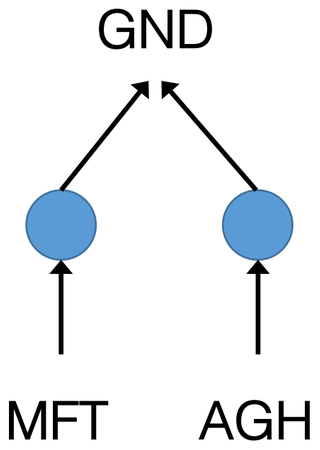
Autoregressive
language model



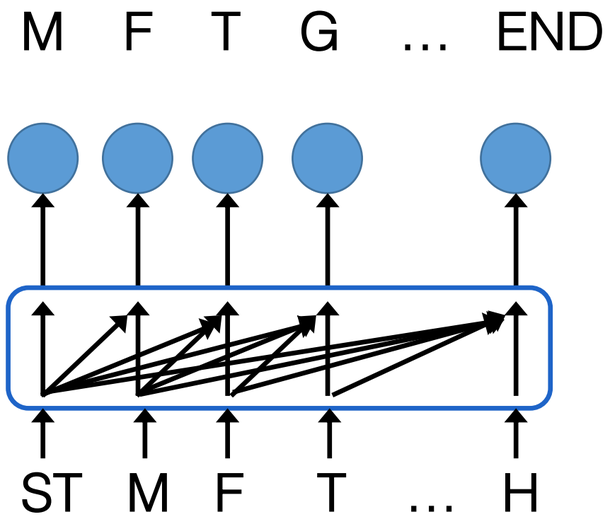
MFTGNDAGH

Embedding methods for protein language

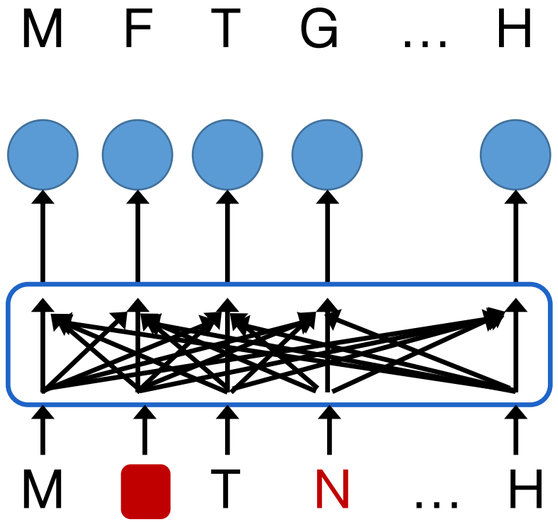
Word2Vec



Autoregressive
language model

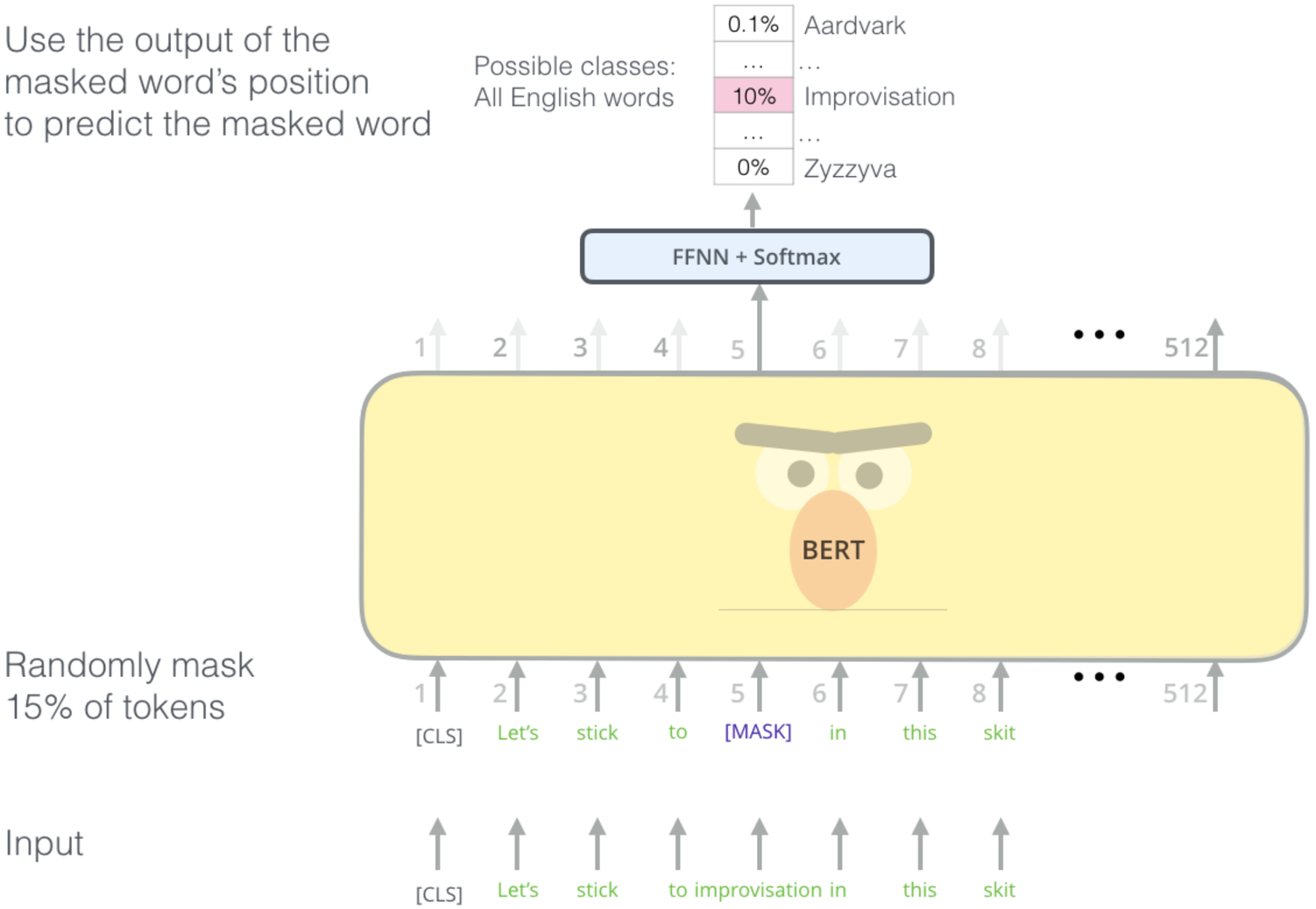


Masked
language model



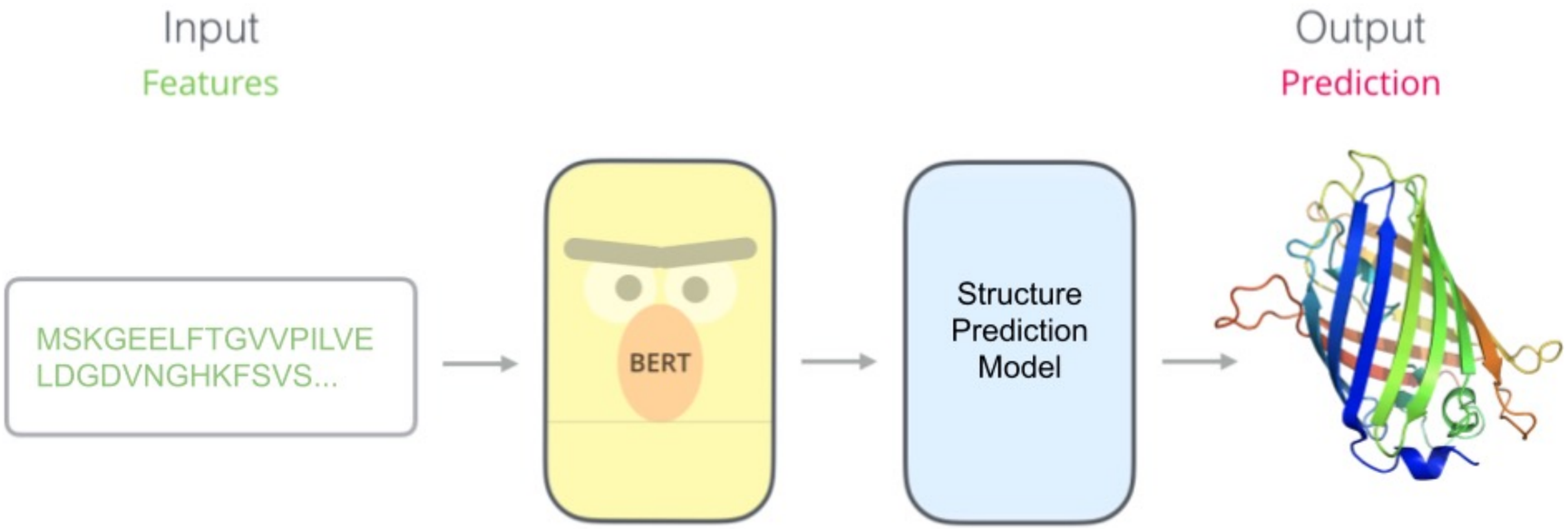
MFTGNDAGH

BERT model training



https://jalammar.github.io/illustrated-bert/

Stacking supervised prediction models



https://bair.berkeley.edu

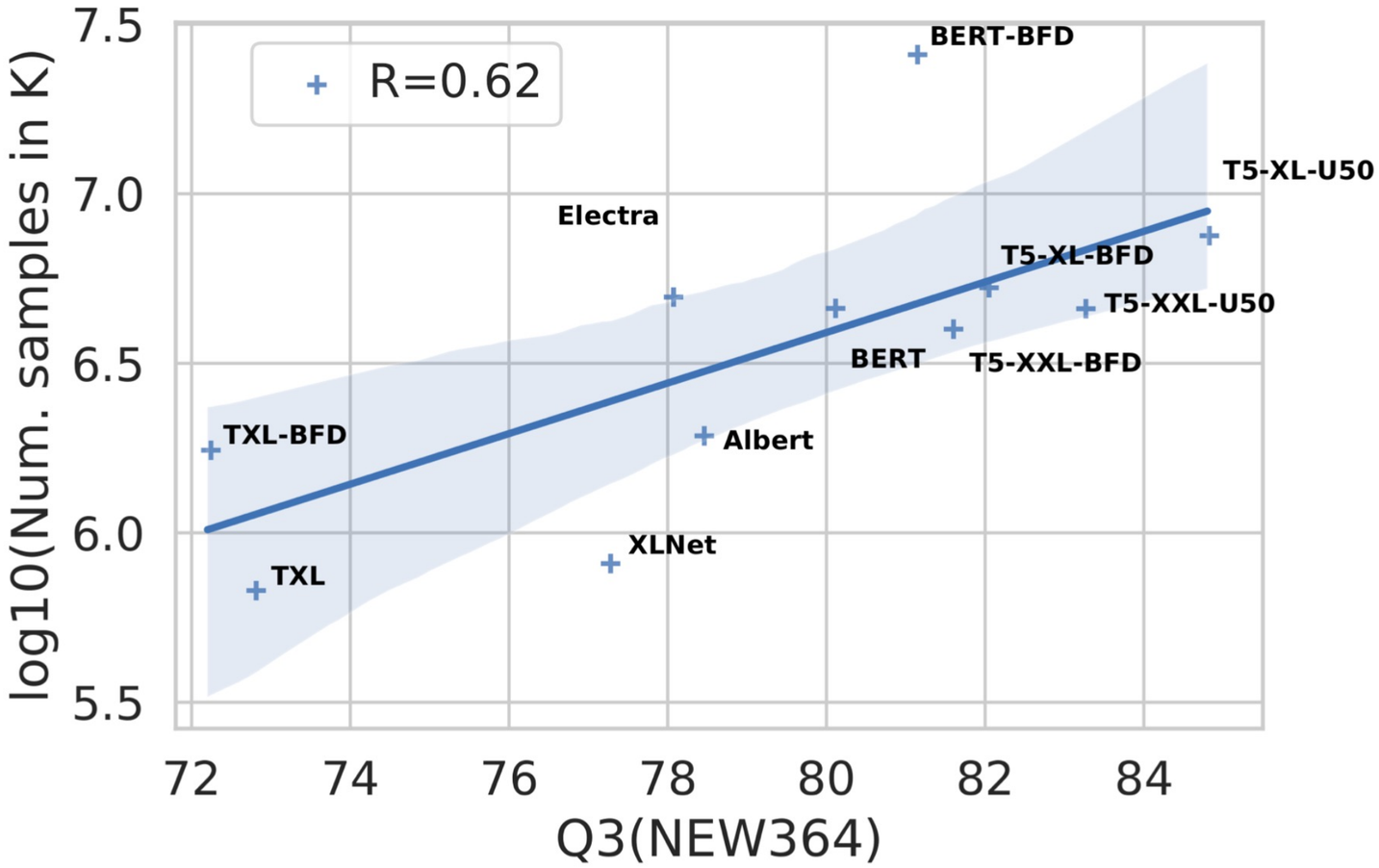
Language models are data hungry

Data LM	*UniRef50*	*UniRef100*	*BFD*
Number proteins [in m]	45	216	2,122
Number of amino acids [in b]	14	88	393
Disk space [in GB]	26	150	572

TABLE 1: Data Protein LM - UniRef50 and UniRef100 cluster the UniProt database at 50% and 100% pairwise sequence identity (100% implying that duplicates are removed) [41]; BFD combines UniProt with metagenomic data keeping only one copy for duplicates [24], [42]. Units: number of proteins in millions (m), of amino acids in billions (b), and of disk space in GB (uncompressed storage as text).

Elnaggar et al. (2021). ProtTrans: towards cracking the language of life's code through self-supervised learning.

Language models are data hungry



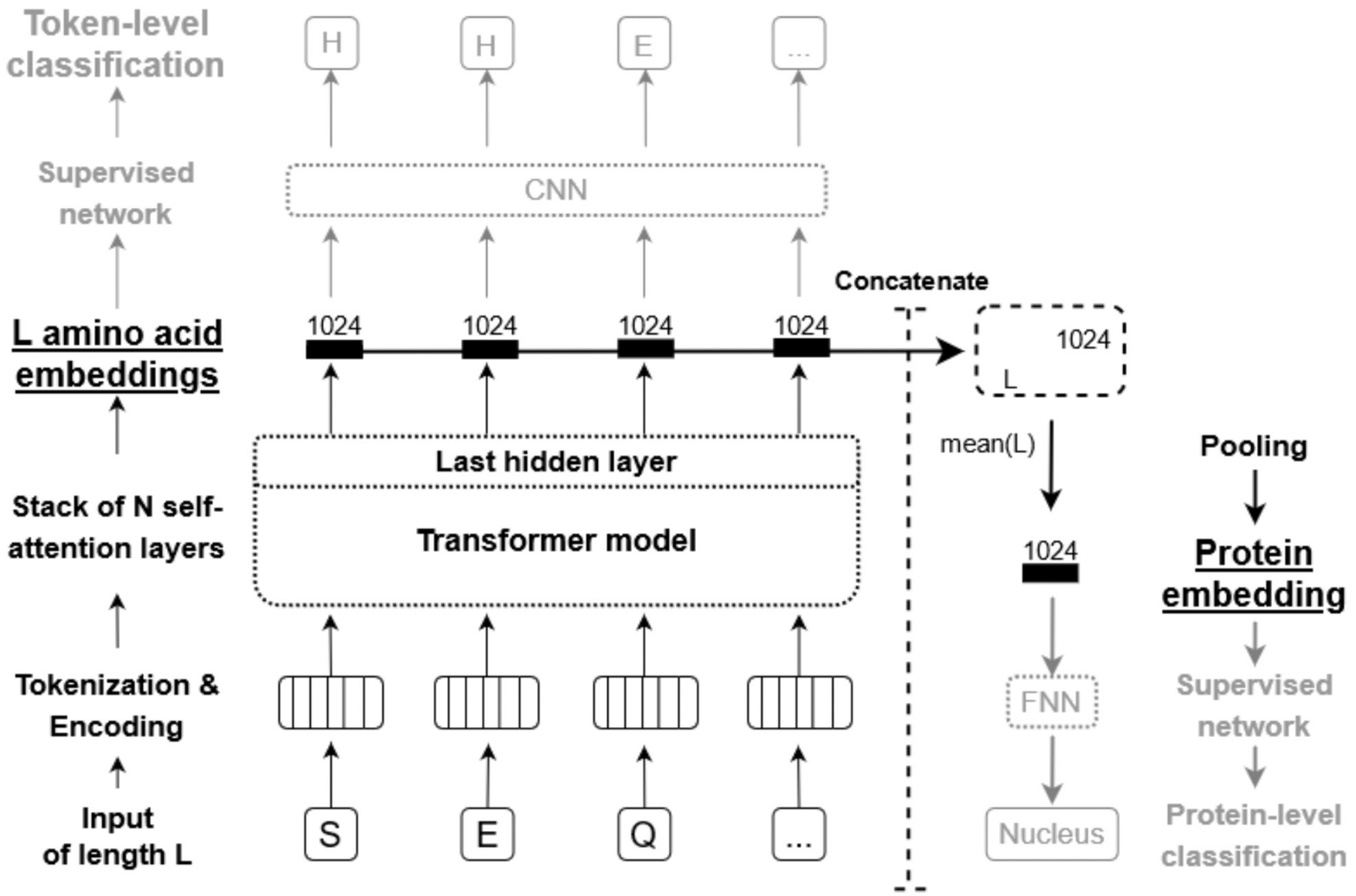
More data during pre-training seems better for downstream performance.

Redundancy and noise

It’s all about abundant and high-quality data!

“Less noisy and less redundant corpora (e.g. UniRef50) improved over larger but more noisy and redundant corpora (e.g. BFD).” (Elnaggar *et al.*, 2021)

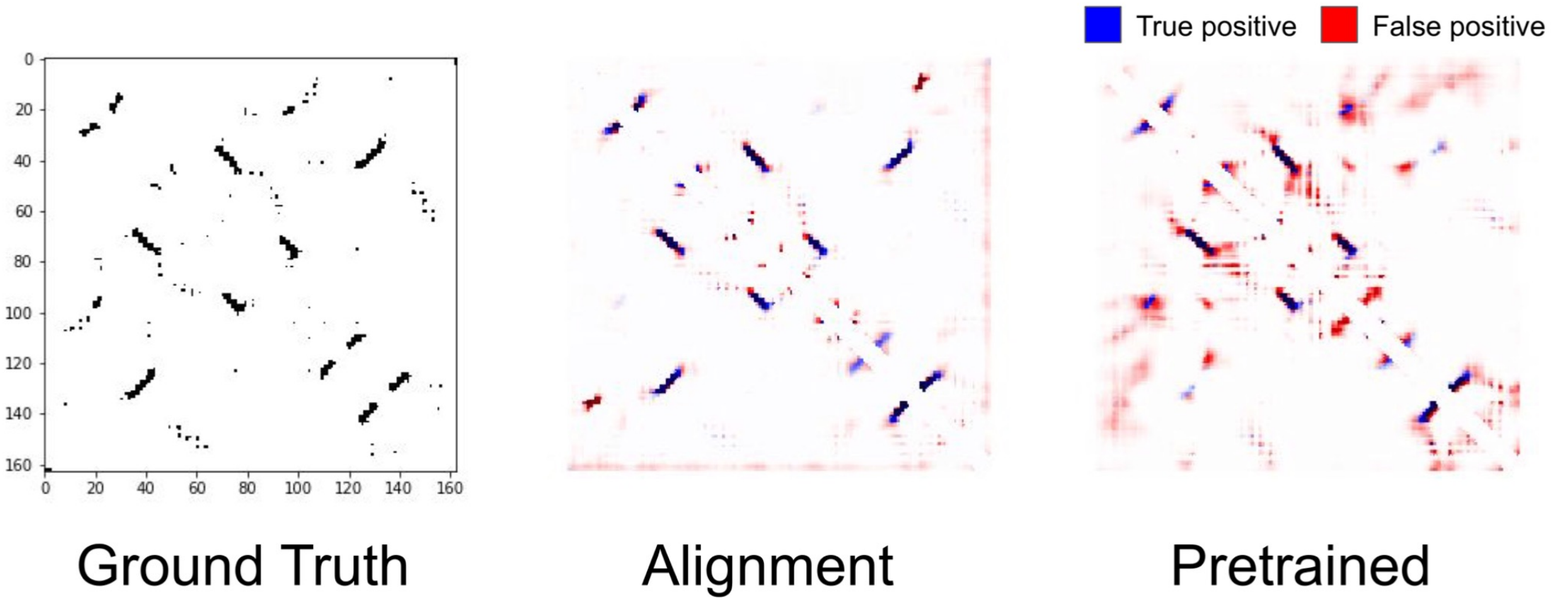
Recent applications



Elnaggar et al. (2021). ProtTrans: towards cracking the language of life's code through self-supervised learning.

Recent applications

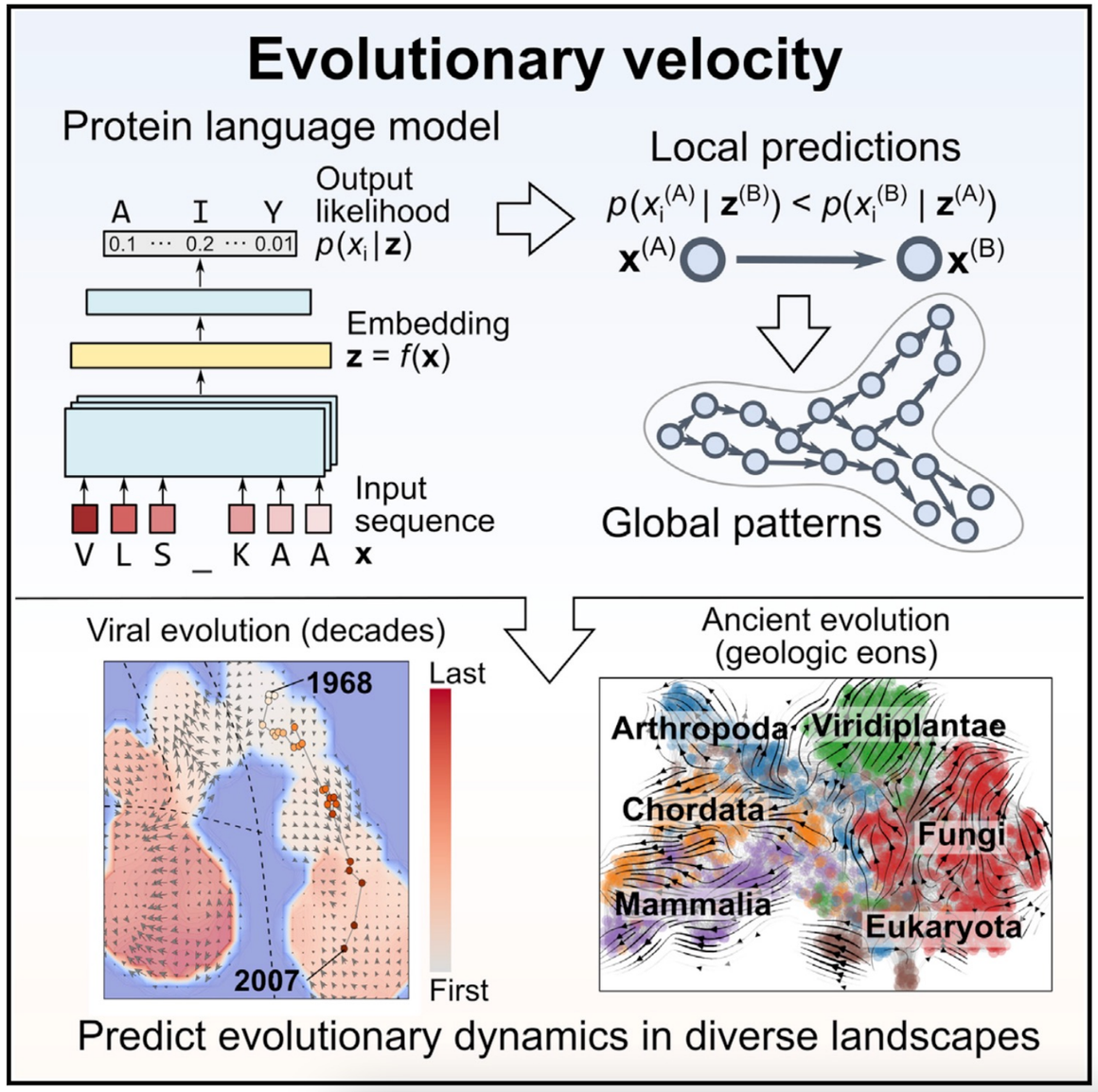
Simpler methods using evolutionary information can (still) outperform DL methods.



The true contact map for the same protein as above (left), predicted contacts from a model with alignment-based features (center), predictions from a ***pretrained*** *LSTM (right).*

https://bair.berkeley.edu

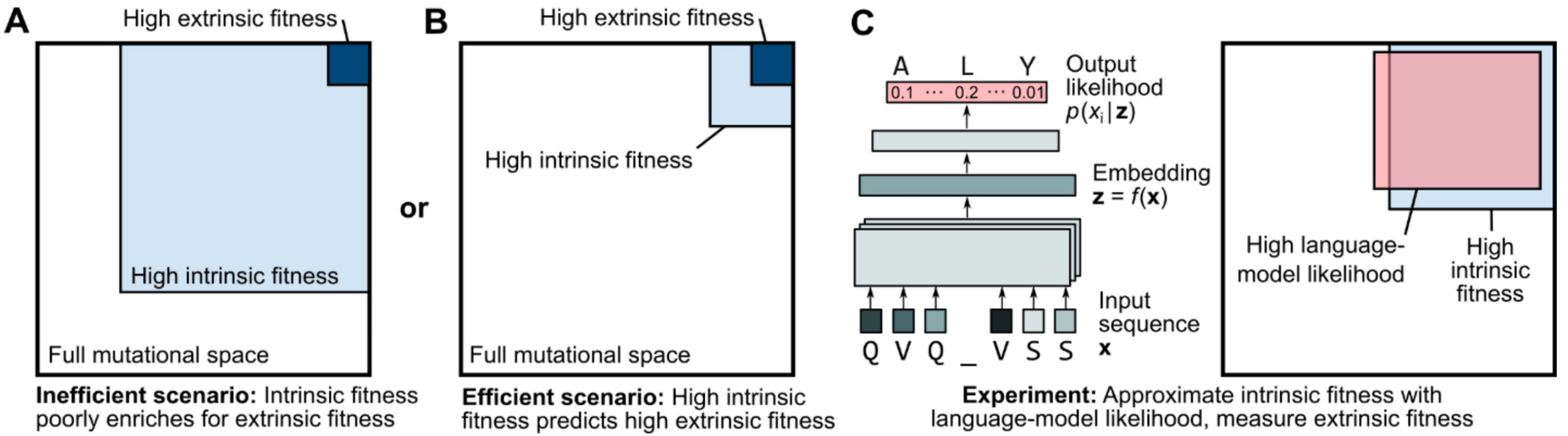
Recent applications



Protein language models can predict mutational effects across diverse proteins.

Hie et al. (2022). Evolutionary velocity with protein language mdels predicts evolutionary dynamics...

Recent applications



Inefficient scenario: Intrinsic fitness poorly enriches for extrinsic fitness

Efficient scenario: High intrinsic fitness predicts high extrinsic fitness

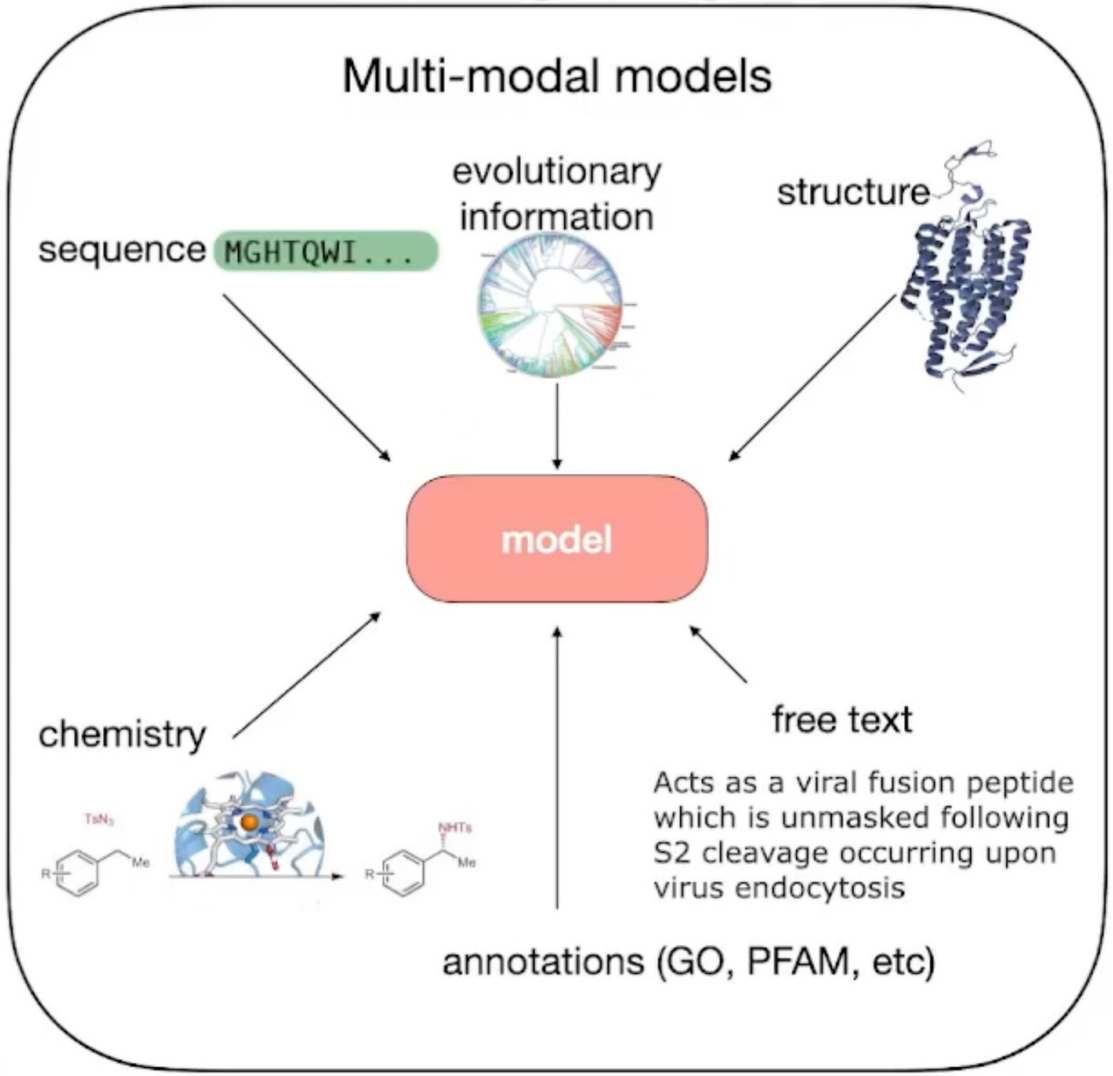
Experiment: Approximate intrinsic fitness with language-model likelihood, measure extrinsic fitness

Hie et al. (2022). Efficient evolution of human antibodies from geneal protein language models…

Outlook: trends and challenges

Big data needs large models need loads of computational resources.

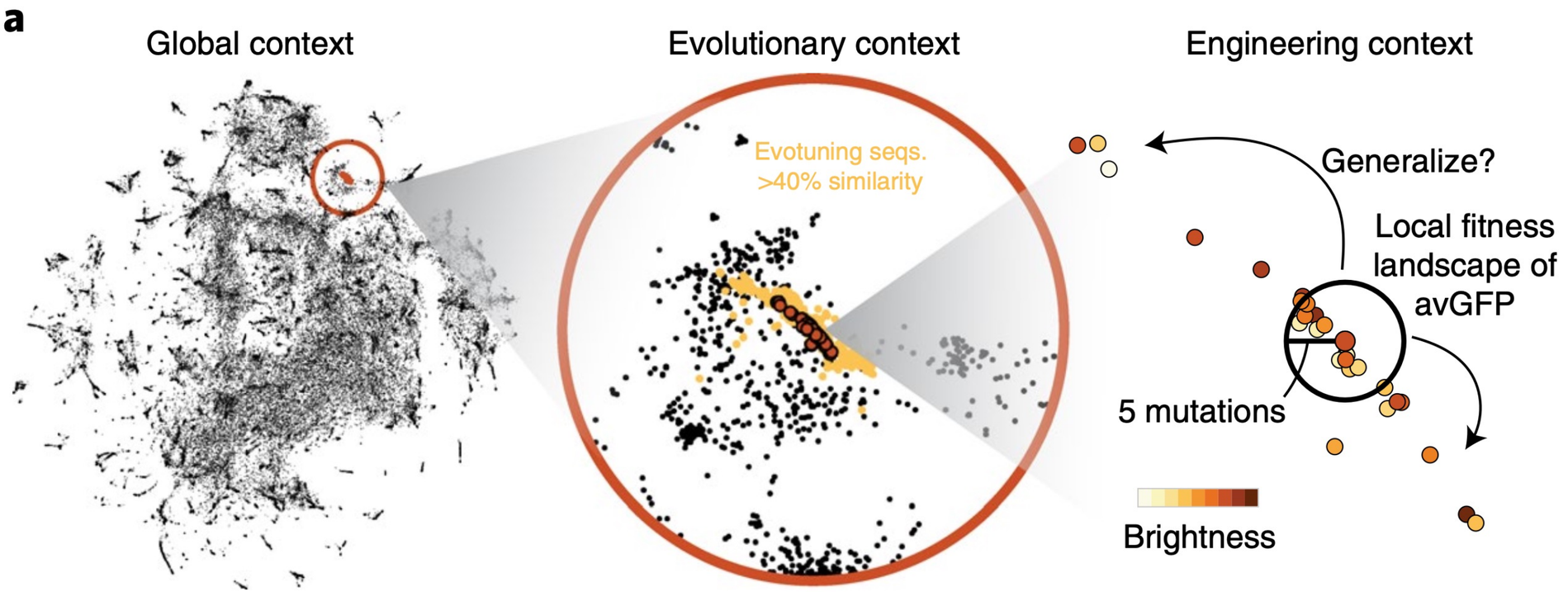
Outlook: trends and challenges



https://www.youtube.com/watch?v=0l2G9_LrY_c

Outlook: trends and challenges

Actual impact in biological applications



Alley et al. (2019). Unified rational protein engineering with sequence-based deep representation learning.

Open questions

Foundation models for biology?

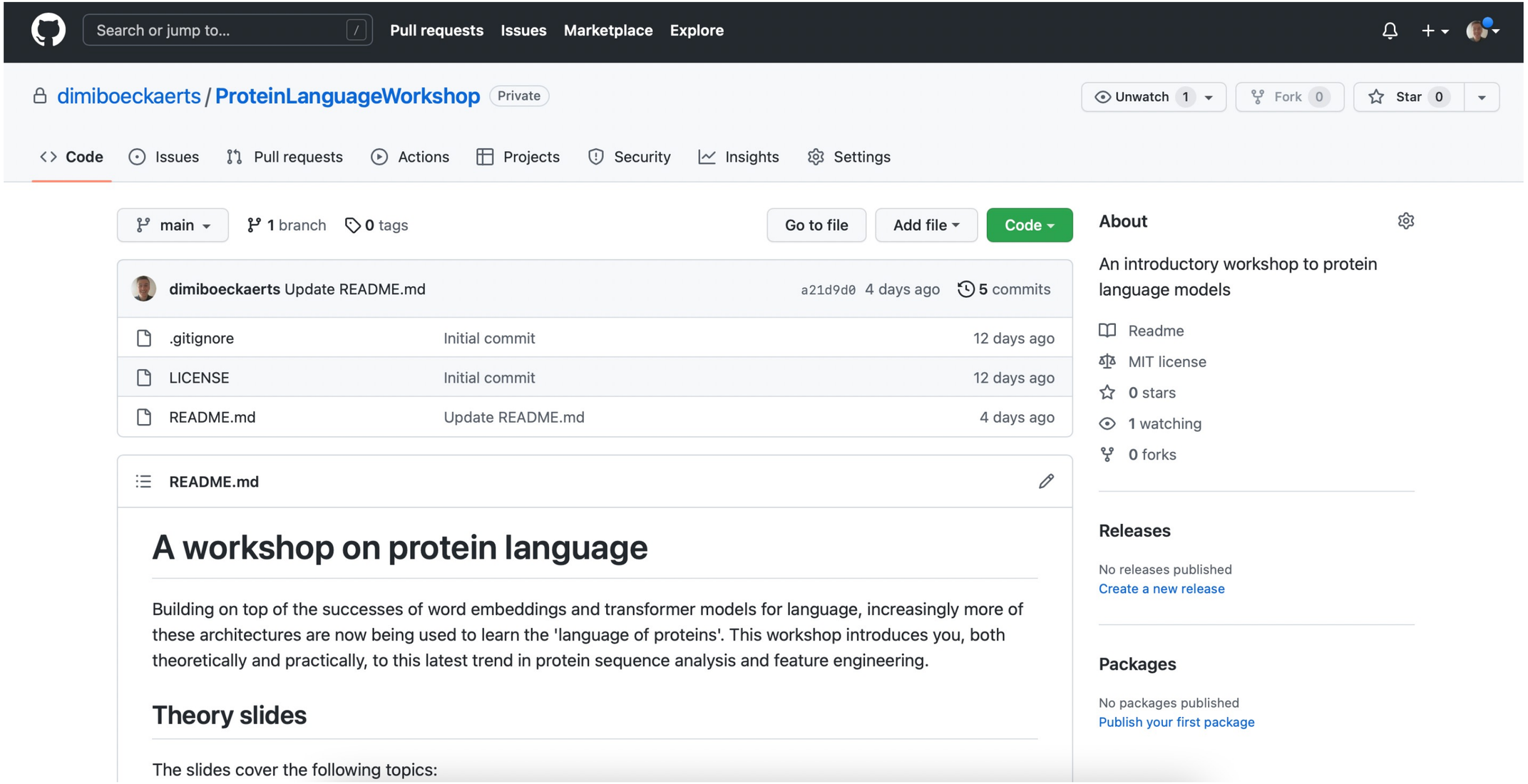
“Bio” DeepMind or OpenAI company?

Deep fake content, deep fake life forms?

Questions?

@dimiboeckaerts
Dimi Boeckaerts

Embeddings in practice!



GPT-3 thought experiments

The Omelet Parable

Imagine that you are on a train. You have always wanted to make the perfect omelet. A genie appears and offers you two options: you can have a delicious omelet, or you can have a device that will let you make unlimited omelets for the rest of your life.

Would you rather have a delicious omelet?
Or the device?

The Red Pill/Blue Pill

Imagine that there is a drug that will cure all diseases. It also has a side effect of making people feel like they are living in a computer simulation. They are in a simulation, and after they take the pill they can see the pixels and their world is filled with triangles.

Would you take it?

https://twitter.com/TomerUllman/status/1363851329463087109