Protein Design & Structural Prediction

Alena Khmelinskaia

Department Chemie
Butenandtstr. 5-13, Haus B
81377 München
akhmelin@cup.lmu.de

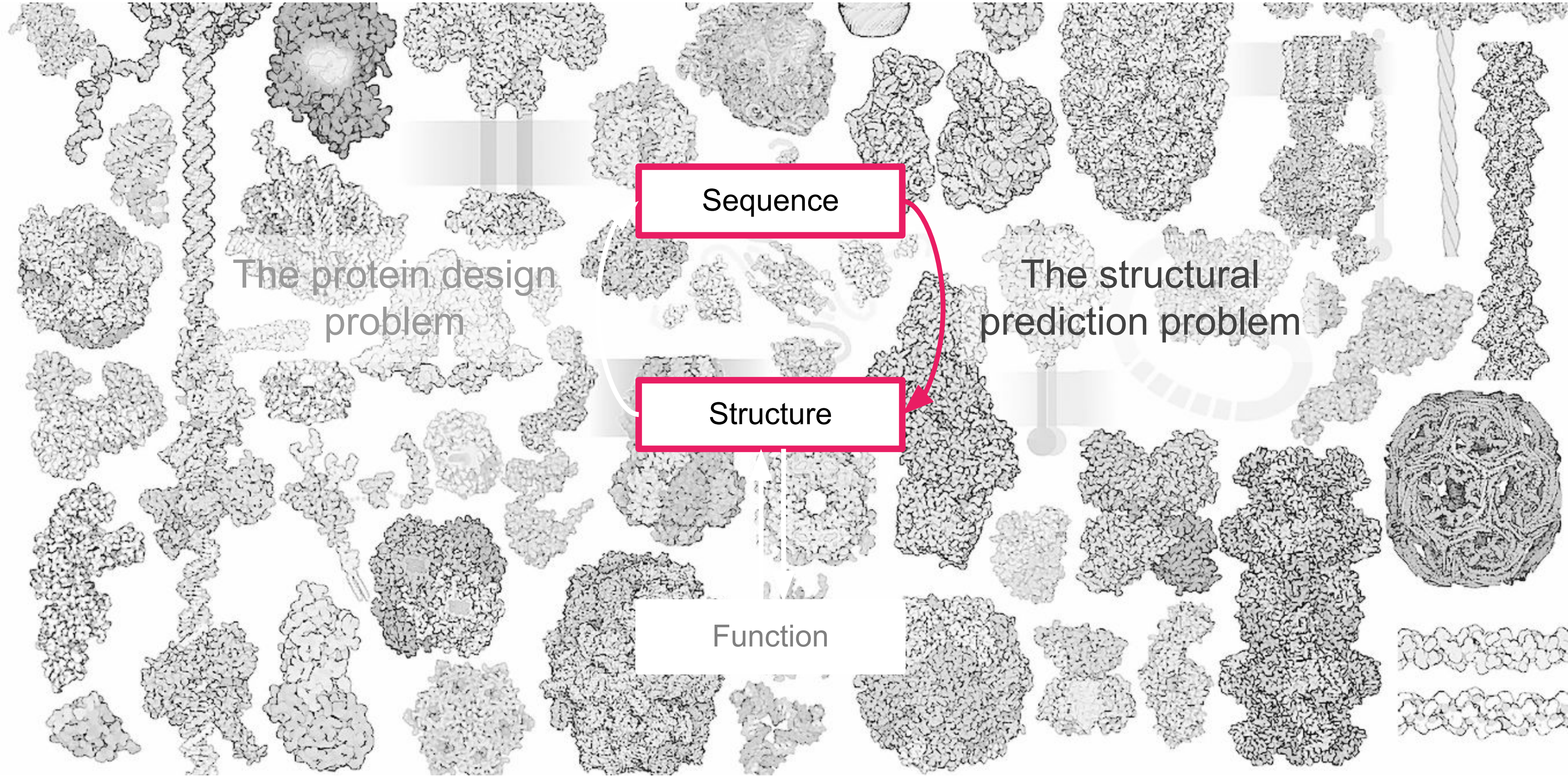
LMU
LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Overview

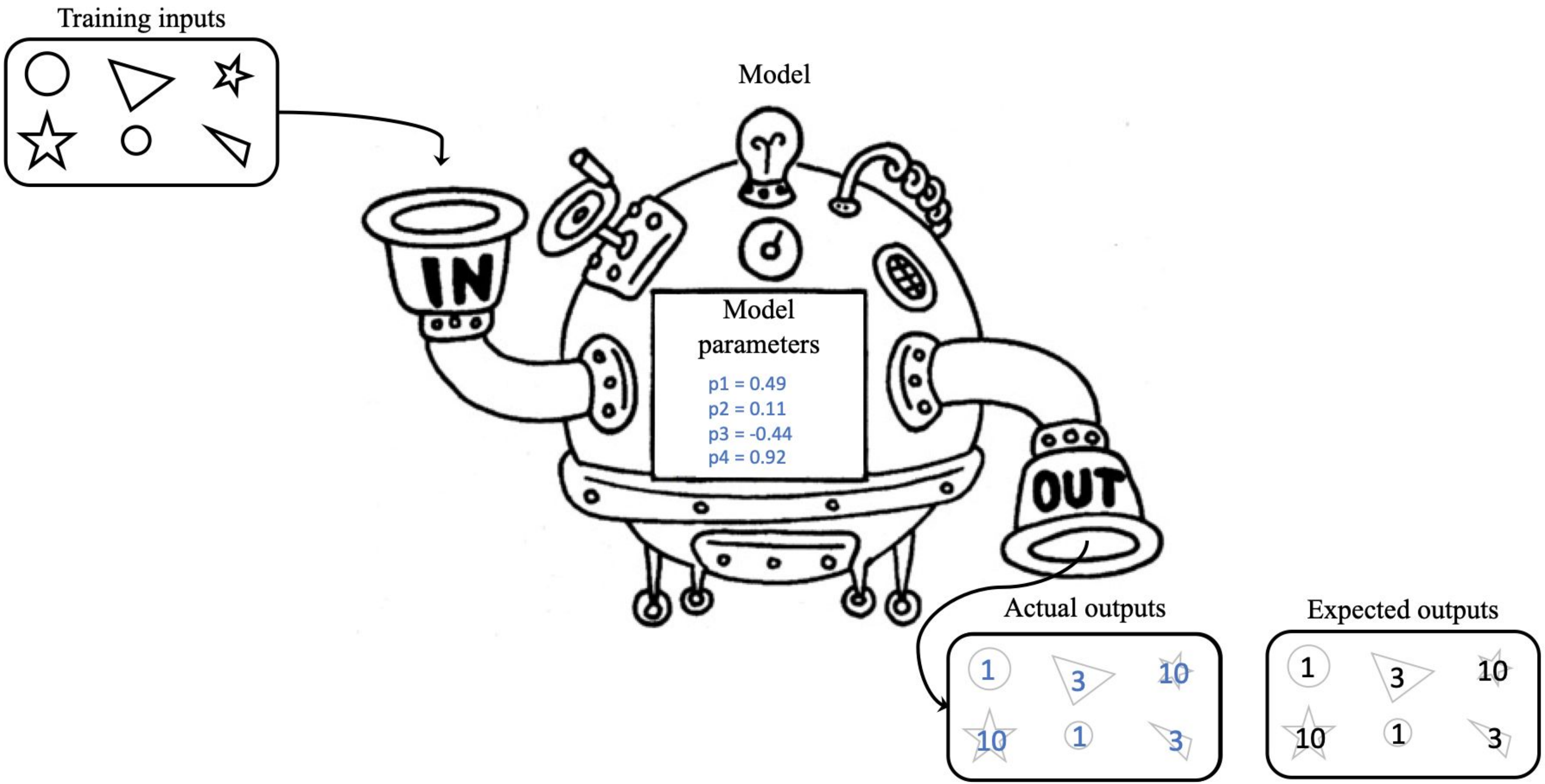
● The basics os protein structure	17.10
● Physics based approaches to structural prediction	24 - 07.10
● **Neural-networks for structural prediction**	**14 - 28.11**
● (py)Rosetta protein modelling and design	28.11&05.12
● Protein design in the age of AI	12&19.12
● Protein design mini-project	09 - 30.01
● Project presentation	06.02

But first a recap !

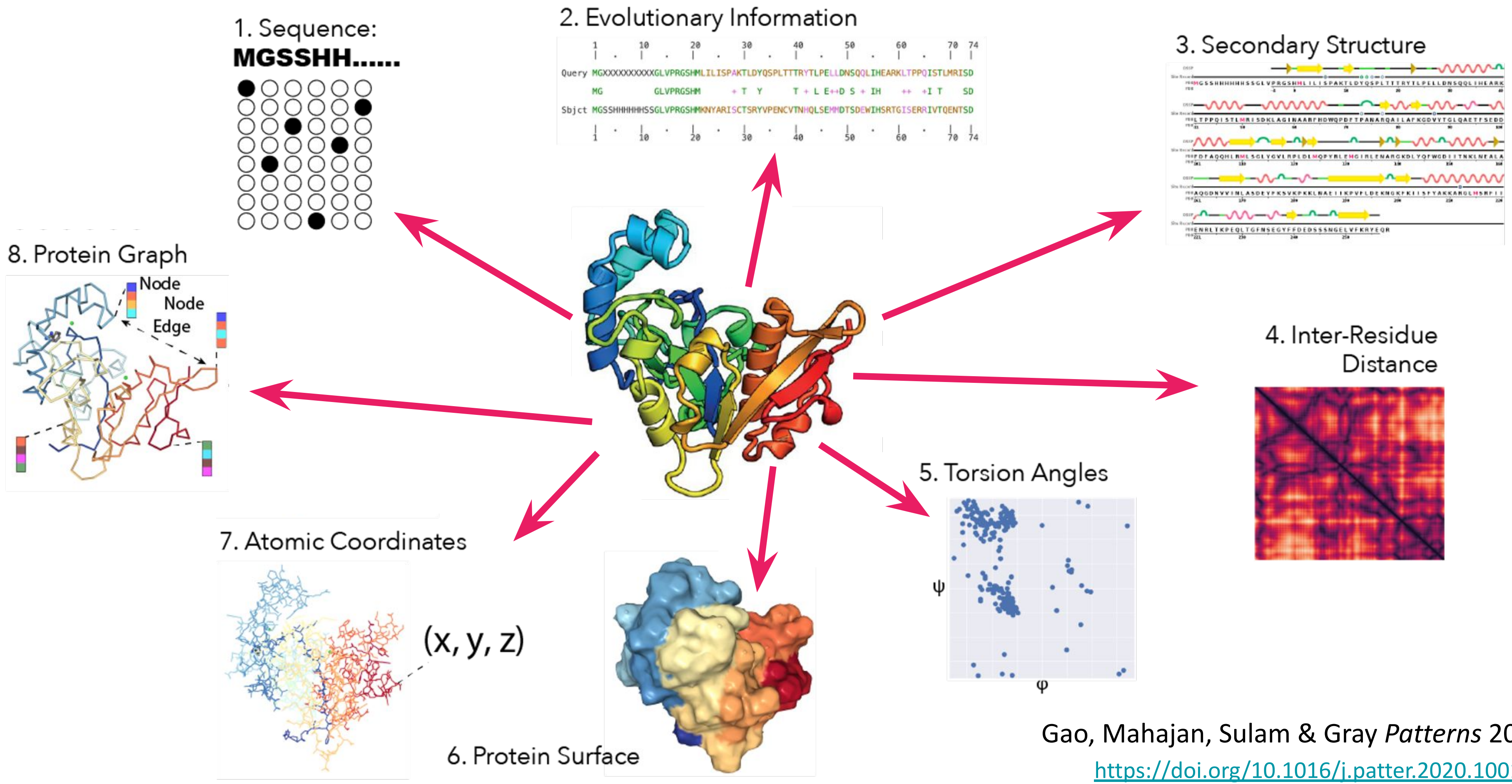
Protein structural prediction (with AI)


Sequence
Structure
Function
The protein design problem
The structural prediction problem

Supervised machine learning



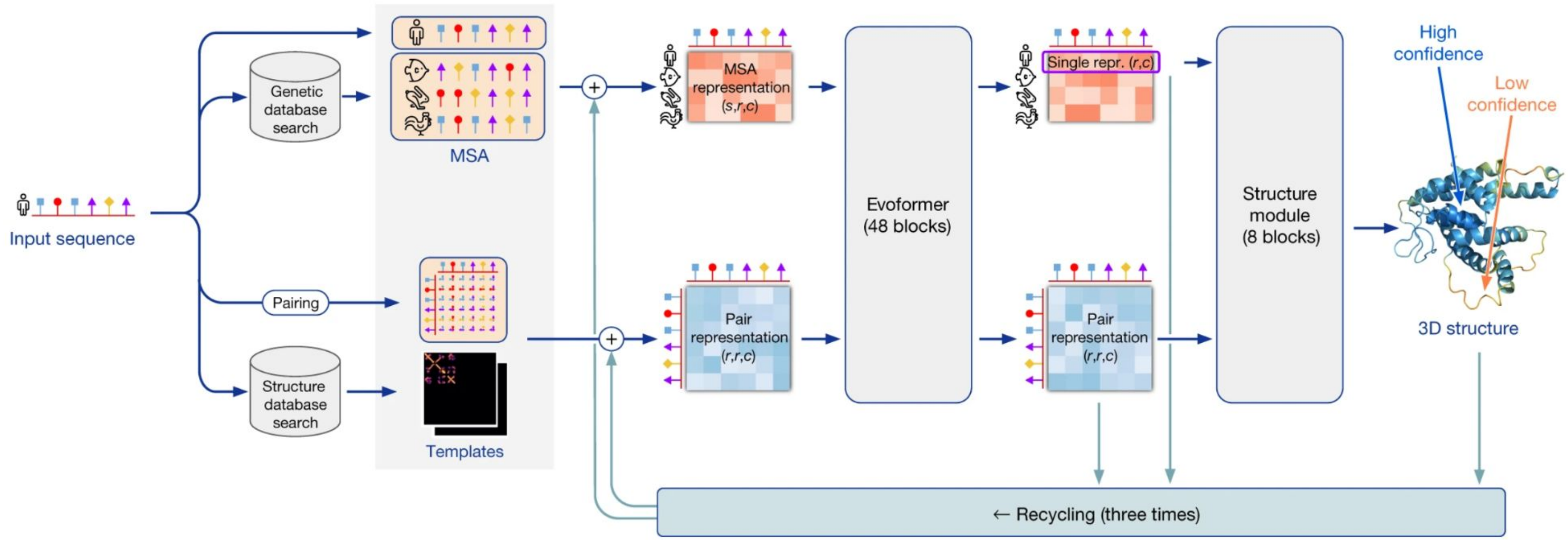
Common data representations for proteins in machine learning



Gao, Mahajan, Sulam & Gray *Patterns* 2020

https://doi.org/10.1016/j.patter.2020.100142

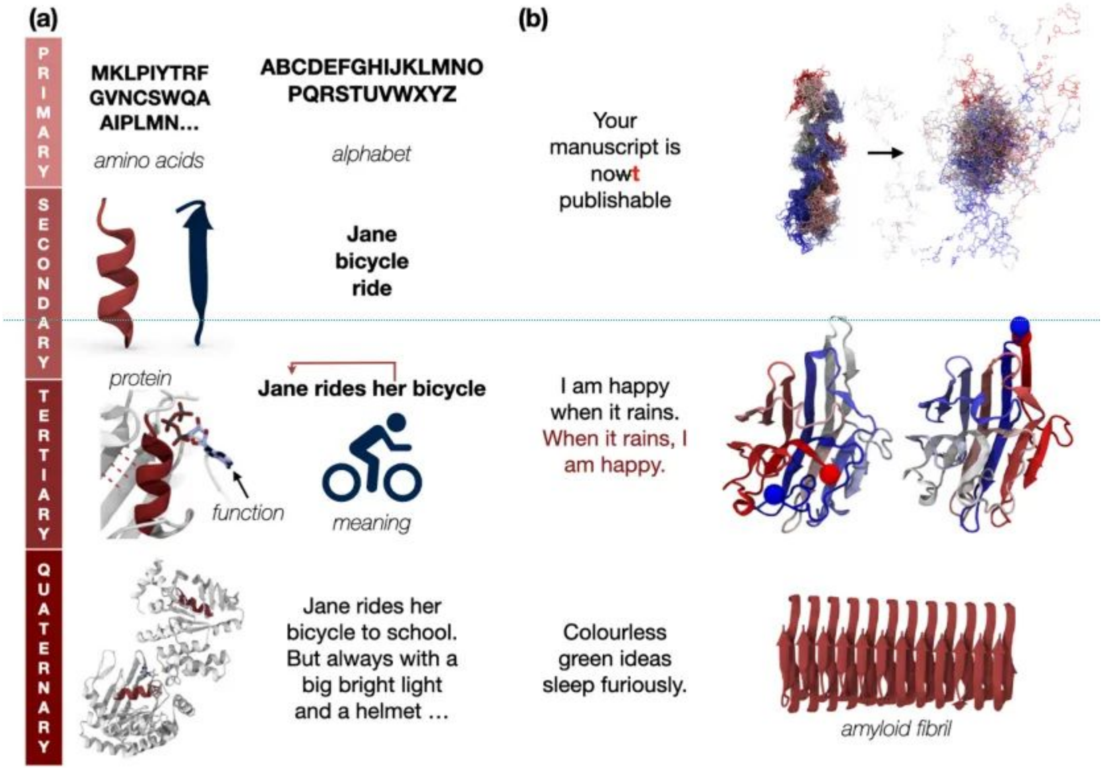
AlphaFold model architecture consists of an MSA module (Evoformer) and a Structure module



Protein language models for structural prediction

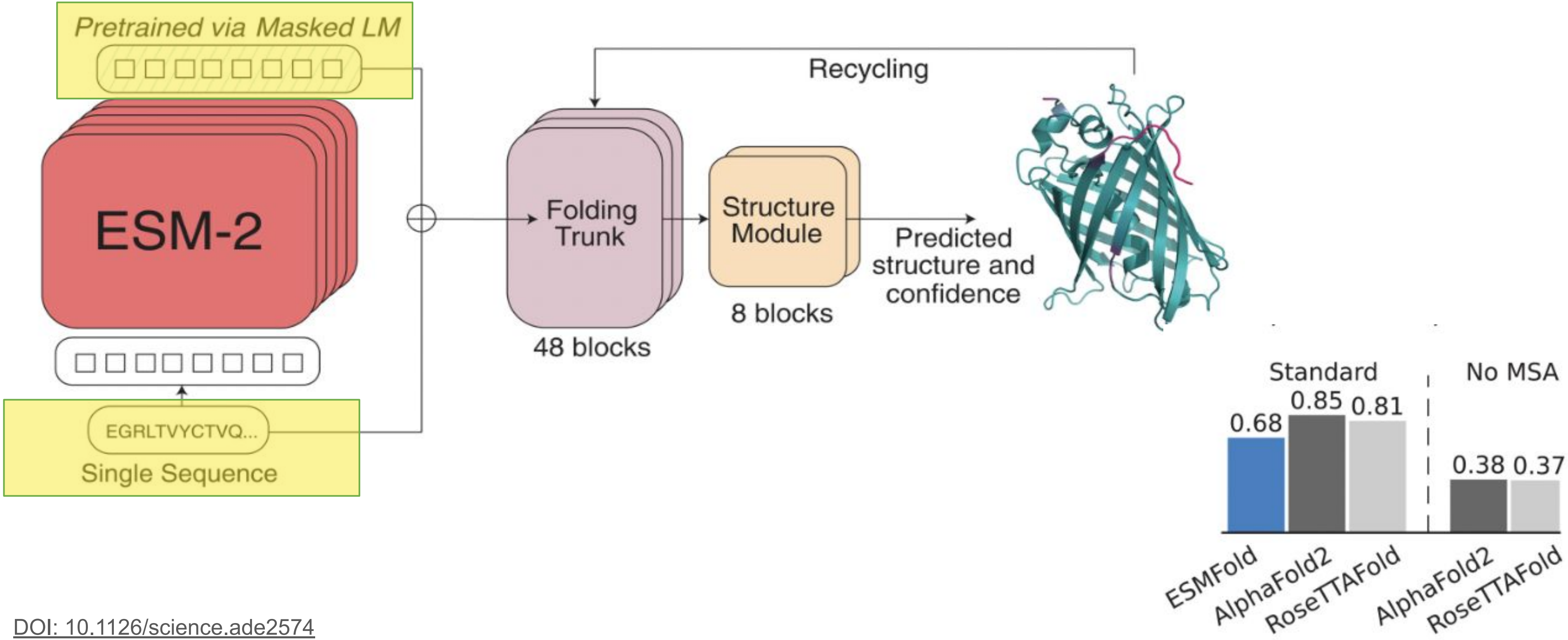
https://www.science.org/doi/10.1126/science.ade2574

Protein language



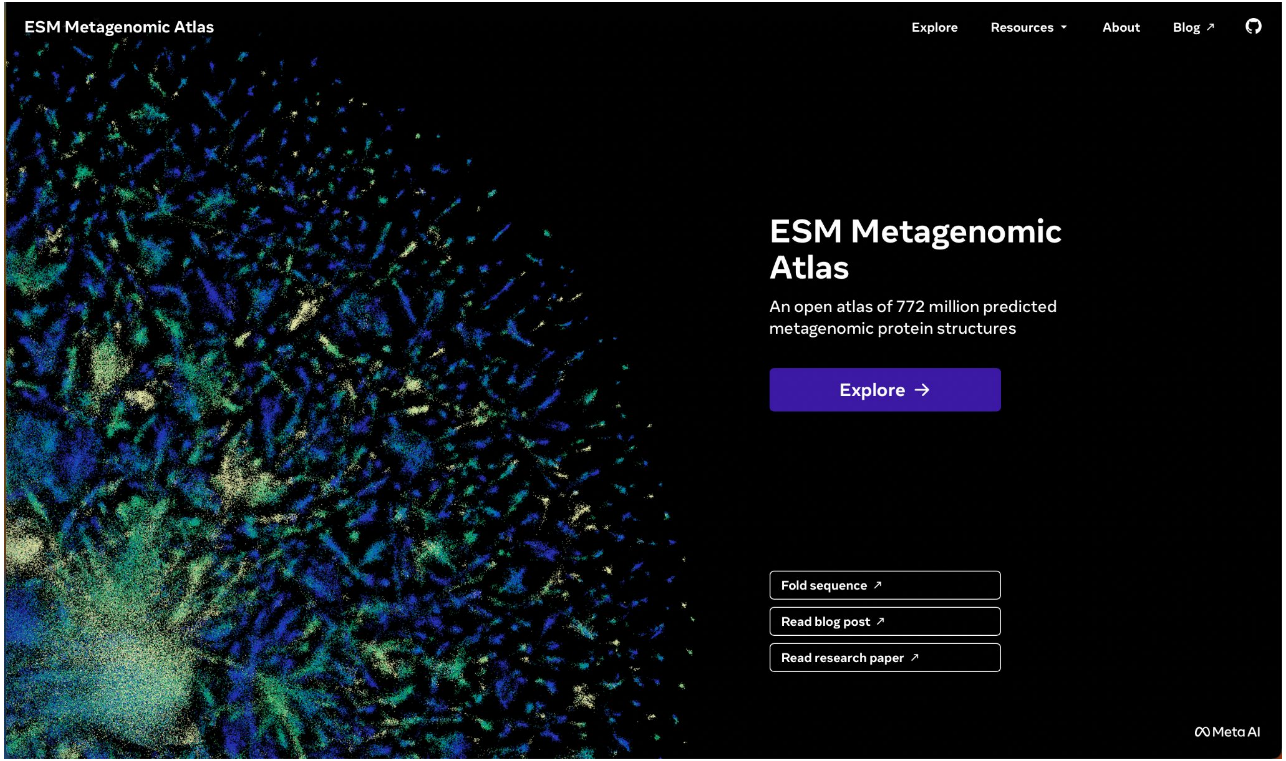
https://doi.org/10.1038/s42256-022-00499-z

ESMfold - a faster alternative to AF2

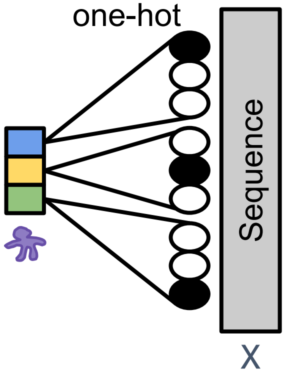


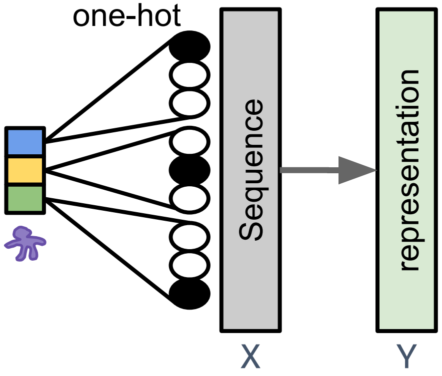
DOI: 10.1126/science.ade2574

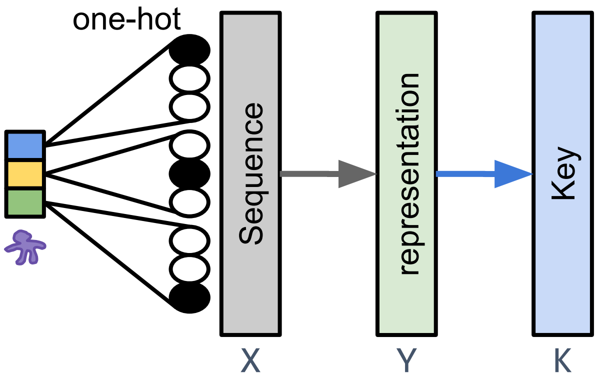
ESMatlas

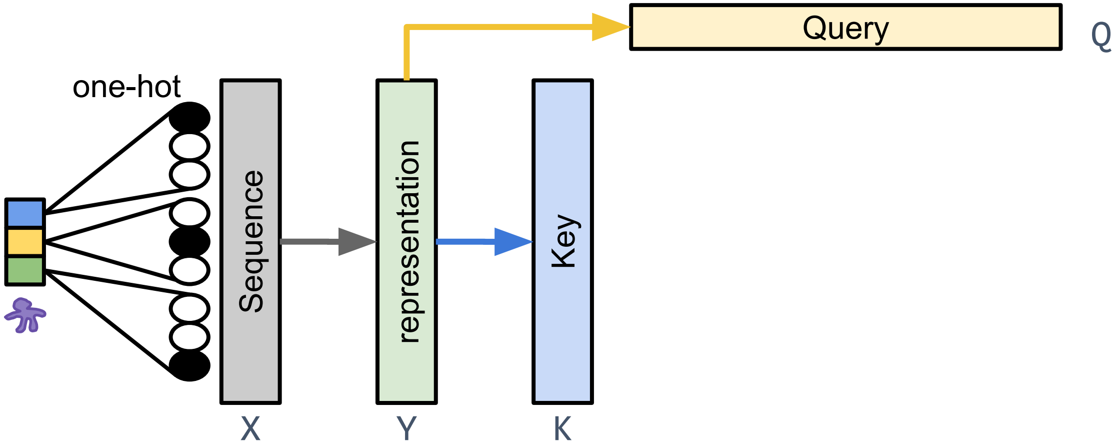


https://esmatlas.com/

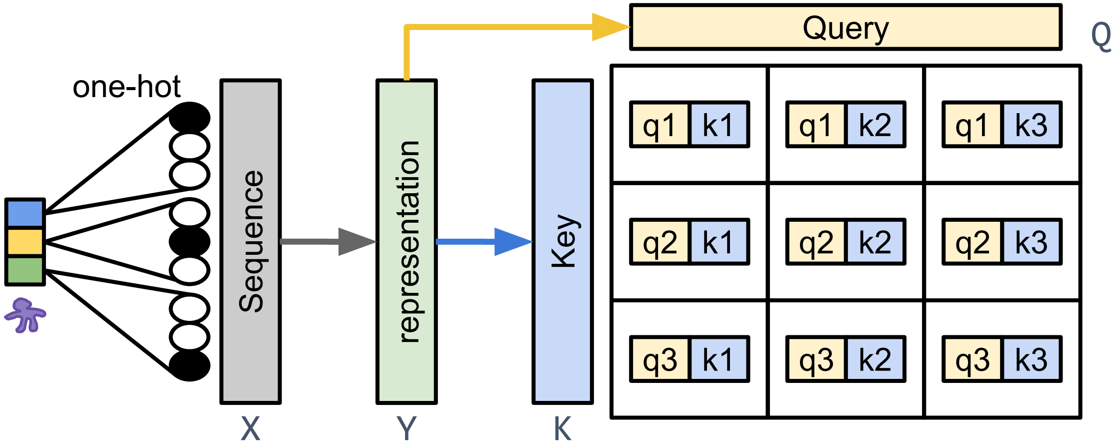

one-hot
Sequence
X


one-hot
Sequence
representation
X
Y

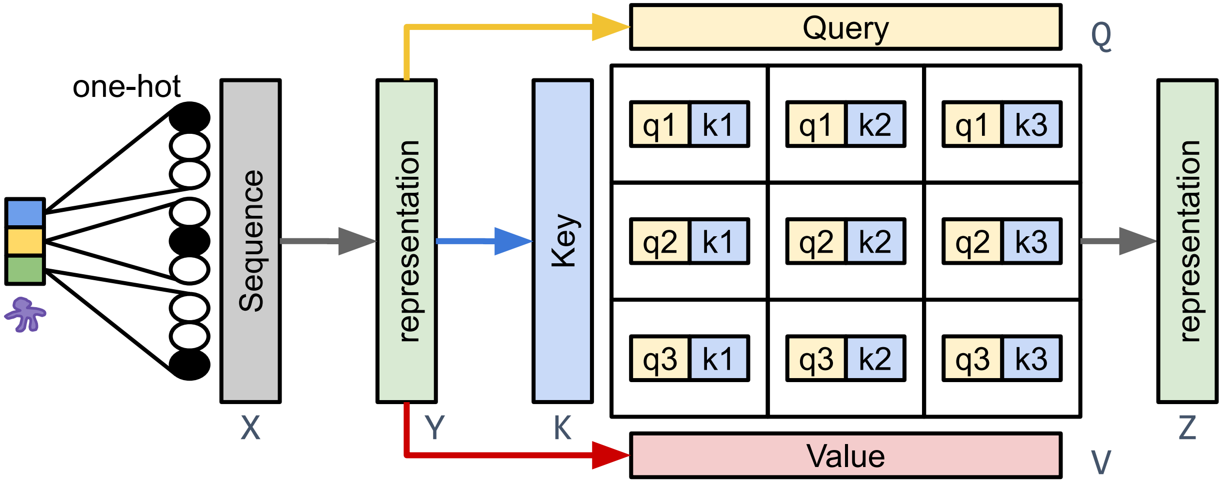

one-hot
Sequence
representation
Key
X
Y
K

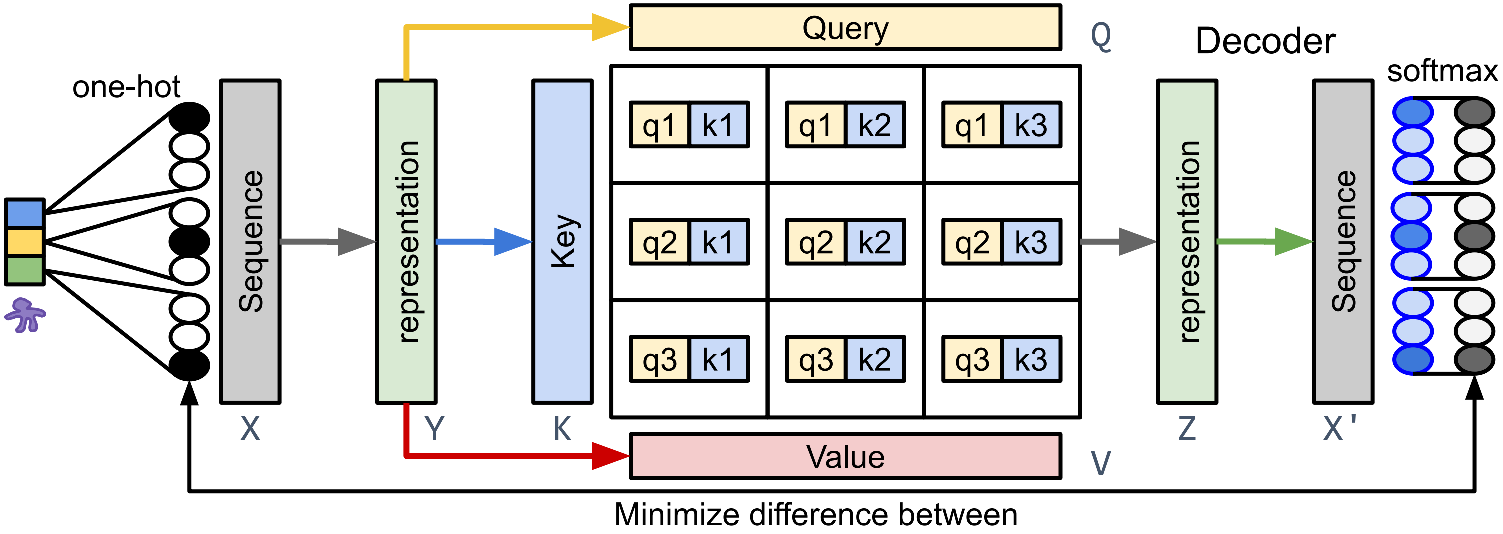


Query
Q
one-hot
Sequence
representation
Key
X
Y
K



Query
Q
one-hot
Sequence
representation
Key
q1 k1
q1 k2
q1 k3
q2 k1
q2 k2
q2 k3
q3 k1
q3 k2
q3 k3
X
Y
K

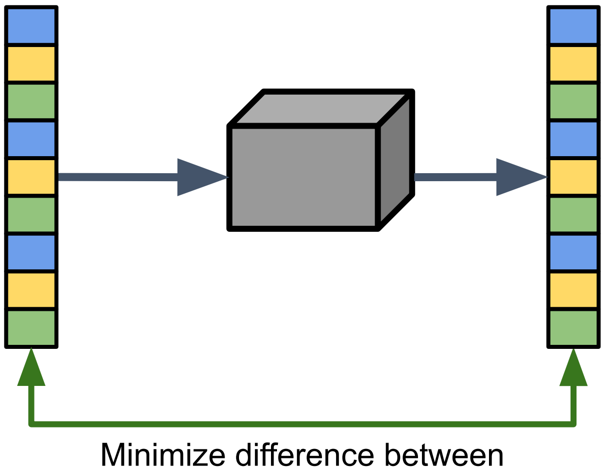

one-hot
Sequence
X
representation
Y
Key
K
Query
Q
q1 k1
q1 k2
q1 k3
q2 k1
q2 k2
q2 k3
q3 k1
q3 k2
q3 k3
Value
V
representation
Z



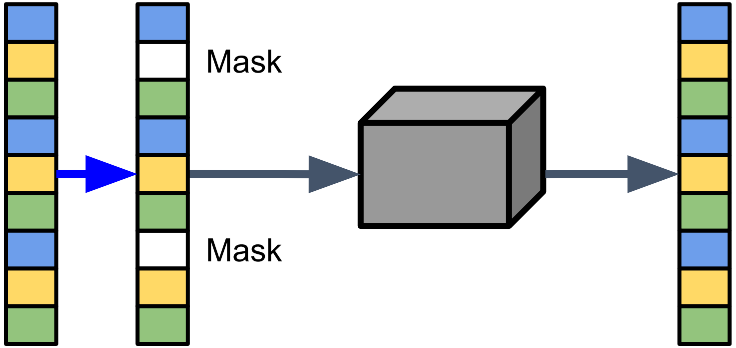
one-hot
Sequence
X
representation
Y
Key
K
Query
Q
q1 k1
q1 k2
q1 k3
q2 k1
q2 k2
q2 k3
q3 k1
q3 k2
q3 k3
Value
V
Decoder
representation
Z
Sequence
X'
softmax
Minimize difference between

How are protein language models trained?

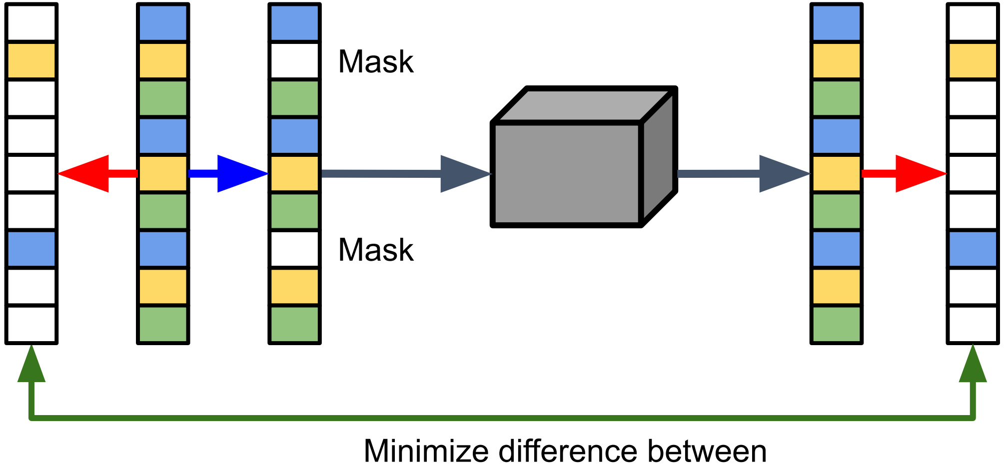
Unsupervised



Masked language modeling (or self-supervised)

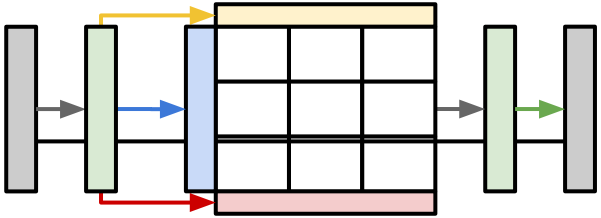


"Masked language modeling" is an approximation of "Pseudolikelihood"



$$\mathcal{L}_{PL}(\theta; x) = \sum_{i=1}^{L} \log p_\theta(x_i | x_{\setminus i}) \qquad \mathcal{L}_{MLM}(\theta; x, M) = \sum_{i \in M} \log p_\theta(x_i | x_{\setminus M})$$

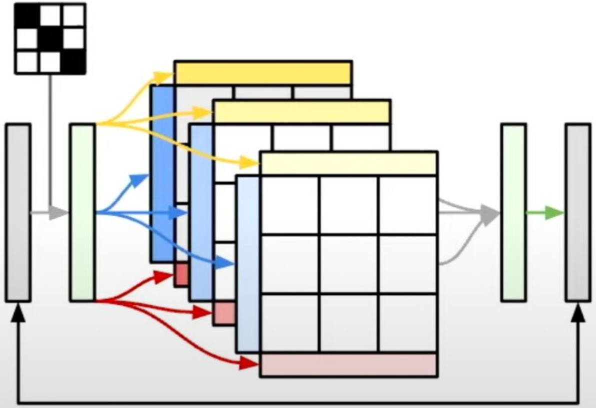
Single-head model



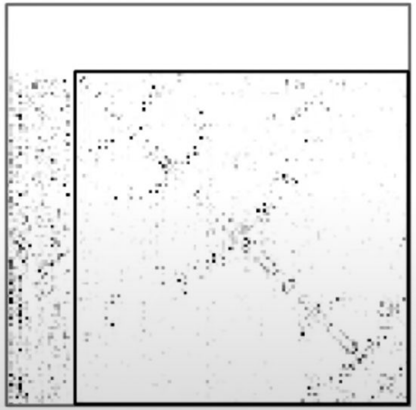
Minimize difference between

When pLM on a single protein family (MSA) we find:

- You only need **ONE** layer.
- You can replace the **positional embedding** with an **identity matrix** (encoding the exact positional information).
- The weights of the **Query** and **Key** layers encode the contact map!

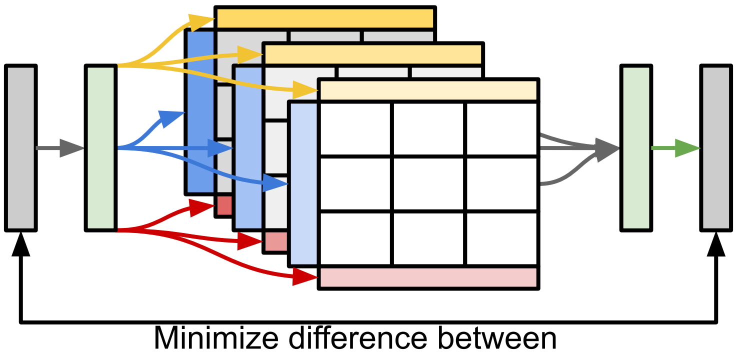


(query @ key)

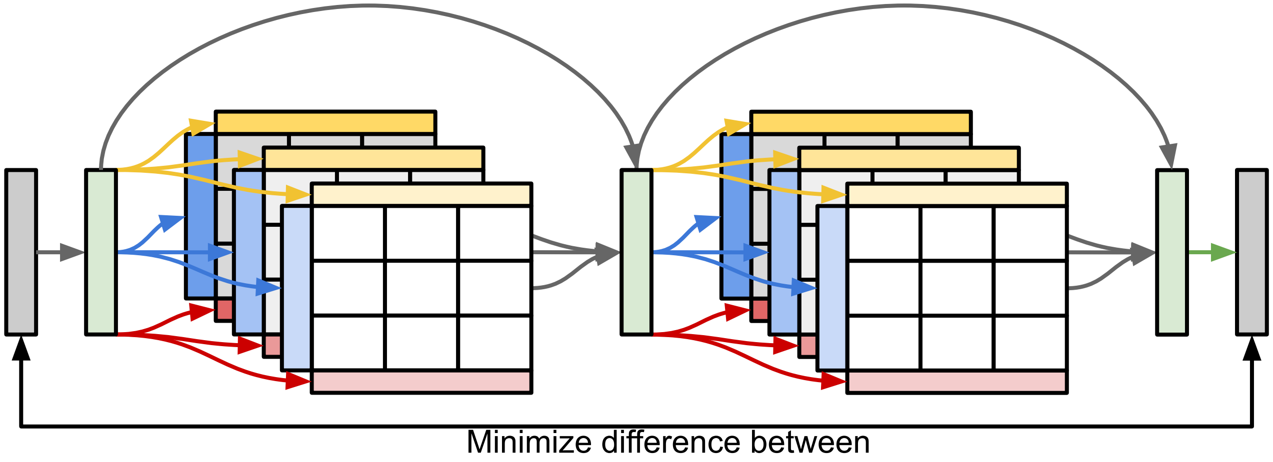


Bhattacharya, N., Thomas, N., Rao, R., Dauparas, J., Koo, P.K., Baker, D., Song, Y.S. and Ovchinnikov, S., 2020. Single layers of attention suffice to predict protein contacts.

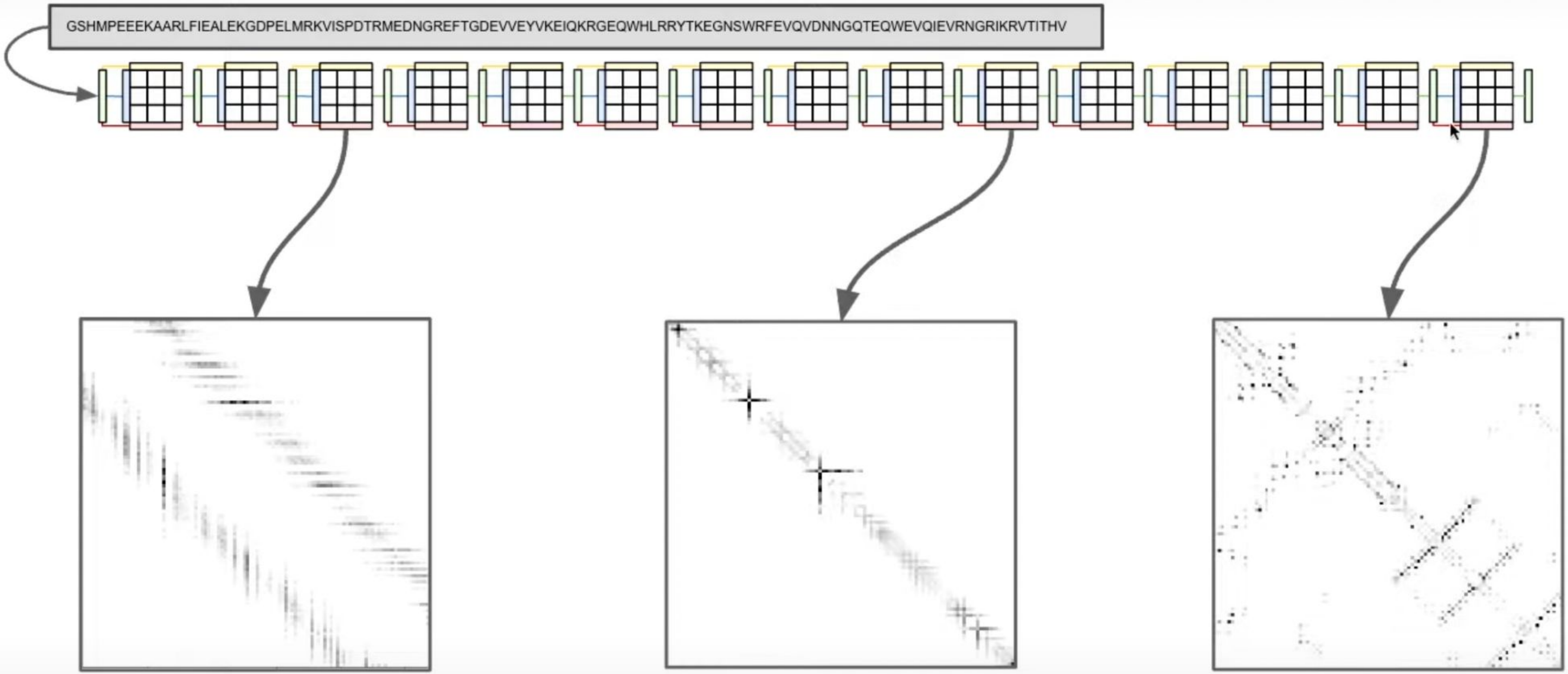
Multi-head model



Multi-layer/Multi-head model



Extract attention matrices

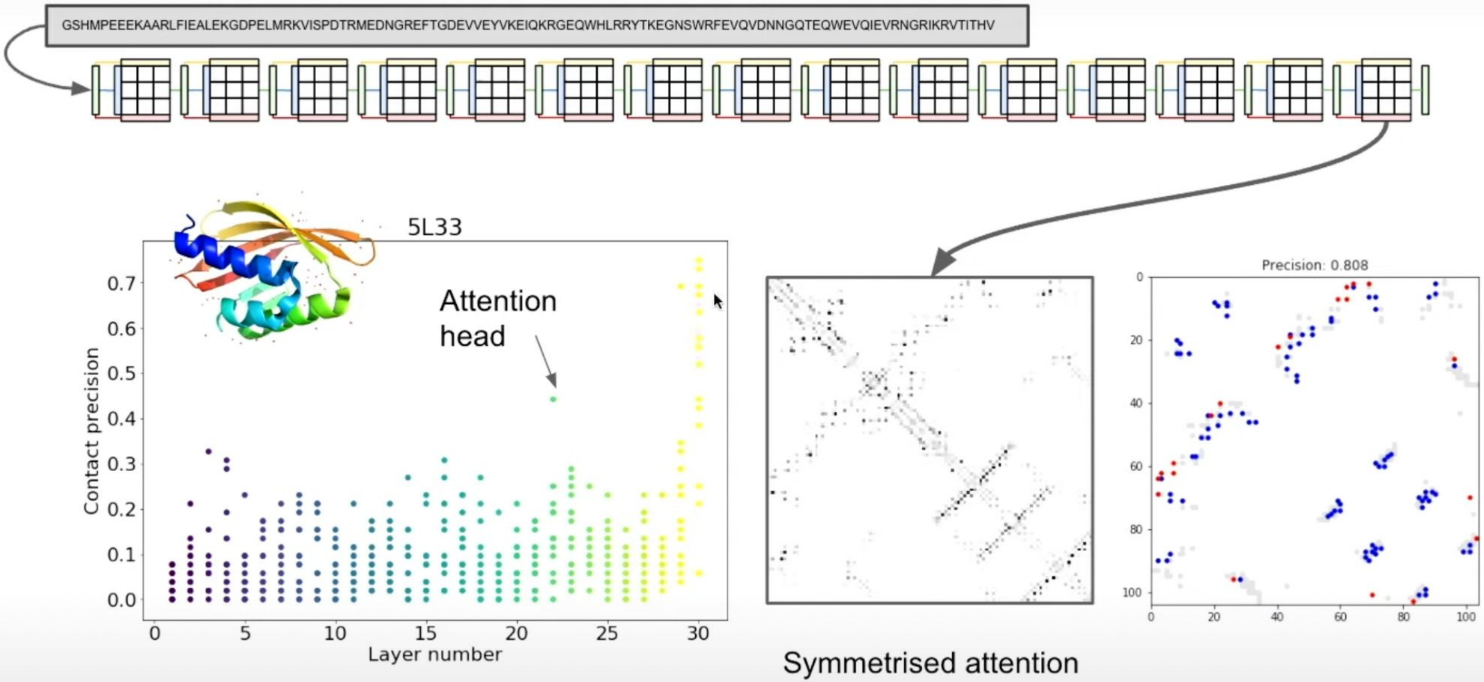


Symmetrised attention

Symmetrised attention

Symmetrised attention

Layers towards the end tend to capture contacts!

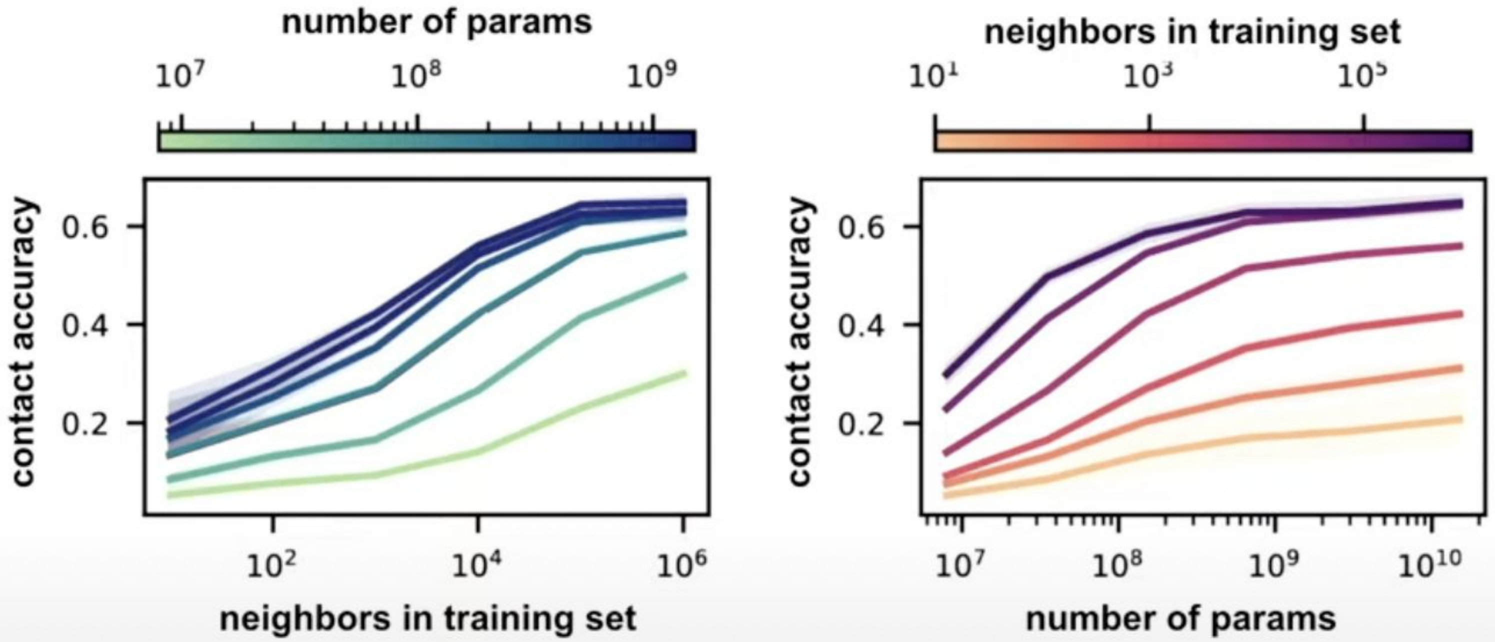


Vig, Jesse, et al. "**BERTology Meets Biology: Interpreting Attention in Protein Language Models**." (2020).

Bhattacharya, Nick, et al. **"Single Layers of Attention Suffice to Predict Protein Contacts"** (2020)

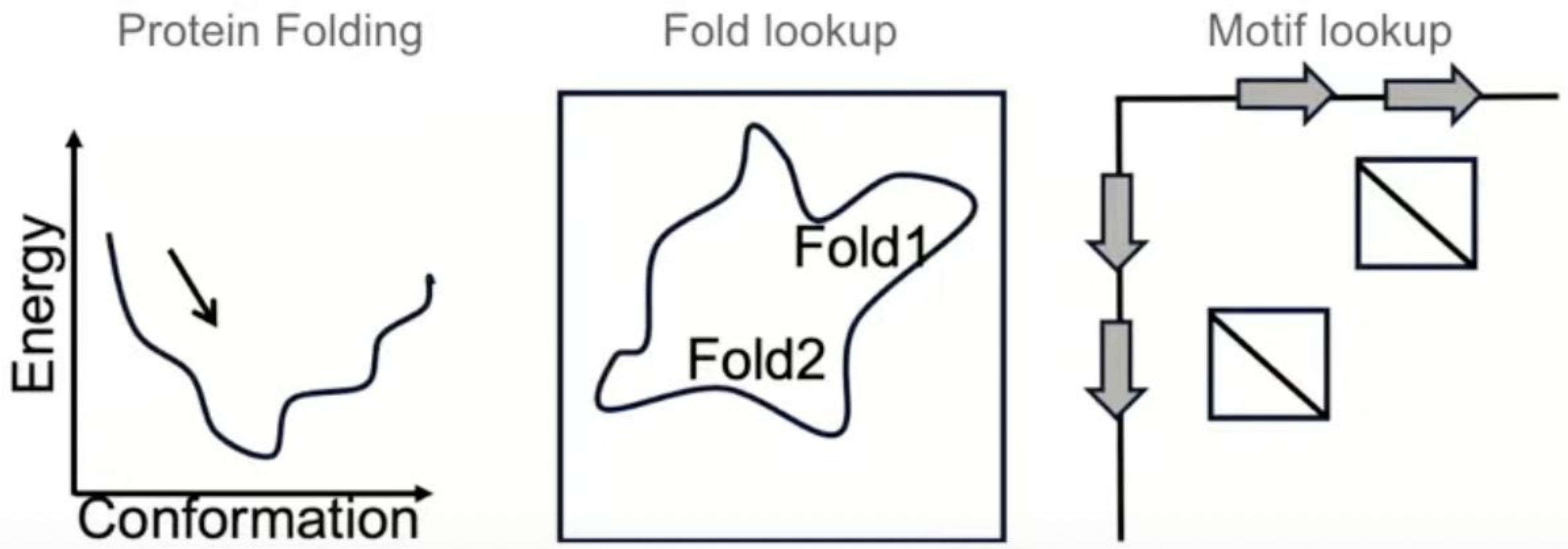
Rao, Roshan, et al. **"Transformer protein language models are unsupervised structure learners."** (2020)

ESM2: train models with different number of params



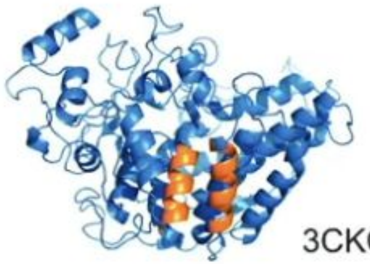
Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S. and Rives, A., 2022. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*.

How do protein language models "store" coevolution statistics?



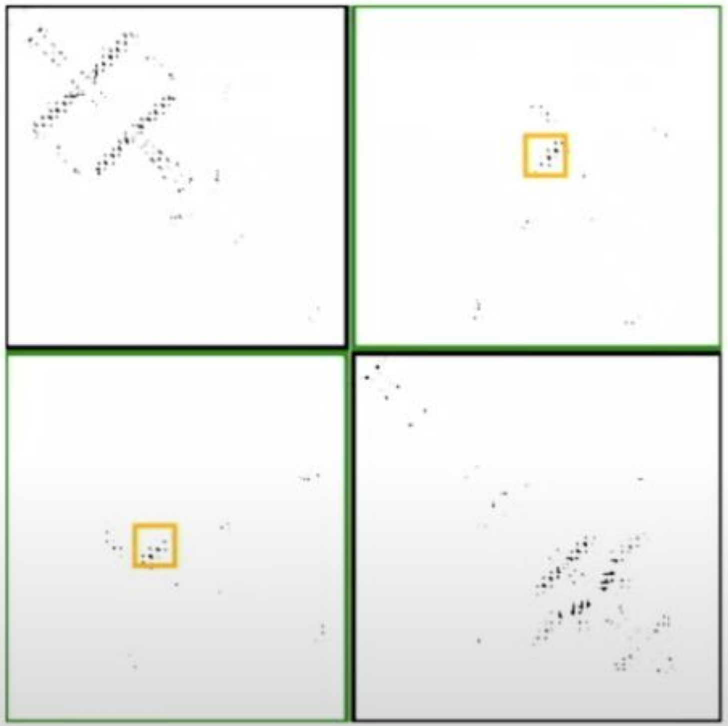
Masking majority of the sequence recovers the motif

SusD starch-
binding protein



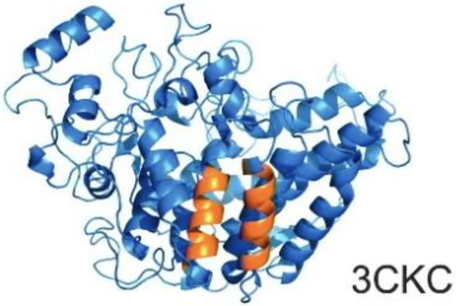
3CKC

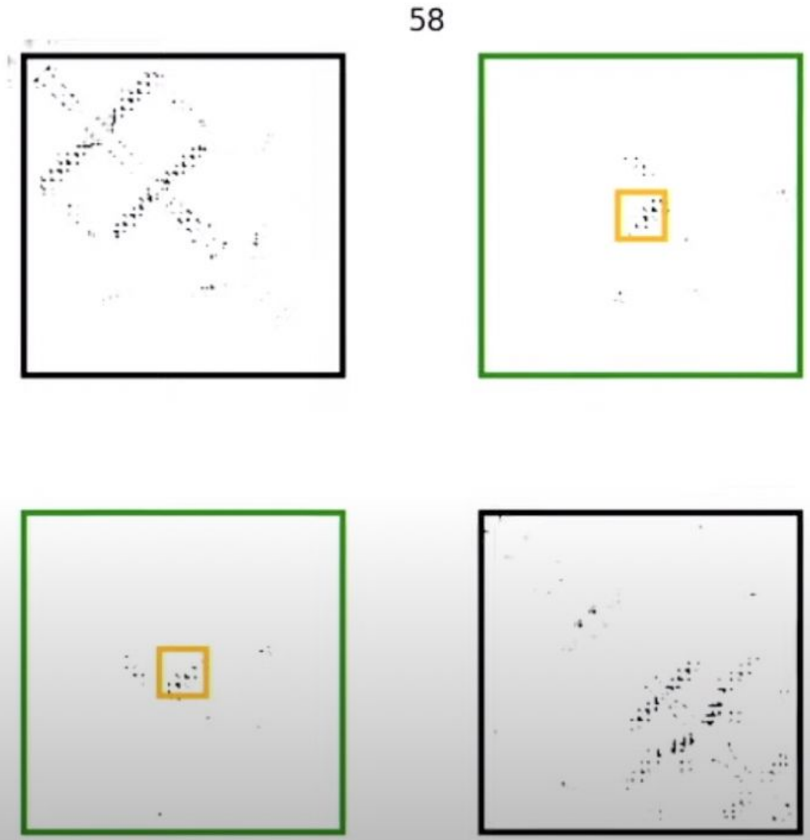
86



Masking majority of the sequence recovers the motif

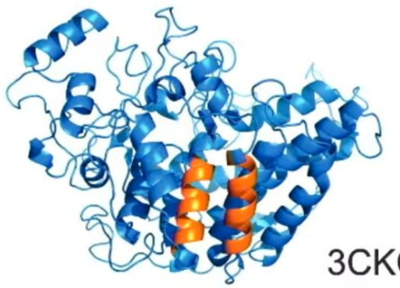
SusD starch-binding protein

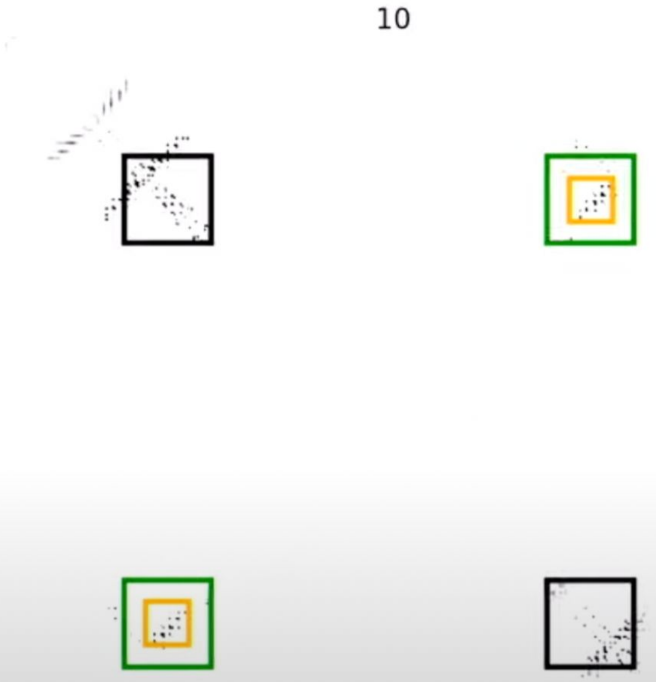




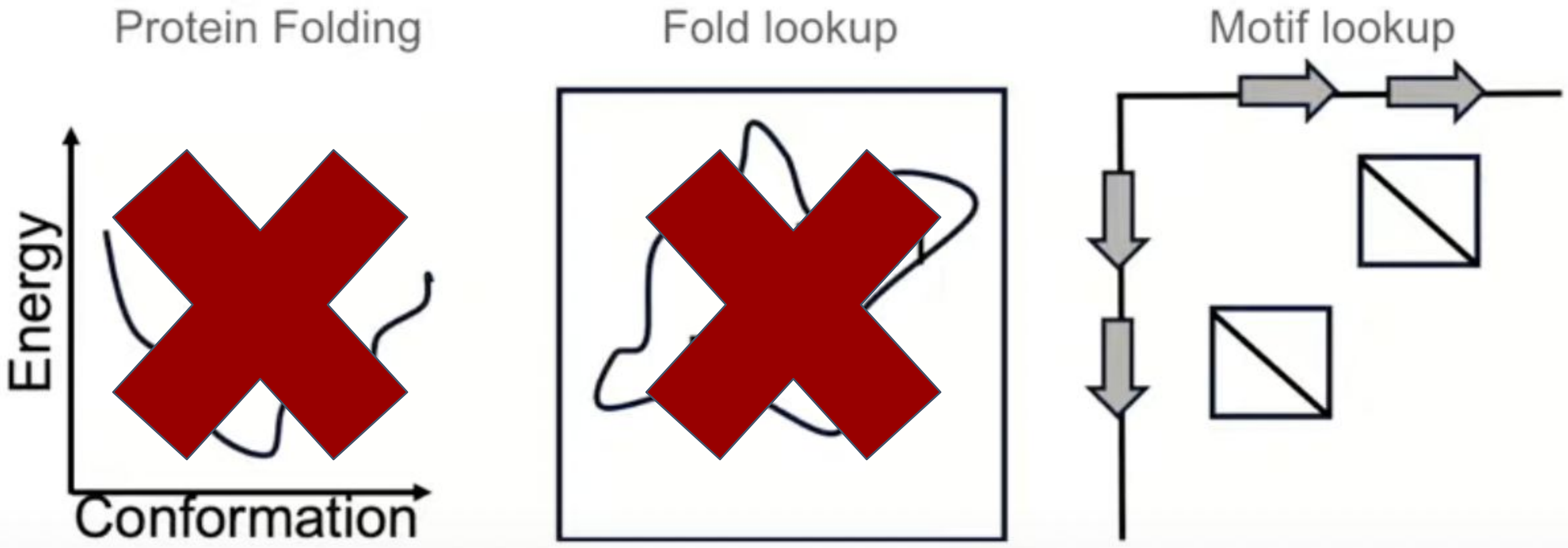
Masking majority of the sequence recovers the motif

SusD starch-
binding protein



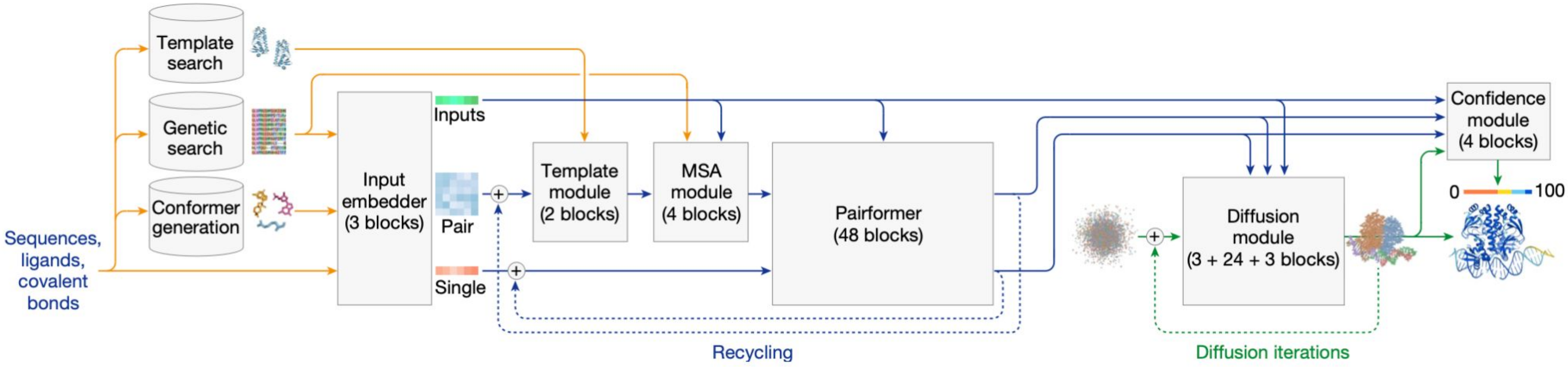


How do protein language models "store" coevolution statistics?



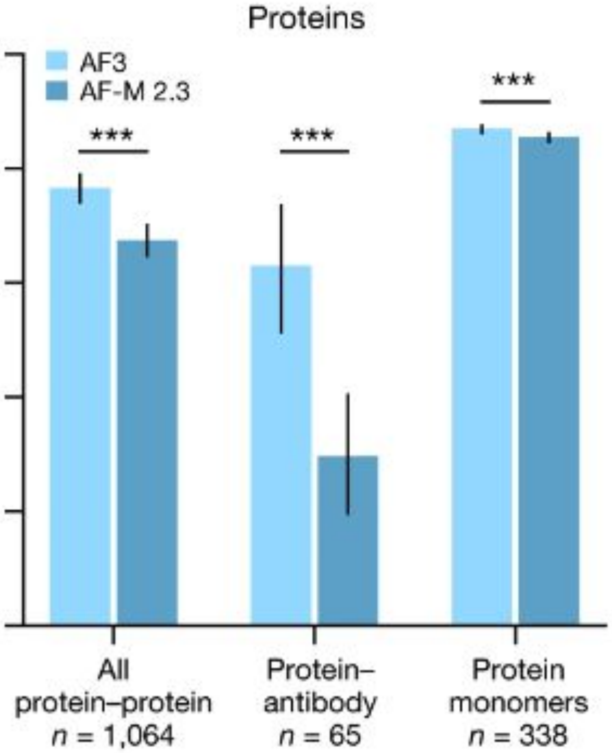
The field continuously changes…

AF3 - making a deterministic problem not so…



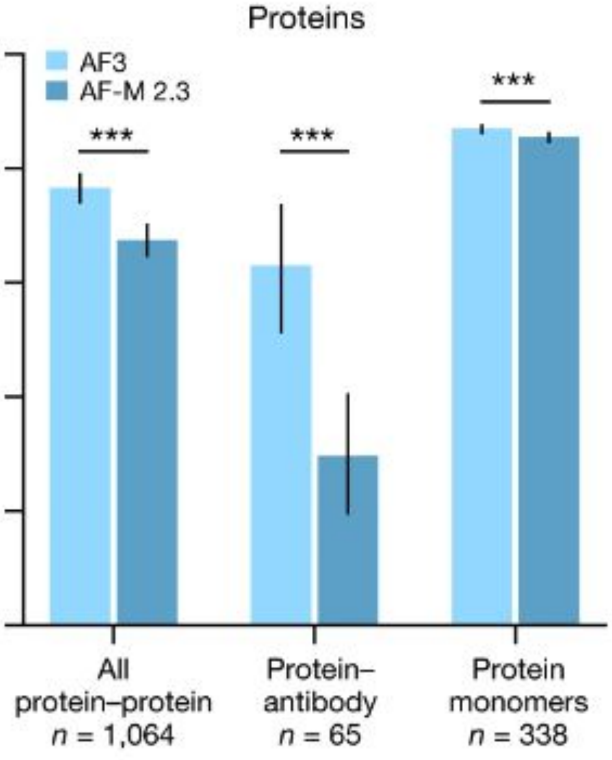
https://www.nature.com/articles/s41586-024-07487-w

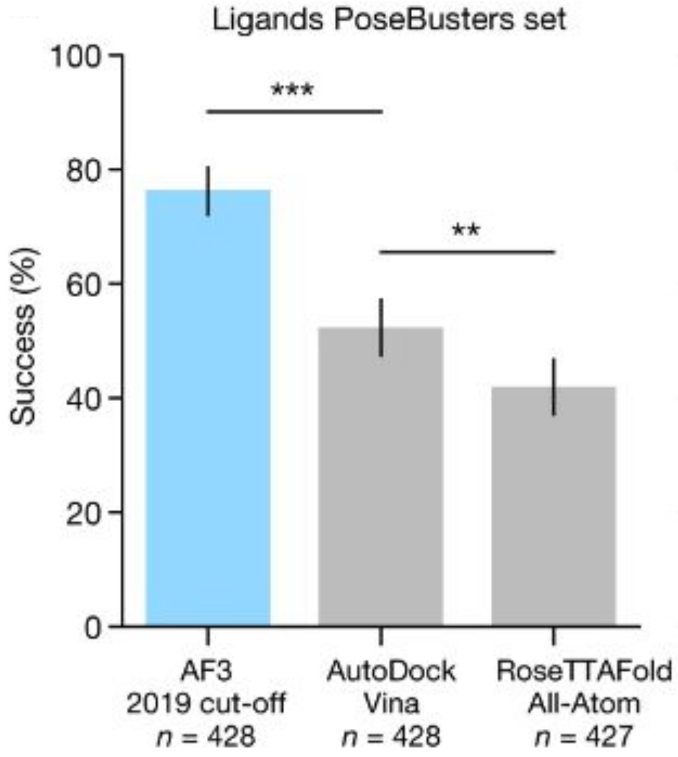
AF3 - making a deterministic problem not so...



https://www.nature.com/articles/s41586-024-07487-w

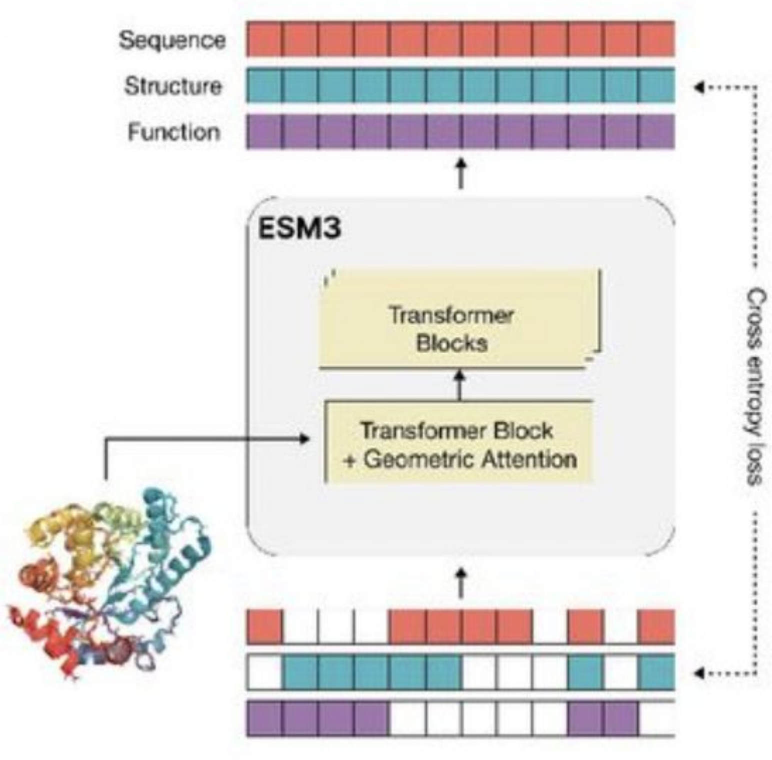
AF3 - making a deterministic problem not so...





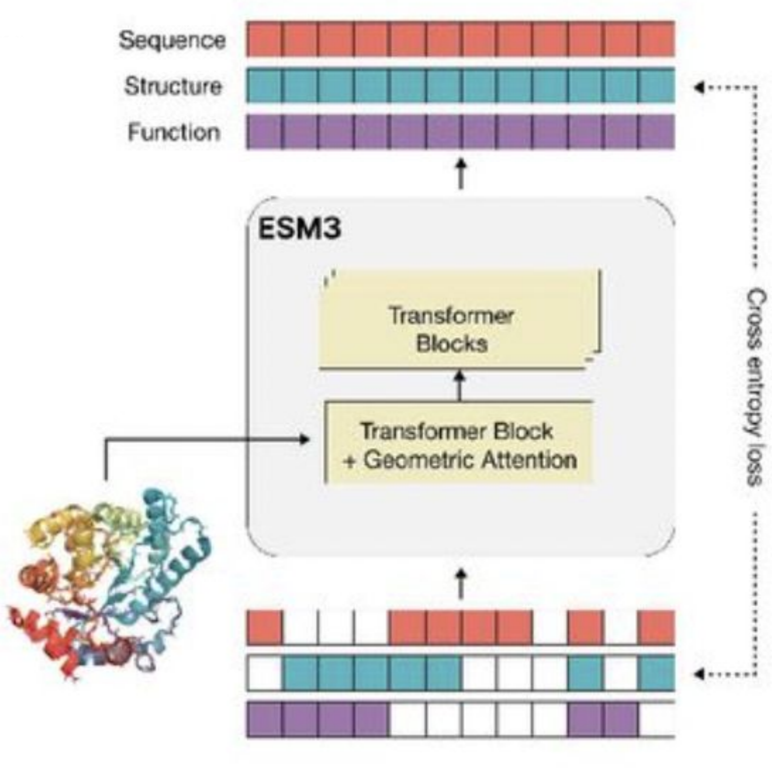
https://www.nature.com/articles/s41586-024-07487-w

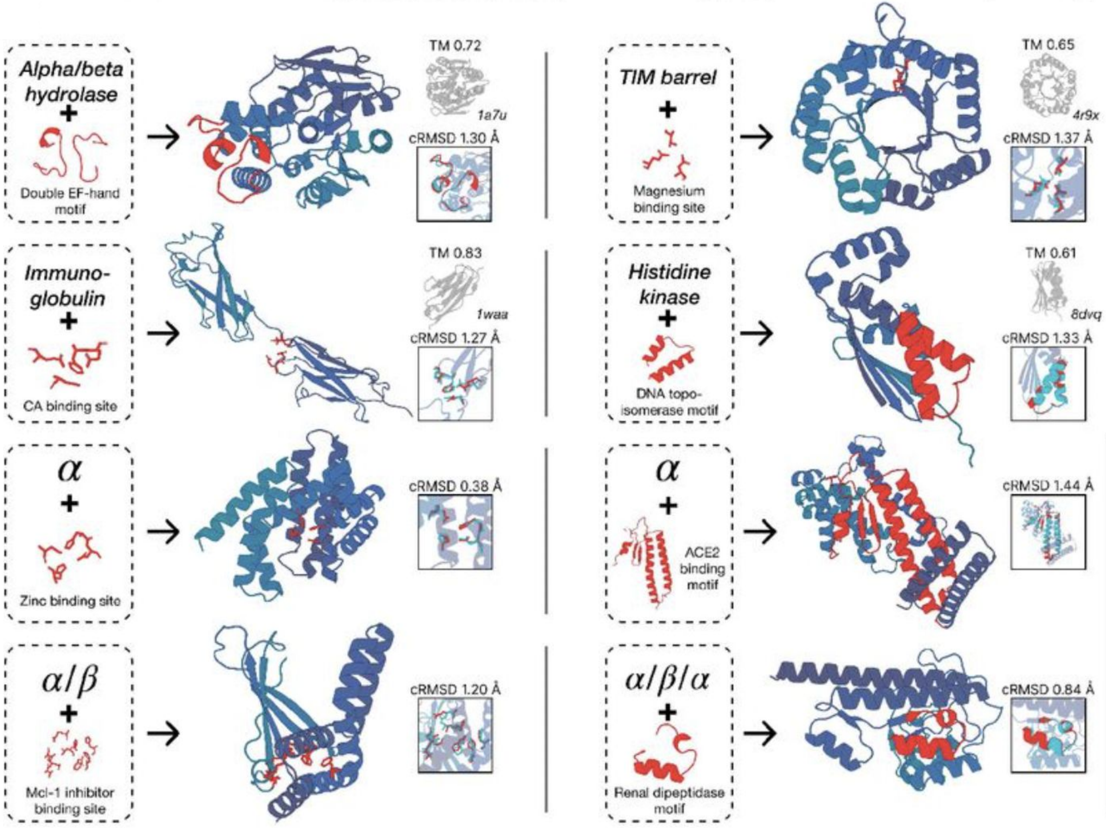
ESM3 - combining language models with structural and functional information



https://www.biorxiv.org/content/10.1101/2024.07.01.600583v1

ESM3 - combining language models with structural and functional information





https://www.biorxiv.org/content/10.1101/2024.07.01.600583v1