Protein Language Models

Anne-Florence Bitbol

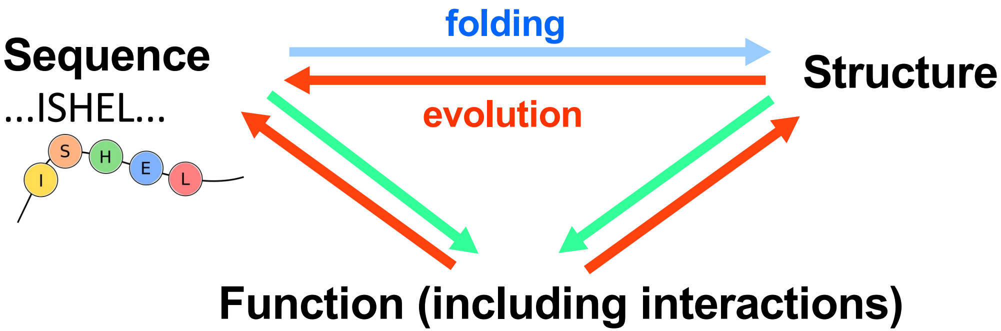
EPFL SIB

Laboratory of Computational Biology and Theoretical Biophysics
Institute of Bioengineering, School of Life Sciences

Reviews in Quantitative Biology, UNIL
November 22, 2024

Introduction: Proteins

Proteins



Sequence

...ISHEL...

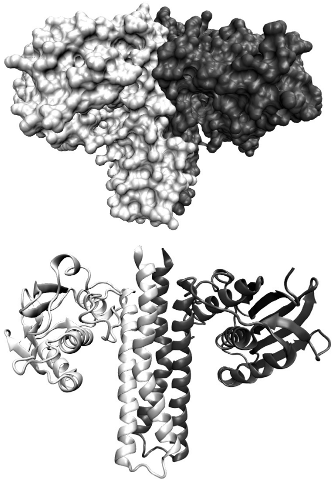
folding

evolution

Structure

Function (including interactions)

- Enzyme
- Molecular motor
- Receptor...



Mutations act on sequences BUT selection acts on function

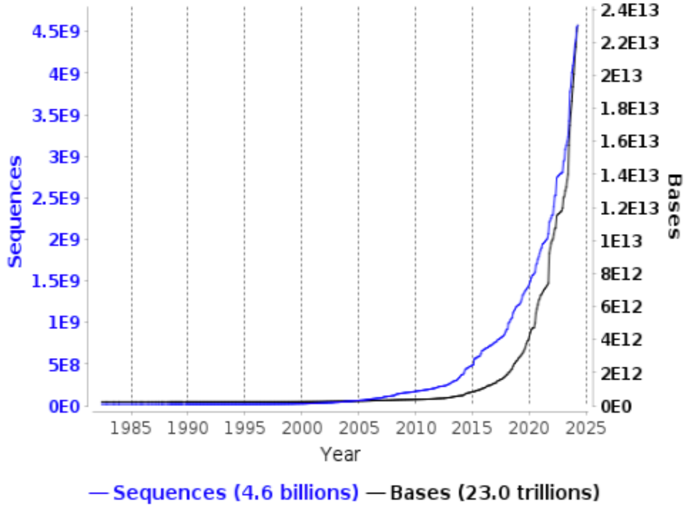
- Heteropolymers made of 20 types of amino-acids (monomers) → $\sim 20^{100}$ possible proteins
- A given natural protein folds into a compact and (almost) unique 3D **structure**
- It has specific **interactions** with other molecules → **function**

- Experiment: random proteins do not fold properly Socolich et al. (2005)

→ Natural proteins are special, due to natural selection for folding and function

Introduction: Protein sequence data

A growing amount of sequence data



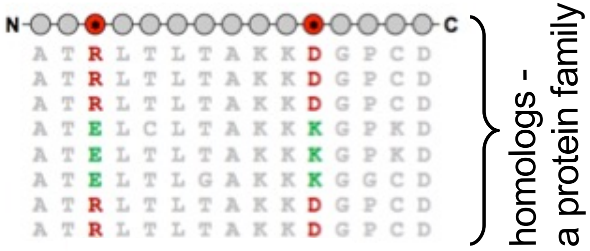
Accumulating sequence data
(currently > 10^9 sequences)
https://www.ebi.ac.uk/ena/browser/about/statistics

Proteins: UniProt (The UniProt Consortium 2021)

→ Great opportunity for machine learning methods to learn about proteins!

Goals: infer structure, function, interactions…

Protein families and multiple sequence alignments (MSAs)

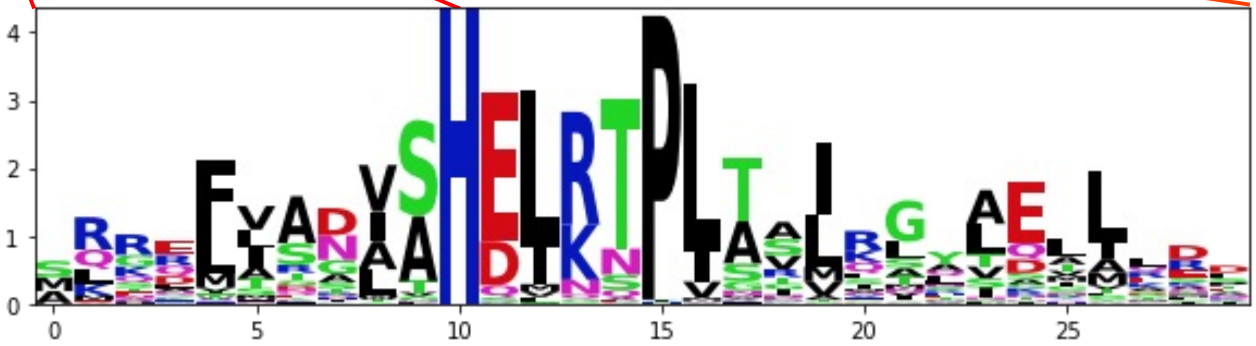


```
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T E L C L T A K K K G P K D
A T E L T L T A K K K G P K D
A T E L T L G A K K K G G C D
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
```

Introduction: Protein sequence data and inference

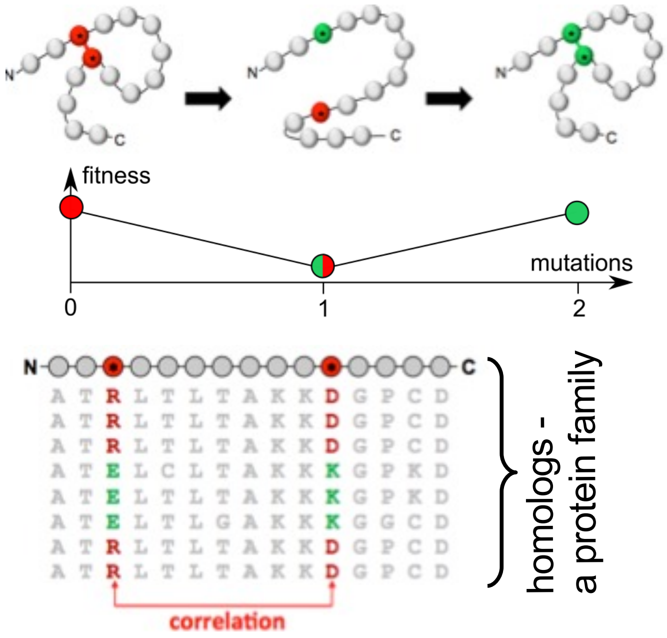
- **Inferring structure and function from sequences – conservation, correlations**

```
-RTEFVSNVSHELRTPLTSIKGYVETLLDEPGVRERFLQVIKDETDRLERLITDLLNLSQLES-
-RTEFVSNVSHELRTPLTSIKGYVETLLDEPGVRERFLQVIKDETDRLERLITDLLNLSQLES-
-QKQFVSDASHELRTPISVIQGYIDLLDRDKEVLEEAIEAIQAETTSMKKLLEQLLFLARSDKG
-RKELIANISHDLKTPITAIKGYVEGIRDSPEKLSRYVDTIYRKILEVDGLIDELFLFSKLD--
-KSEIIAMVSHELKTPLTSILAFGEILLALLPWQKEYLEDIMESGQELLKQIETLLTMAKIEAG
-----LHSLVHDLKTPLMTIQGLSSLIGLDSPKLQEYVQKIEQAVENVNKMISEIL--------
-RREFLANVSHELRTPLTIIQGYTEALLDTDEKIREHLKNILQEAERLKAMANELLDLASIEEG
-LGLLAAGVAHEINNPLATVSAYAEDLLERSGELARYLQVIGKQIERCKKITGSLLNFARQPA-
MRSEFIANVSHELRTPLTSIKGFLETLLDDKTIAKHFLQIMNSETERLTRLIDDLLSLSKIEA-
-RRQMIADIAHELRTPLSILQGNFELLLEVIEADEETLRSLAEEVKRLSRLVEELRELSLAEAG
-QKEFFANVSHELRSPATAILGEAQITLRSDDEYRQTLLRISESAEQLAFRIEDLLMLIRHDE-
```



Introduction: Protein sequence data and inference

- **Inferring structure and function from sequences – conservation, correlations**



```
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T E L C L T A K K K G P K D
A T E L T L T A K K K G P K D
A T E L T L G A K K K G G C D
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
```

Evolutionary coupling between interacting residues
→ correlations in MSAs inform us about
structure and function

Several approaches exploit these signatures to understand protein structure, interactions and function
de Juan et al, 2013

- **Simple data-driven approach: retain some statistics**

One- and two-body frequencies; (generalized) covariances

```
...ISHEL...
...VSHDI...
...VSHEL...
```

$$\rightarrow \begin{cases} f_i(\alpha) & i \in \{1, .., L\} \\ f_{ij}(\alpha, \beta) & \alpha \in \{A_1, .., A_{20}, A_{21} = -\} \end{cases}$$

$$C_{ij}(\alpha, \beta) = f_{ij}(\alpha, \beta) - f_i(\alpha) f_j(\beta)$$

1. A few words about language models

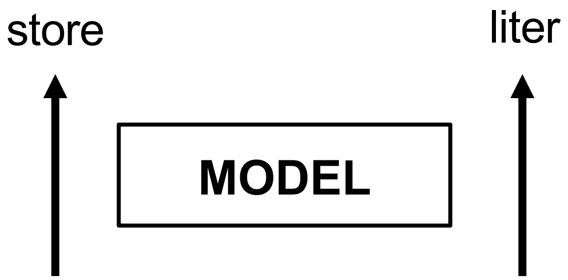
2. Protein language models based on single sequences

3. Protein language models based on multiple sequence alignments

Masked Language Modeling in NLP

- **Masked Language Modeling objective: self-supervised learning**

Randomly **mask** a fraction of the **words** and train the model to predict them using the surrounding **context**



The man went to the **[MASK]** and bought a **[MASK]** of milk.

The model is trained to minimize a pseudo-likelihood loss:

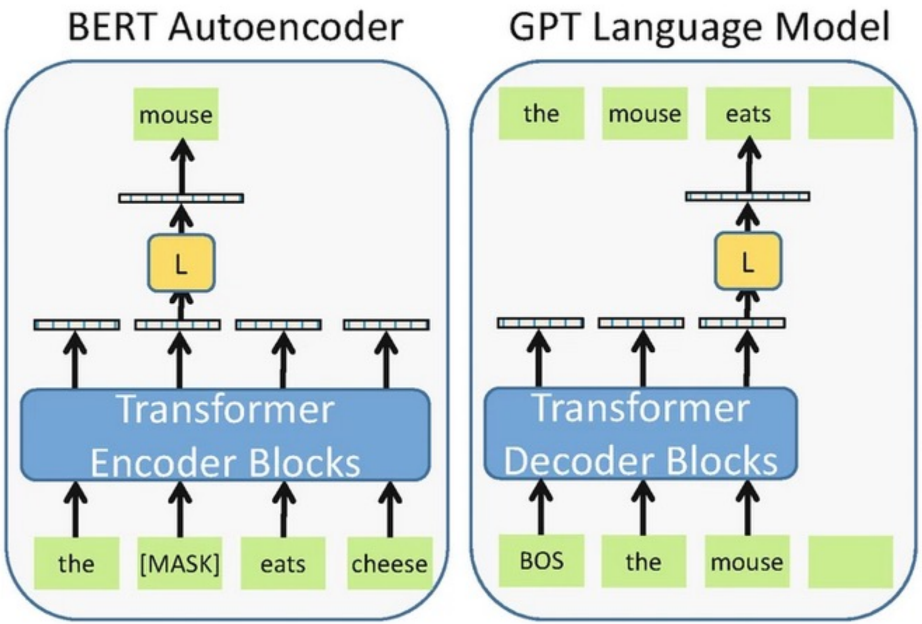
$$L_{MLM}(x,\theta) = -\sum_{m \in mask} \log p(x_m \mid \tilde{x};\theta) \qquad with \ \ \tilde{x} : masked\ sentence$$

Attention: Bahdanau et al 2014; transformer: Vaswani et al 2017

Masked Language Modeling in NLP

- **Two types of objectives in NLP**

MLM: predict masked words using the surrounding context (left and right) → BERT, Devlin et al 2018
Autoregressive: predict next word using previous ones (left only) → GPT, Radford et al 2018

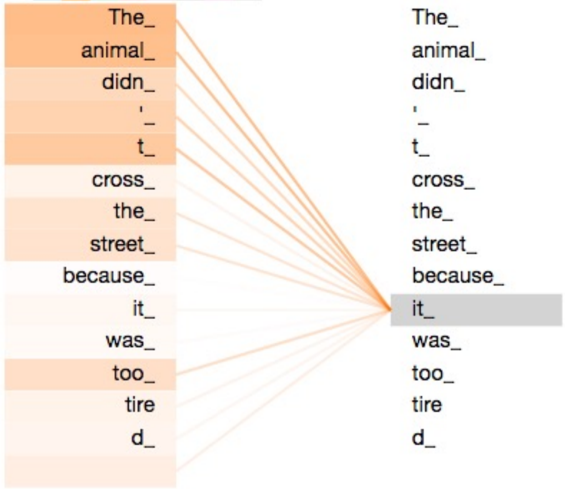


BERT: Bidirectional Encoder Representations from Transformers
GPT: Generative Pre-trained Transformer
Both are deep learning models relying on the transformer architecture (Vaswani et al 2017)

Transformers in NLP

- **Transformer architecture**

One attention head



M tokens → $M \times M$ softmax values

The Illustrated Transformer, Alammar

L layers
A heads per layer

Full architecture

M tokens → LA matrices, each of size $M \times M$

BERT$_{\text{BASE}}$: L = 12, A = 12
(Total parameters = 110M)

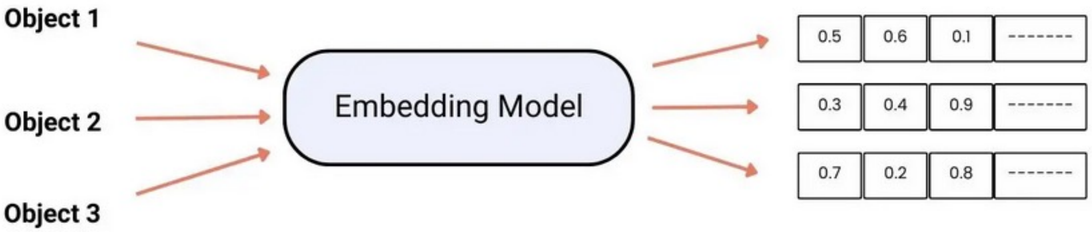
Embeddings in NLP

Representation of data

- Each word is represented by a real-valued vector: “embeddings”

- But words can have different meaning depending on context:
  I sent a letter to my friend. versus This is a list of four-letter words.

- Context-dependent embeddings: each occurrence of a word has its own embedding
- Such embeddings are learned



0.5	0.6	0.1	-------
0.3	0.4	0.9	-------
0.7	0.2	0.8	-------

1. A few words about language models

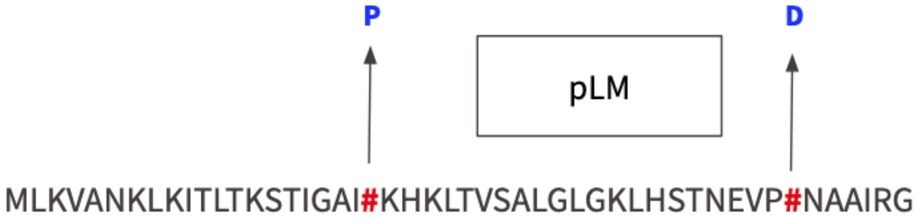
2. **Protein language models based on single sequences**

3. Protein language models based on multiple sequence alignments

Using NLP methods for protein sequences

- **Protein language models (pLMs): Direct use of NLP models, on a different kind of data**

- Sentence → protein sequence; word → amino acid (small vocabulary)



- Models trained using MLM, e.g. ESM2 (Lin et al 2022)
- Or autoregressive modeling, e.g. ProtGPT2 (Ferruz et al 2022)

Limitations of each approach:

- AR models only benefit from partial information about the sequence; no natural temporal order in protein sequences, vs. language
- MLM are not ideal for generation

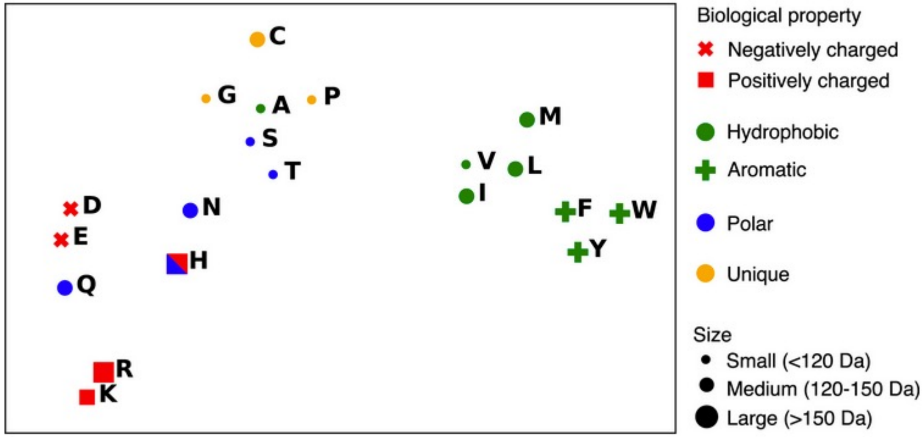
What do these models capture?

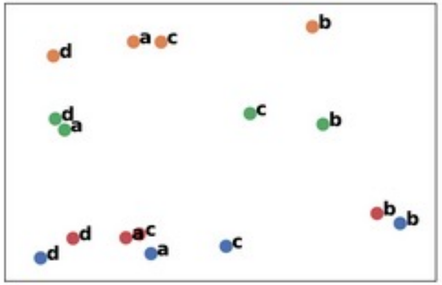
pLMs:
ProtVec (Asgari et al 2015)
SeqVec (Heinzinger et al 2019)
ESM1, 1b, 2, 3
MSA Transformer
ProtBERT
ProtT5
ProtGPT
ProGen
Ankh
Tranception
PoET
ProstT5
PST
…

Data representation in protein language models

- **Protein language models learn important features of protein sequence data**

Embeddings of ESM-1b – BERT model with 670M parameters (Rives et al 2021):
“Through unsupervised learning, residues are clustered into hydrophobic, polar, and aromatic groups and reflect overall organization by molecular weight and charge” (left)
“Protein sequence representations encode and organize biological variations” (right)





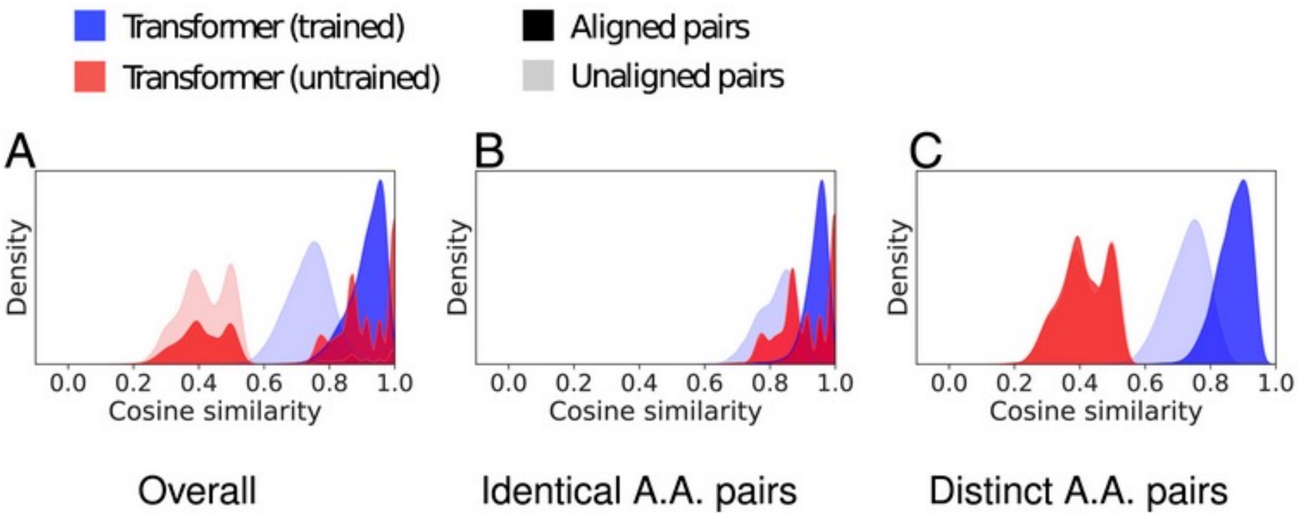
Transformer (trained)

Genes are colored by their orthologous group,
and their species are indicated by a character label

Data representation in protein language models

- **Protein language models learn important features of protein sequence data**

Embeddings of ESM-1b – BERT model with 670M parameters (Rives et al 2021):
“Final representations from trained models implicitly align sequences”



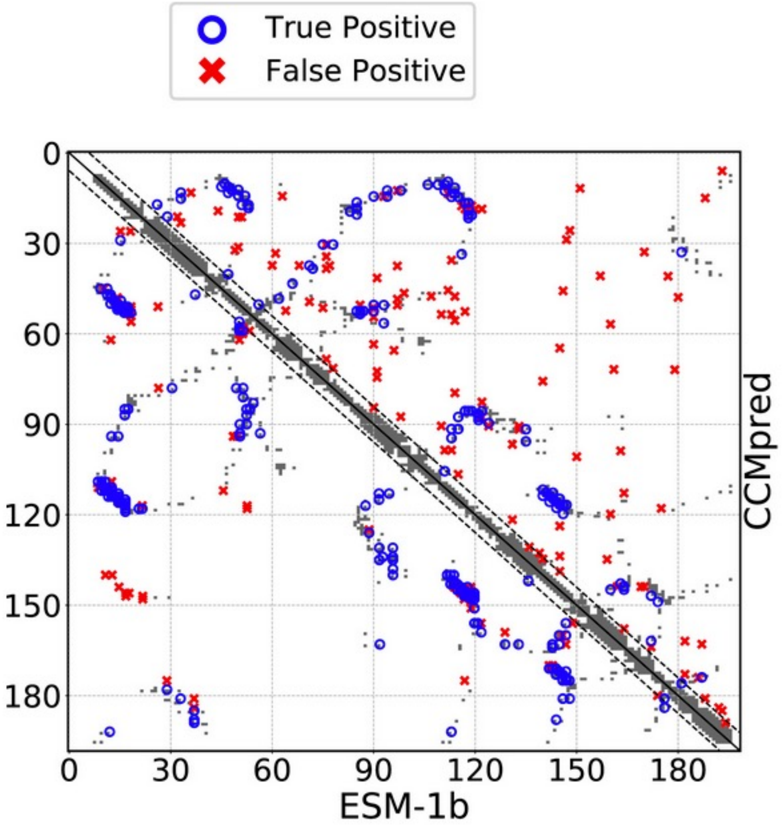
Overall

Identical A.A. pairs

Distinct A.A. pairs

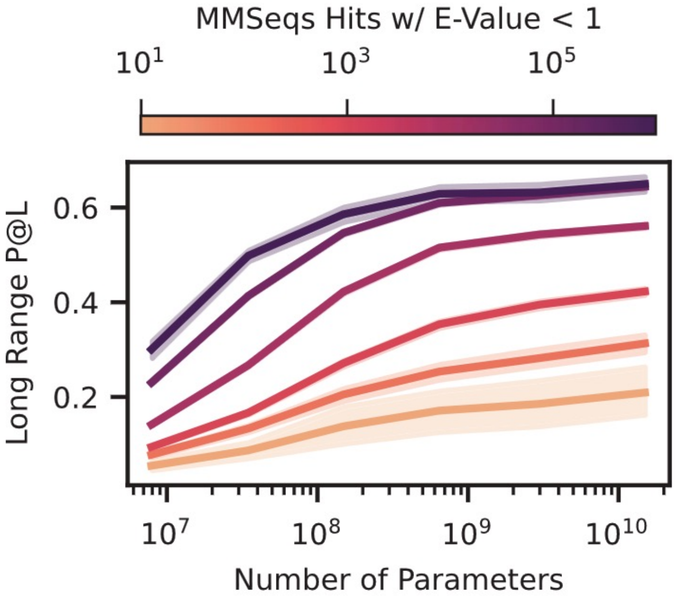
Some applications of protein language models

- **Structure prediction based on single-sequence language models**



Left - ESM-1b (Rives et al 2021):
Attention coefficients capture structural contacts

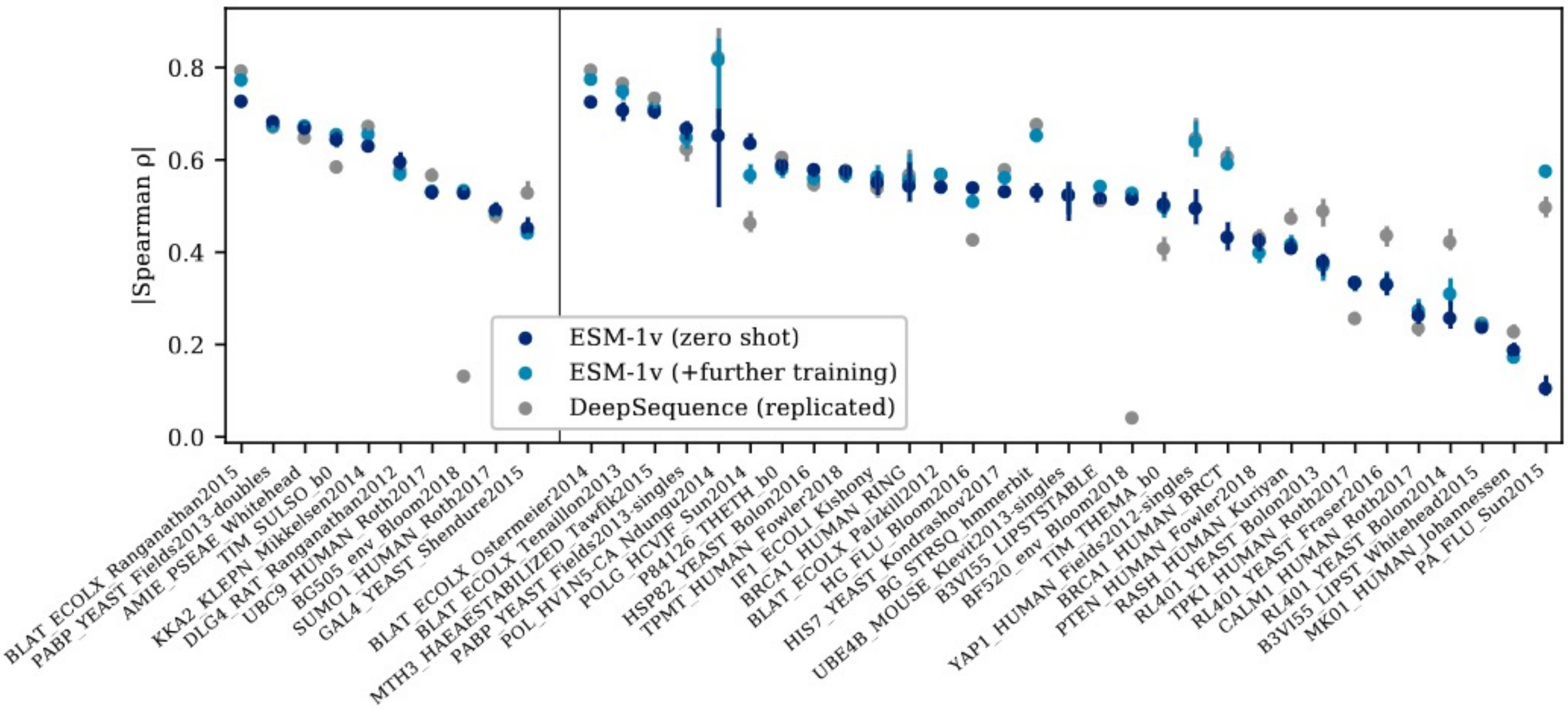
Below - ESM-2 (Lin et al 2023):
Larger model with better performance
(Unsupervised) contact prediction is
strongly affected by the number of existing homologs



Some applications of protein language models

- **Predicting the effect of mutations**

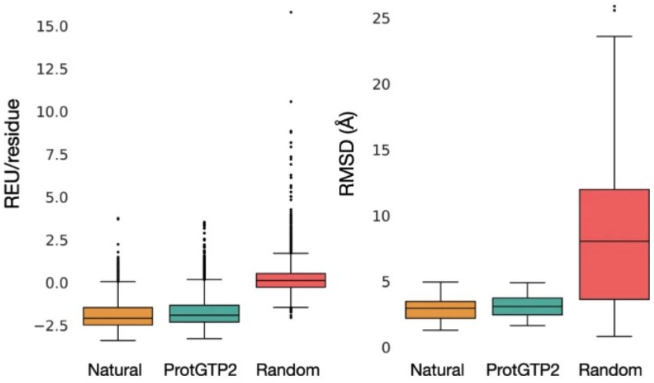
Ground truth: experimental deep mutational scans
Predictions: ESM-1v single-sequence protein language model (Meier et al 2021)



Some applications of protein language models

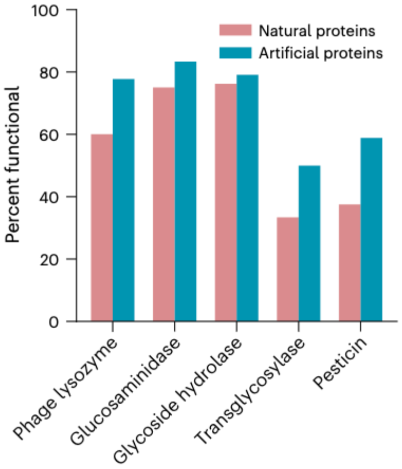
- **Designing new protein sequences**

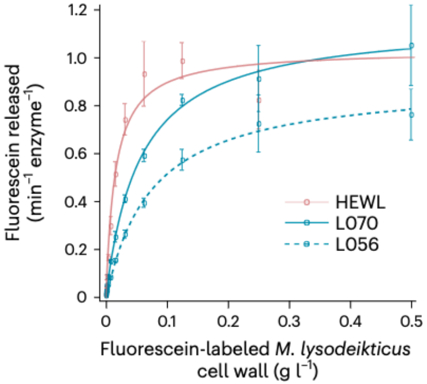
ProtGPT2 (Ferruz et al 2022): autoregressive transformer



Rosetta energy and flexibility patterns (from MD) similar to those of natural proteins

ProGen (Madani et al 2023): decoder transformer for *conditional* autoregressive generation





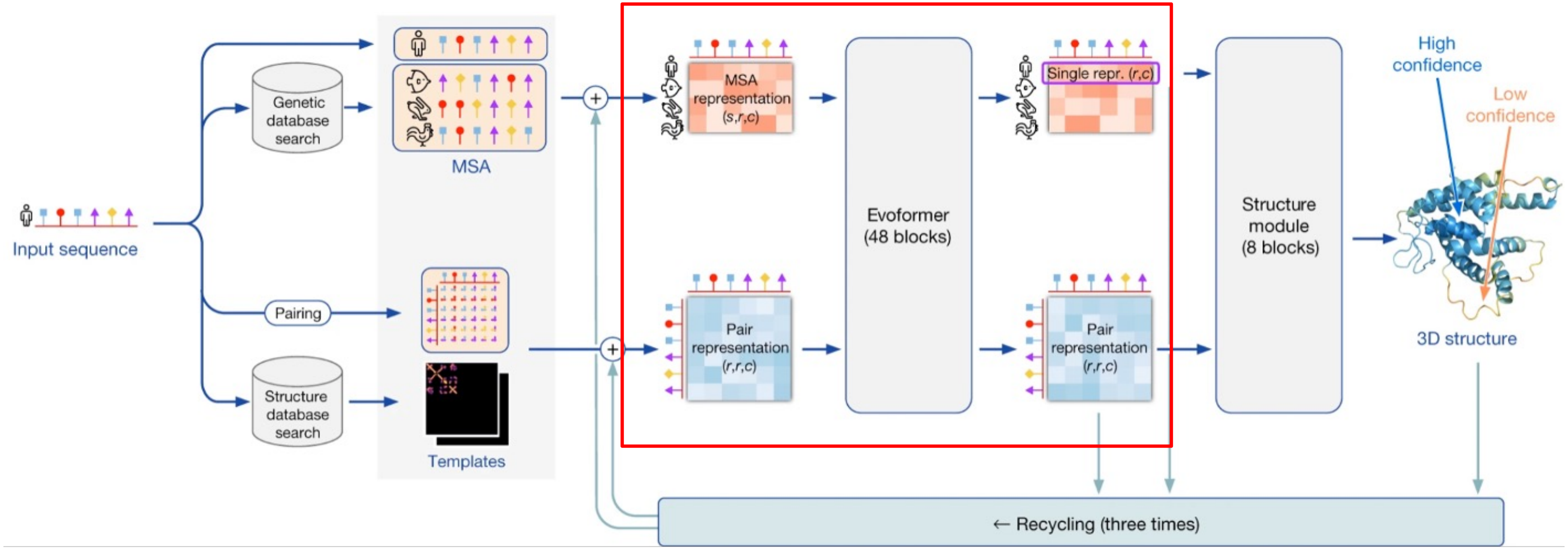
1. A few words about language models

2. Protein language models based on single sequences

3. **Protein language models based on multiple sequence alignments**

A few words about AlphaFold

- **Recent developments in protein structure prediction – Jumper et al 2021** (chemistry Nobel prize 2024)
  - **Supervised** deep learning approaches – AlphaFold, AlphaFold2 – won CASP13 and **CASP14**
    Other model: RoseTTAFold (Baek et al 2021); open retraining: OpenFold (Ahdritz et al 2024)
  - Part of AlphaFold is a **protein language model trained on MSAs**


Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
+
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

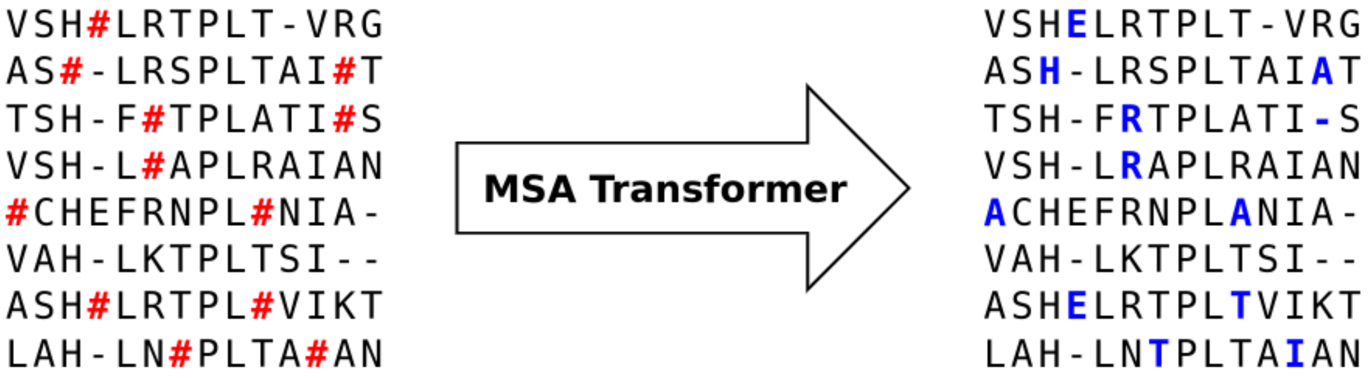
Jumper et al 2021

- AlphaFold3 (Abramson et al 2024): PairFormer module

MSA Transformer

- **Masked Language Modeling (MLM) objective on protein MSAs –** Rao et al 2021

Randomly mask (#) a fraction of the amino acids and train the model to predict them, using the surrounding context



```
VSH#LRTPLT-VRG        VSHELRTPLT-VRG
AS#-LRSPLTAI#T        ASH-LRSPLTAIAT
TSH-F#TPLATI#S        TSH-FRTPLATI-S
VSH-L#APLRAIAN        VSH-LRAPLRAIAN
#CHEFRNPL#NIA-        ACHEFRNPLANIA-
VAH-LKTPLTSI--        VAH-LKTPLTSI--
ASH#LRTPL#VIKT        ASHELRTPLTVIKT
LAH-LN#PLTA#AN        LAH-LNTPLTAIAN
```

The model is trained to minimize a pseudo-likelihood loss:

$$\mathcal{L}_{\text{MLM}}(\mathcal{M}, \widetilde{\mathcal{M}}; \theta) = -\sum_{(m,i)\in \text{mask}} \log p(x_{m,i} \mid \widetilde{\mathcal{M}}; \theta)$$

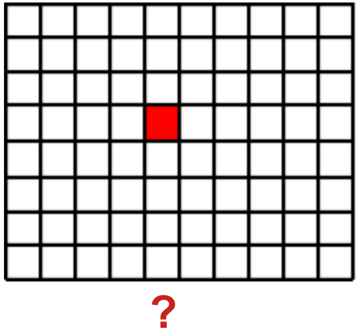
$\mathcal{M}$ MSA

$\widetilde{\mathcal{M}}$ masked MSA

MSA Transformer is similar to AlphaFold’s EvoFormer, but it is self-supervised
Here we focus on a model that works on MSAs – other ones work on single sequences

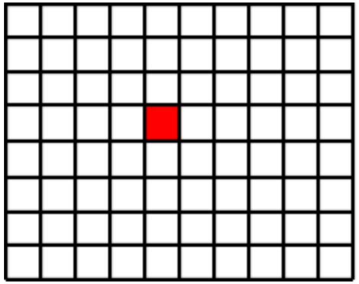
Architecture of MSA Transformer

- **Adapting the transformer architecture to protein MSAs –** Rao et al 2021

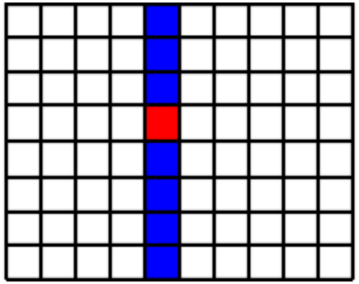


Architecture of MSA Transformer

- **Adapting the transformer architecture to protein MSAs –** Rao et al 2021



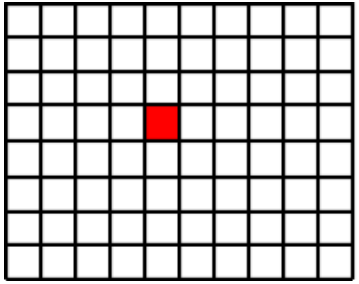
?



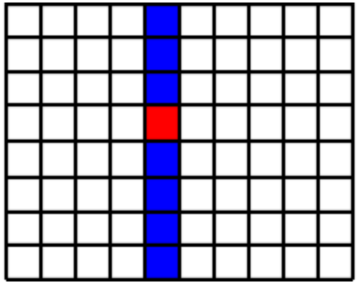
→ column attention

Architecture of MSA Transformer

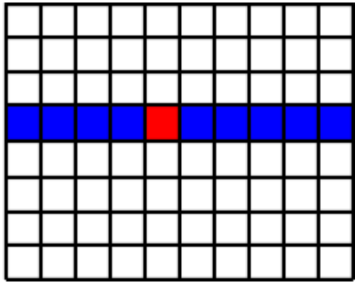
- **Adapting the transformer architecture to protein MSAs –** Rao et al 2021



?



→ column attention

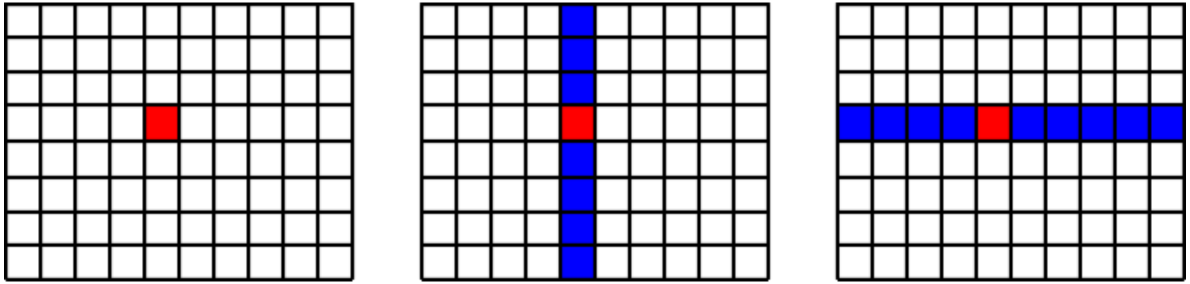


→ row attention

Context for an amino acid is both its column and its row (“axial attention” – Ho et al 2019)

Architecture of MSA Transformer

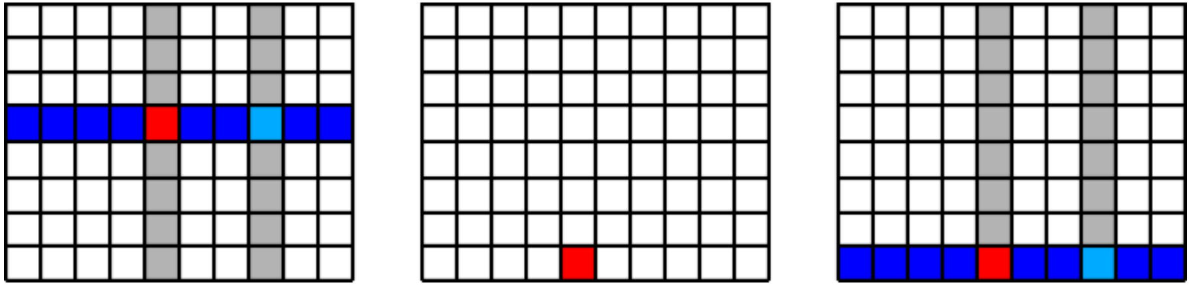
- **Adapting the transformer architecture to protein MSAs –** Rao et al 2021



? → column attention → row attention

Context for an amino acid is both its column and its row ("axial attention" – Ho et al 2019)

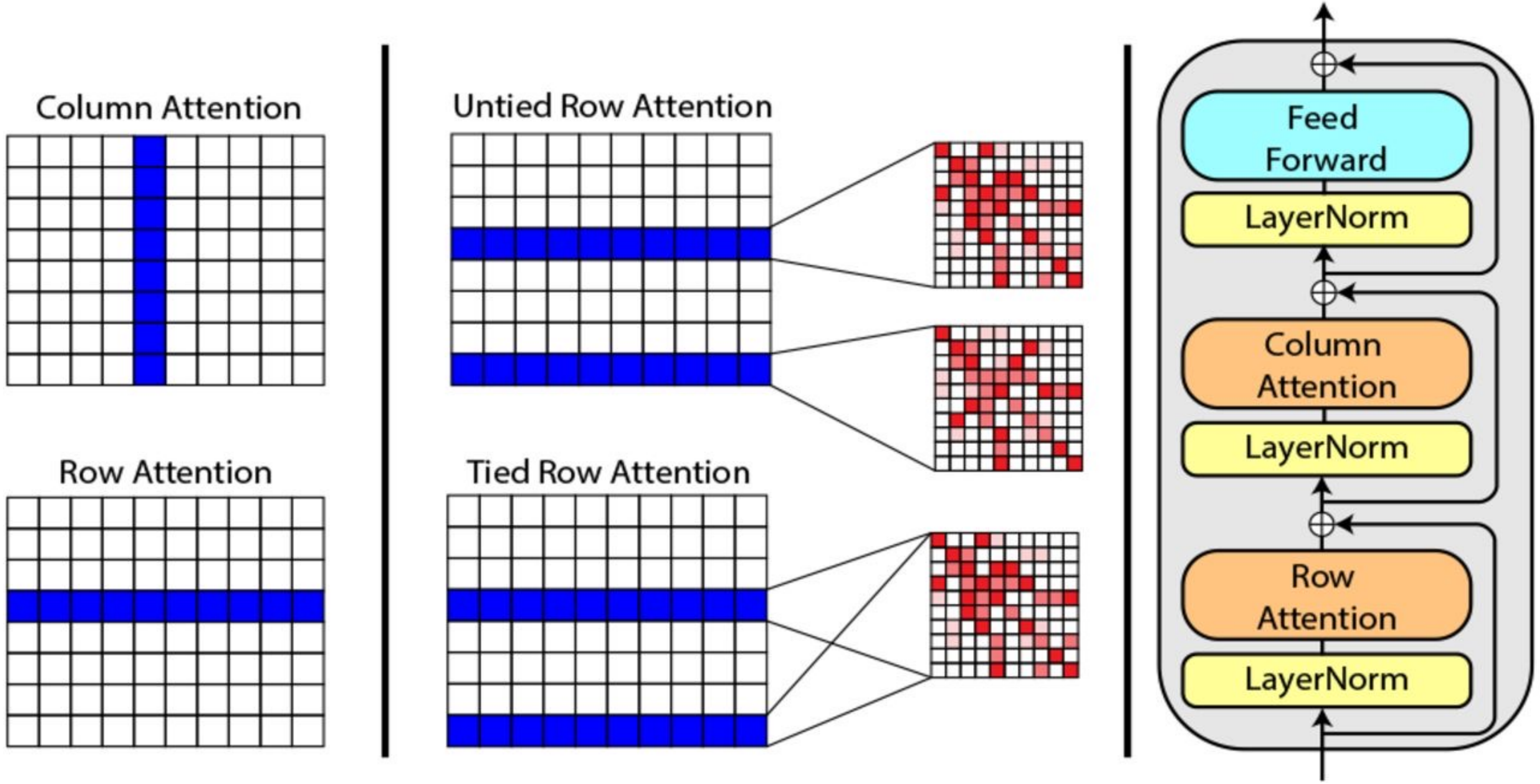
Coevolution → row attention should be the same for all rows



12 (layers) × 12 (heads) tied row attention units
12 × 12 independent column attention units
100M total parameters

Architecture of MSA Transformer

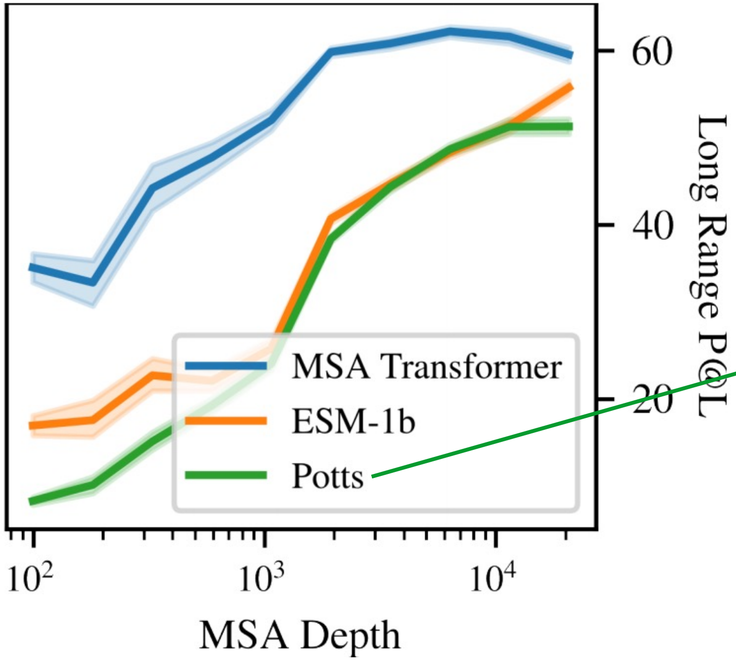
- **Adapting the transformer architecture to protein MSAs –** Rao et al 2021



Training set:
- 26M MSAs corresponding to UniRef50 clusters
- average depth of MSAs: 1192

MSA Transformer's data representation

- **(Tied) row attentions capture structural contacts –** Rao et al 2021
  - Simple combinations of the row attention softmax matrices allow contact prediction
  - State-of-the-art unsupervised contact prediction



Contact prediction performance

Potts model: pairwise maximum entropy model / DCA [Weigt, White et al 2009]

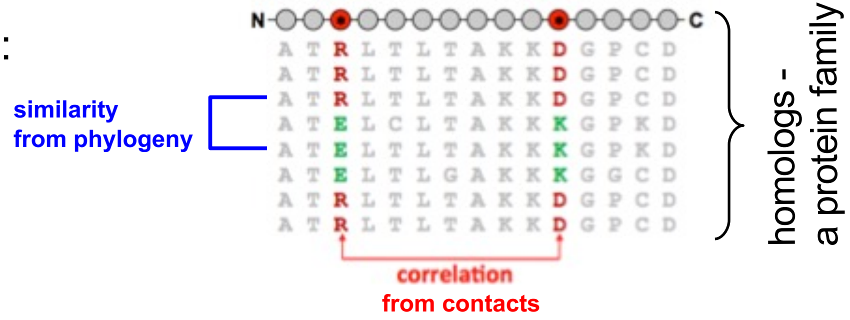
$$P(\alpha_1, ..., \alpha_L) = \frac{1}{Z} \exp\left\{ - \left[\sum_{i=1}^{L} h_i(\alpha_i) + \sum_{i<j} e_{ij}(\alpha_i, \alpha_j) \right] \right\}$$

One model per family
(vs. language models trained on many families)

What kind of information is encoded in column attentions?

MSA Transformer's data representation

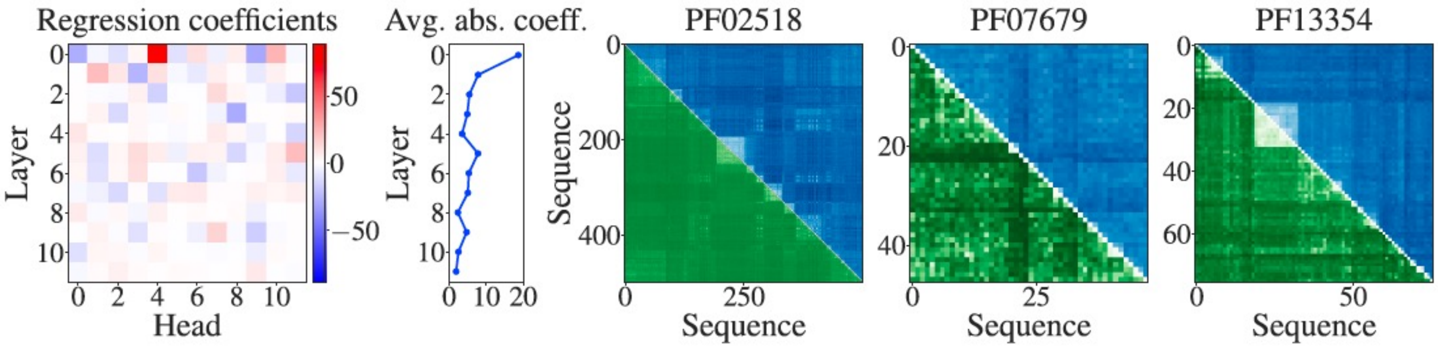
- **Column attentions encode phylogenetic relationships –** Lupo et al 2022
  - Motivation:



```
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T E L C L T A K K K G P K D
A T E L T L T A K K K G P K D
A T E L T L G A K K K G G C D
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
```

Casari et al 1995

- We fit a logistic model of the column attention matrices (averaged over columns) to predict the matrix of pairwise Hamming distances between sequences in MSAs
- Training: seed MSAs of 12 Pfam protein families; test: seed MSAs of 3 other Pfam families

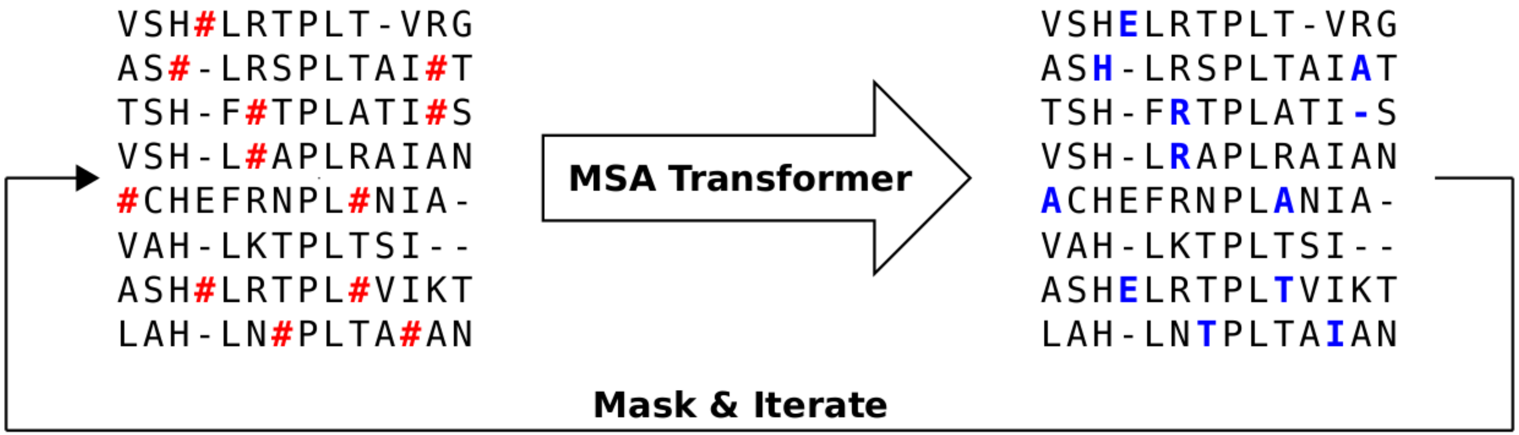


R^2=0.60	R^2=0.28	R^2=0.67
ρ=0.90	ρ=0.85	ρ=0.92

→ A simple combination of column attention heads “implements” Hamming distance

Generating sequences with MSA Transformer

- **Iterative masking algorithm based on MLM –** Sgarbossa et al 2023



```
VSH#LRTPLT-VRG
AS#-LRSPLTAI#T
TSH-F#TPLATI#S
VSH-L#APLRAIAN
#CHEFRNPL#NIA-
VAH-LKTPLTSI--
ASH#LRTPL#VIKT
LAH-LN#PLTA#AN
```

MSA Transformer

```
VSHELRTPLT-VRG
ASH-LRSPLTAIAT
TSH-FRTPLATI-S
VSH-LRAPLRAIAN
ACHEFRNPLANIA-
VAH-LKTPLTSI--
ASHELRTPLTVIKT
LAH-LNTPLTAIAN
```

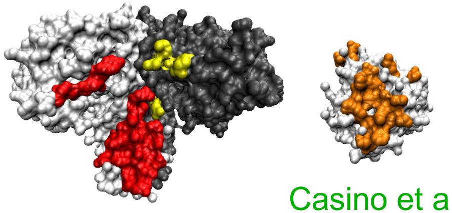
Mask & Iterate

Run iteratively this masking process on the same MSA → generate sequences

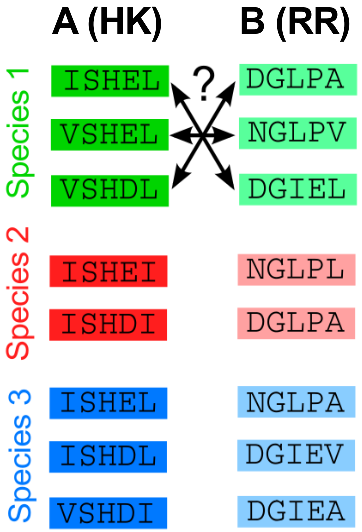
- Characterization of these sequences
- Comparison to sequences generated by a Potts model, using Metropolis-Hastings MCMC sampling (bmDCA Potts models are good generative models – Figliuzzi et al 2018, experimental validation Russ et al 2020)
- **Results:** Generated sequences are different from natural ones and score well for homology, coevolution and structural scores. Particularly promising for small protein families where Potts models overfit.

Predicting interaction partners with MSA Transformer

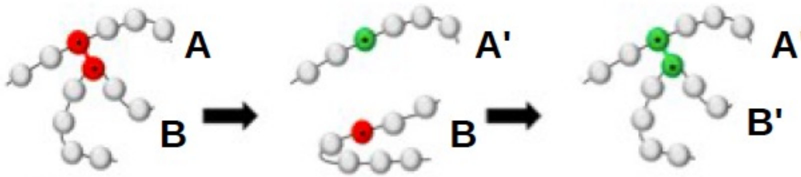
- **Coevolution can be used to infer interaction partners from sequences**

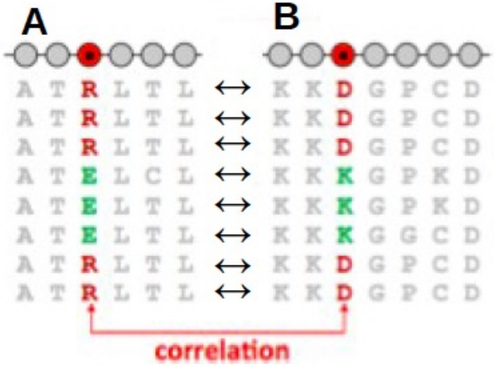


Casino et al. (2009)



	A (HK)	B (RR)
Species 1	ISHEL	DGLPA
	VSHEL	NGLPV
	VSHDL	DGIEL
Species 2	ISHEI	NGLPL
	ISHDI	DGLPA
Species 3	ISHEL	NGLPA
	ISHDL	DGIEV
	VSHDI	DGIEA





```
A T R L T L ↔ K K D G P C D
A T R L T L ↔ K K D G P C D
A T R L T L ↔ K K D G P C D
A T E L C L ↔ K K K G P K D
A T E L T L ↔ K K K G P K D
A T E L T L ↔ K K K G G C D
A T R L T L ↔ K K D G P C D
A T R L T L ↔ K K D G P C D
```

Within a species, which A interacts with which B?

→ Use correlations from coevolution to infer interaction partners (i.e. match paralogs):
Bayesian tree (Burger & van Nimwegen 2009),
Potts models (Bitbol et al 2016; Gueudre et al 2016)
Mutual Information (Bitbol 2018)
Potts or MI + phylogeny (Gandarilla-Pérez et al 2023)
MLM loss from MSA Transformer (Lupo, Sgarbossa et al 2024)

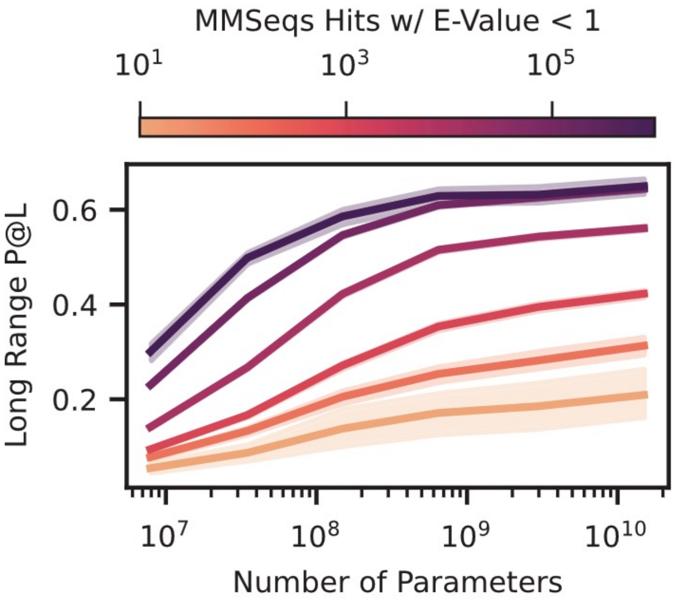
Are MSAs really necessary?

- **Structure prediction based on single-sequence language models**

Motivations: - Some proteins have few homologs
- MSA construction is imperfect and slow
- Predicting structure from a single sequence = closer to “understanding protein folding”

Strategy: - Train language models on large ensembles of non-aligned single sequences
- Add a structure module inspired by the one of AlphaFold2

AminoBERT → RGN2 (Chowdhury et al 2021); OmegaPLM → OmegaFold (Wua et al 2022); ESM-2 → ESMFold (Lin et al 2023)



ESM-2 & ESMFold (Lin et al 2023):

(Unsupervised) contact prediction:

- slightly less good than with MSA Transformer, even with many more parameters (15B vs. 100M)
- strongly affected by the number ofexisting homologs!

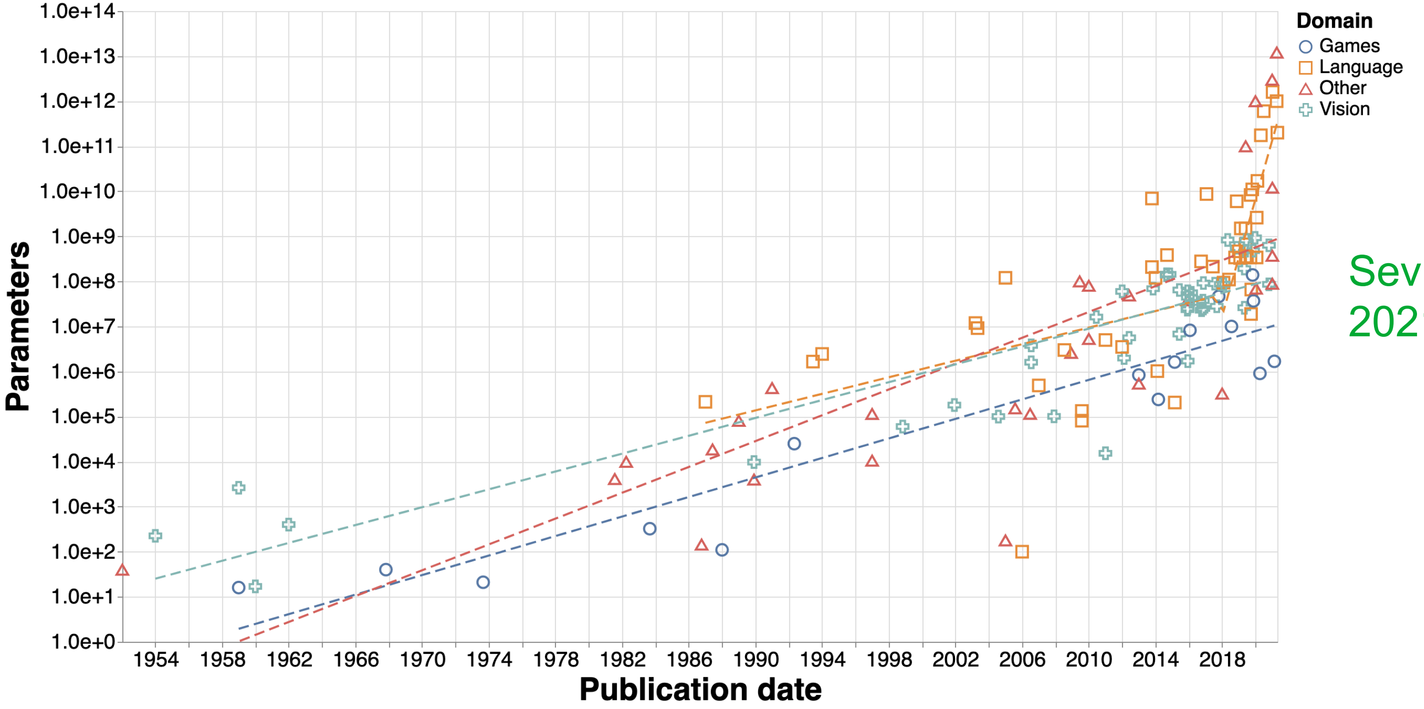
(Supervised) structure prediction:

- less good than AlphaFold2
- much faster → structure prediction at metagenomic scale

Are MSAs really necessary?

- **As of now, best performance for structure prediction requires MSAs**

Optimistic take for single-sequence LMs: we just need more parameters (Lin et al 2023)
“Our current models are very far from the limit of scale in parameters, sequence data, and computing power that can in principle be applied. We are optimistic that as we continue to scale, there will be further emergence. Our results showing the improvement in the modeling of low depth proteins point in this direction.”

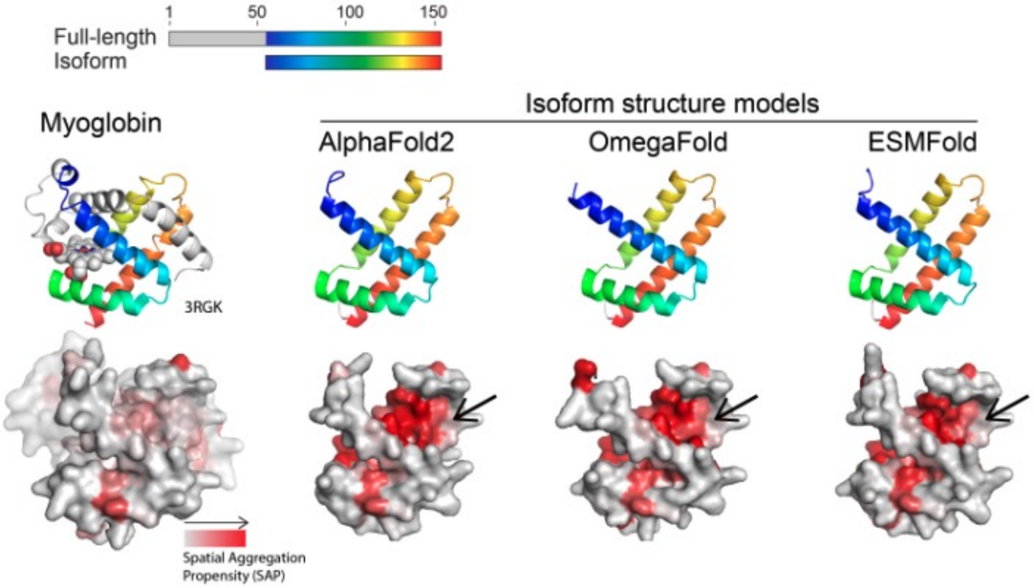


Sevilla et al, 2021

Are MSAs really necessary?

- **As of now, best performance for structure prediction requires MSAs**

Pessimistic take for single-sequence LMs: evolutionary information is crucial (Zhang et al 2024)
“Some have wondered if pLMs have finally solved the “protein folding problem”, given their accurate structure prediction from single sequences and no supplied co-evolutionary signal in an input multiple sequence alignment. This should have been quickly debunked, as the accuracy of models was found to be highly correlated to the number of related proteins in the training set, indicating that the models store evolutionary information in their parameters”



Isoform structure prediction is a challenge

Providing local windows of sequence information allows ESM-2 to best recover predicted contacts → pLMs may predict contacts by storing motifs of pairwise contacts (Zhang et al 2024)

Some recent developments

- **An alternative to single sequences / MSAs: use homology but not MSAs**

PoET (Truong et al 2023)
Transformer model trained on non-aligned homologs – uses both per-sequence attention and attention across sequences
However: limitations in context length & expensive to train

ProtMamba (Sgarbossa, Malbranke et al 2024)
Uses state-space model (Mamba) architecture, which can handle very long contexts
Starts from concatenated homologous sequences
Combines autoregressive modeling and fill-in-the-middle objective (~MLM)

- **Structure-aware models; multi-modal models**

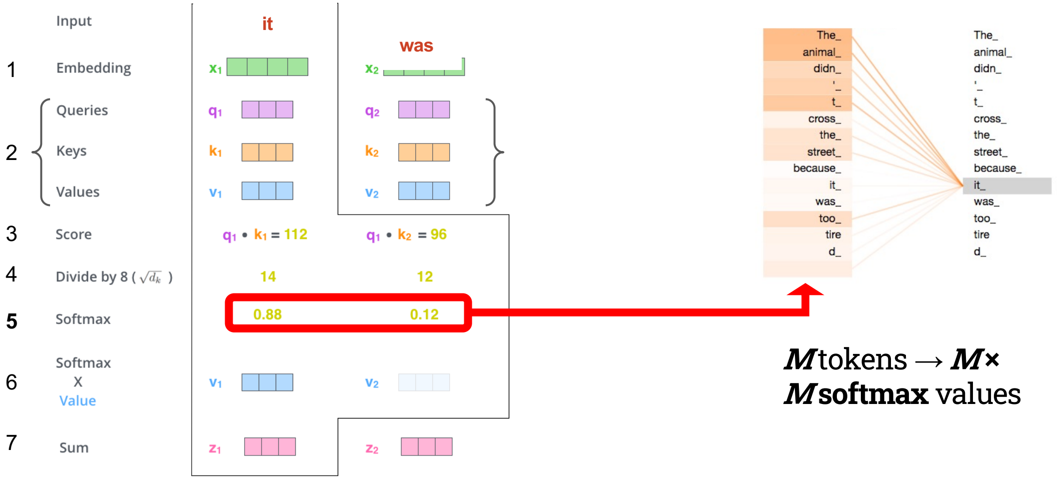
Structure-aware models relying on 3Di alphabet of FoldSeek (van Kempen et al 2023)
ProstT5 (Heinzinger et al 2023), SaProt (Su et al 2023), ProSST (Li et al 2024)

Multi-modal models: ESM3 (Hayes et al 2024)

Thanks!

Self-attention and the Transformer

A computational unit that models “focussing on what’s most relevant”

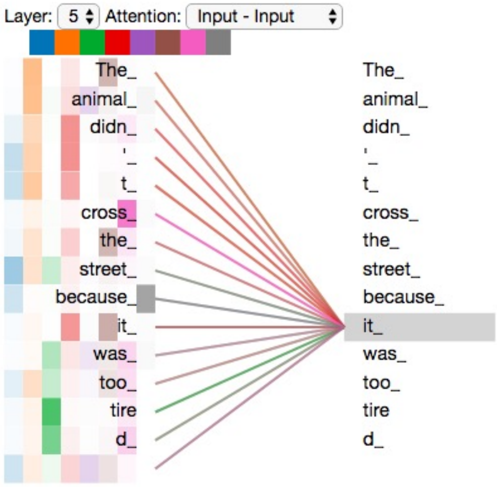


$\boldsymbol{M}$ tokens → $\boldsymbol{M}\times$ $\boldsymbol{M}$ **softmax** values

(Adapted from The Illustrated Transformer by Jay Alammar)

Multi-headed attention

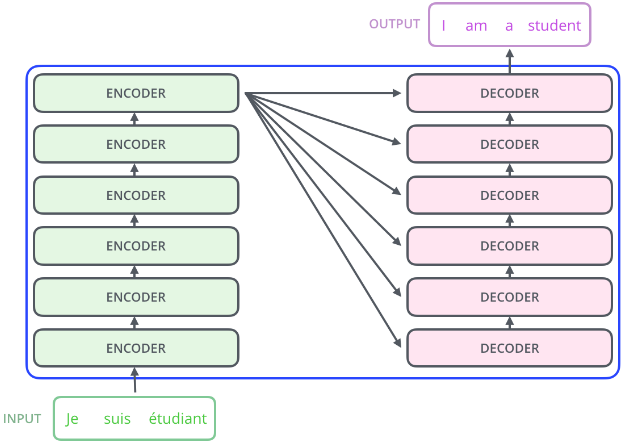
Many **independent attention "heads"** for specialized focus

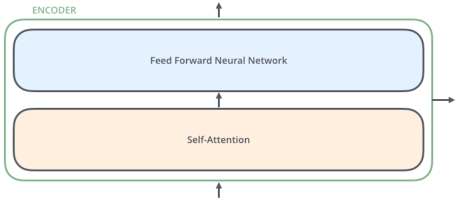


Layer: 5 Attention: Input - Input

Stack many layers!

Hierarchical learning: each layer processes the previous layer's output.





No decoders in masked language modelling (Devlin et al, 2018)