

Pattern recognition in big data - sequence analysis in the era of LLMs

Valentyn Bezshapkin
(@valentynbez)
Sunagawa lab
ETH Zurich



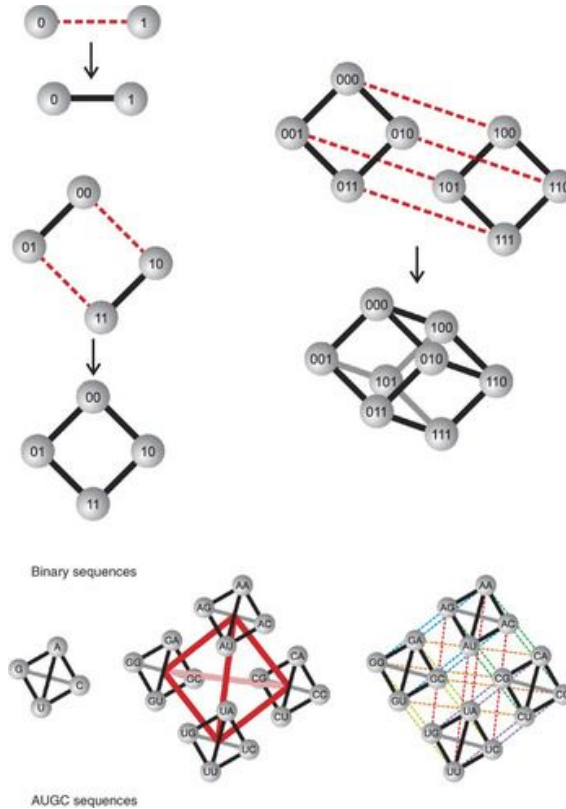
Definition of a sequence with examples

- A sequence is a **set of objects which is listed in a specific order**, one after another.

Examples of sequence data		
Speech recognition		→ "The quick brown fox jumped over the lazy dog."
Music generation		→ 
Sentiment classification	"There is nothing to like in this movie."	→ 
DNA sequence analysis	AGCCCCCTGTGAGGAACTAG	→ AGCCCCCTGTGAGGAACTAG
Machine translation	Voulez-vous chanter avec moi?	→ Do you want to sing with me?
Video activity recognition		→ Running
Name entity recognition	Yesterday, Harry Potter met Hermione Granger.	→ Yesterday, Harry Potter met Hermione Granger .

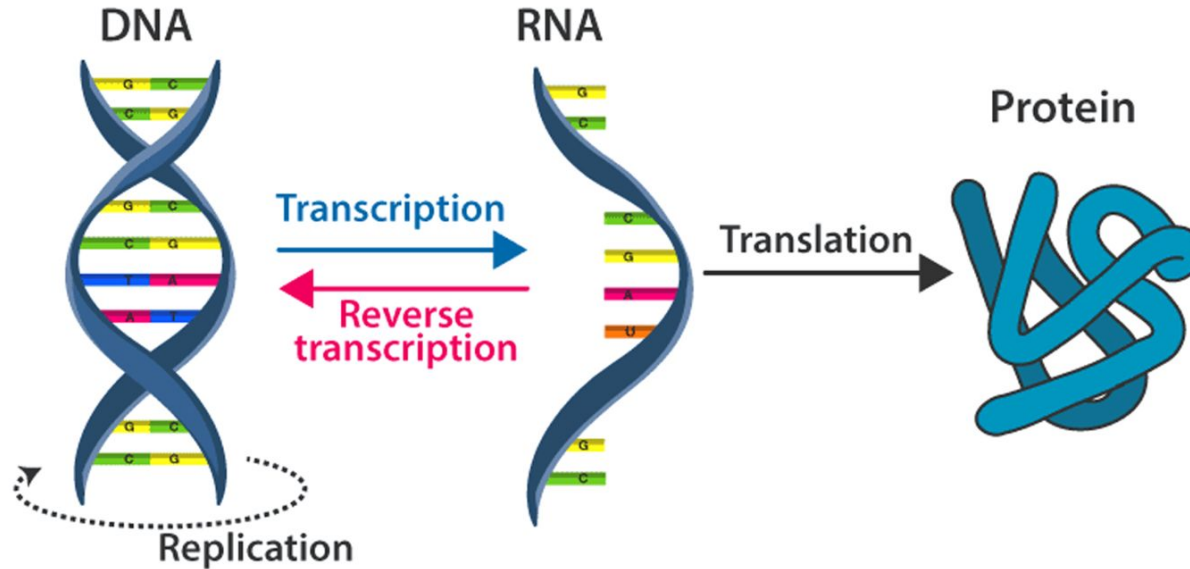
Andrew Ng

Defining a sequence space



Sequence is a core concept in molecular biology

CENTRAL DOGMA : DNA TO RNA TO PROTEIN



Pairwise sequence alignment - a core method for comparison

Pairwise Sequence Alignment

EEELTKPRLWALYFNMRDALSSG

VEKPRILYALYFNMRDSSDE



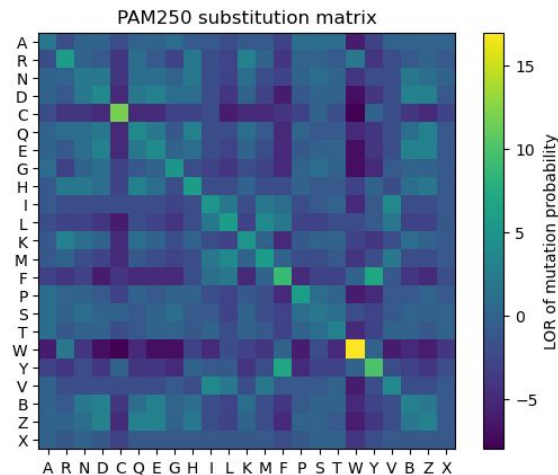
Assuming evolution is non-random, how
can we utilise knowledge about
sequence space for analysis?

Substitution matrix

- **Introduced by Dayhoff et al. (1978)**
- **Log-likelihood** **calculation**
Estimates the probability of mutation from globally aligned sequences with $\geq 85\%$ identity
- **Evolutionary** **model** **assumption**
Relies on a theoretical framework to understand sequence relationships
- **Short-term** **to** **long-term** **extrapolation**
Infers long-term evolutionary patterns from short-term mutation observations
- **Core** **concept**
Models sequence relationships from observed "protein sequence space" to inform future analyses



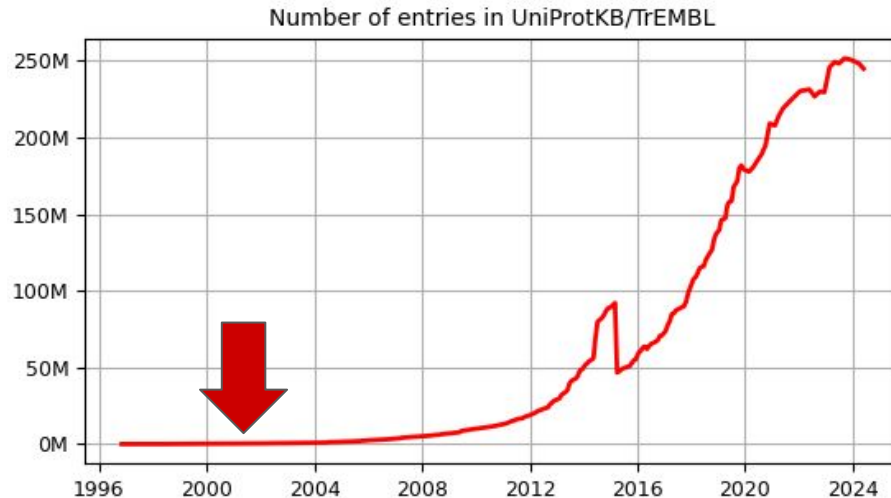
Margaret Dayhoff
(1925-1983)



<https://github.com/althonos/scoring-matrices>

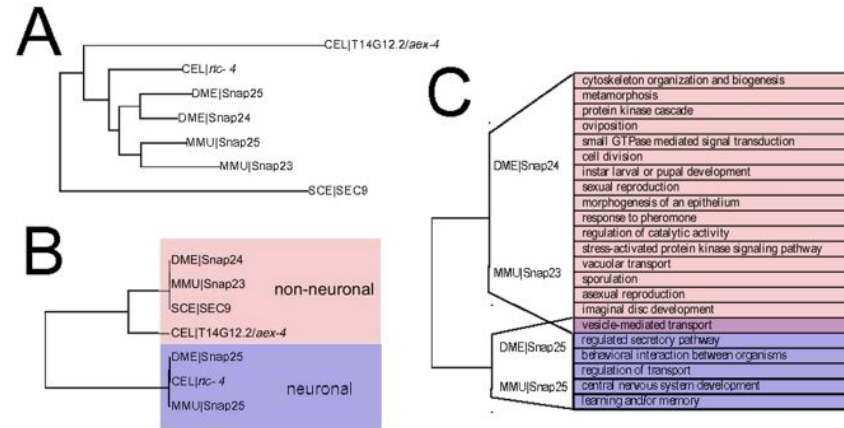
Exponential growth of available sequence data

- The amount of available sequencing data continues to grow exponentially
- Sequence analysis is done via alignment
- The most recent substitution matrix for alignment was constructed almost 25 years ago (VTML, Muller et al., 2001)



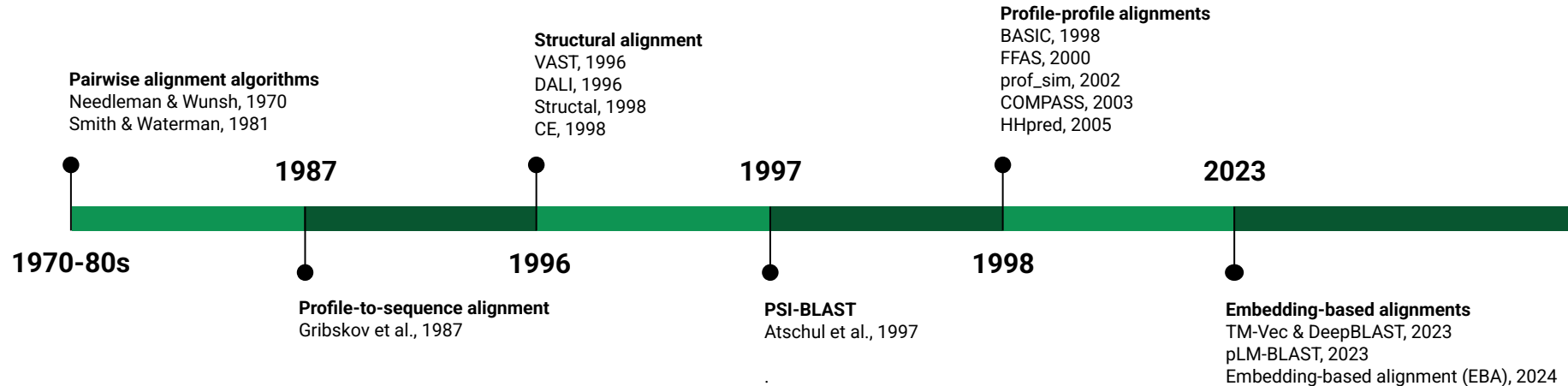
Limitations of the sequence alignment

- Sequence similarity $\neq$ homology
- Homology $\neq$ functional analogy
- Sequence alignment may lead to an incorrect identification of protein relationships



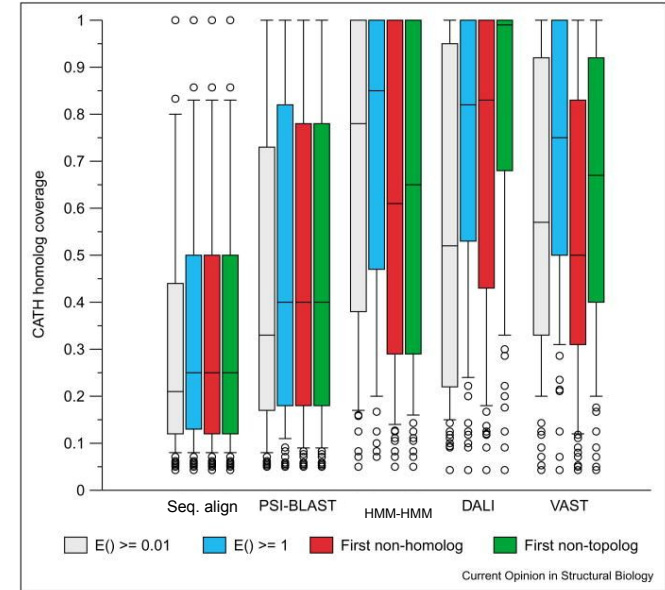
Difference between sequence similarity
and GO annotation of sequences
Chikina & Troyanskaya, 2011

A timeline of protein comparison methods



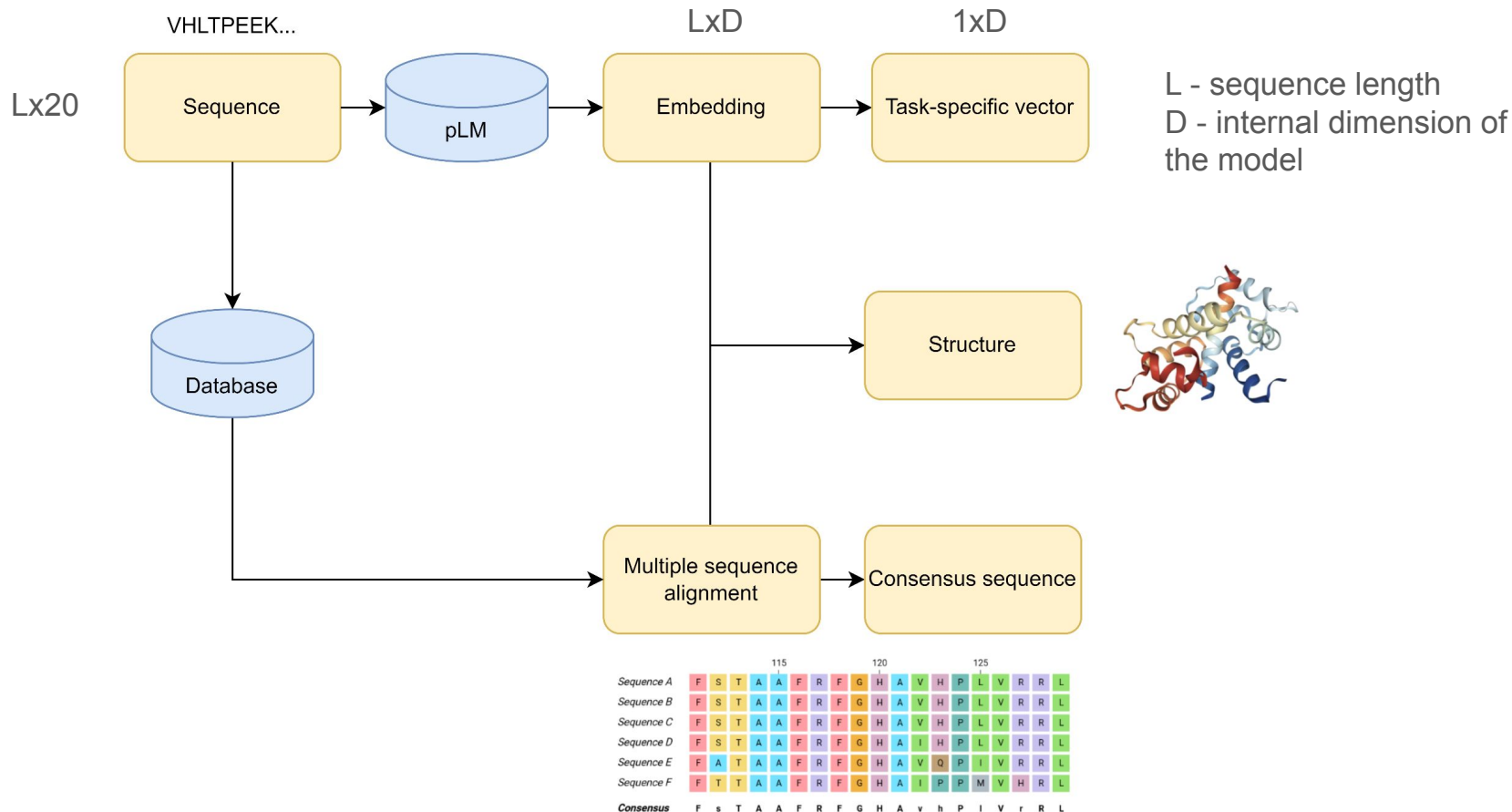
Performance of different sequence comparison methods

- Profile/HMM/PSSM methods are more sensitive than single-sequence comparison methods
- Profile-profile methods identify more homologs than PSI-BLAST
- Structure alignment identifies more homologs than profile-profile methods

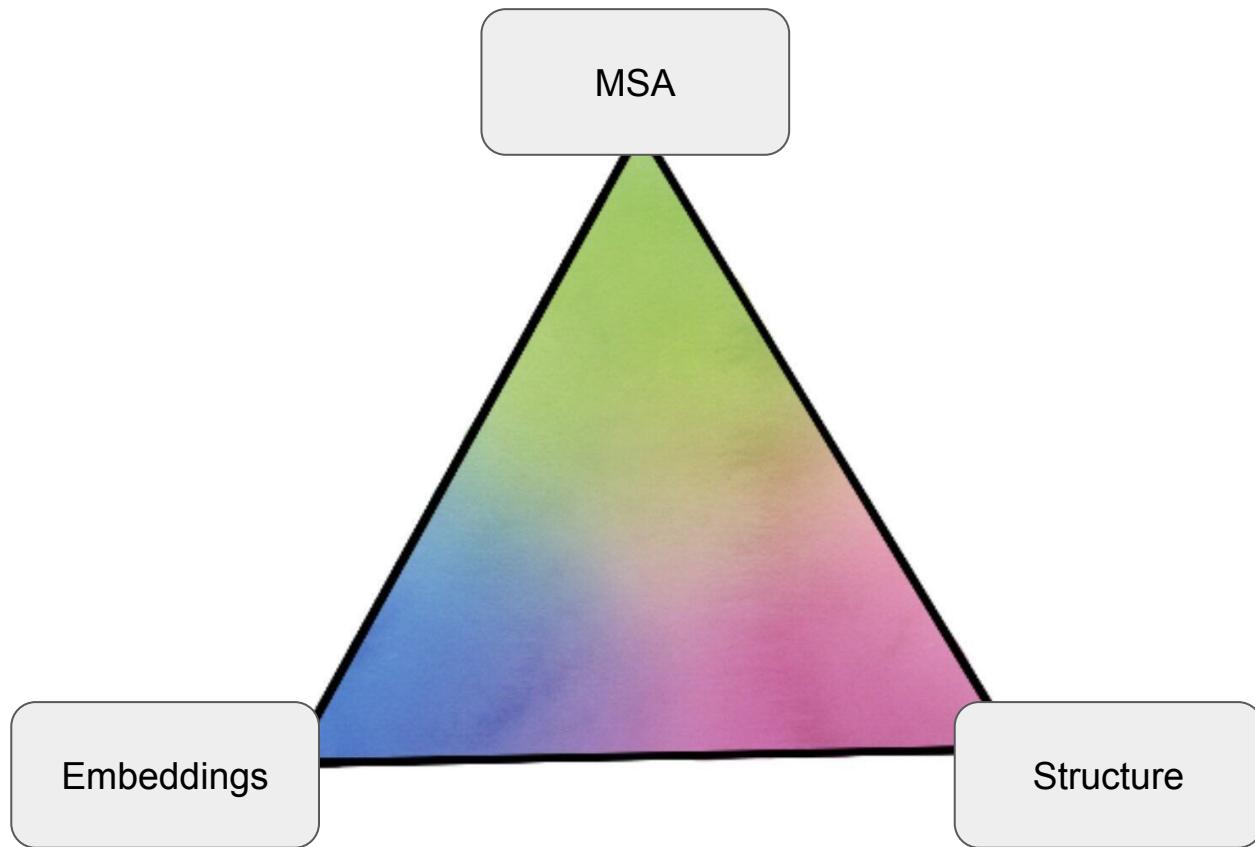


Pearson & Sierk, 2005

Modern computational representations of a protein



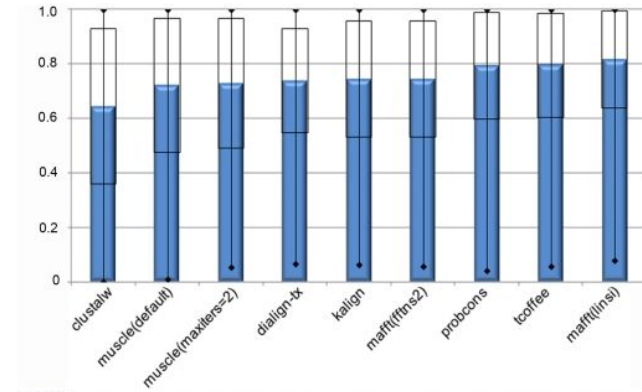
Computational representations with prior knowledge



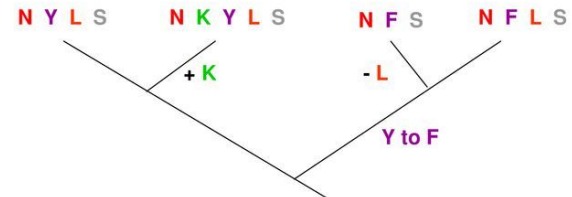
Multiple sequence alignment

Multiple sequence alignment

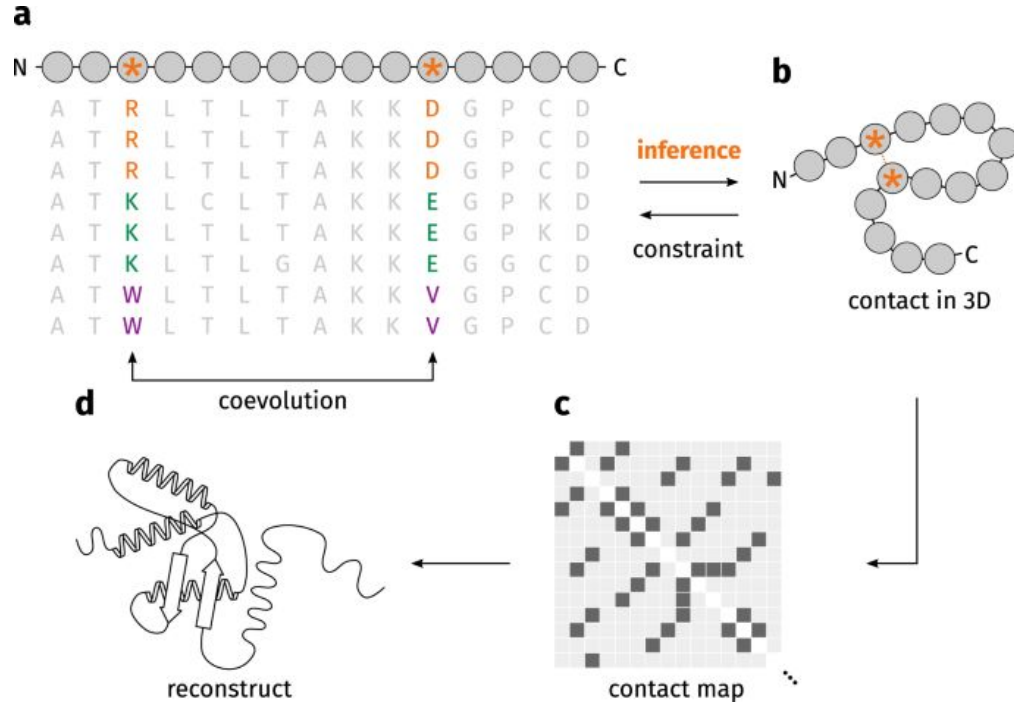
- Computationally challenging
 - Uses heuristics
 - Alignments are usually suboptimal
- Can be manually curated
 - Not scalable
 - Citizen-science was used (Borderlands Science by Knight Lab)
- MSA helps to identify conserved and coevolving residues
 - Can be used for phylogenetics



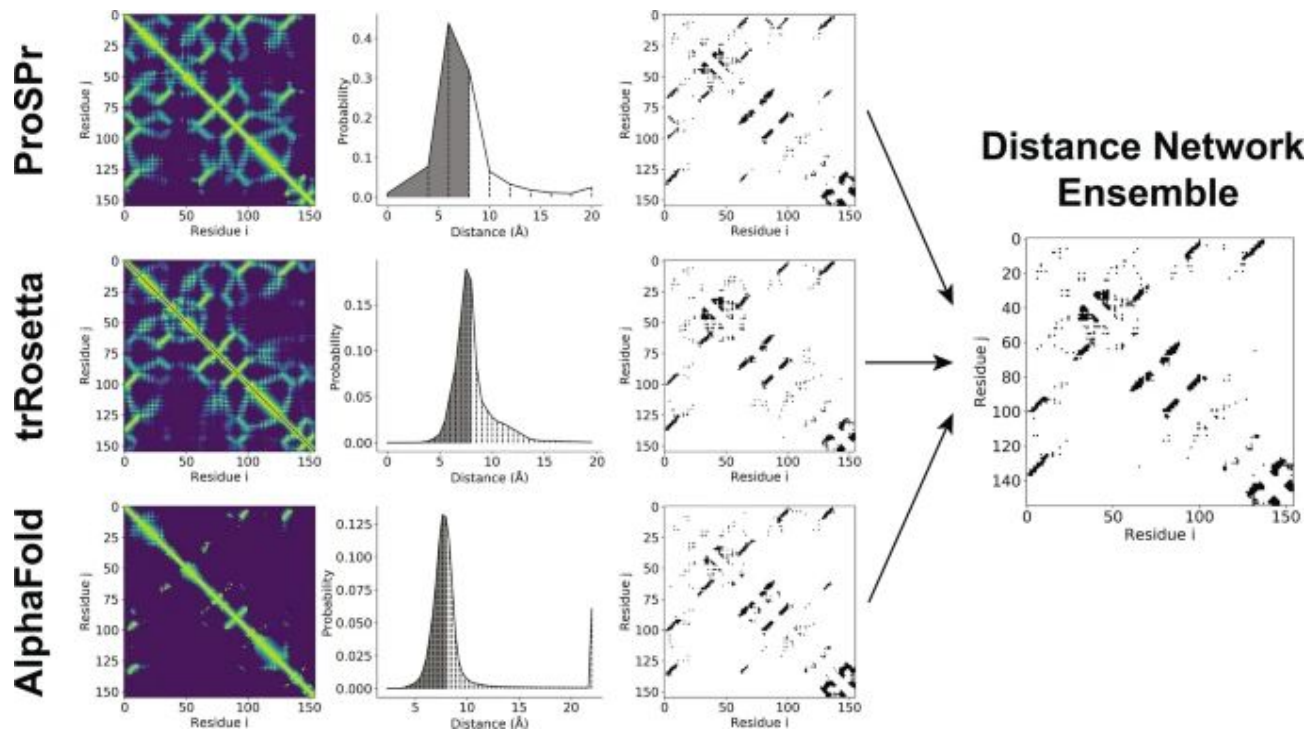
seqA	N	-	F	L	S
seqB	N	-	F	-	S
seqC	N	K	Y	L	S
seqD	N	-	Y	L	S



Coevolutionary information predicts protein structure

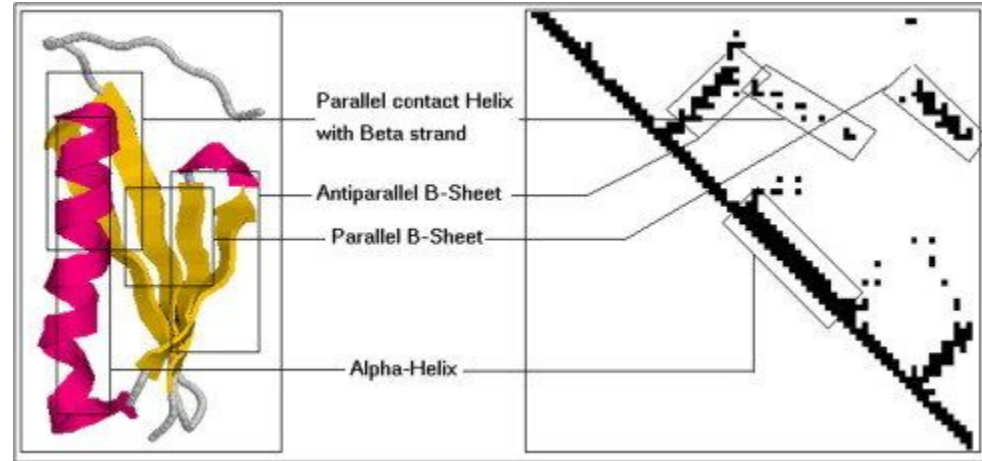


Extracting contacts is fundamental for structure prediction

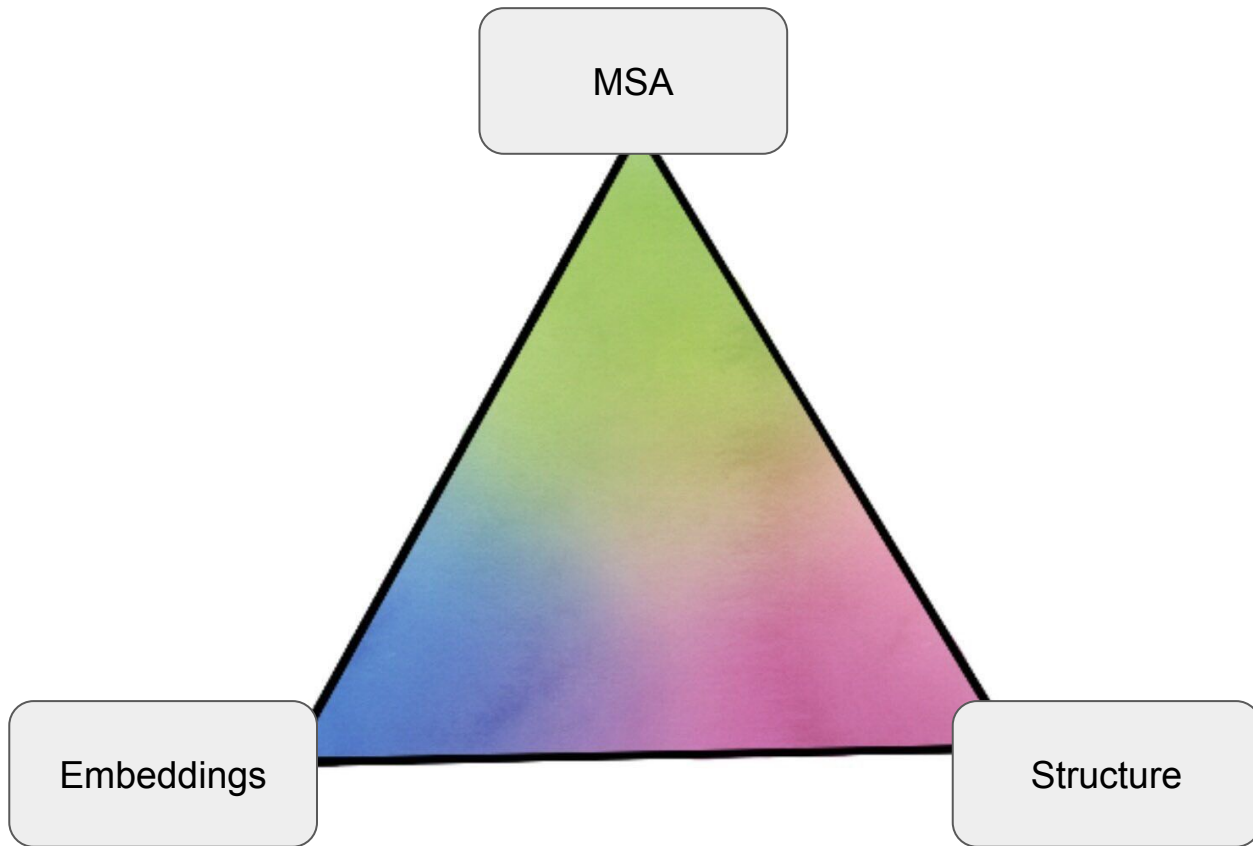


From contacts to protein structure

- Contacts map are 2D representation of a 3D protein structure
- 3D is predicted from 2D map

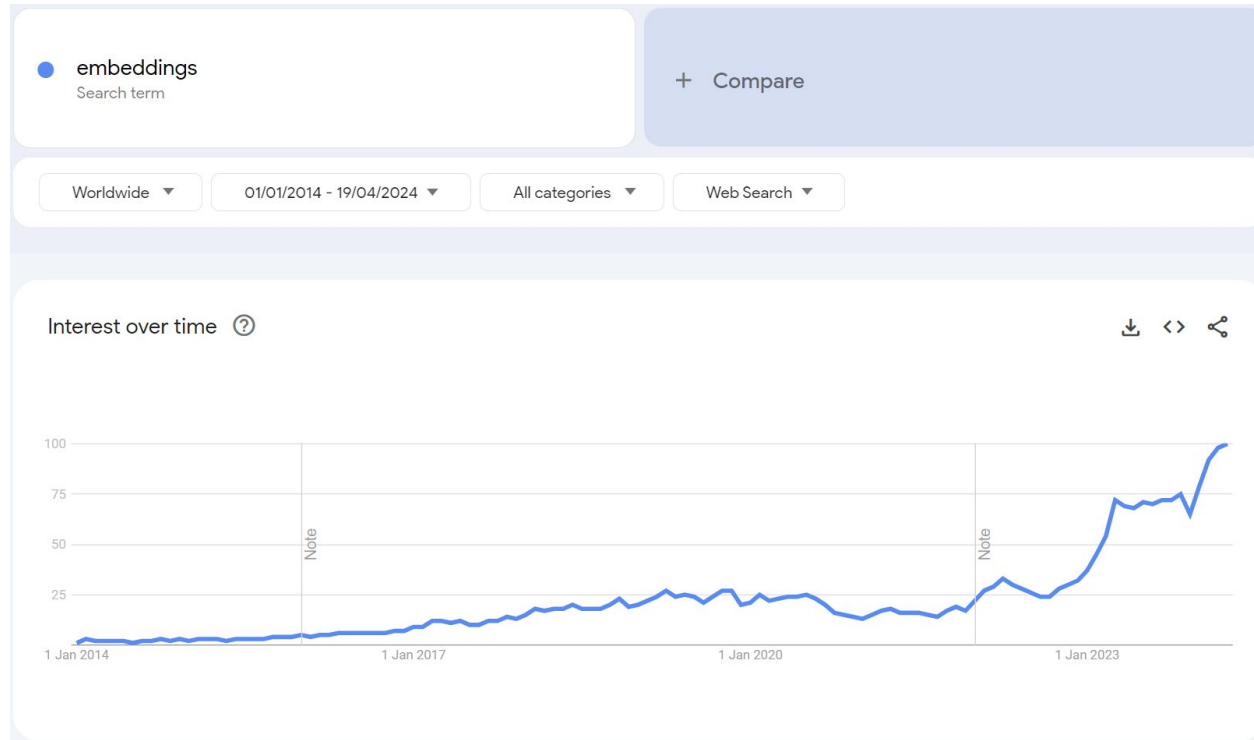


Computational representations with prior knowledge



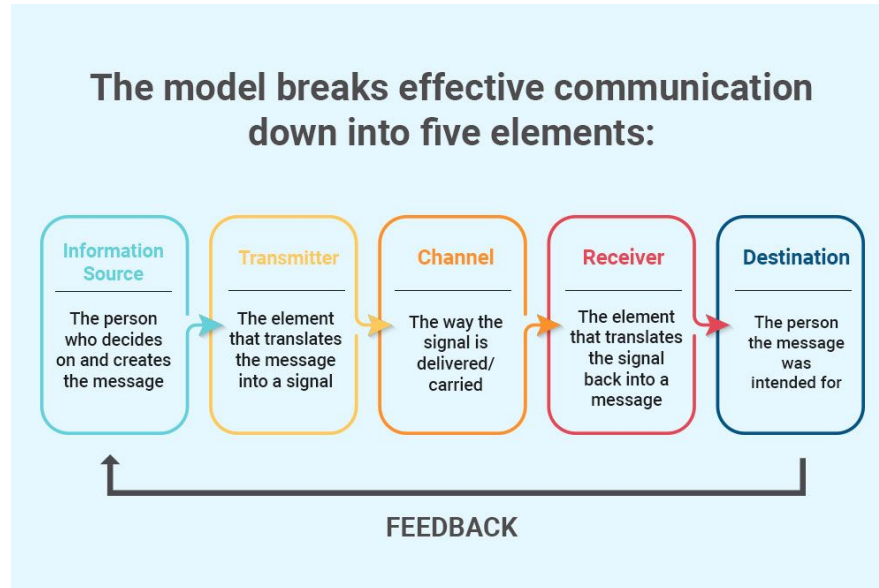
Embeddings

Embeddings are at the peak of popularity



Distributional semantics - perfect theory of language

- A word is characterized by the company it keeps (Firth, 1950)
- Words that occur in the same contexts tend to have similar meanings (Harris, 1954)
- Meaning is learned by inferring how words co-occur in natural language (Kumar 2021; Sahlgren 2008)
- This can be done by next-token prediction (Shannon 1948; Shannon 1951)



Word embeddings

- **Definition:** Dense, distributed, fixed-length word vectors, built using word co-occurrence statistics as per the distributional hypothesis.

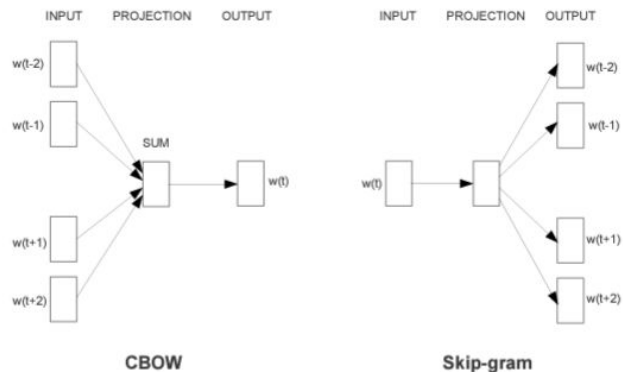


Figure 1: New model architectures. The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word.

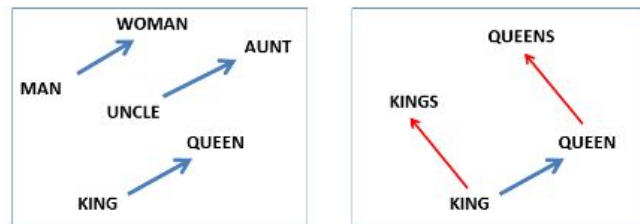
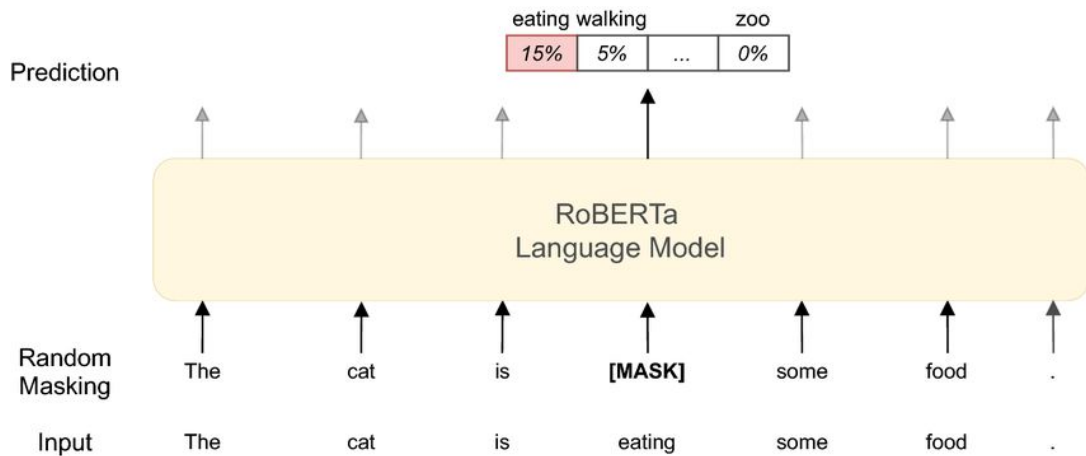


Figure 2: Left panel shows vector offsets for three word pairs illustrating the gender relation. Right panel shows a different projection, and the singular/plural relation for two words. In high-dimensional space, multiple relations can be embedded for a single word.

Mikolov et al., 2013

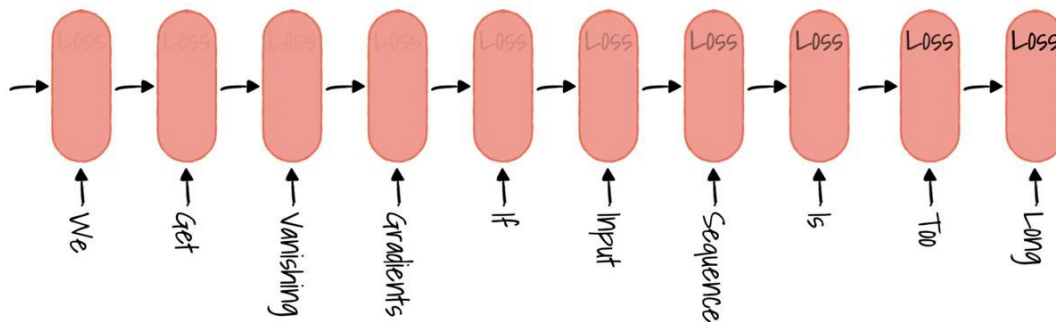
Learning co-occurrence statistics via next token prediction



Quick history of transformers

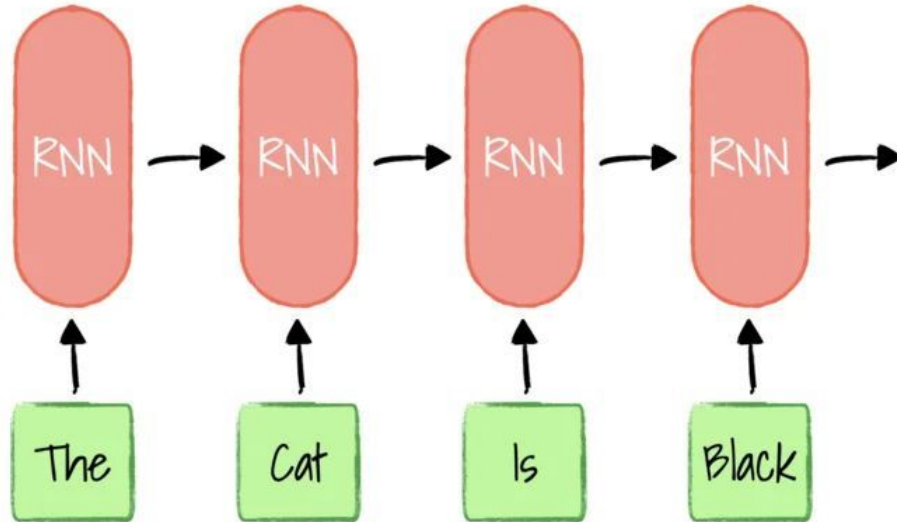
Recurrent neural networks dominated the field since 90's

- Sequential input
- To evaluate the prediction the whole sentence had to be fed word by word
- Vanishing gradients

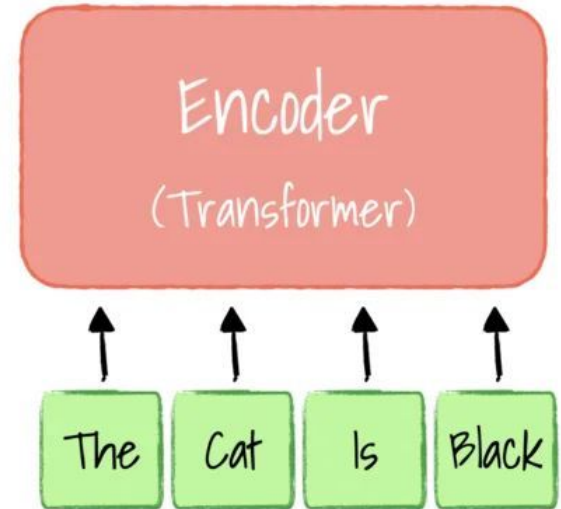


Transformers address all problems of RNN

RNN based Encoder

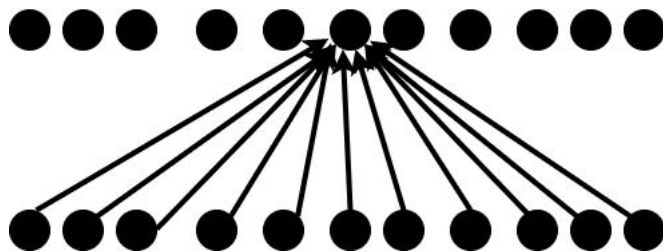


Transformer's Encoder

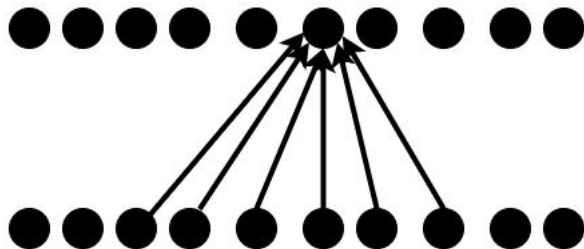


Preserving context

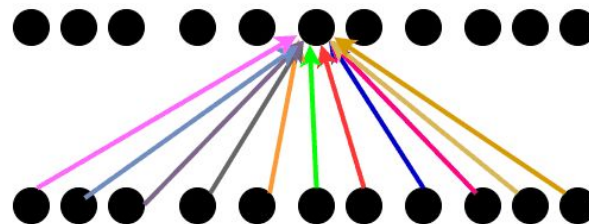
Fully Connected layer



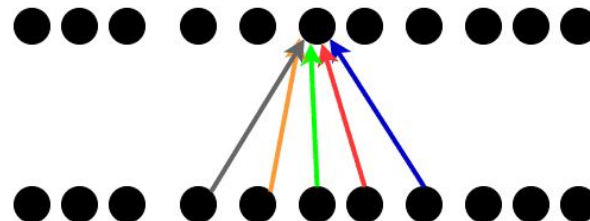
Convolution



Global attention



Local attention



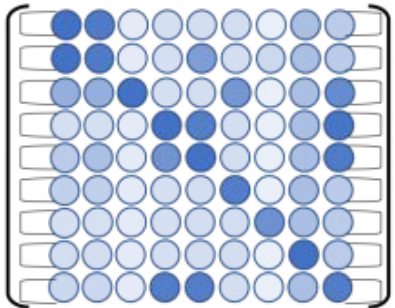
Preserving non-linear relationships

Can you me help this sentence to translate
↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑
Kannst du mir helfen diesen Satz zu uebersetzen ?

Can you help me to translate this sentence
↑ ↑ × × × ×
Kannst du mir helfen diesen Satz zu uebersetzen ?

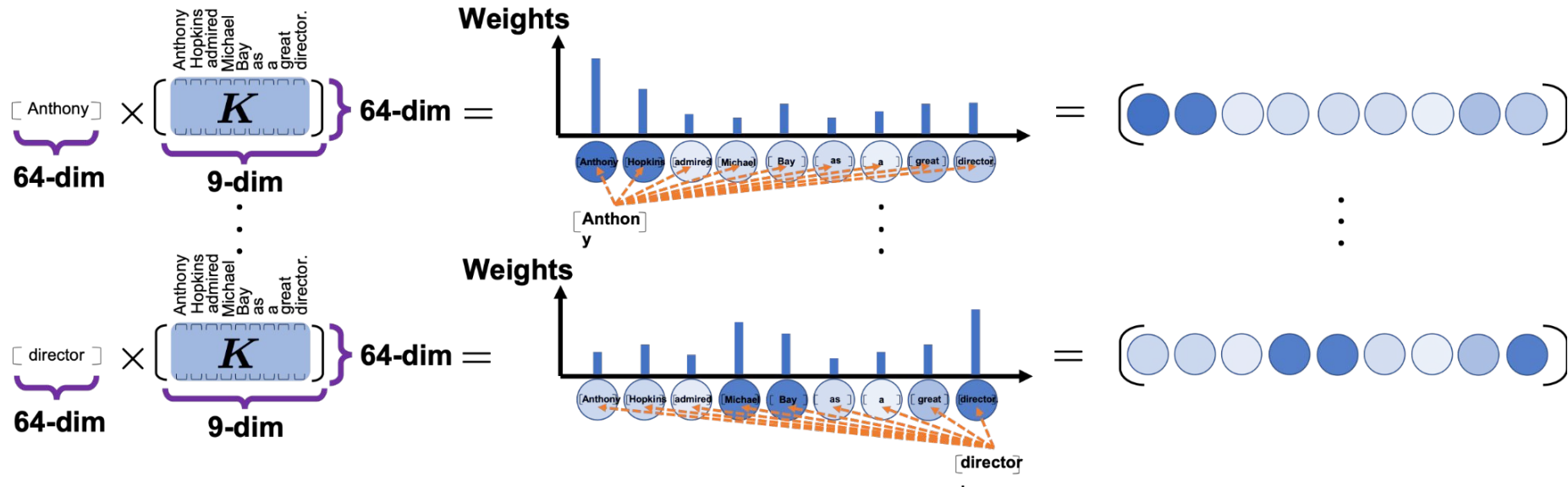
Attention is all you need

softmax $\left(\frac{\begin{matrix} \text{Anthony} \\ \text{Hopkins} \\ \text{admired} \\ \text{Michael} \\ \text{Bay} \\ \text{as} \\ \text{a} \\ \text{great} \\ \text{director.} \end{matrix} \begin{bmatrix} Q \end{bmatrix} \times \begin{matrix} \text{Anthony} \\ \text{Hopkins} \\ \text{admired} \\ \text{Michael} \\ \text{Bay} \\ \text{as} \\ \text{a} \\ \text{great} \\ \text{director.} \end{matrix} \begin{bmatrix} K \end{bmatrix}}{\sqrt{d_k}} \right) =$

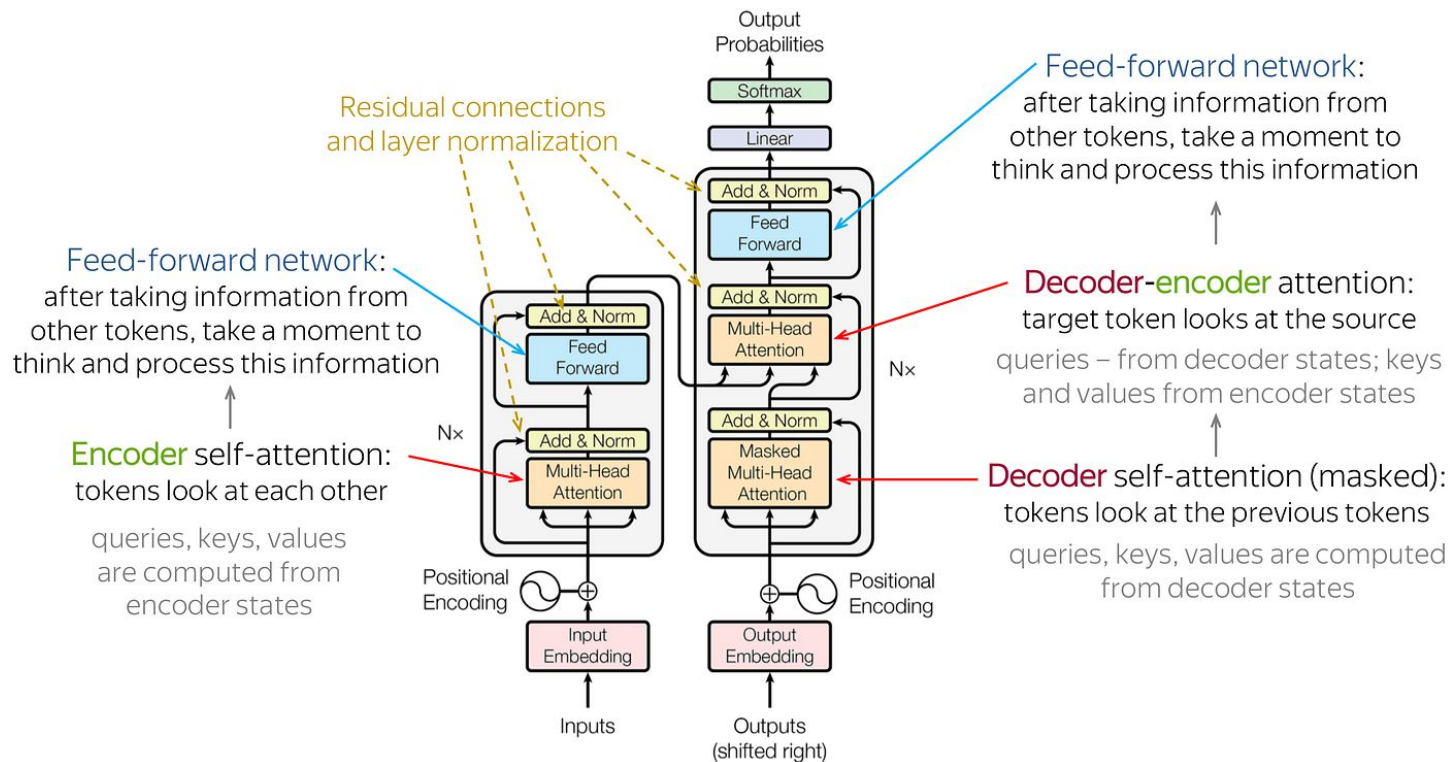


Importantly, the sum of each row is 1.

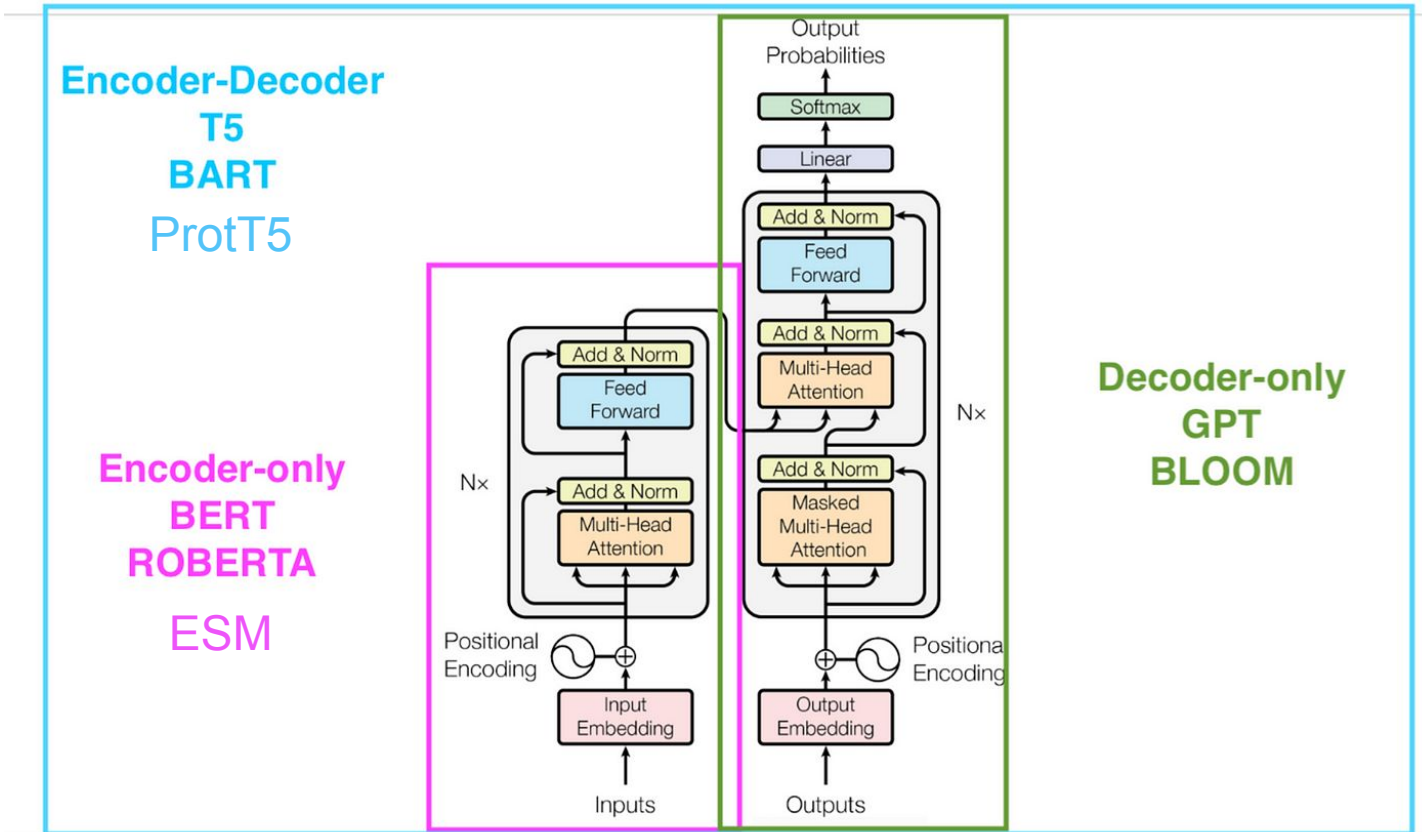
Attention is learnable



Transformer - a cornerstone of modern sequence modeling

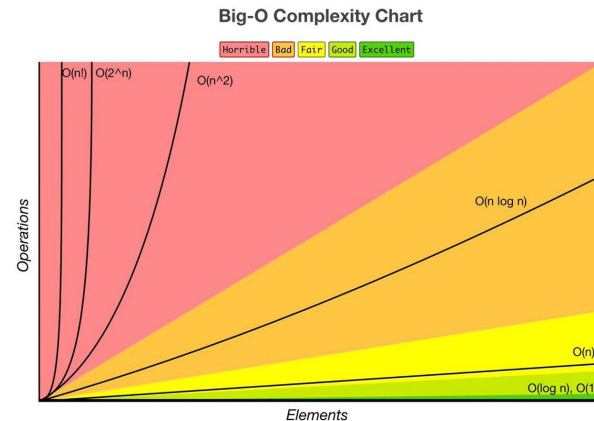
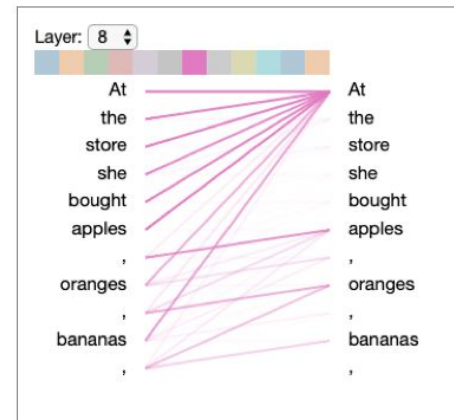


Encoder-decoder LLMs



Attention has quadratic complexity

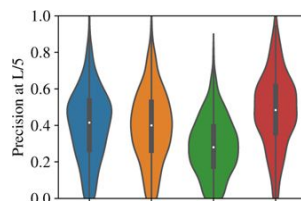
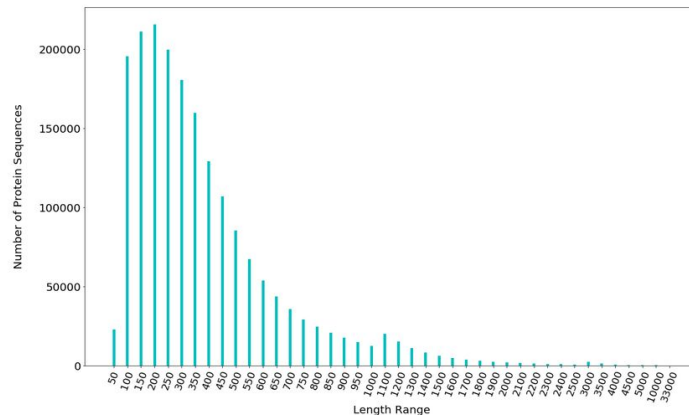
- Dimensionality of attention matrix grows
- Quadratic complexity is the worst complexity encountered in practice
- First implementations (BERT) had a limit of 512 tokens
- Various optimisations have been proposed since (see FlashAttention)



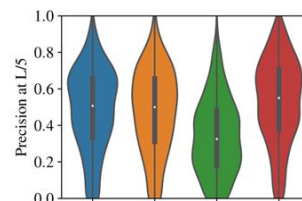
Bridging information theory, linguistics and biology

Protein sequence length does not limit LLMs

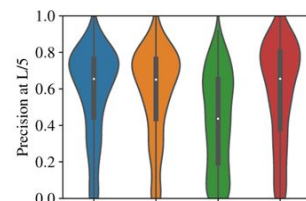
- 512 token limit captures 90% of known sequence space
- Long distance interactions are not solved problem within LLMs



(a) Short-range



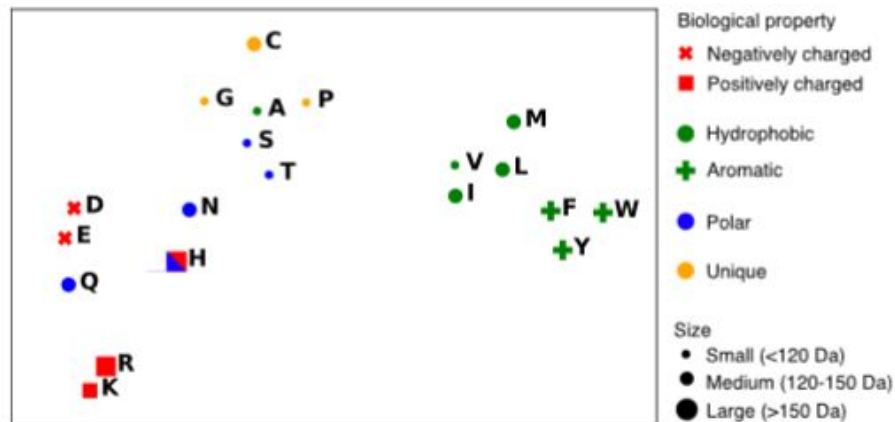
(b) Medium-range



(c) Long-range

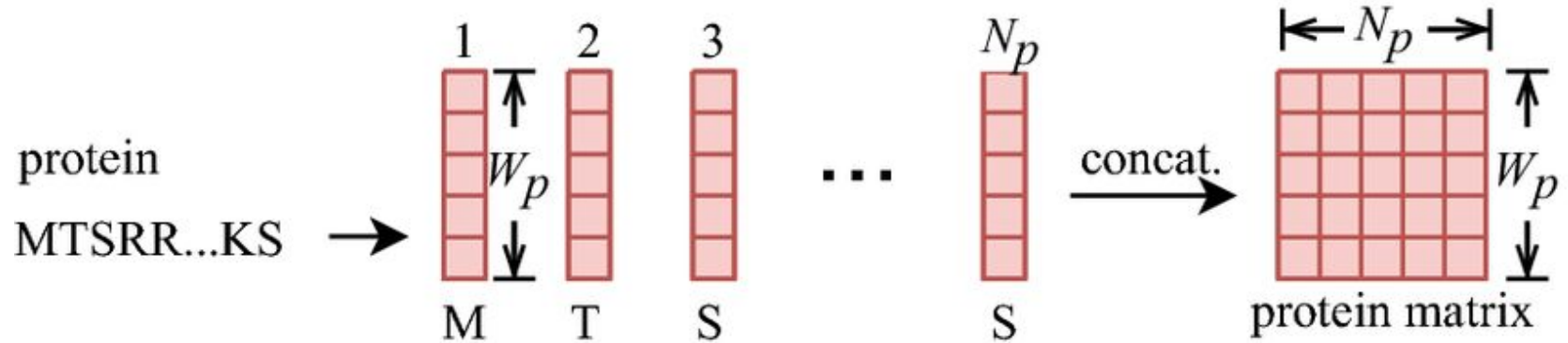
Protein language models extract amino acid features

- Unsupervised training provides meaningful amino acid representations
- They correlate with physicochemical qualities of the residues



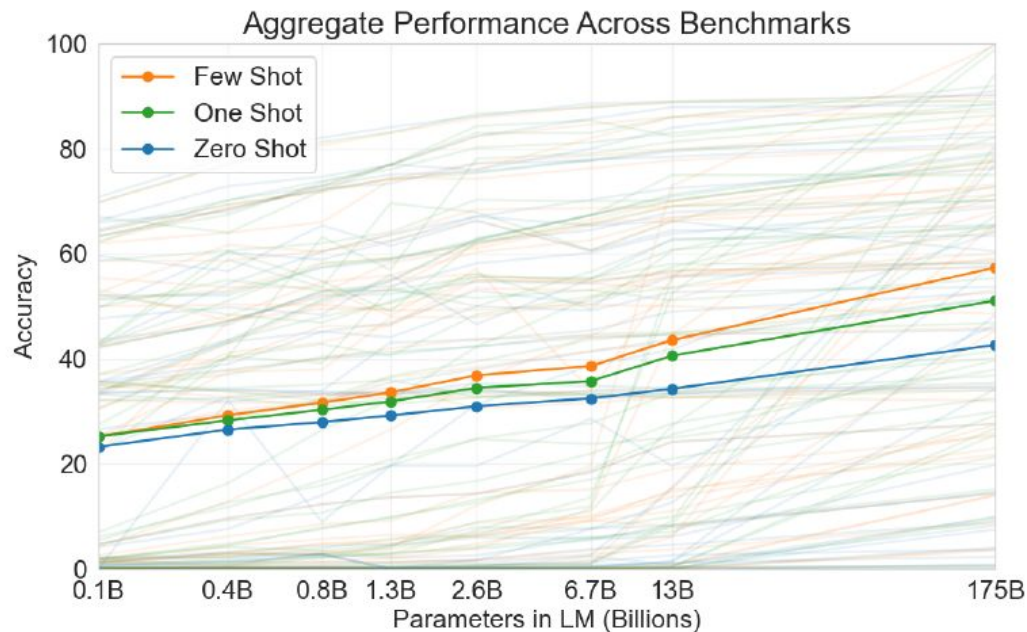
Rives et. al 2022

Forward pass through pLM generates a protein matrix



Embeddings are fundamental for good performance of LLMs

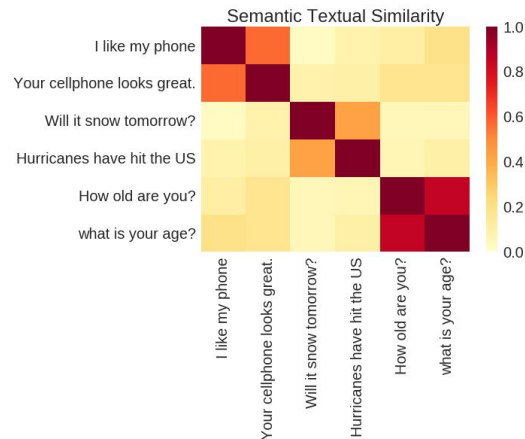
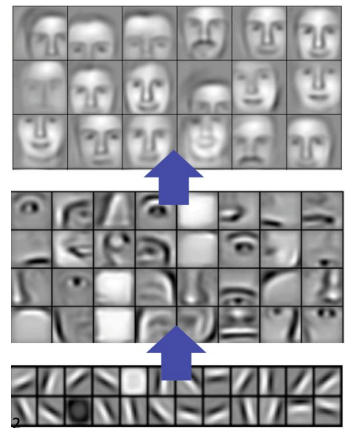
- Unsupervised embeddings allow LLMs to perform variety of tasks without fine-tuning
- Number of parameters allows for more precise fit to the distribution in sequence space and prediction of the next token (perplexity)



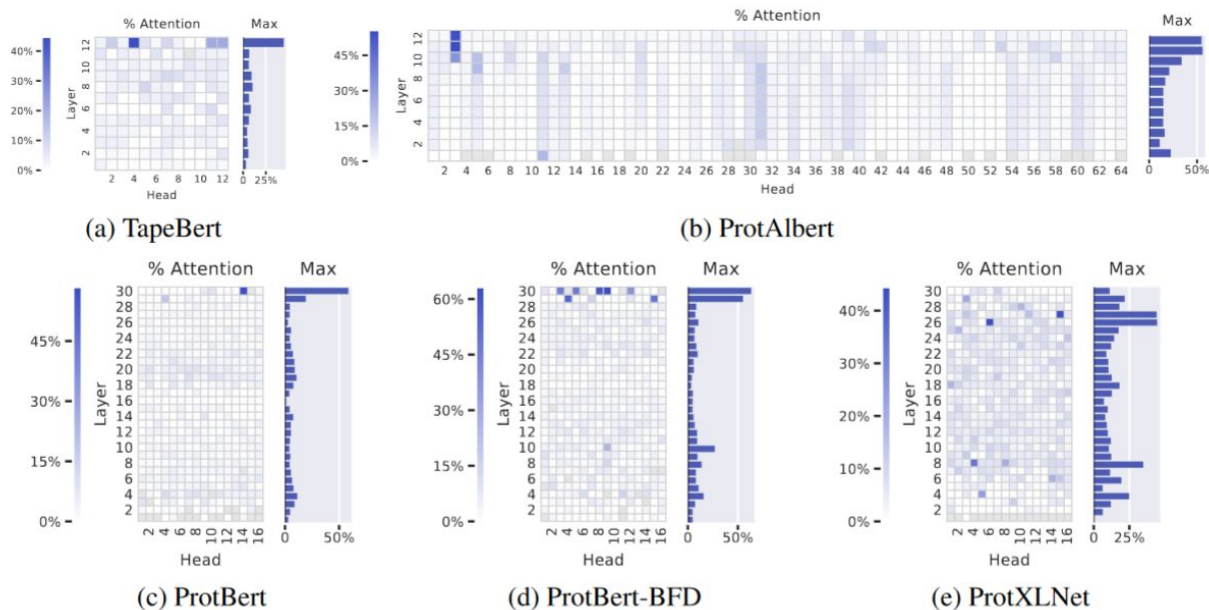
Brown et al. 2020

Neural networks compress data via unifying features

- Neural networks can be seen as data compression algorithms
- Models learn unifying (compressed) features for multiple different objects



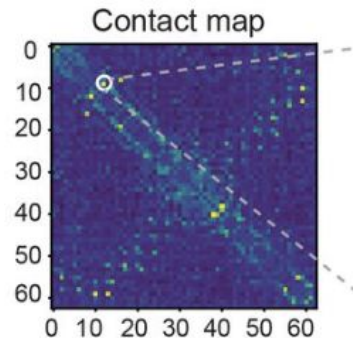
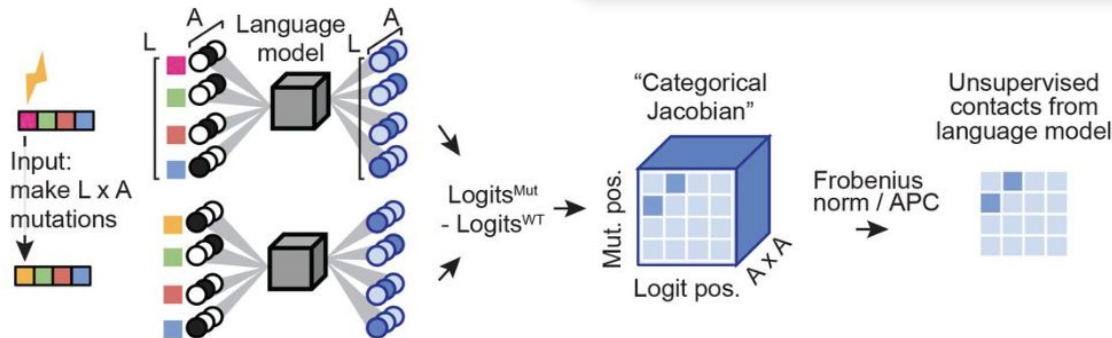
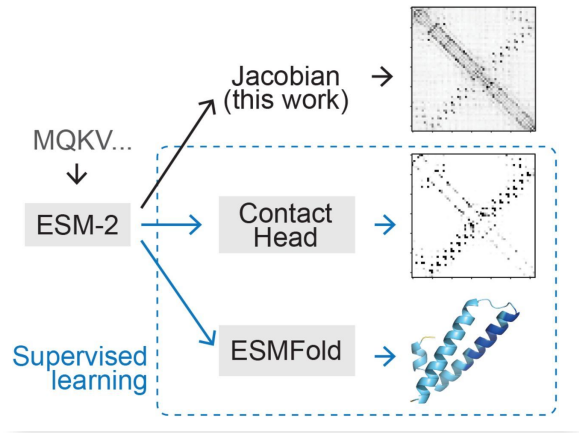
Models learn contact maps as a form of data compression



Vig et al., ICLR 2021

pLMs encodes structural features without supervision

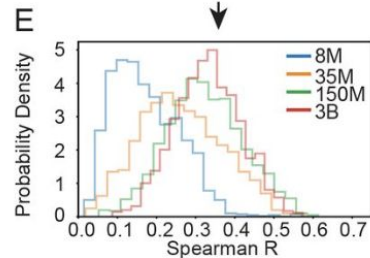
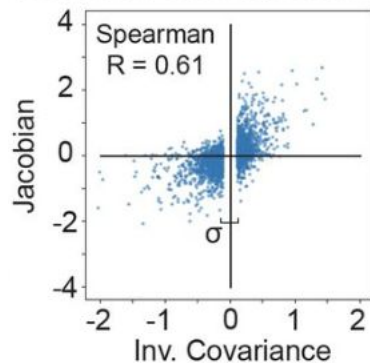
- Contact map for the protein can be extracted via perturbation of the input (Zhang et al. 2024)



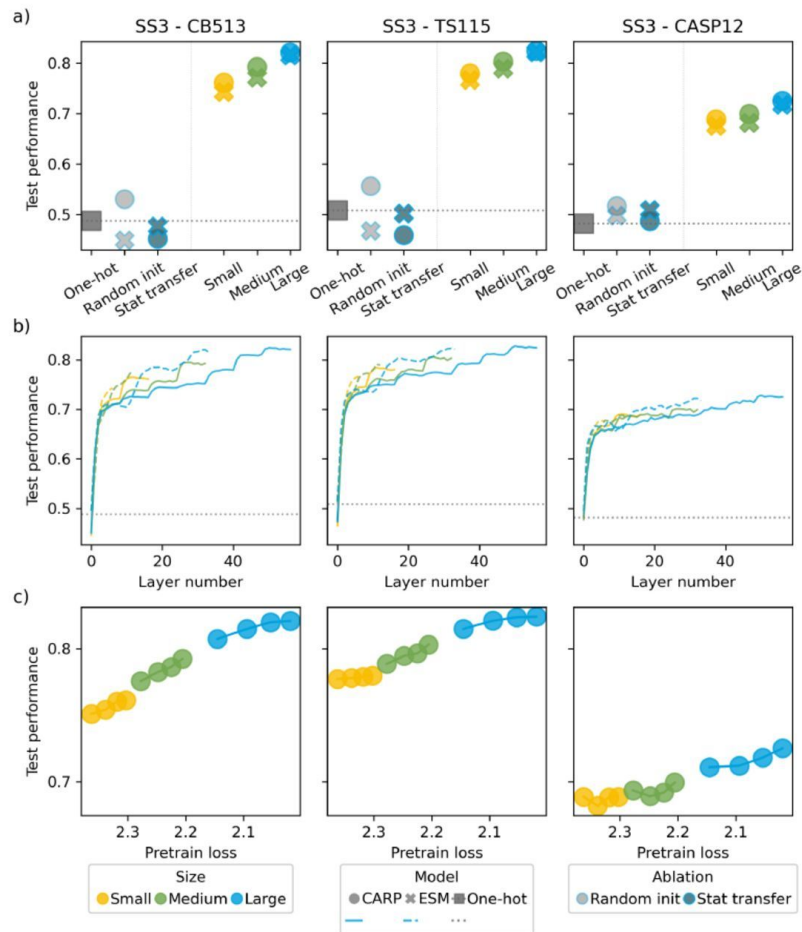
Embeddings similarly to MSA incorporate evolutionary information

- Capture amino acid conservation and variant effects (Marquet et al., 2022)
- A high correlation of encoded signal with signal extracted from MSA (Zhang et al., 2024)

Compare pairwise potentials:
Ribosomal subunit protein RL29



Structure learning is well explained by model investigation



Other downstream tasks for protein embeddings

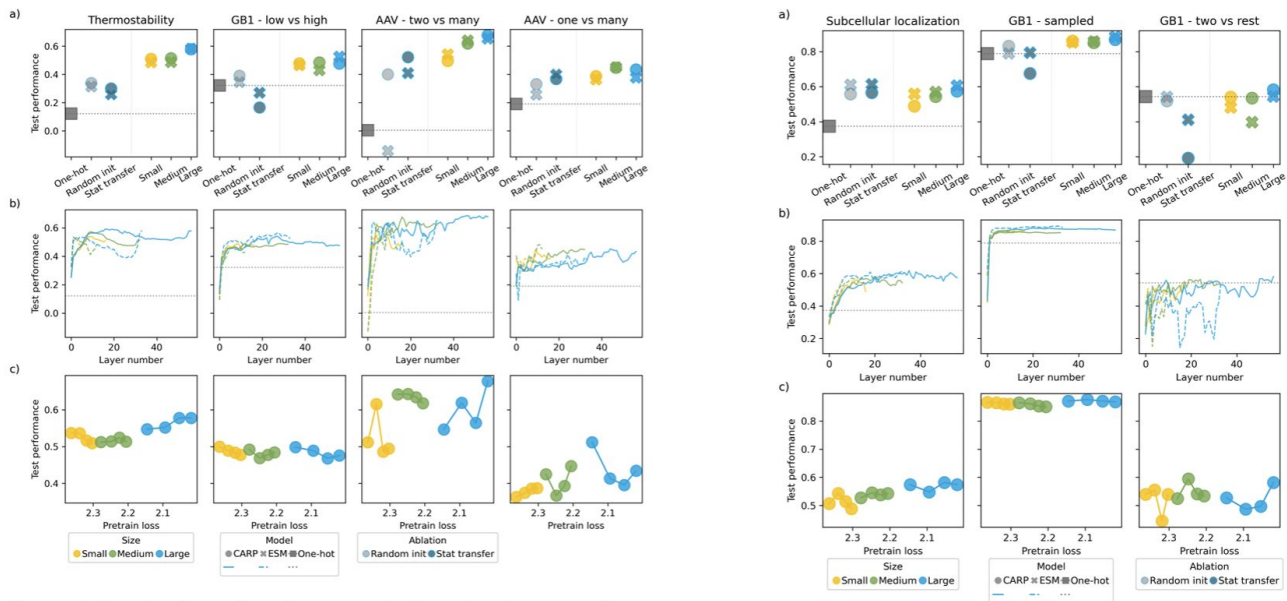
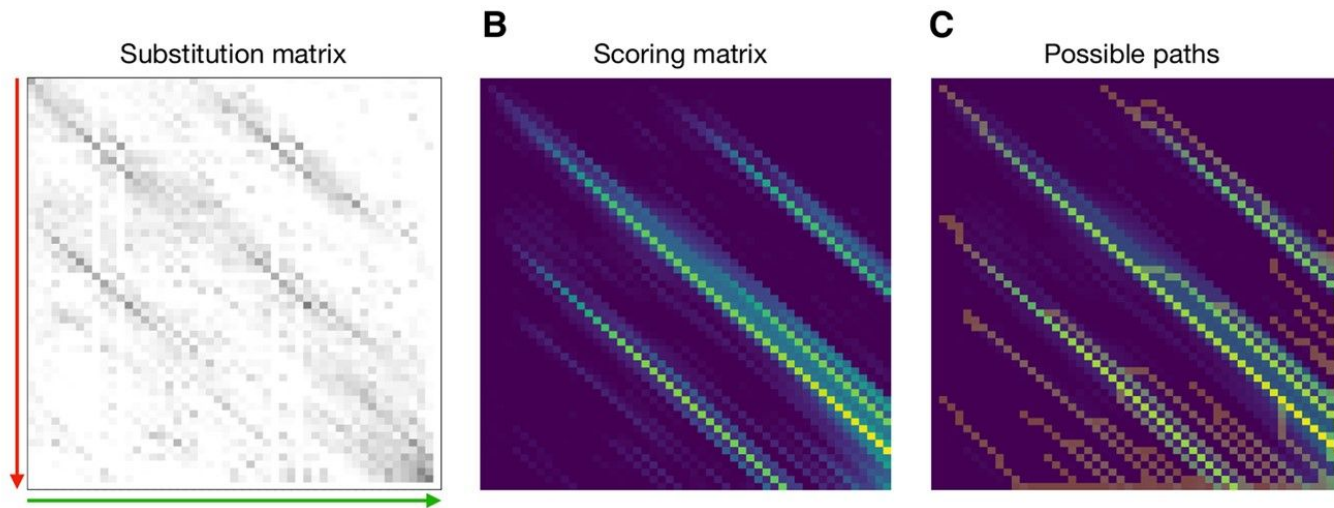


Figure 4. Results for other downstream tasks for learning from sequence data.

Li et al., 2024

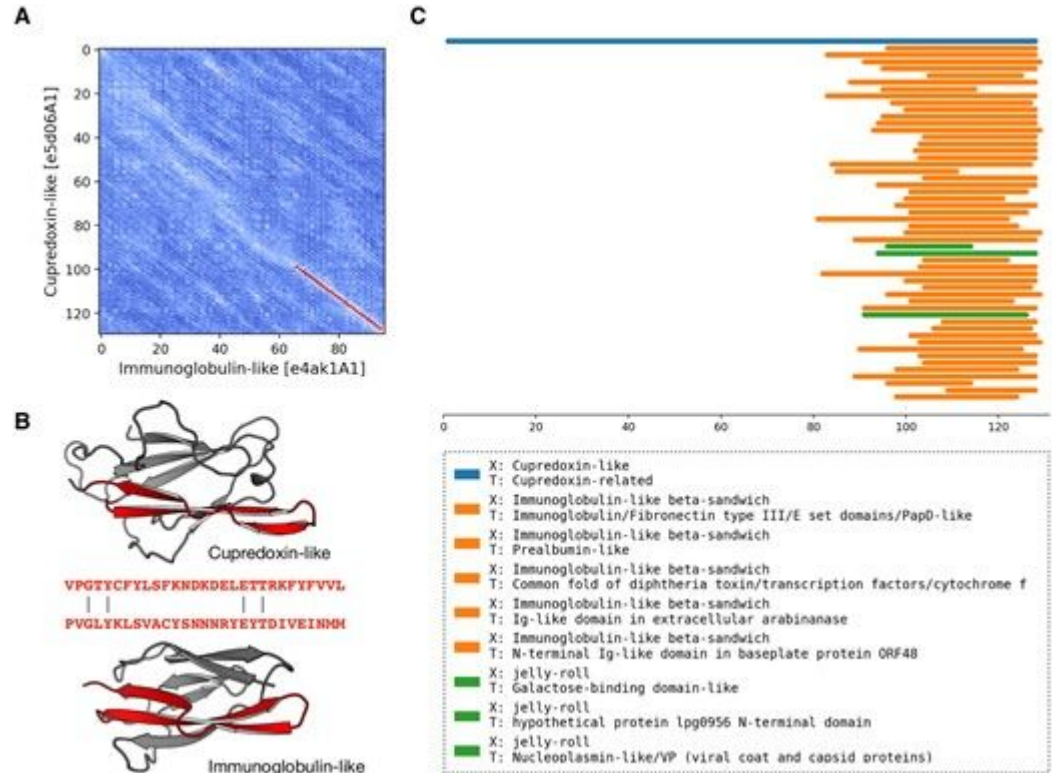
Substitution matrices from embeddings

- From raw embeddings
 - pLM-BLAST (Kamiński et al, 2023)
 - EBA (Pantolini et al, 2024)
 - PEba (Iovino & Ye, 2024)
 - PLMSearch (Liu et al., 2024)



pLM-BLAST highlights conserved structural motifs

- Alignment of structurally similar cupredoxin and IgG reveals conserved beta-hairpin
- Highlights the hypothesis of the common origin of the proteins, “decorated” with secondary structural elements



DEDAL - coevolution informed substitution matrix

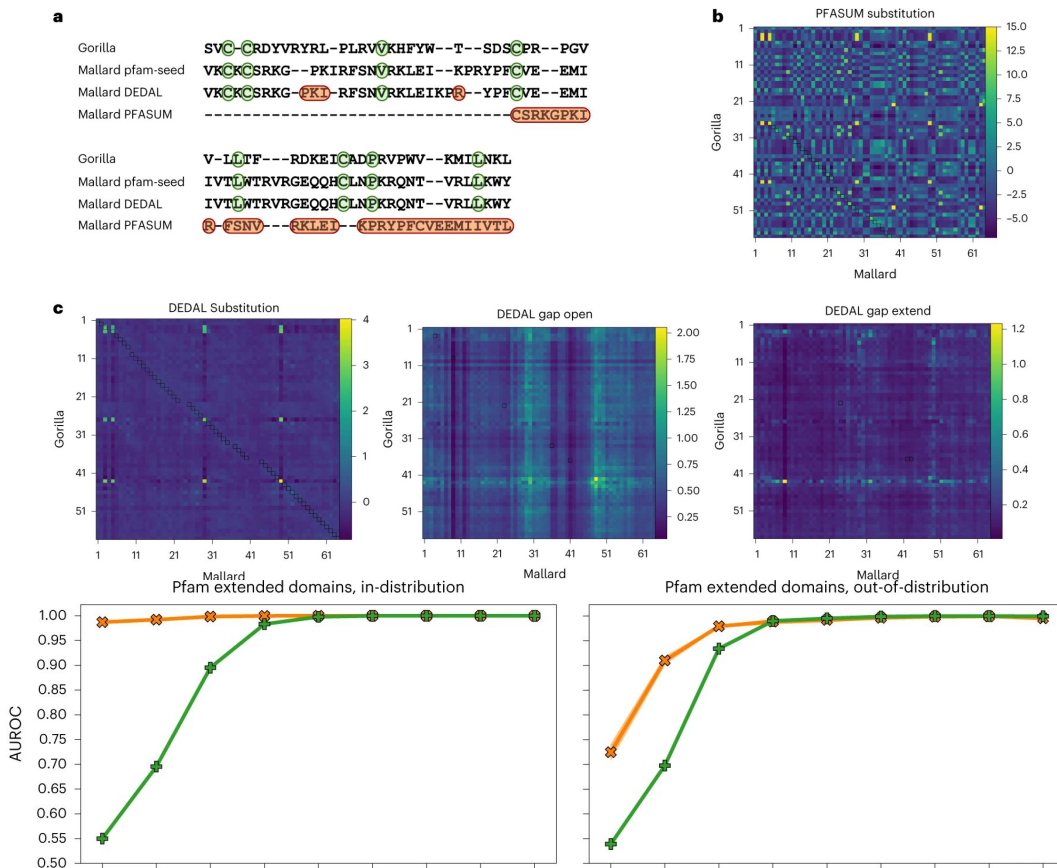
Correspondence | Published: March 2008

BLOSUM62 miscalculations improve search performance

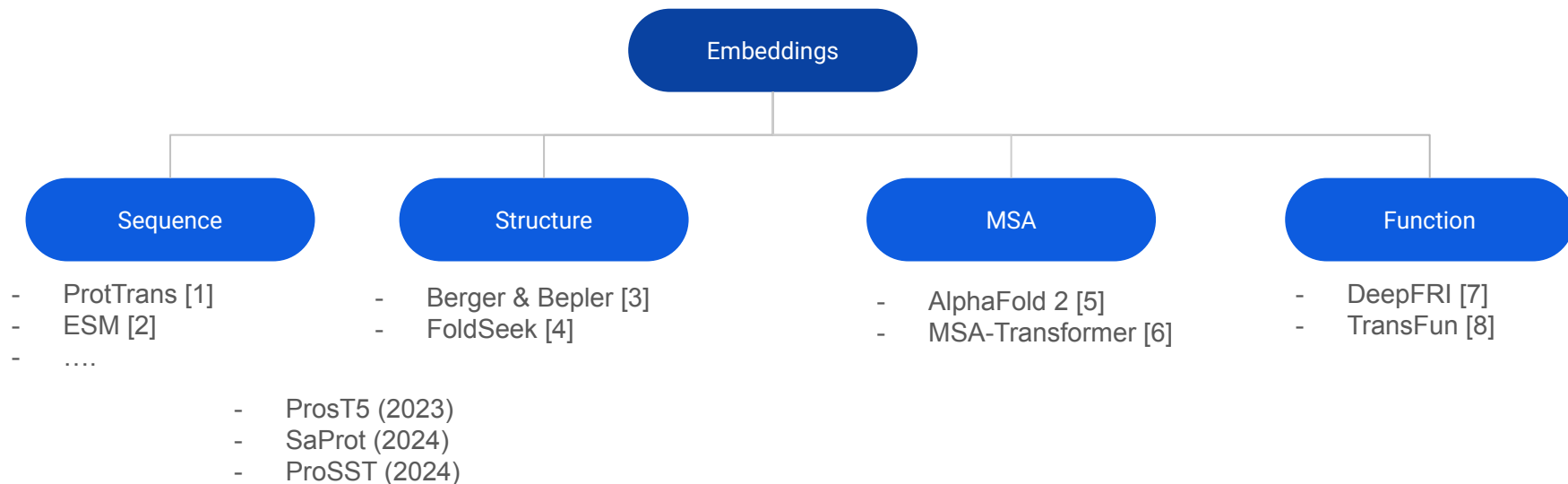
[Mark P Styczynski](#), [Kyle L Jensen](#), [Isidore Rigoutsos](#) & [Gregory Stephanopoulos](#)

[Nature Biotechnology](#) 26, 274–275 (2008) | [Cite this article](#)

- Selecting a proper substitution matrix is hard because evolutionary relationships are unknown
- We can condition substitution matrix on coevolutionary information of input sequences
- Significantly improves alignment of distant proteins

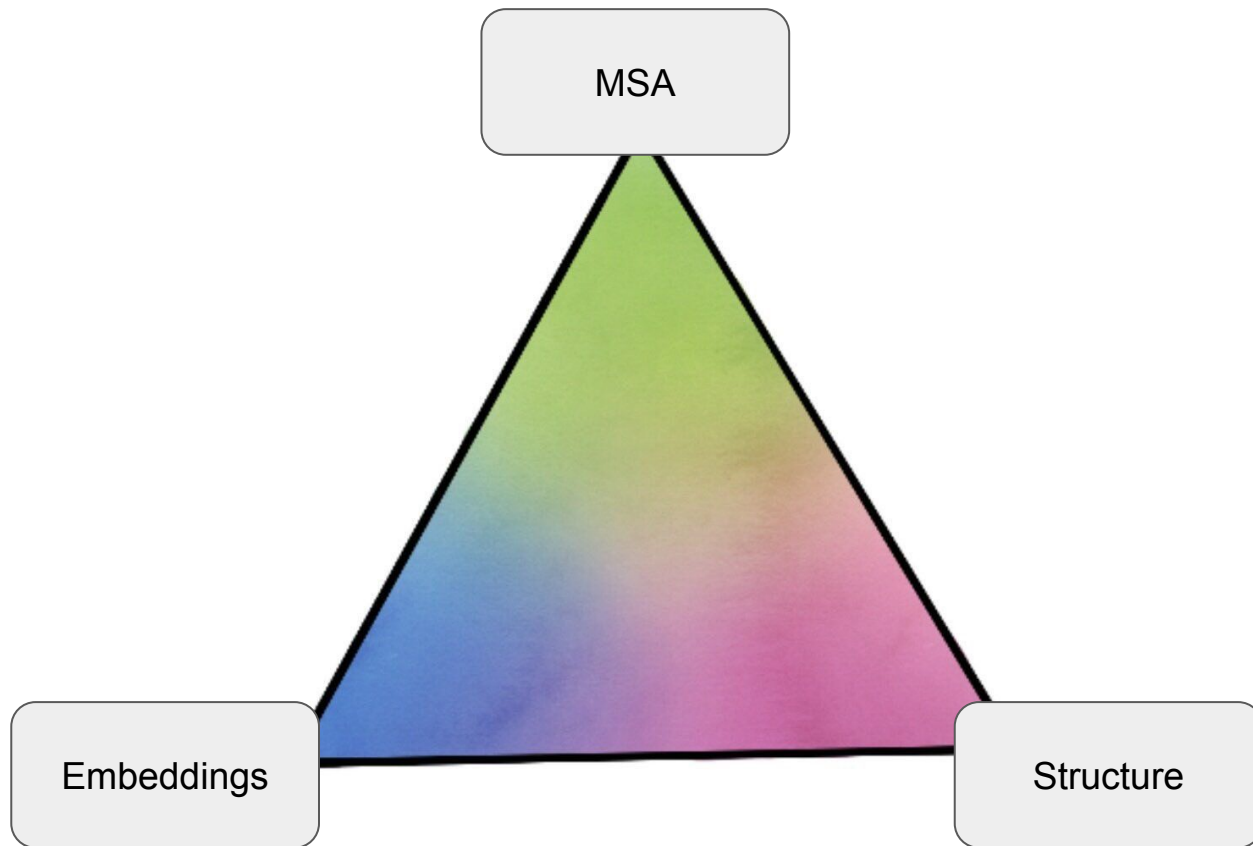


Other types of protein embeddings



1. Elnaggar et al., 'ProtTrans'.
2. Rives et al., 'Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences'.
3. Bepler and Berger, 'Learning Protein Sequence Embeddings Using Information from Structure'.
4. van Kempen et al., 'Fast and Accurate Protein Structure Search with Foldseek'.
5. Jumper et al., 'Highly Accurate Protein Structure Prediction with AlphaFold'.
6. Rao et al., 'MSA Transformer'.
7. Gligorić et al., 'Structure-Based Protein Function Prediction Using Graph Convolutional Networks'.
8. Boadu, Cao, and Cheng, 'Combining Protein Sequences and Structures with Transformers and Equivariant Graph Neural Networks to Predict Protein Function'.

Computational representations with prior knowledge



Structure

Embeddings are predictive of structure

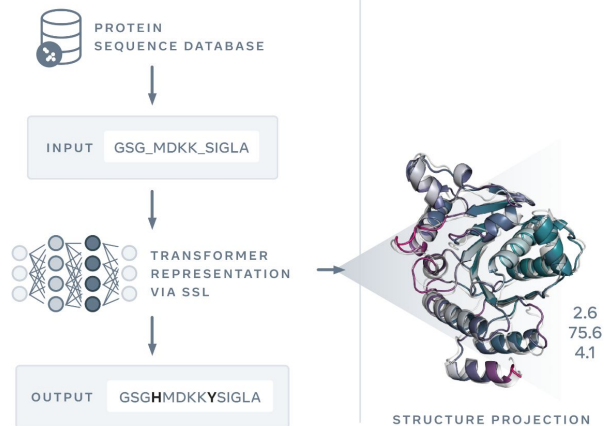
- ProtTrans (Elnaggar et. al, 2022)
- OmegaFold (Wu et al., 2022)
- AminoBERT (Chowdhury et al., 2022)
- ESM (Rives et al. 2022; Lin et al. 2023)

...

Protein language modeling

Pretraining self-supervision on sequences only.

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

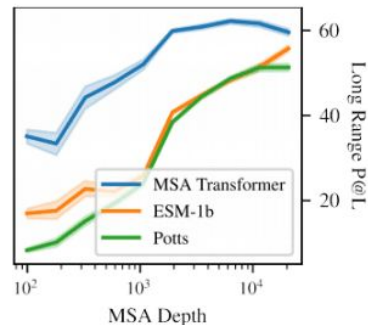
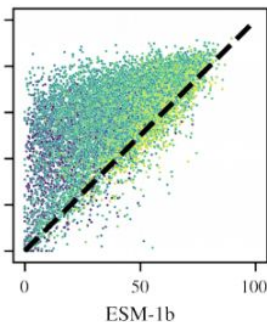
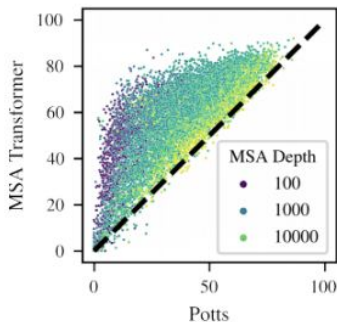
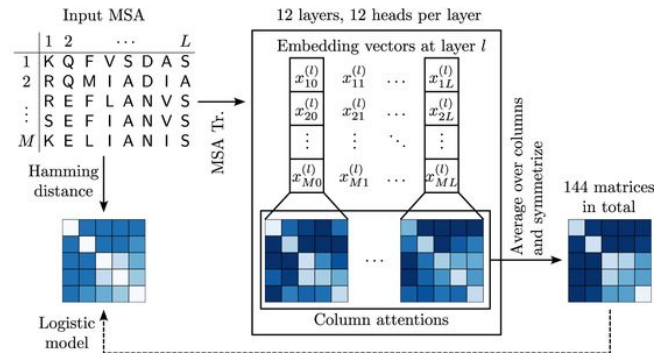


MSA can be embedded - MSA Transformer

(a)

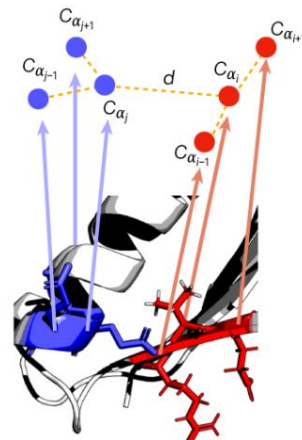


(b)



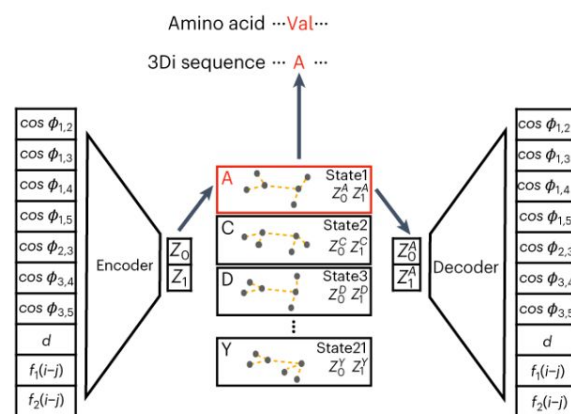
FoldSeek

- TM-align is basically Needleman-Wunsch algorithm ran over C-alpha atoms
- FoldSeek is an embedding of protein structures (C-alpha atoms + other features into 20 states = 3Di alphabet
- 3Di alphabet is aligned using MMSeqs2 internals analogously to raw protein sequences
- Structure = vector of 20 states



(2) Extract features

(4) (Discretization) conversion to 3Di sequence

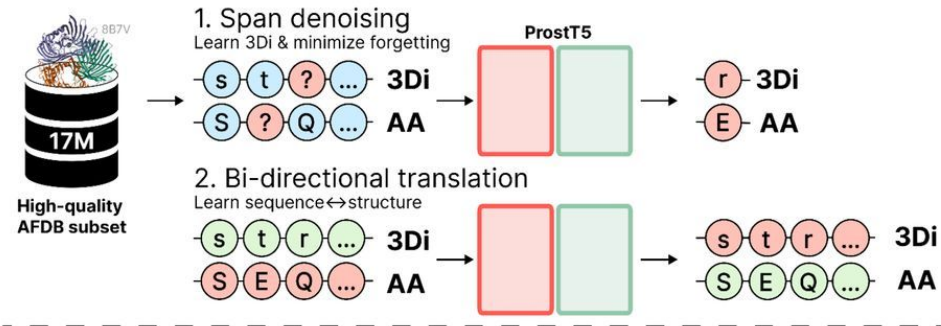


(3) Search 3Di state library

(4) (Training) predict features

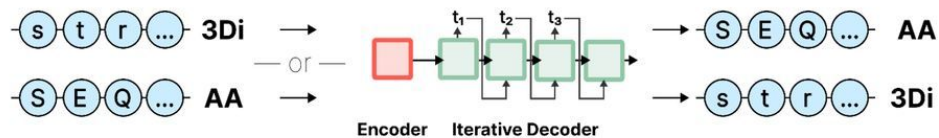
ProstT5

ProstT5 Pre-training

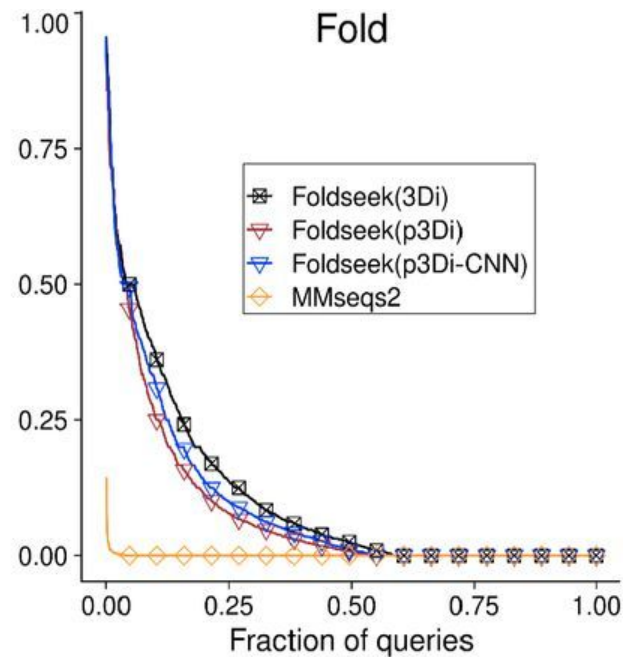
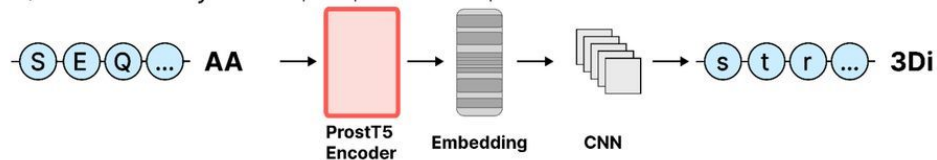


ProstT5 Inference

A) Encoder-Decoder Bi-directional translation

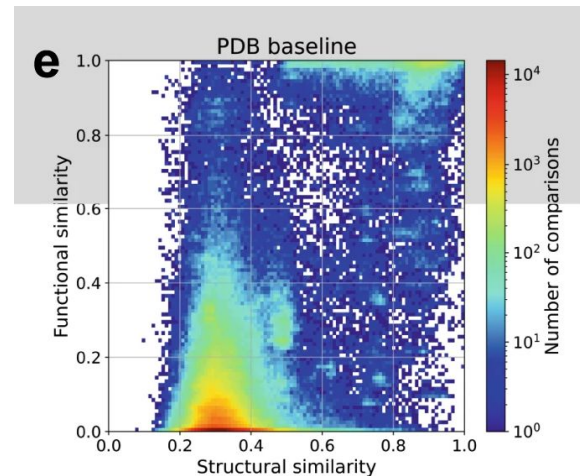


B) Encoder-only ~4000x speedup over structure prediction

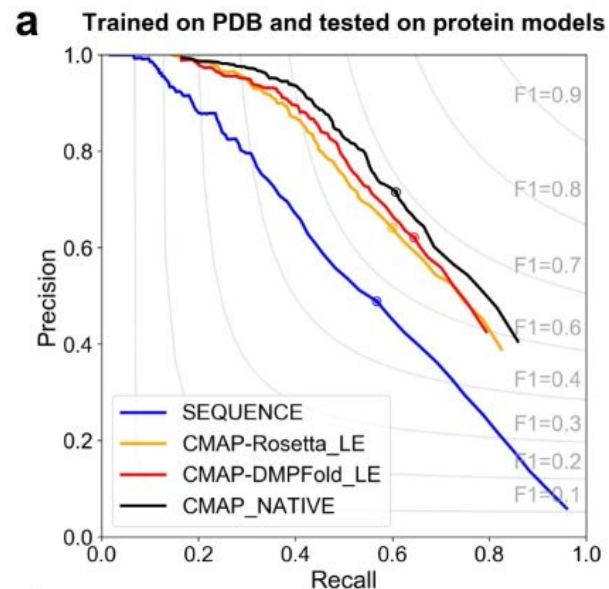
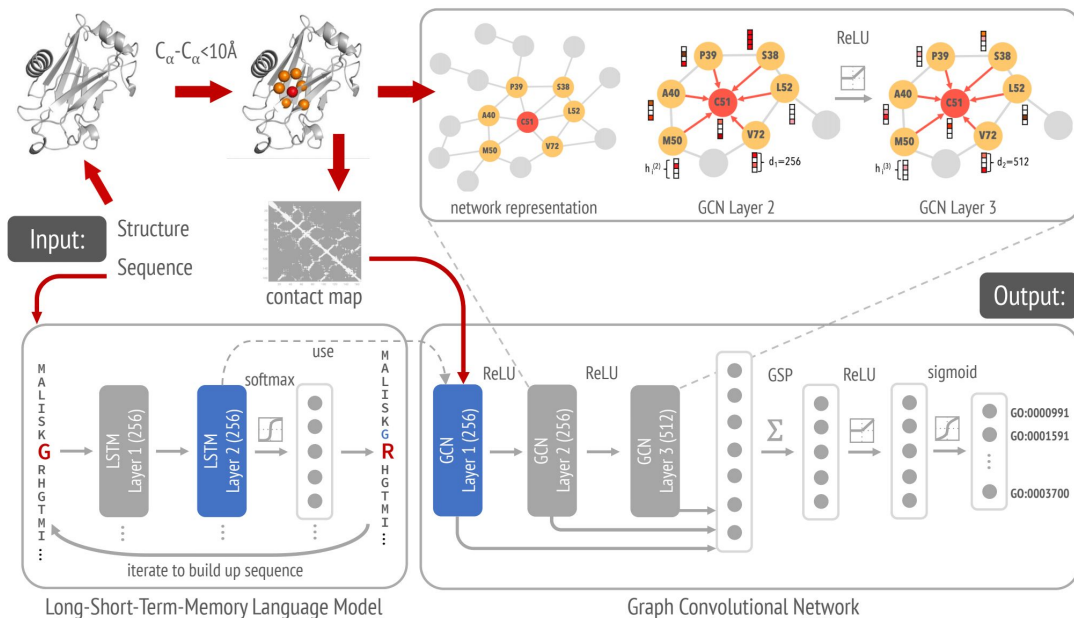


Task-specific vectors

- DeepFRI (Gligorijević et al., 2021; Koehler-Leman et al., 2023)
 - vector distance = functional similarity (GO ontology)
- TM-Vec (Hamamsy et al., 2023)
 - vector distance = structural similarity (TM-score)
- Protein-Vec (Hamamsy et al., 2023)
 - Aspect-Vec models for different ontologies (GO, EC, Pfam, TM-score...)
 - Protein-Vec derives unified protein vector from multiple Aspect-Vecs

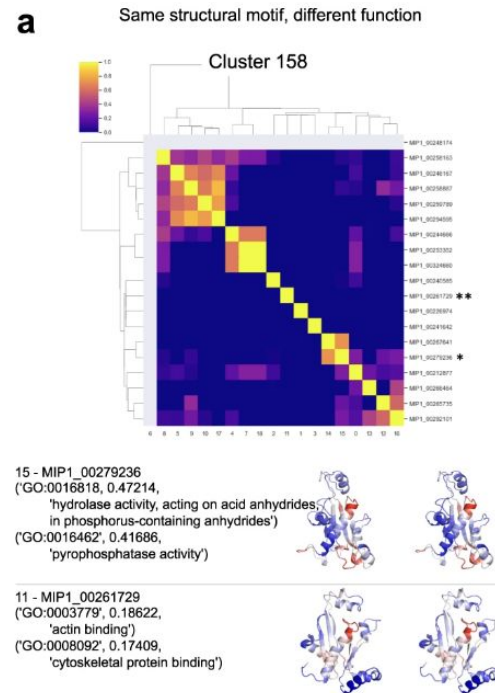
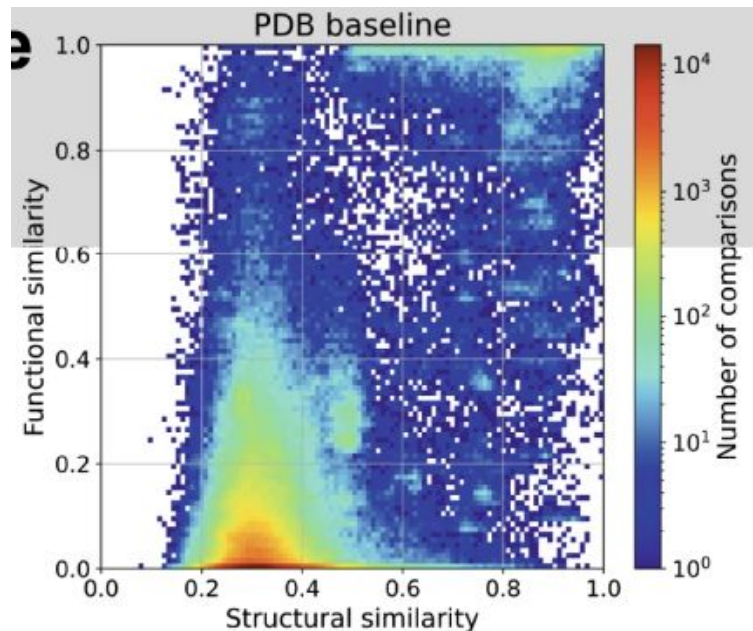


DeepFRI - a function embedding from sequence and structure



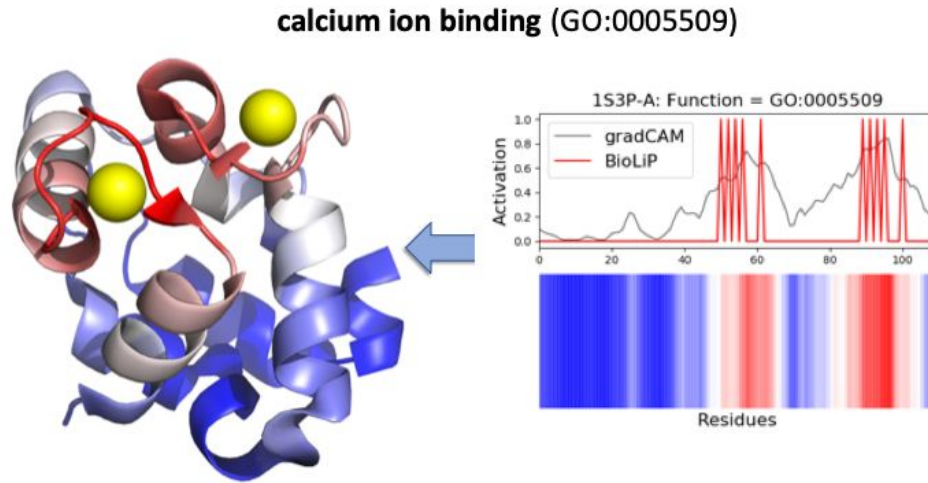
Gligorijevic et al., 2021

Sequence-structure-function relationship of proteins



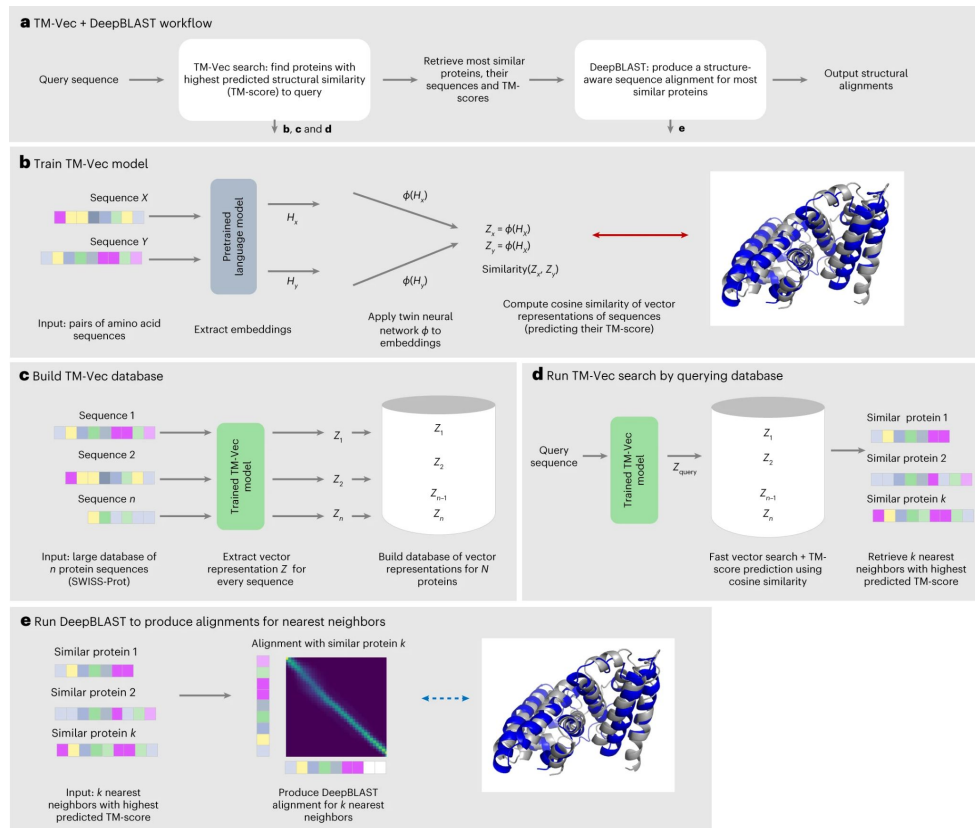
Koehler-Leman et al., 2023

Interpretability of DeepFRI

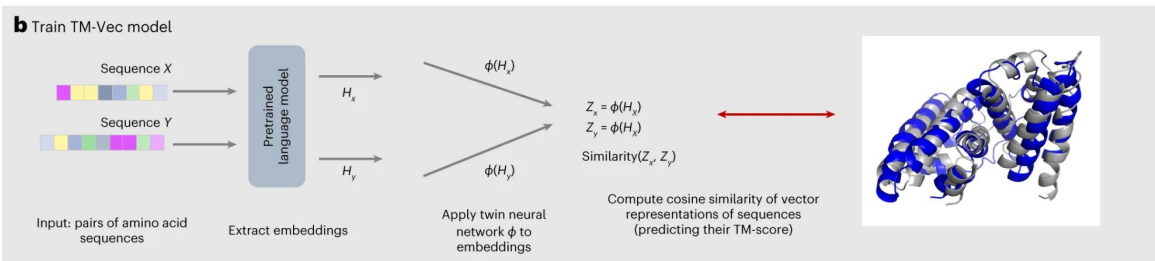


- GradCAM allows to visualize regions important for prediction
- GradCAM correlates with BioLiP activation function
- Precise mostly for binding functionality

TM-Vec & DeepBLAST



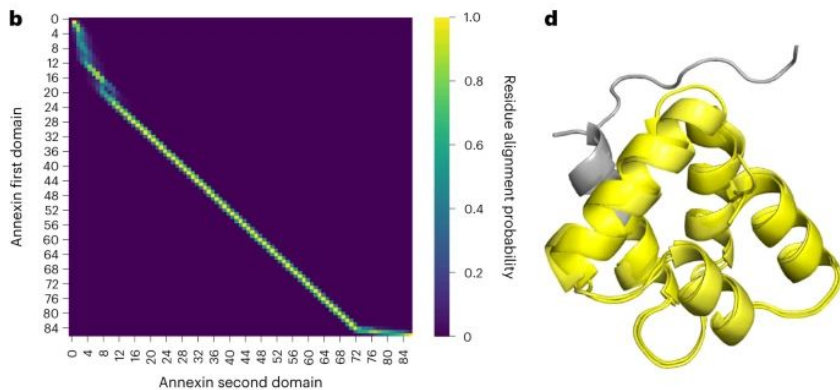
TM-Vec (Hammamsy et. al, 2023)



- Take embeddings from pLM
- Approximate structural similarity from embeddings without folding
- Outperforms FoldSeek on homology identification task

	CATH ID (Homology)
TM-Vec (CATH)	88%
FoldSeek	85%
TM-Vec (SwissProt)	71%
ProtTucker (EAT)	71%
HHBlits	49%
MMSeq2 (E value of 1000)	36%
MMSeq2 (E value < .01)	27%
Diamond	26%

DeepBLAST (Hammamsy et. al 2023)



c

KYTRGTVTAFSPFDARADAEALRKAMKMGMDDEETILKILTSRNNAQRQEIASAFKTLFGRDLVDDLKSELTGKFETLMVSLMRP
 SV----P-A-----YFA-E----TLYYSMKGAGTDDDTLIRVMVSRSEIDLLDIRHEFRKNFAKSLYQMIQKDTSGDYRKALLLCGG

	Malidup		Malisam	
Method	F1 score	Number detected	F1 score	Number detected
BLAST	0.019 ± 0.019	5	0.000 ± 0.000	2
HMMER	0.020 ± 0.02	8	0.020 ± 0.020	3
Needleman-Wunsch	0.098 ± 0.010	234	0.025 ± 0.003	129
Smith-Waterman	0.114 ± 0.010	234	0.031 ± 0.003	129
DeepBLAST	0.265 ± 0.020	234	0.066 ± 0.009	129
Mammoth-local	0.483 ± 0.020	234	0.187 ± 0.017	129
FAST	0.569 ± 0.026	234	0.300 ± 0.030	129
TM-align	0.576 ± 0.024	234	0.393 ± 0.031	129
Dali	0.791 ± 0.014	234	0.619 ± 0.029	129

- Complements TM-Vec
- Aligns pre-selected proteins
- Outperforms sequence aligners, but falls short from structure alignment
- Partially alleviates the need to compute structures
- Properly classifies bacteriocins

Bonus: LLMs for other sequence modalities

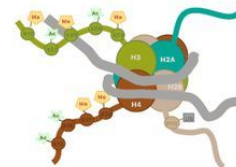
DNA-BERT

- Identification of promoter regions
- Identification of TF binding sites
- Prediction of functional genetic variants
-

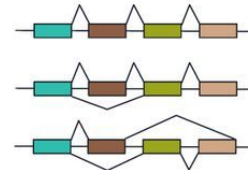
B Promoter activity prediction



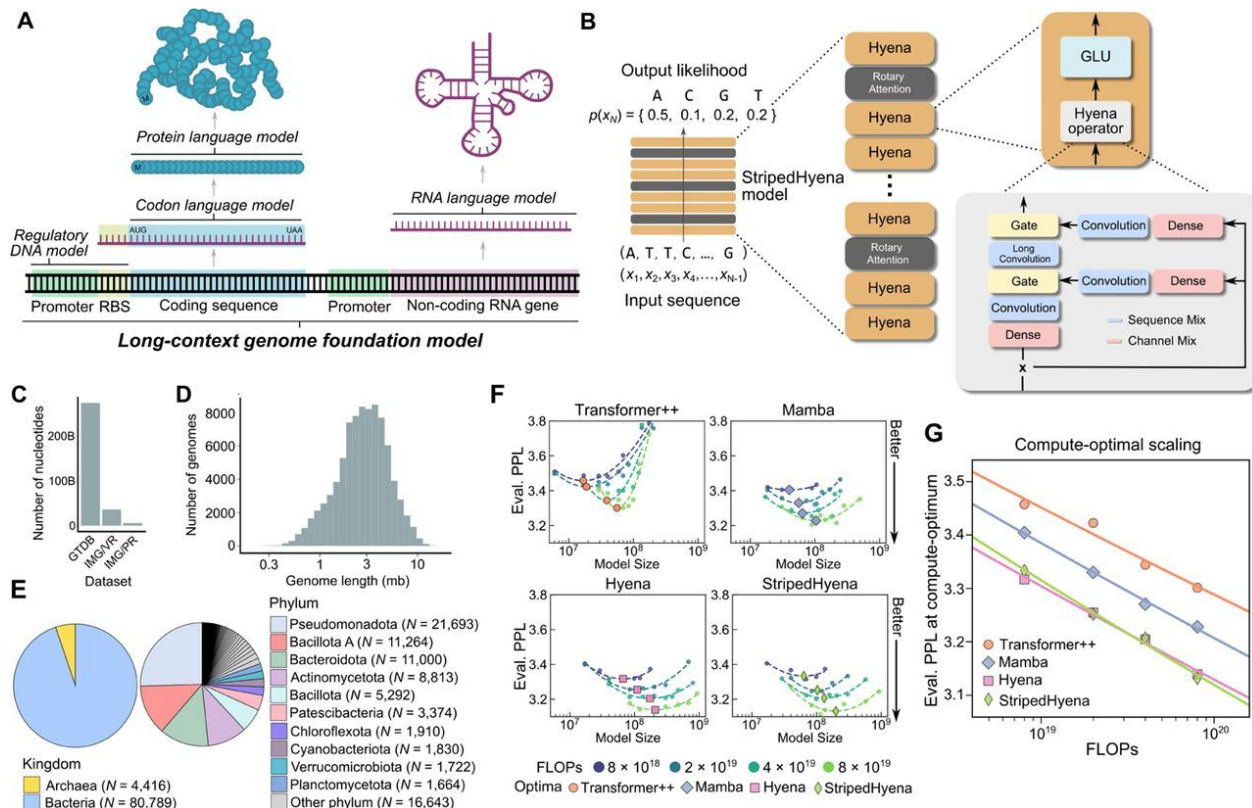
Prediction of chromatin profiles



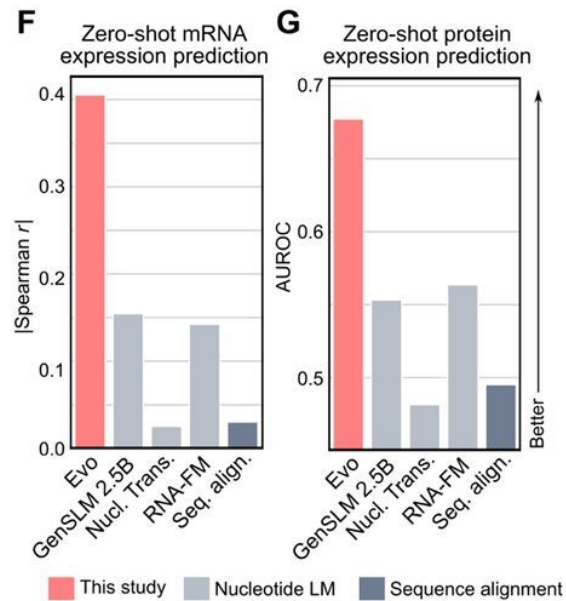
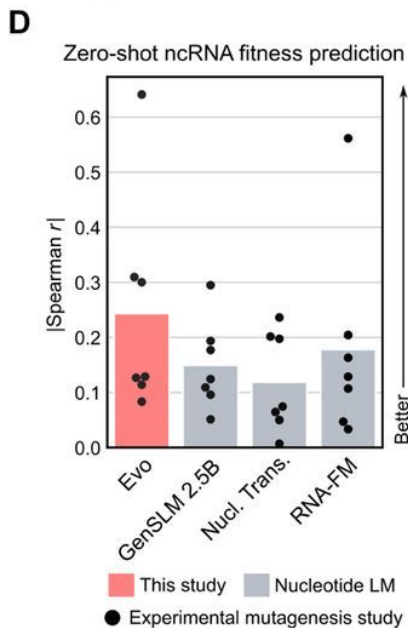
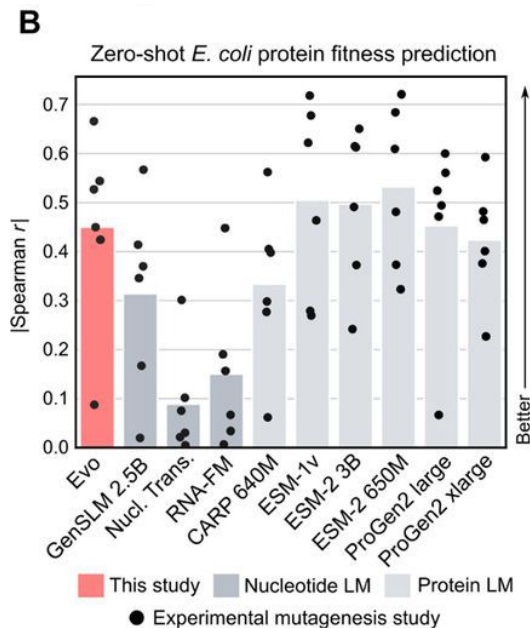
Splice sites annotation



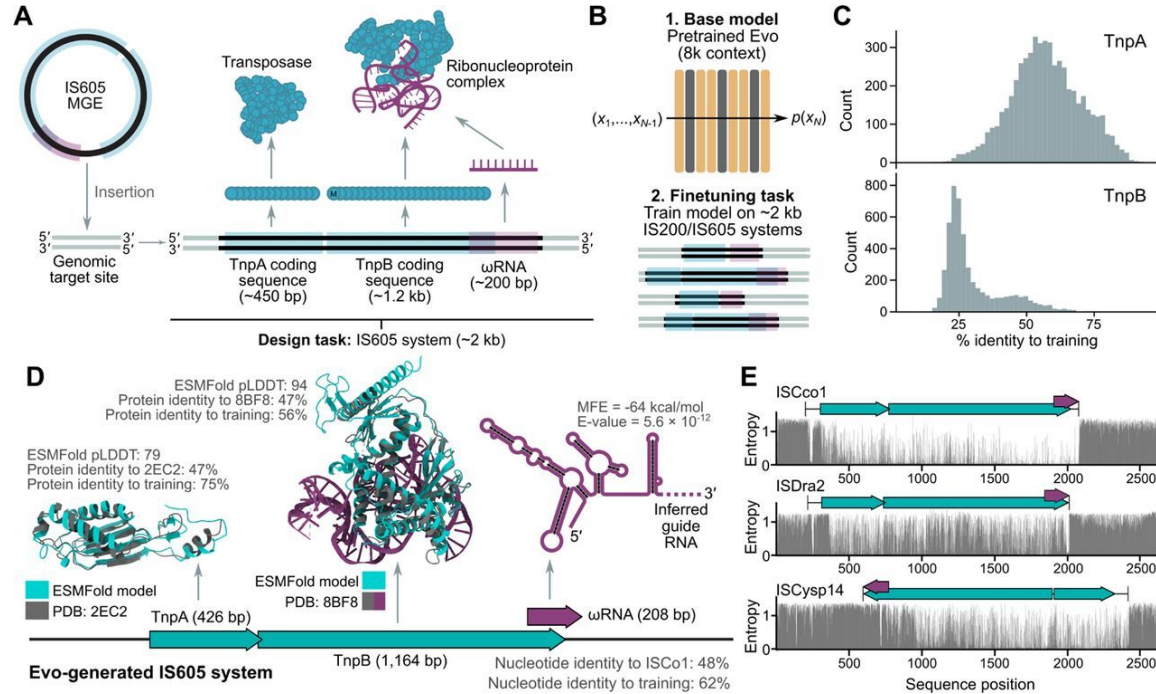
Evo: foundational model for prokaryotic life



Evo executes a lot of downstream tasks



Evo can act as a generative model



Disease and druggability prediction

Model	Accuracy	Precision	Recall	F1 Score	MCC	AUROC	AUPRC
OneHot-WT	0.657	0.653	0.990	0.786	0.135	0.722	0.832
OneHot-PTM	0.656	0.652	0.990	0.786	0.13	0.722	0.832
ESM-2-650M	0.662	0.667	0.936	0.78	0.178	0.737	0.836
ESM-2-3B	0.593	0.681	0.676	0.68	0.122	0.531	0.656
PTM-Mamba	0.692	0.696	0.914	0.792	0.276	0.753	0.847

Model	Accuracy	Precision	Recall	F1 Score	MCC	AUROC	AUPRC
OneHot-WT	0.767	0.777	0.978	0.866	0.067	0.686	0.84
OneHot-PTM	0.767	0.777	0.978	0.866	0.067	0.687	0.84
ESM-2-650M	0.855	0.858	0.975	0.912	0.55	0.921	0.976
ESM-2-3B	0.835	0.878	0.879	0.891	0.545	0.851	0.942
PTM-Mamba	0.861	0.886	0.947	0.913	0.581	0.905	0.967

pTM prediction

