

UNIVERSITÄT
HEIDELBERG



Evolution and Bioinformatics

L4, Structural Bioinformatics

WiSe 2023/24, Heidelberg University











Overview

1. **Protein Evolution**
2. **Language Modelling**
3. **Protein Linguistics: Language Models in Biology**
4. **Practical Considerations**
5. **Current Research**

1. Protein Evolution

How do we get function from sequence?

Compare similar proteins across species via alignments

	WALRKTRKRLEEPFGGVKVL... 180
	R R++ +PFGG+++++ GD QL PV G + F FQ+ W+
	AVARAVRQQ-NKPFGGIQLI... 168
	-----RFCFQSKSWKRCV
	-VALRVHRLWESQRQREDPL... 237
	V L + ++W ++ D F LL+ +R G + L A G + L
	PVTLELTKVW-----RQAD... 224
	QTFISLLQAVRLGRCSDEV... 224
	PRRKEADALNLKRLEALPG... 293
	+ + N +RL+ LPGK ++A E A T P L LK GAQV+L++N
	THQDDVALTNERRLQELPG... 284
	KVHRFEAMDSNPELASTLD... 284
	DPLGE-YFNGDLGWVEDLE... 350
	+ NG G V EAE + R G VI W T + ++ +
	LSVSRGLVNGARGVVVGFE... 339
	AEGRGLPQVRFLCGVTEVI... 339
	VVGTFRQVPVRLAWALTVH... 406
	+Q+P++LAWA+++HK+QG... 406
	-----QQLPLQLAWAMSI... 389
	HKSQGMTLDCVEISLGR-V... 389

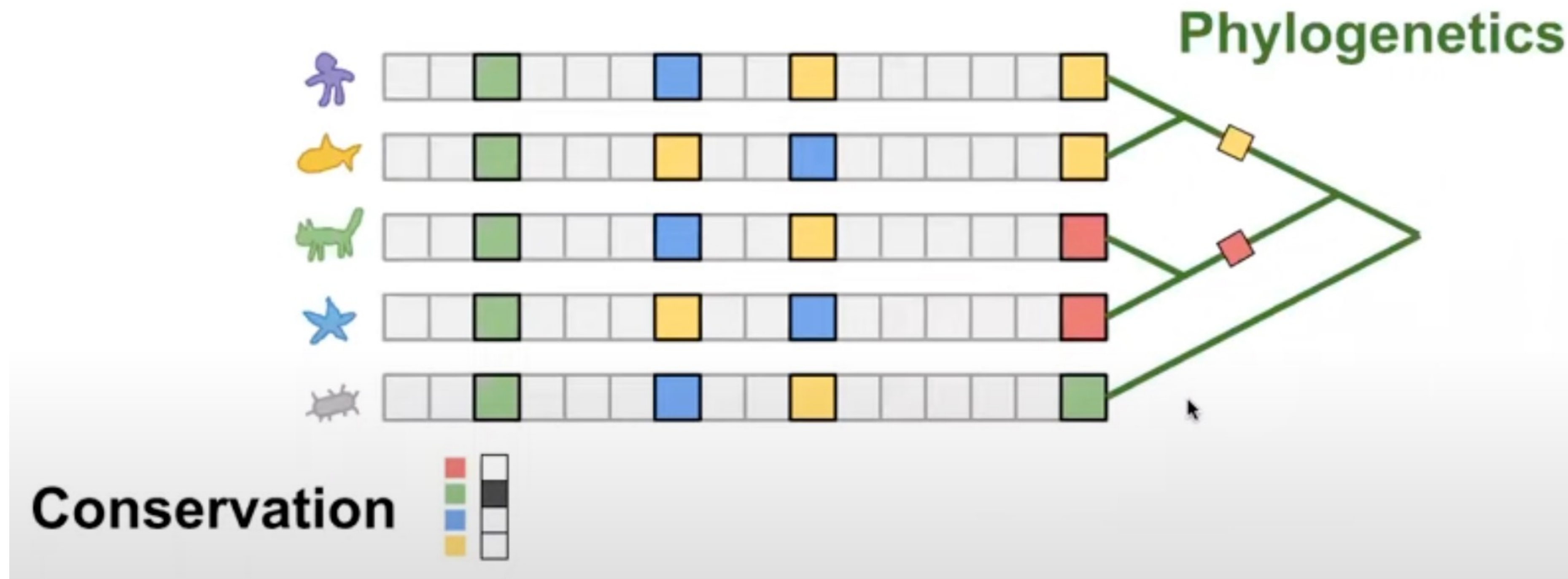
Evolution can give us hints about function

Phylogenetics: the study of evolutionary history



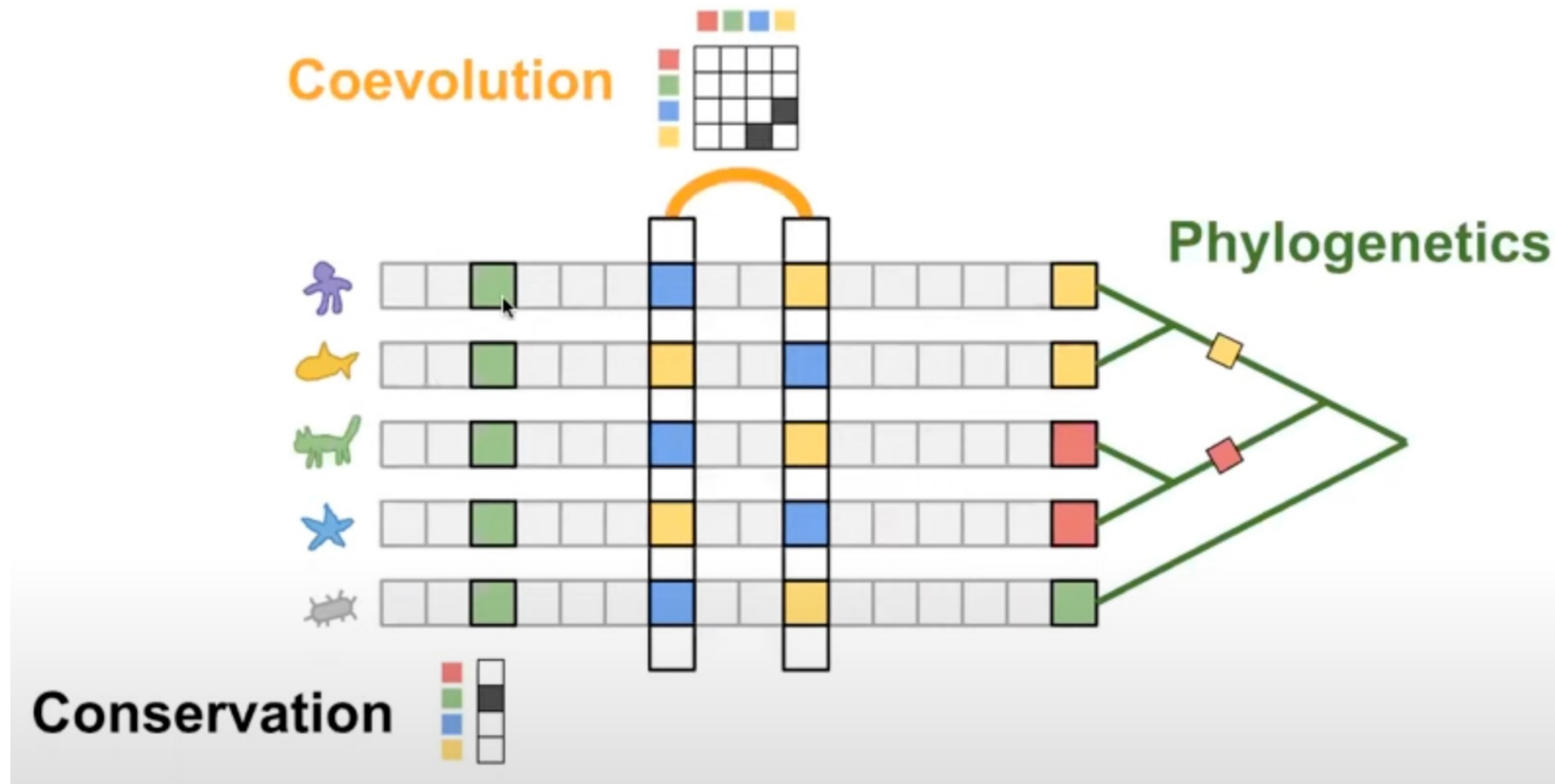
Evolution can give us hints about function

Conservation: which residues are important?



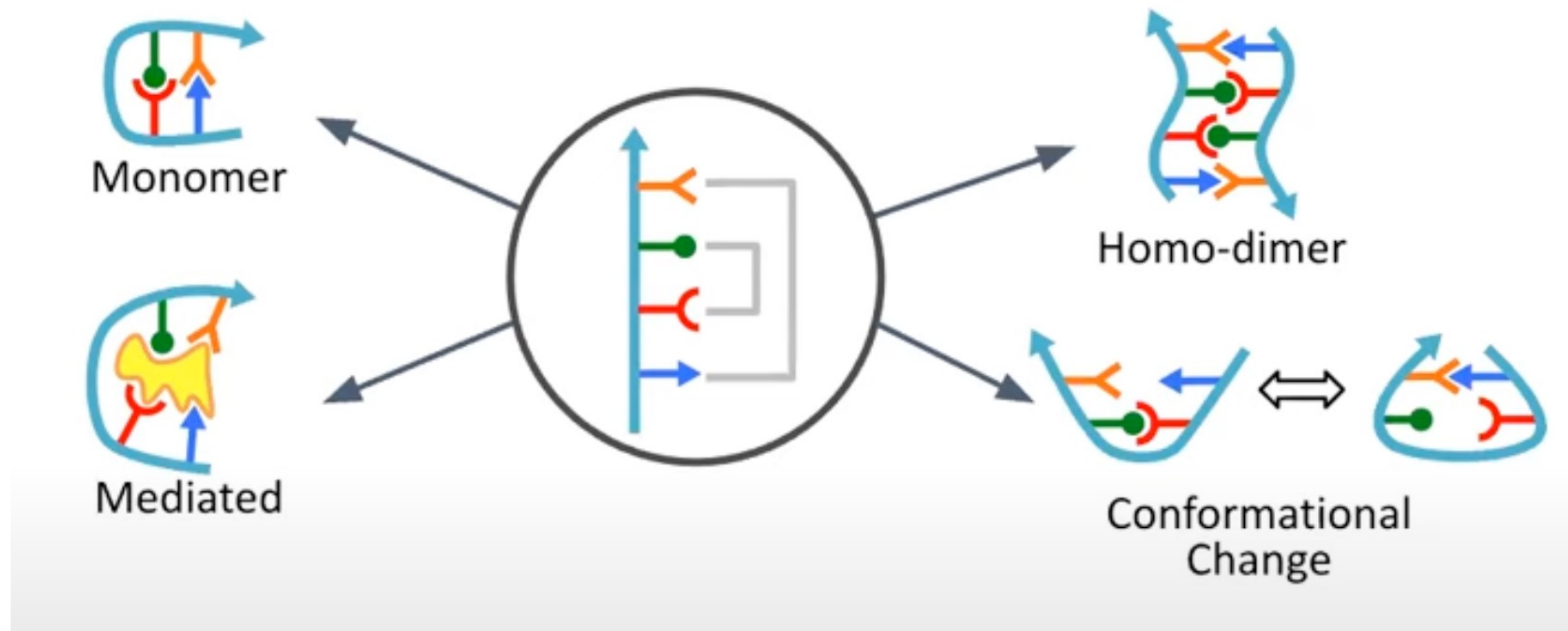
Evolution can give us hints about function

Coevolution: which residues interact?



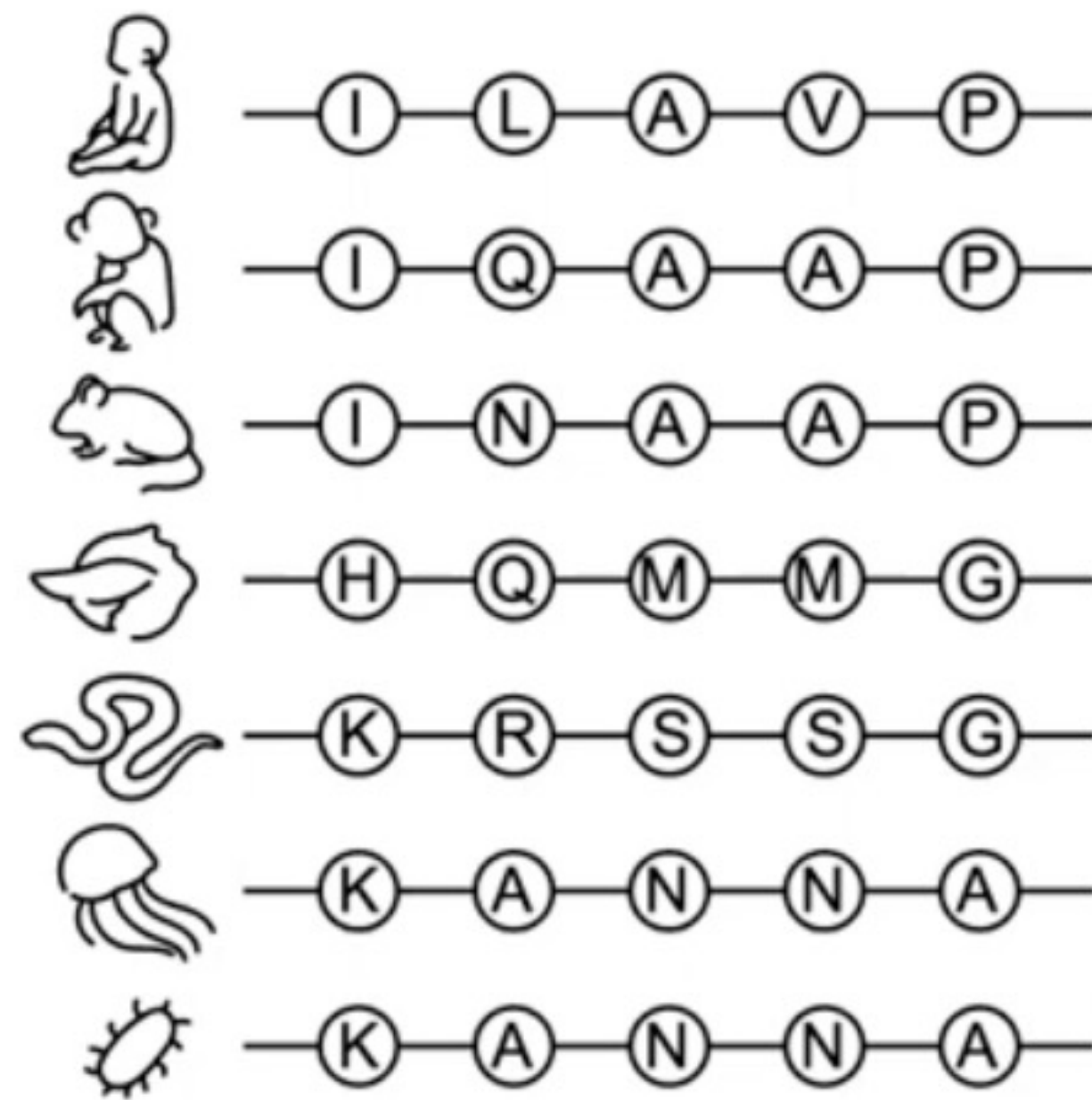
Evolution can give us hints about function

Coevolution: many types of interactions possible

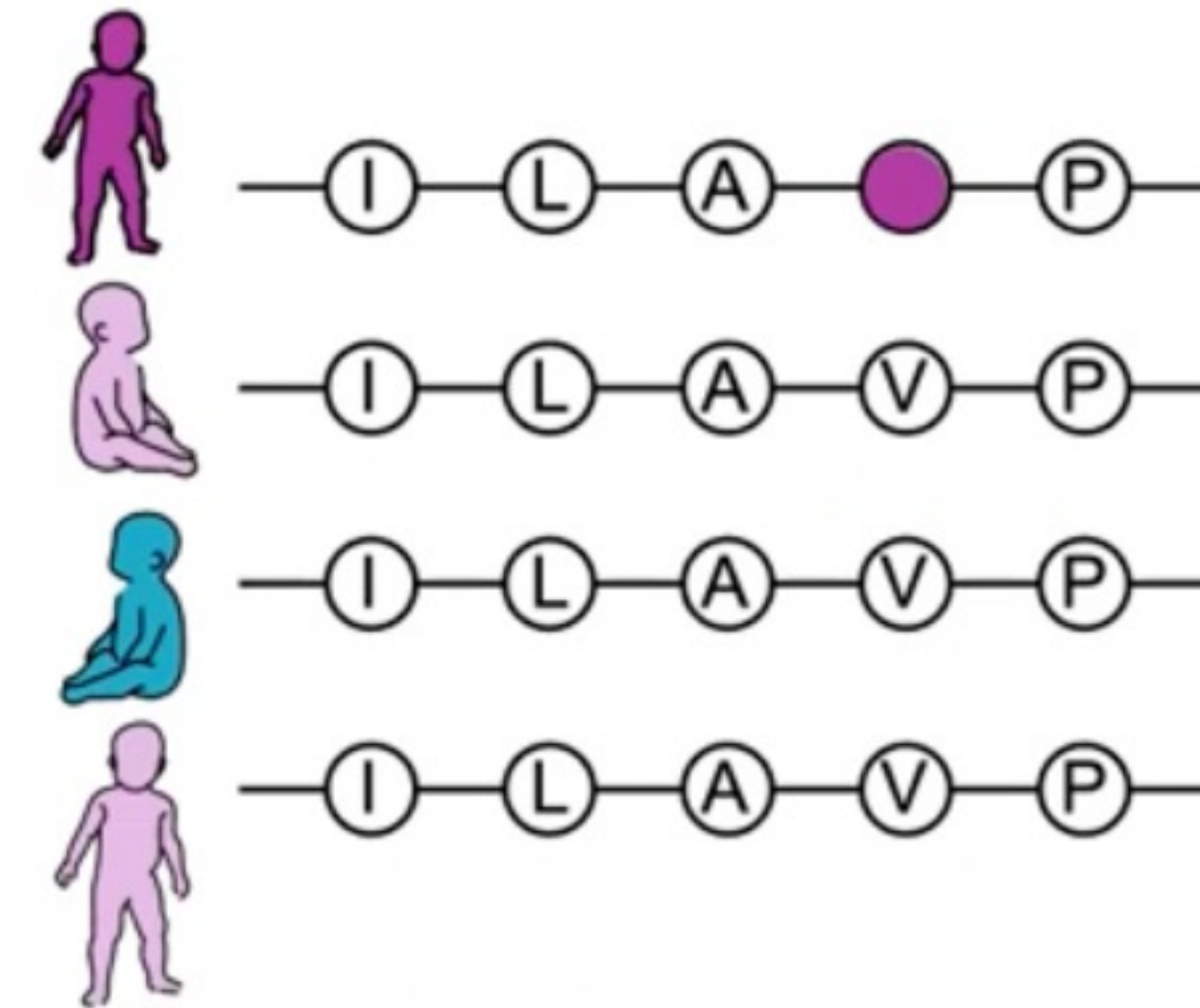


Fitness Prediction of Variants

How do we model this problem?



$$p(\mathbf{x}|\theta)$$

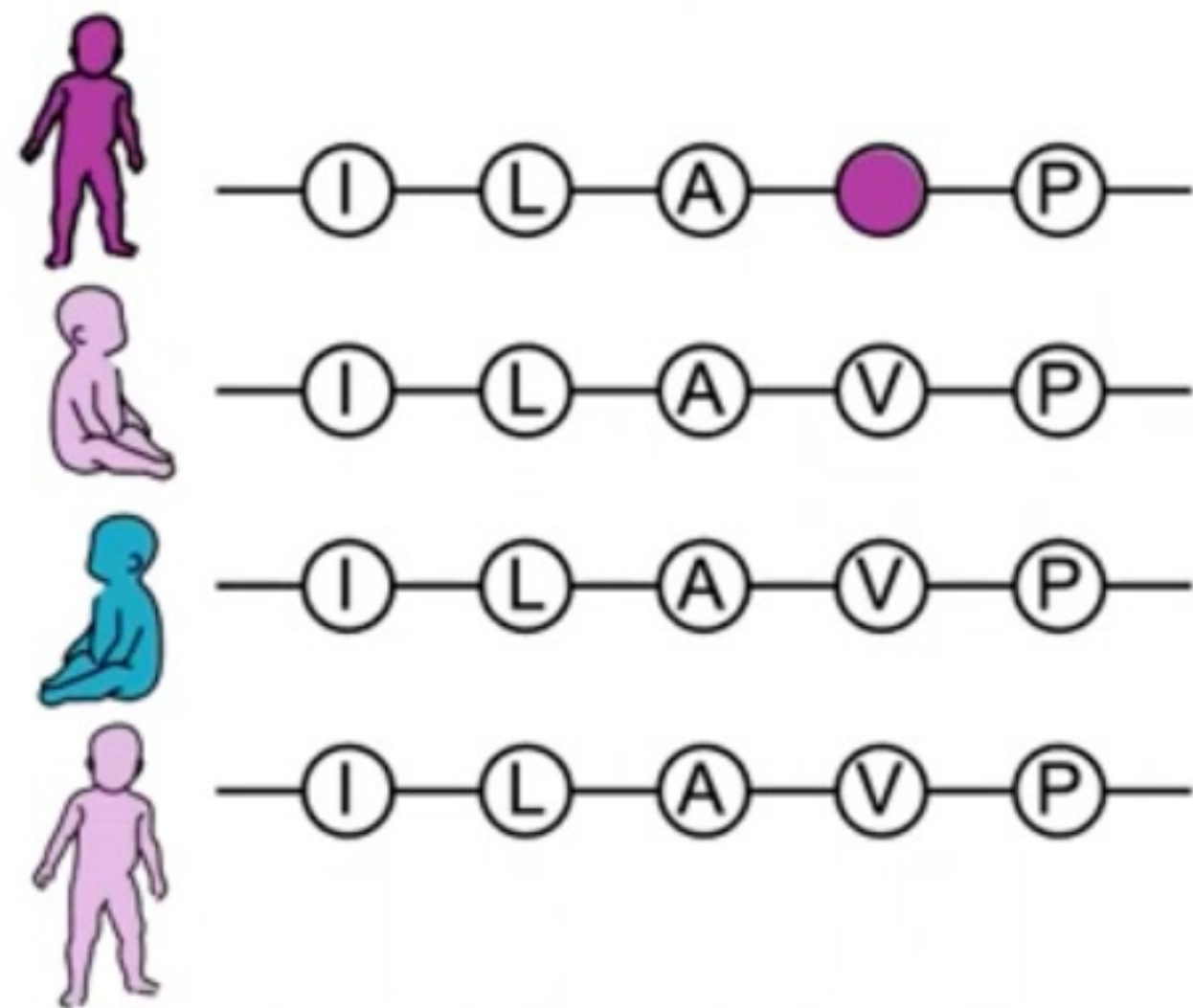


$$\log \left(\frac{p(\mathbf{x}_v|\theta)}{p(\mathbf{x}_{\text{ref}}|\theta)} \right)$$

less probable \rightarrow less fit

General Substitution Rules

Context-independent scores



$$\log \left(\frac{p(\mathbf{x}_v | \theta)}{p(\mathbf{x}_{\text{ref}} | \theta)} \right)$$

less probable \rightarrow less fit



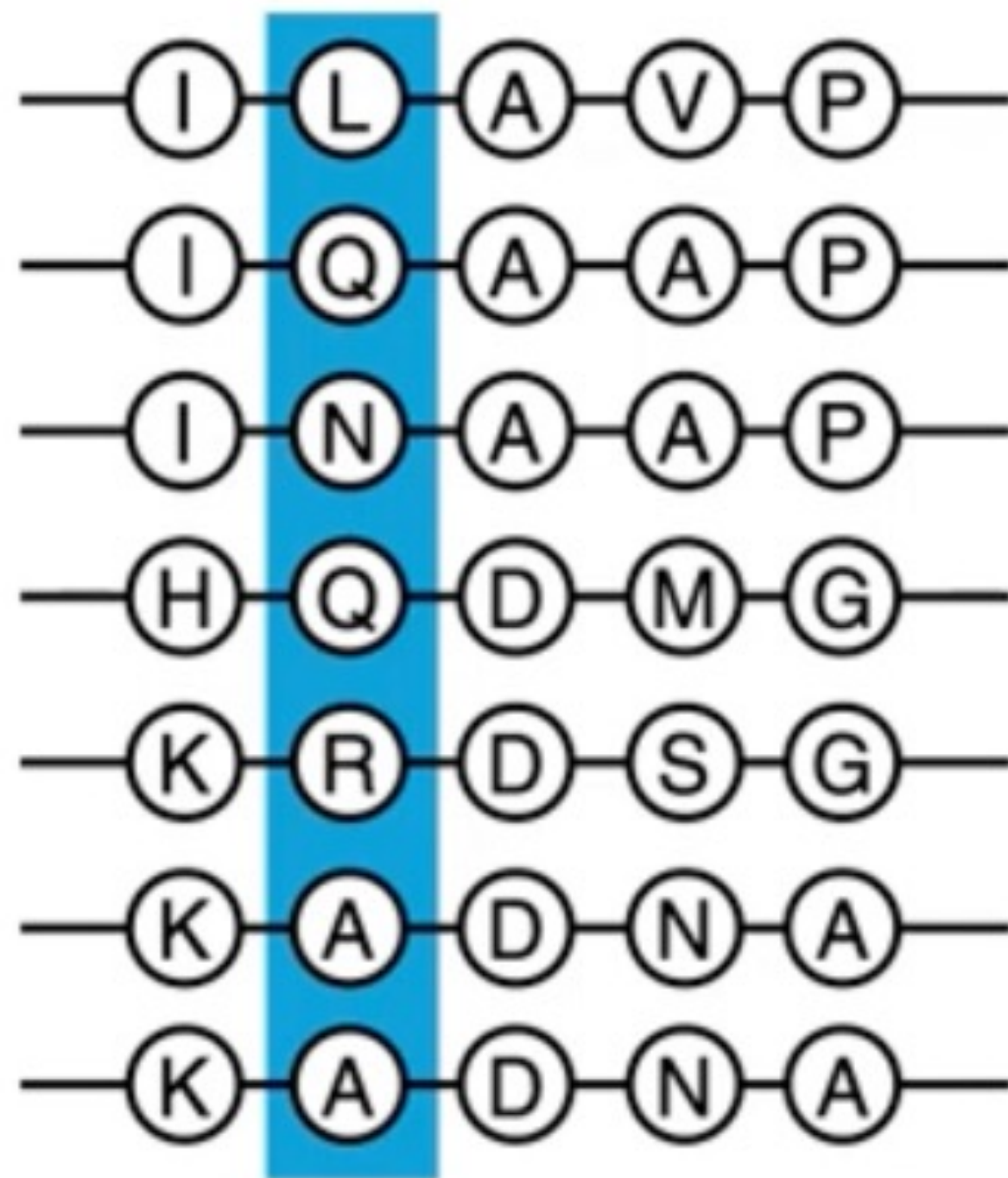
e.g., BLOSUM62

Henikoff and Henikoff, *PNAS*, 1992

Single-site models: Conservation

Site-independent scores based on single sites

column conservation



$$E(\mathbf{x}) = \sum_{ij} h_i^j x_i^j$$

indicator function

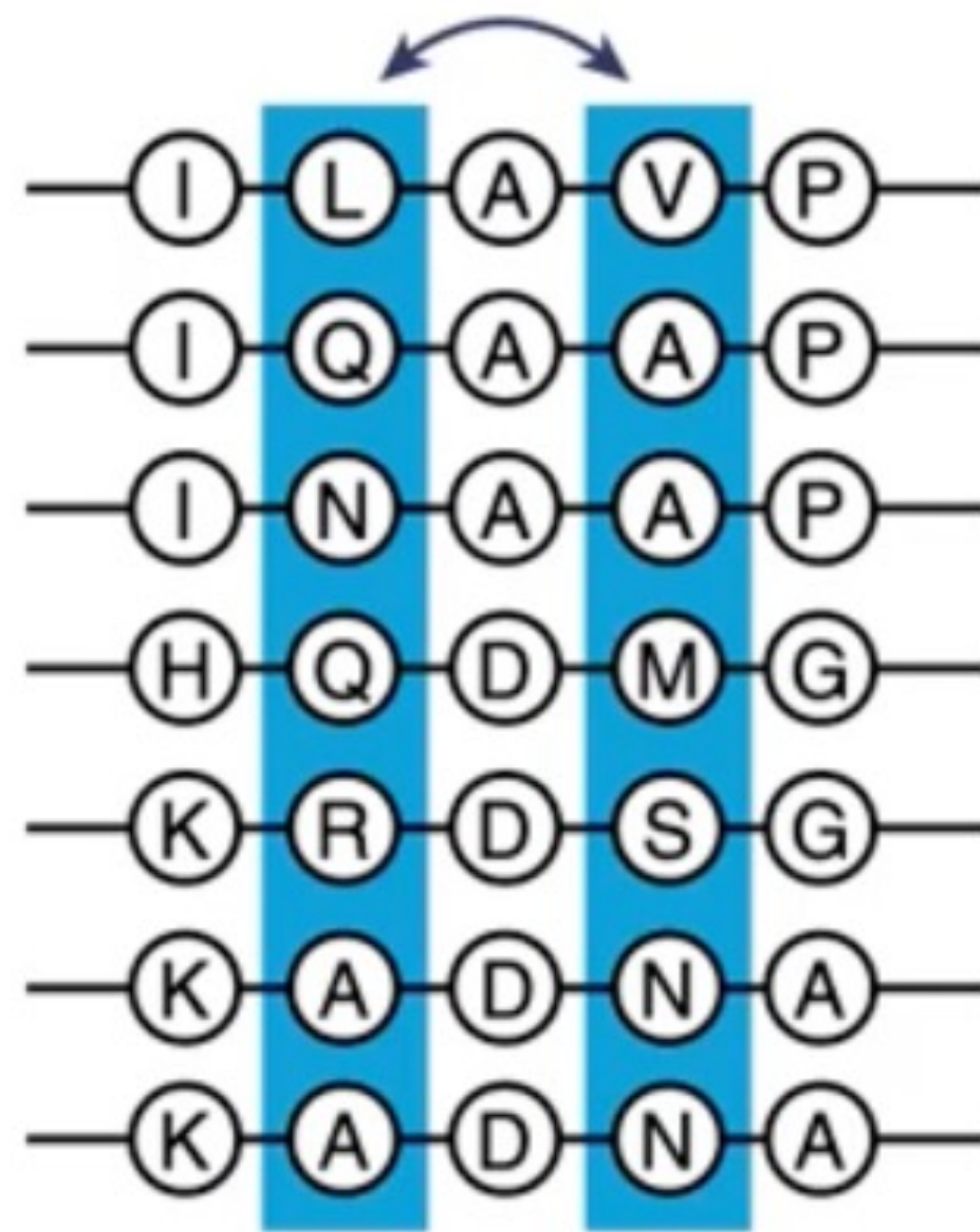
amino acid position

$$p(\mathbf{x}) = \frac{1}{Z} e^{E(\mathbf{x})}$$

Pairwise models: Coevolution

Interactions captured, but still site-independent

pairwise interactions
(Potts model)



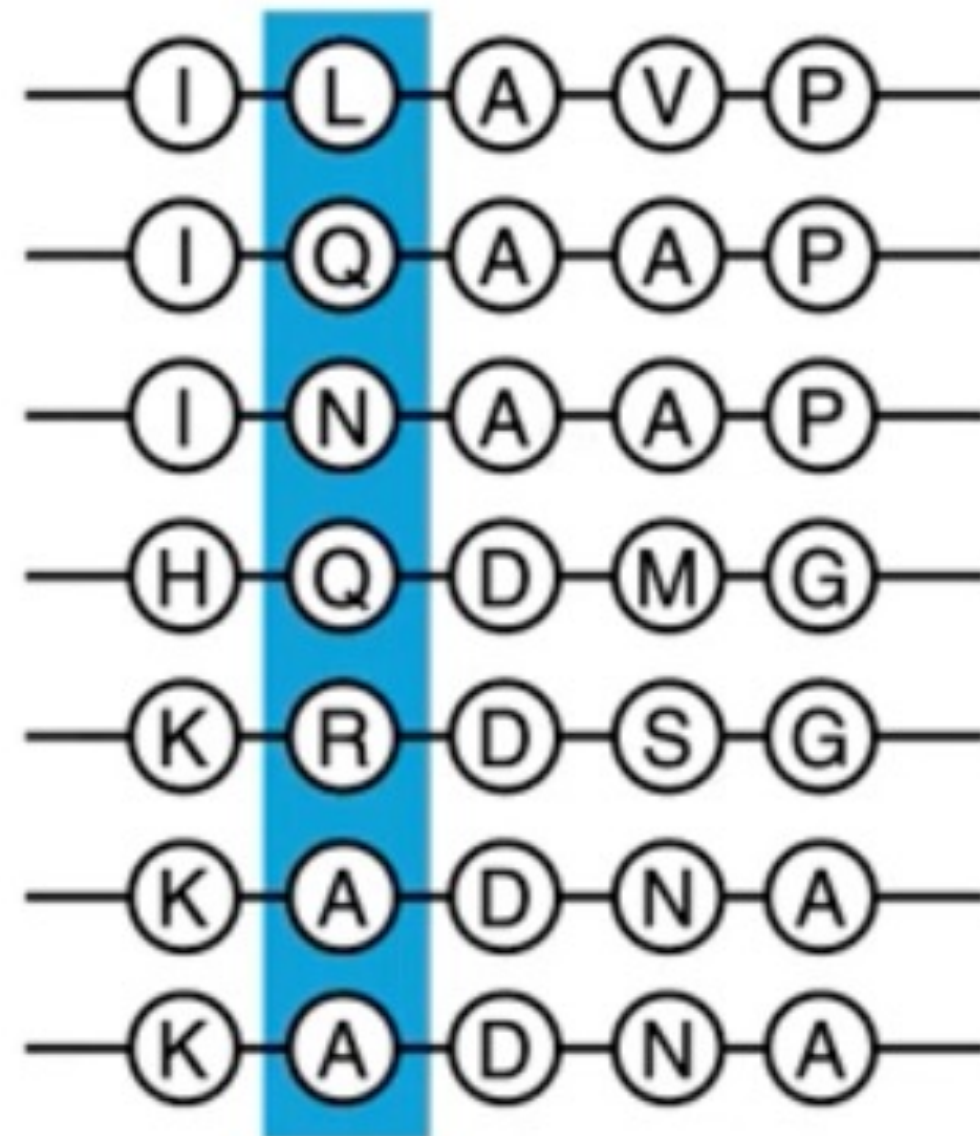
$$E(\mathbf{x}) = \sum_{ij} h_i^j x_i^j + \sum_{ijkl} J_{ij}^{kl} x_i^k x_j^l$$

$$p(\mathbf{x}) = \frac{1}{Z} e^{E(\mathbf{x})}$$

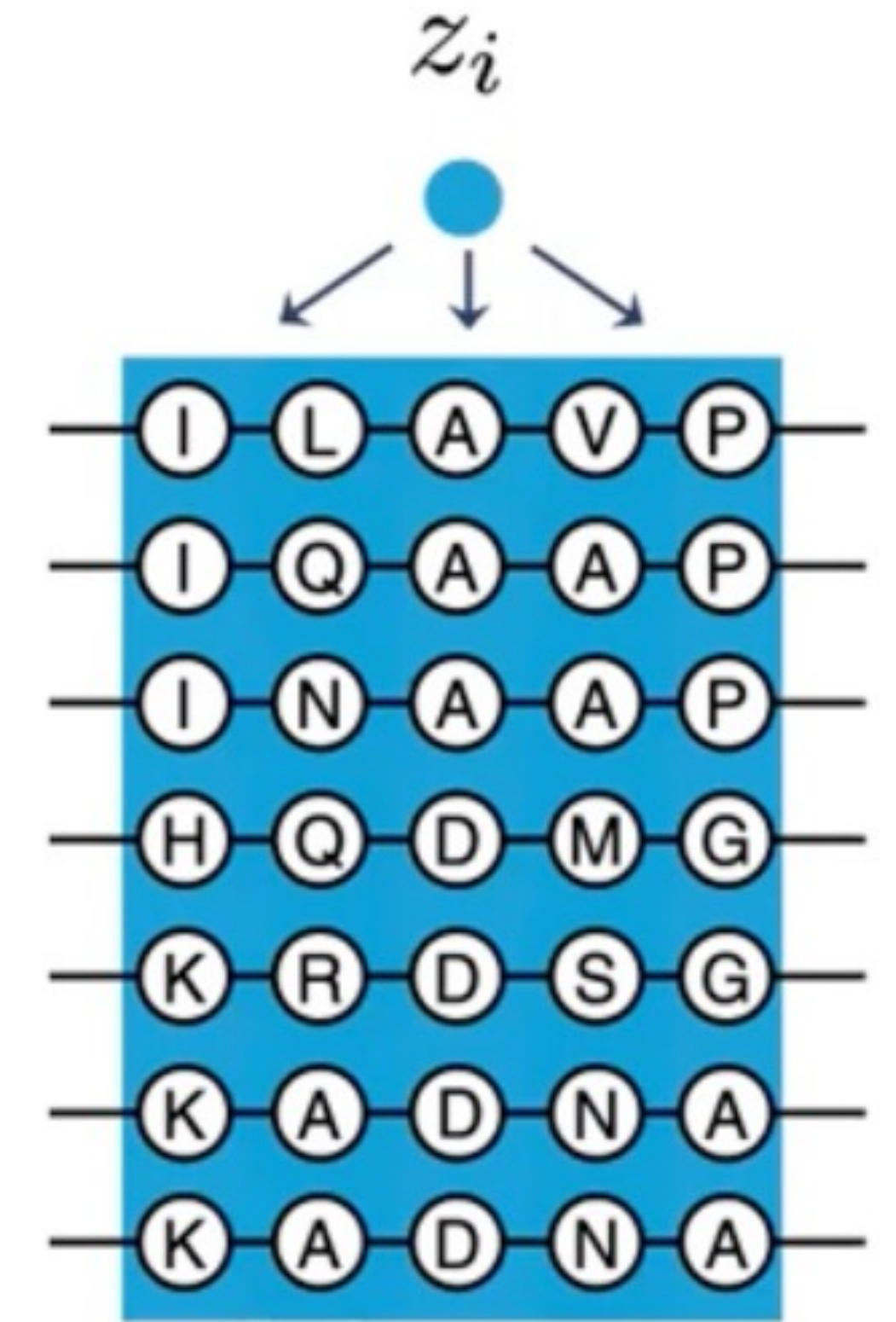
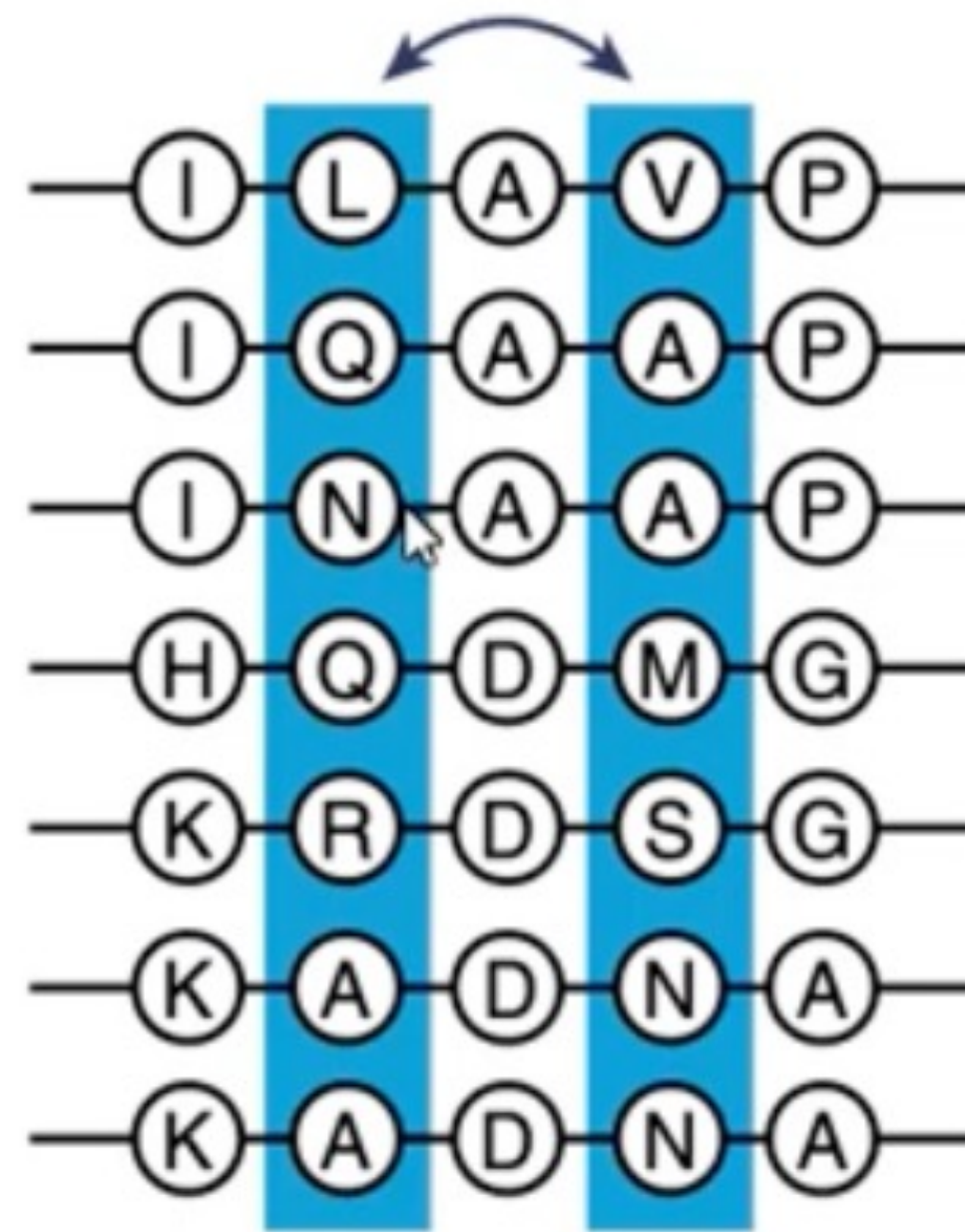
Higher-order models

Taking context into account

column conservation

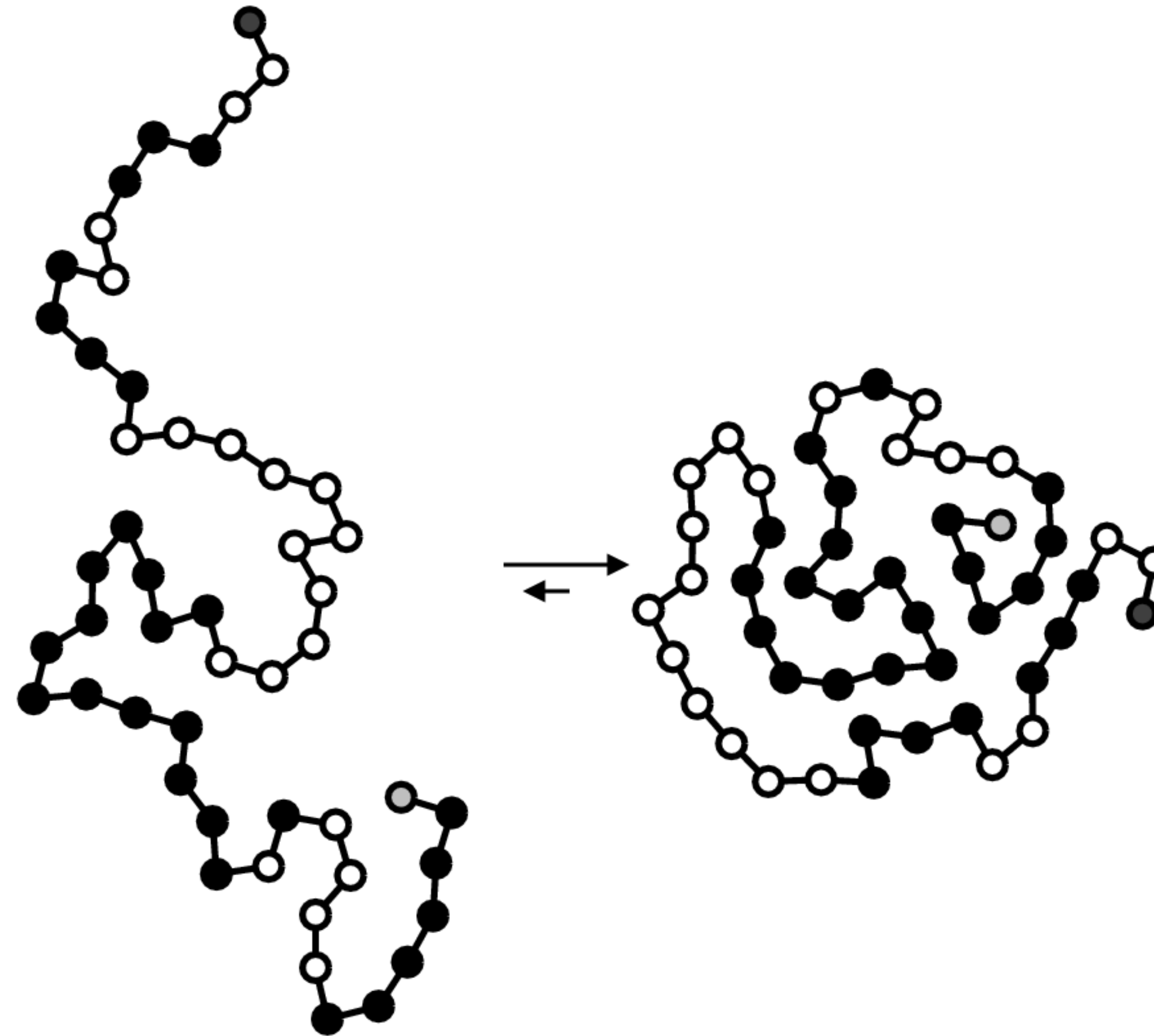


pairwise interactions
(Potts model)



Context is everything

Functions can vary depending on the local environment



2. Language Modelling

Evolution of language modelling

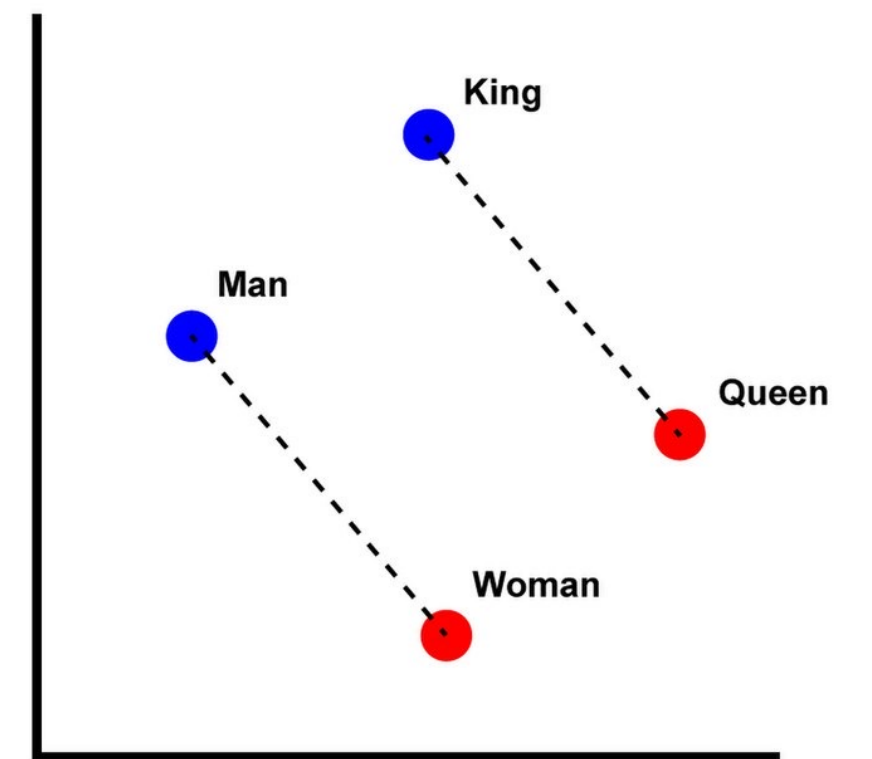
A similar story: we start with site-independent models

And the king...

$[0.2, 0.03, -0.4, \dots]$

However, the queen...

$[0.1, -0.51, 0.2, \dots]$



Evolution of language modelling

A similar story: site-independent models (Word2Vec, GloVe, ...)

Along the river **bank** ...

[0.1 0.03, -0.5, ...]



The **bank** robber ...

[0.1, 0.03, -0.5, ...]



?

Evolution of language modelling

Solution: contextual representations

Along the river **bank** ...

[0.1 0.03, -0.5, ...]

The **bank** robber ...

[0.2, -0.35, 0.4, ...]

?

Evolution of language modelling

Solution: contextual representations

Semi-supervised Sequence Learning

Andrew M. Dai
Google Inc.
adai@google.com

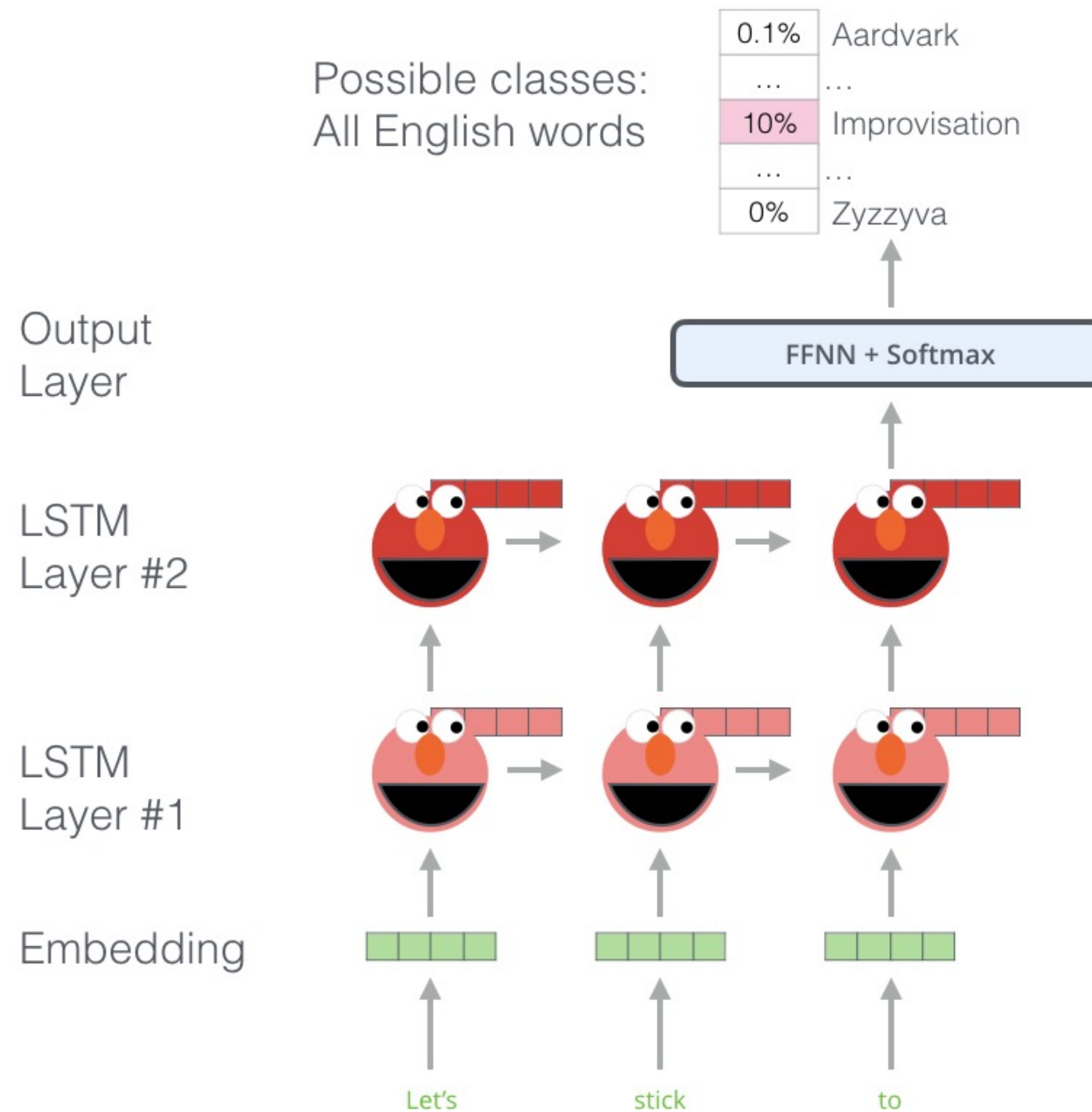
Quoc V. Le
Google Inc.
qvl@google.com

Abstract

We present two approaches that use unlabeled data to improve sequence learning with recurrent networks. The first approach is to predict what comes next in a sequence, which is a conventional language model in natural language processing. The second approach is to use a sequence autoencoder, which reads the input sequence into a vector and predicts the input sequence again. These two algorithms can be used as a “pretraining” step for a later supervised sequence learning algorithm. In other words, the parameters obtained from the unsupervised step can be used as a starting point for other supervised training models.

ELMo: next-word prediction

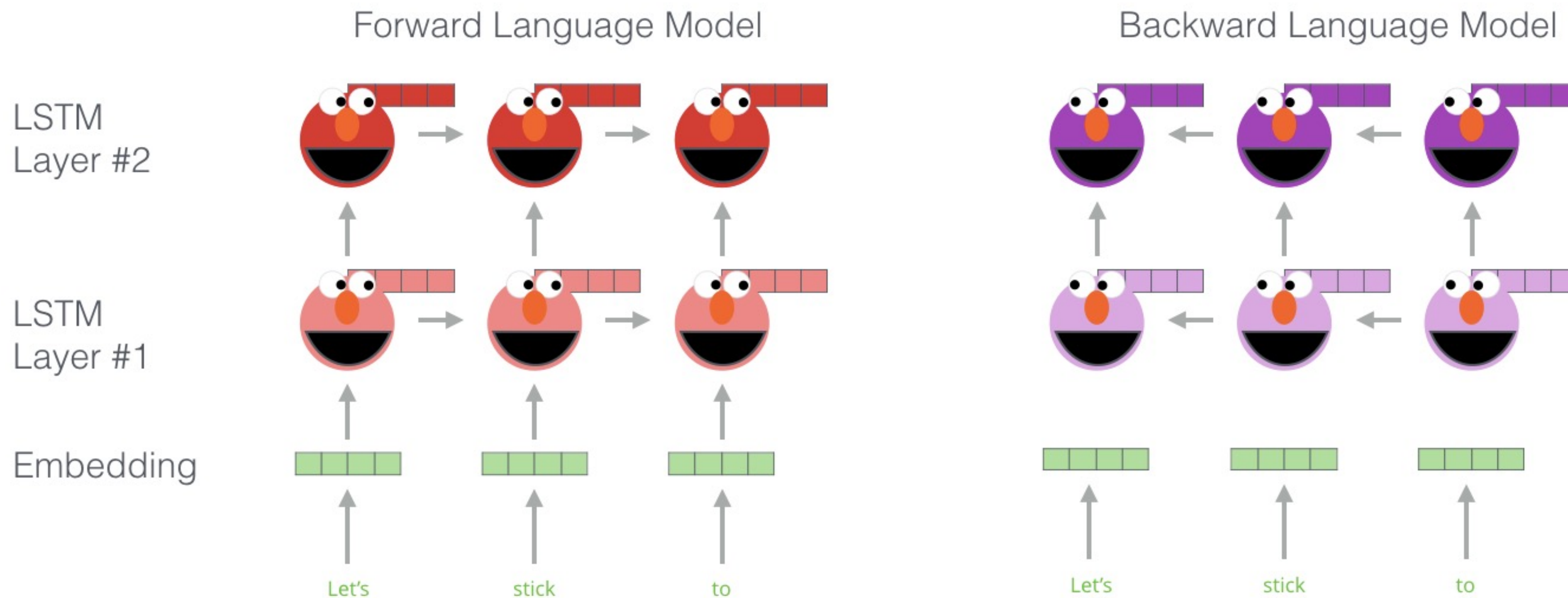
Solution: contextual representations



ELMo: forward and backward

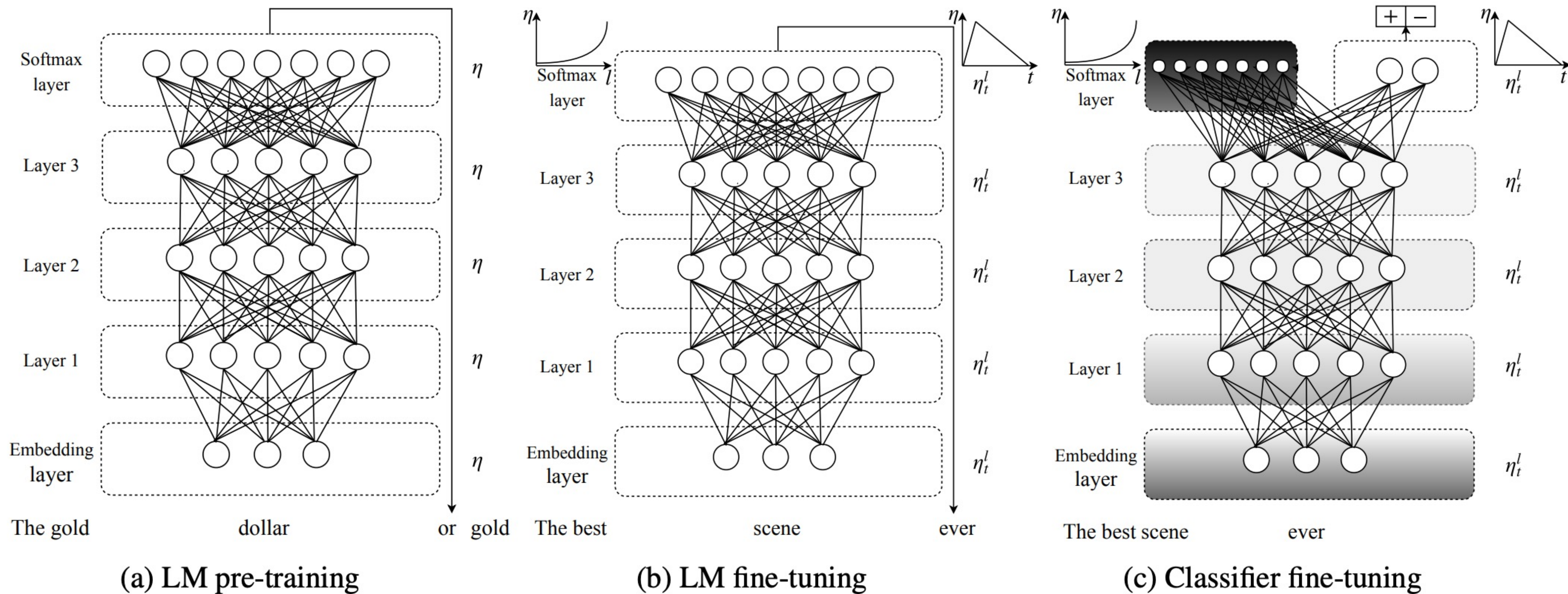
Look at text from both ways

Embedding of “stick” in “Let’s stick to” - Step #1



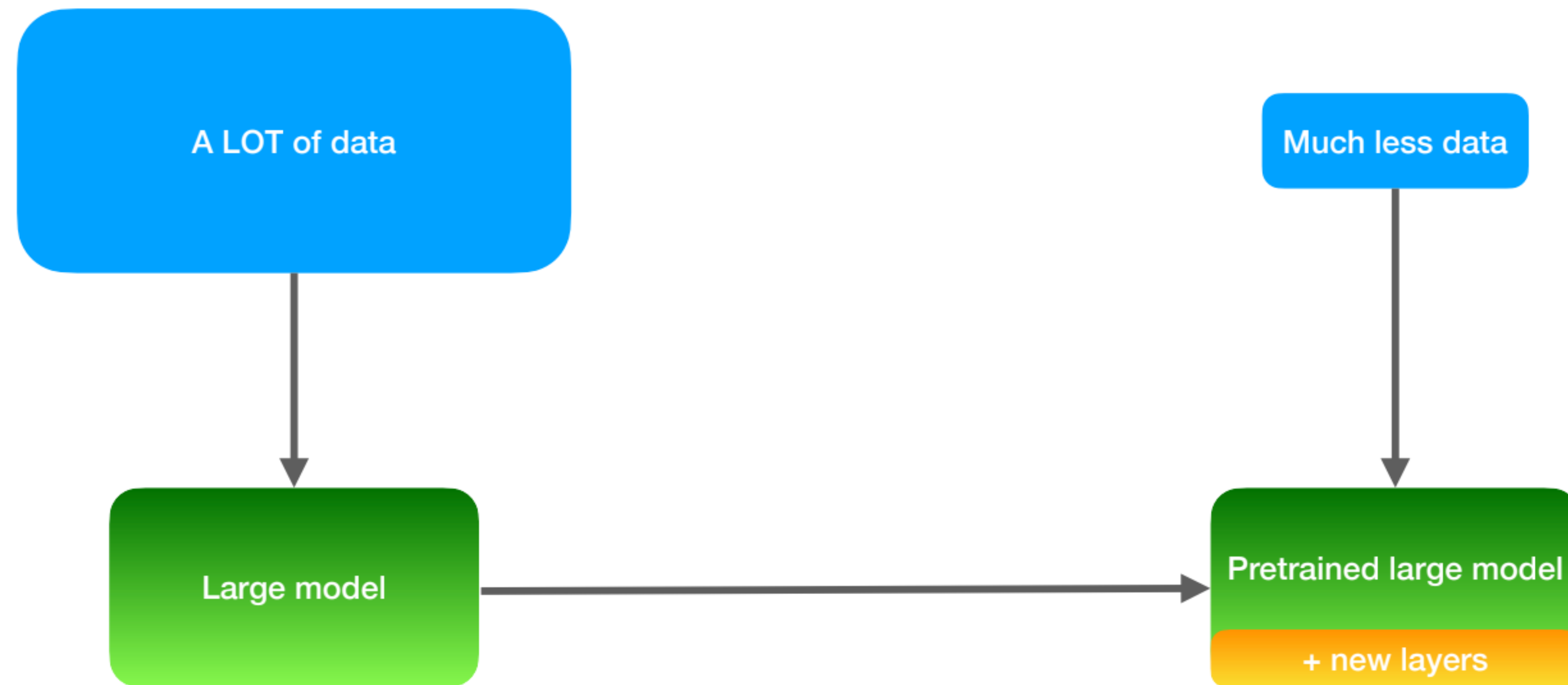
ULM-FiT: how to do transfer learning in NLP

Use more than embeddings: finetuning for transfer learning



Transformers are good at Transfer Learning

Use unlabeled data to get better on specific tasks

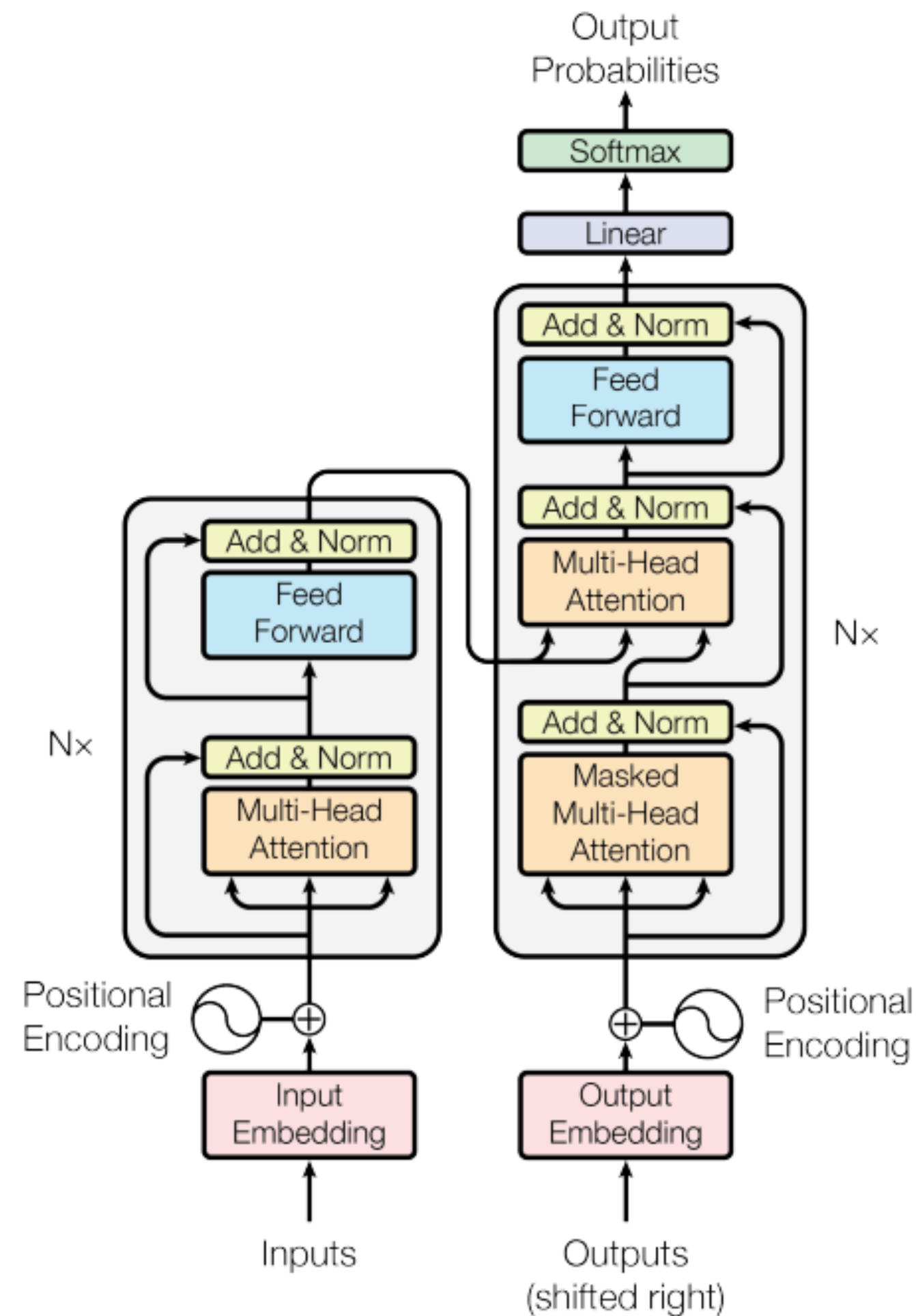


Traditional Machine Learning:
slow training on a lot of data

Transfer learning:
fast training on a little data

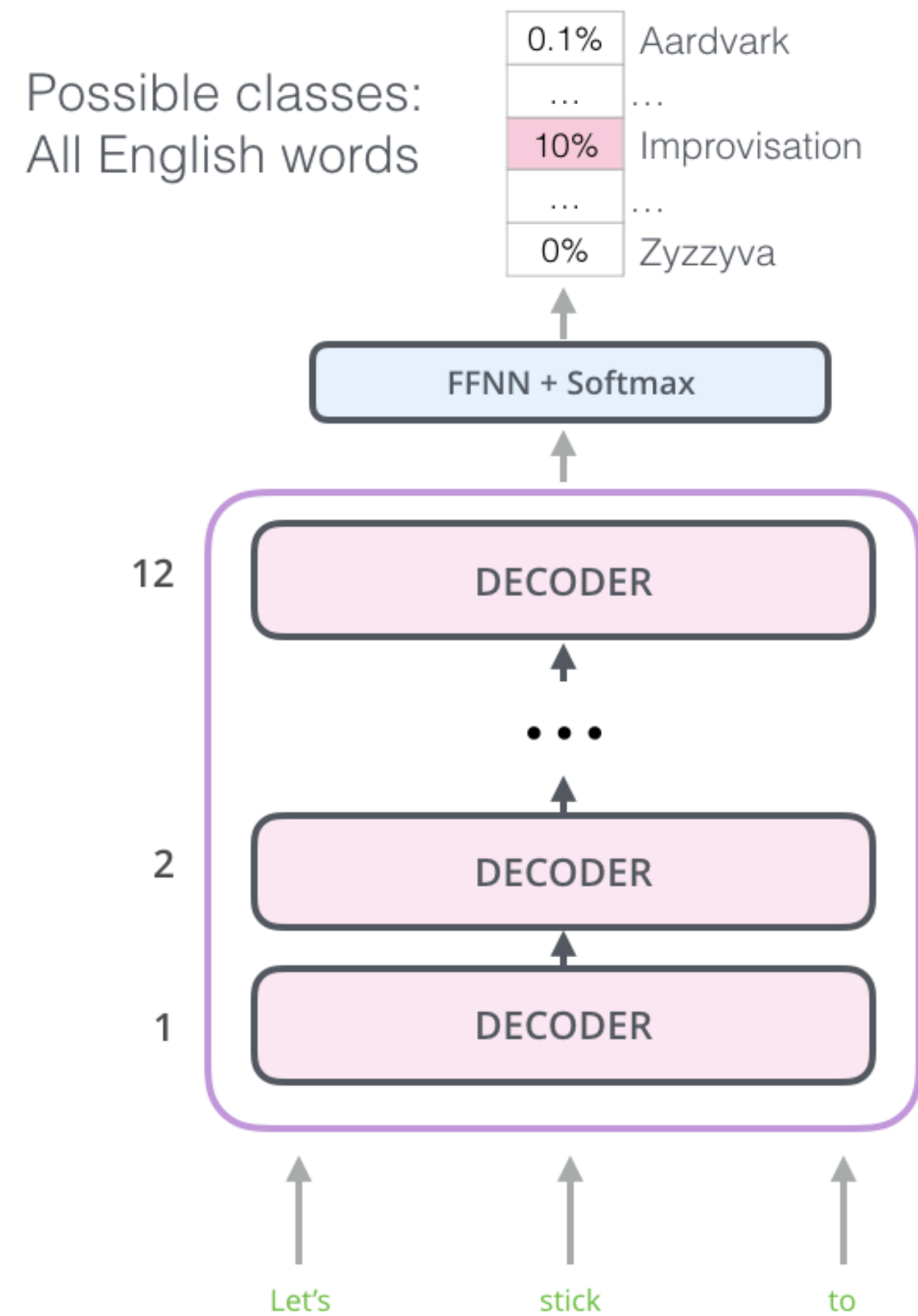
OpenAI Transformer

Train Deep Transformer LM and fine-tune on final task



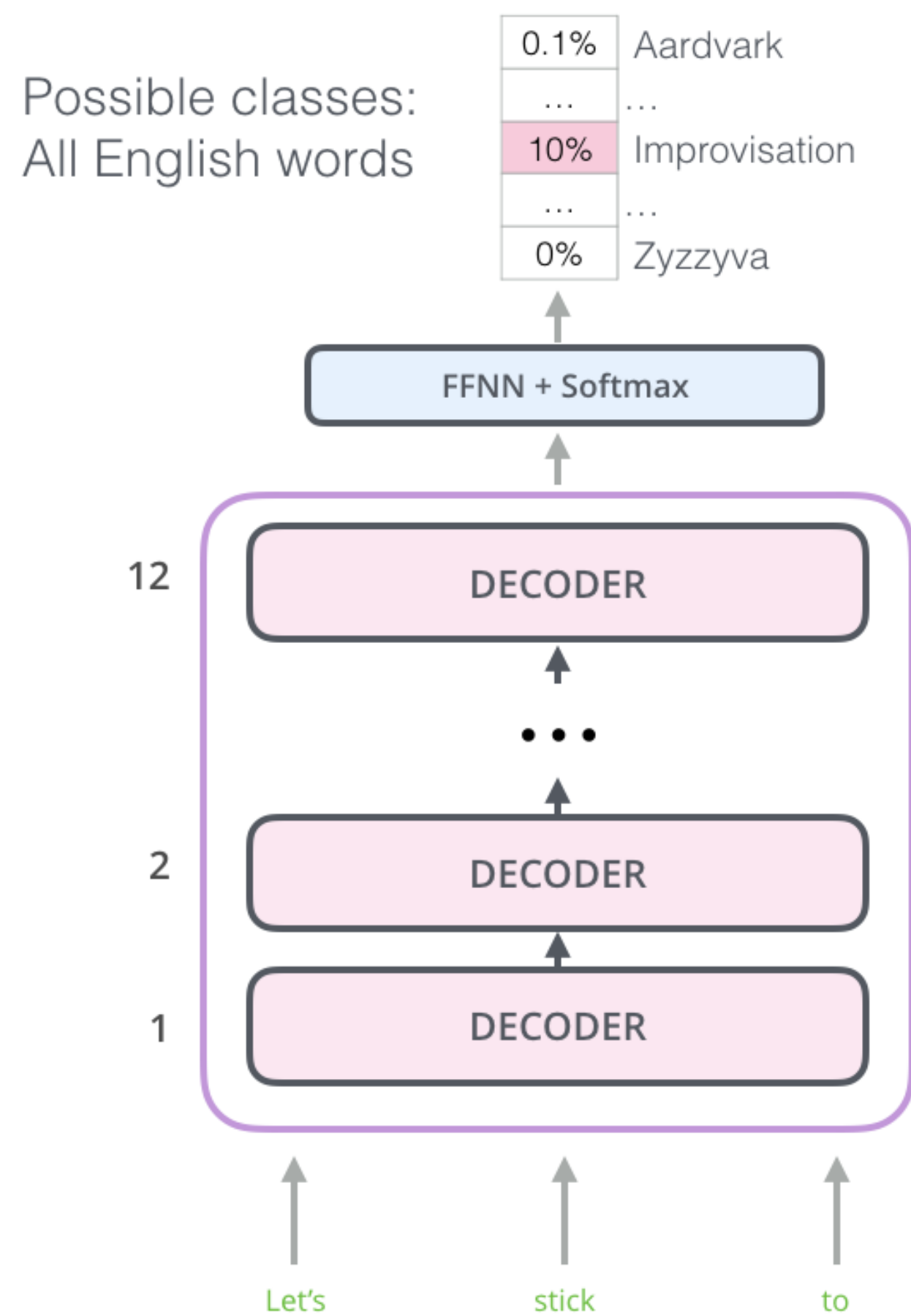
OpenAI Transformer

Train Deep Transformer LM and fine-tune on final task

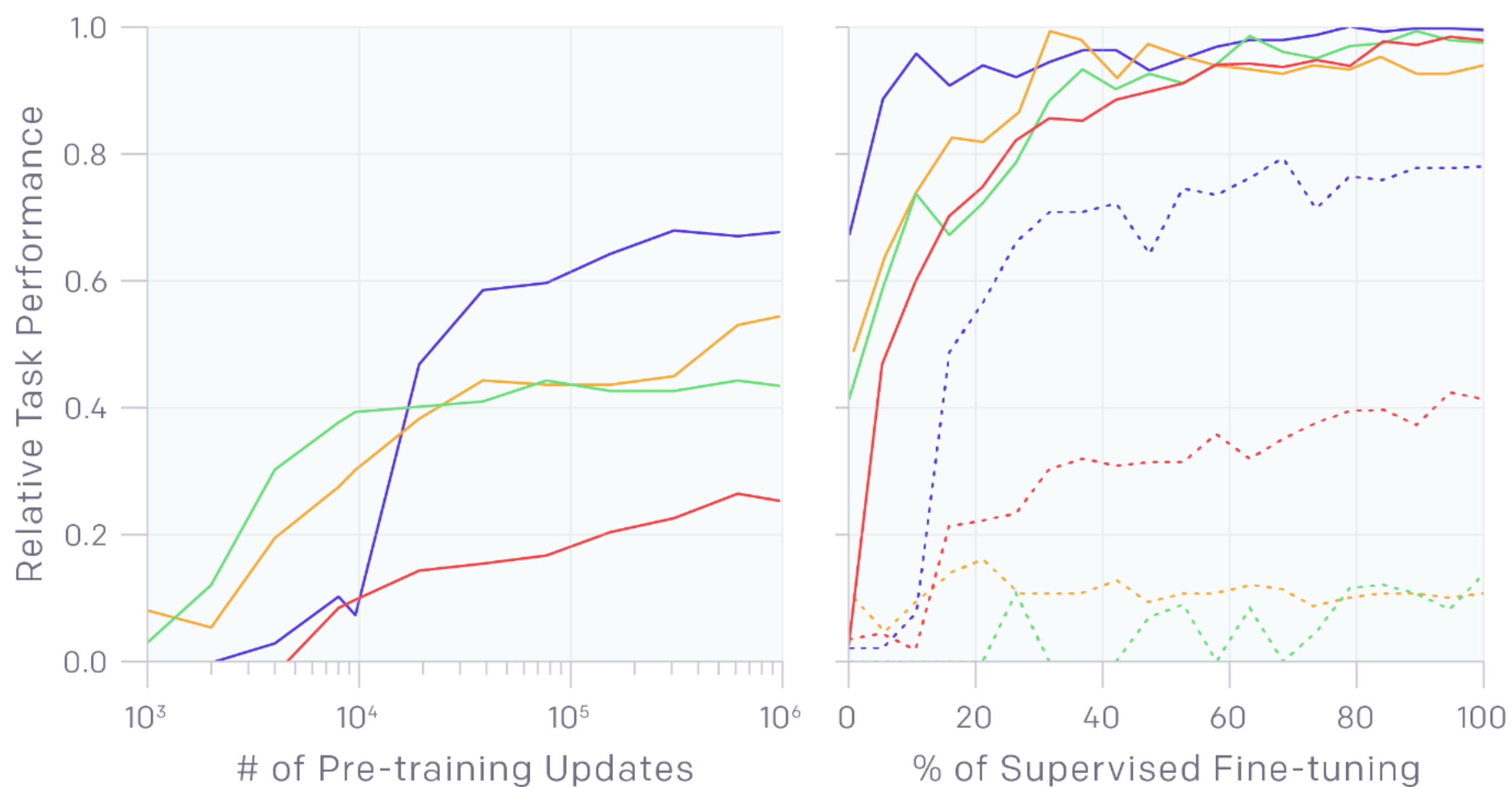


OpenAI Transformer

Train Deep Transformer LM and fine-tune on final task



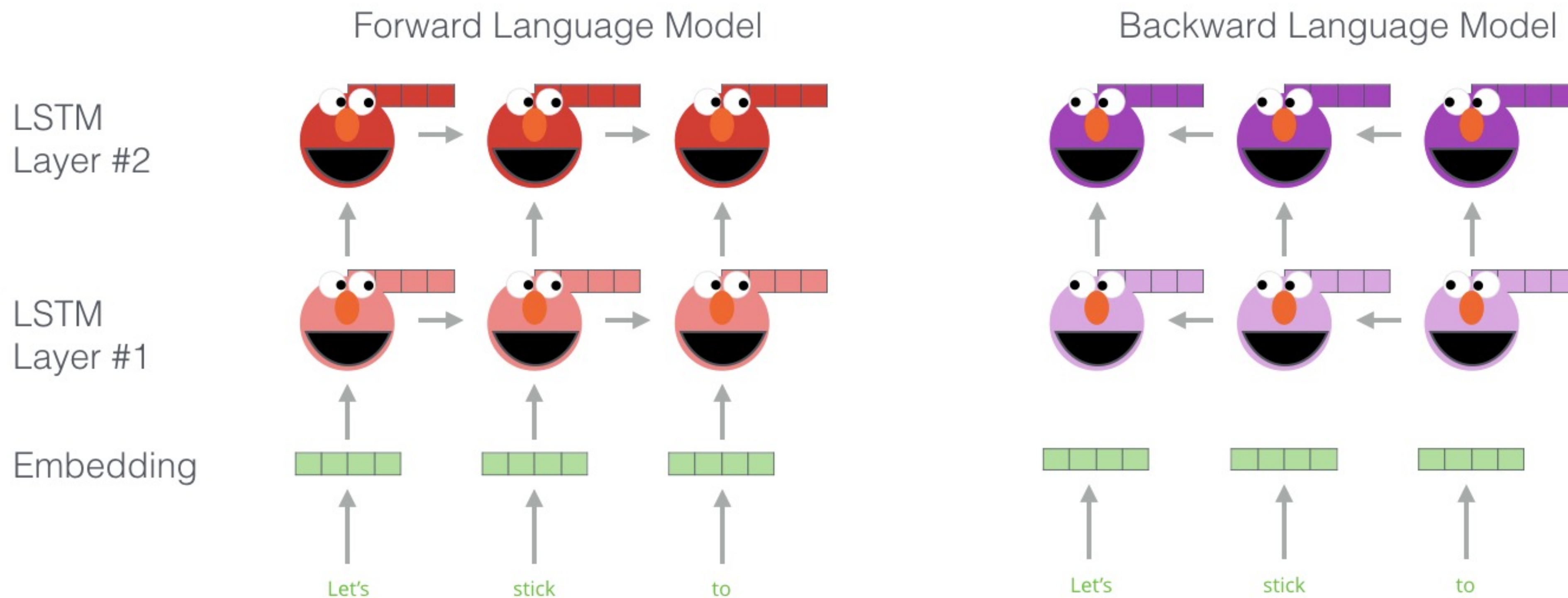
Zero-shot Transfer Can Directly Accelerate Supervised Fine-tuning



Problem: Decoder only

How does our model learn to think backward?

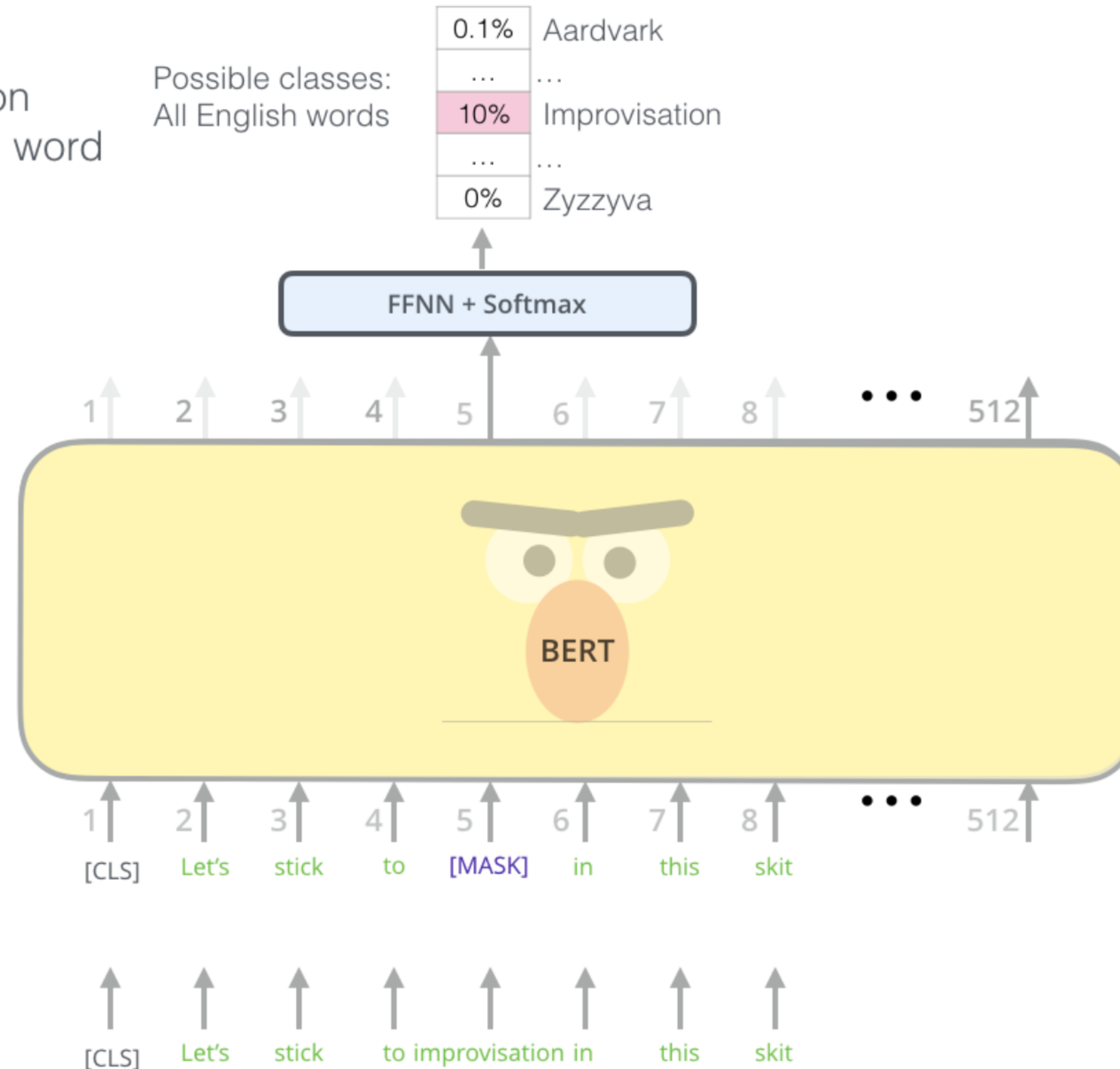
Embedding of “stick” in “Let’s stick to” - Step #1



BERT

Just use encoders and mask random tokens

Use the output of the masked word's position to predict the masked word



Randomly mask 15% of tokens

Input

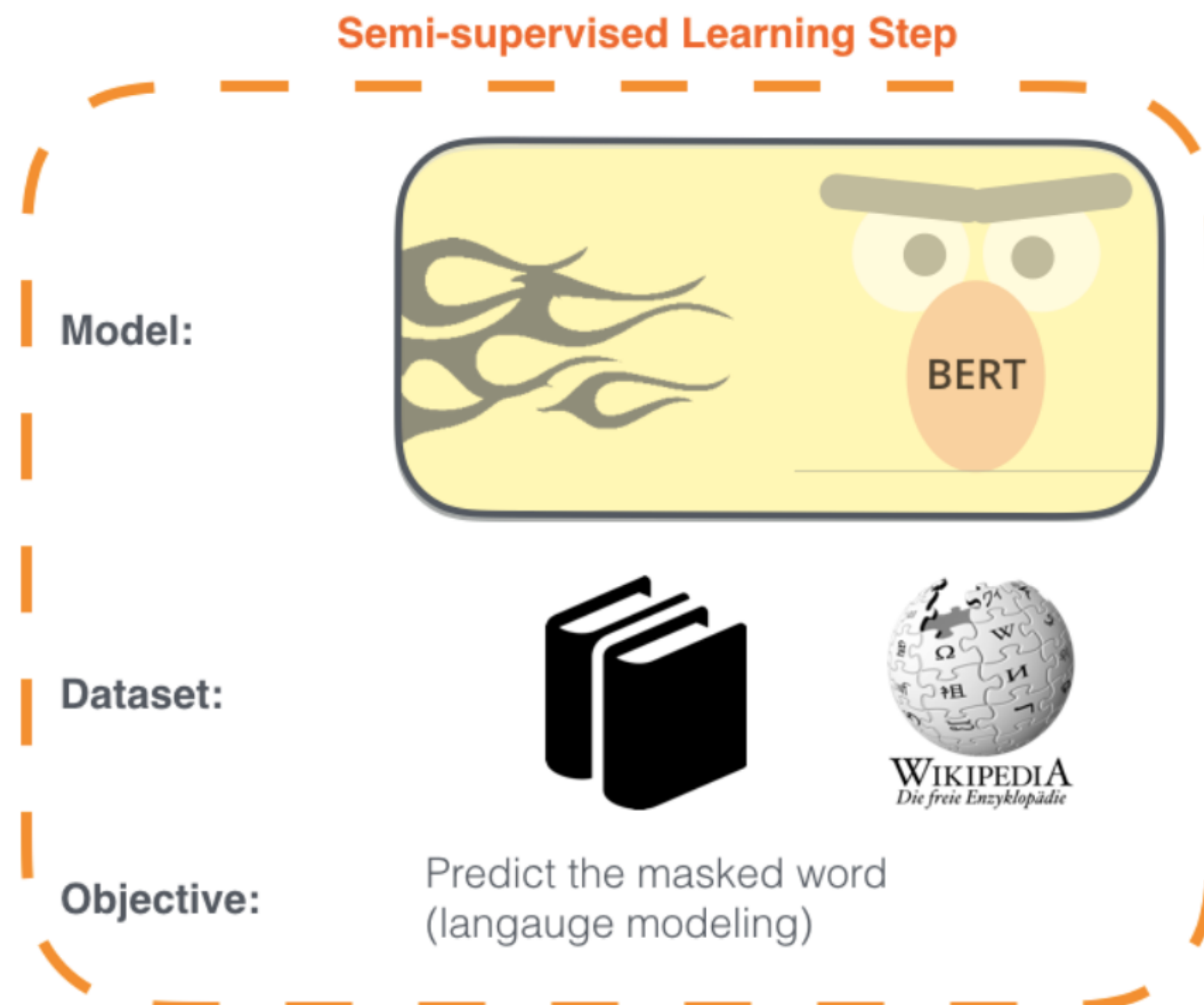
BERT's clever language modeling task masks 15% of words in the input and asks the model to predict the missing word.

Language Models

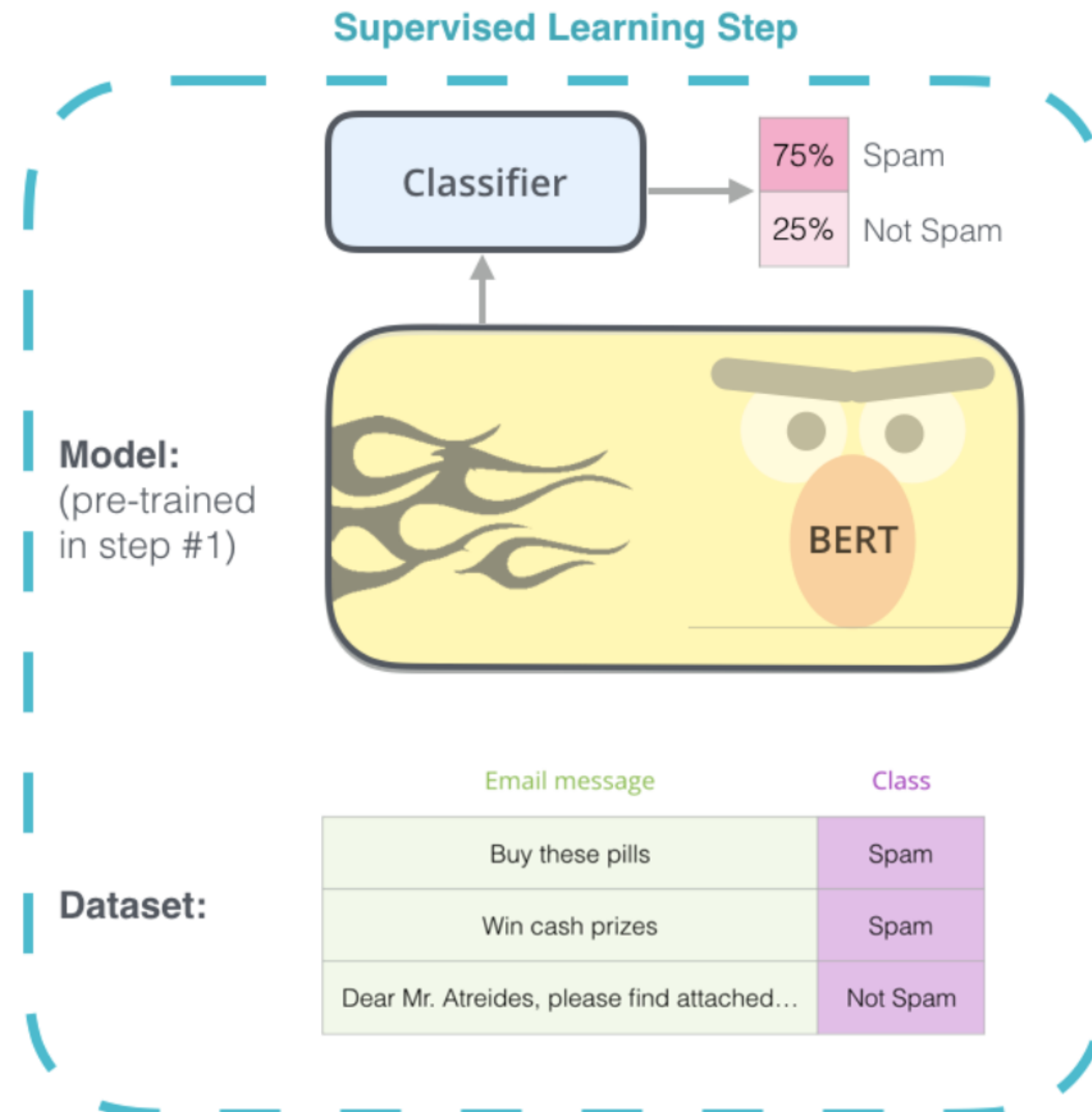
Bigger = Better ?

1 - **Semi-supervised** training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.



2 - **Supervised** training on a specific task with a labeled dataset.



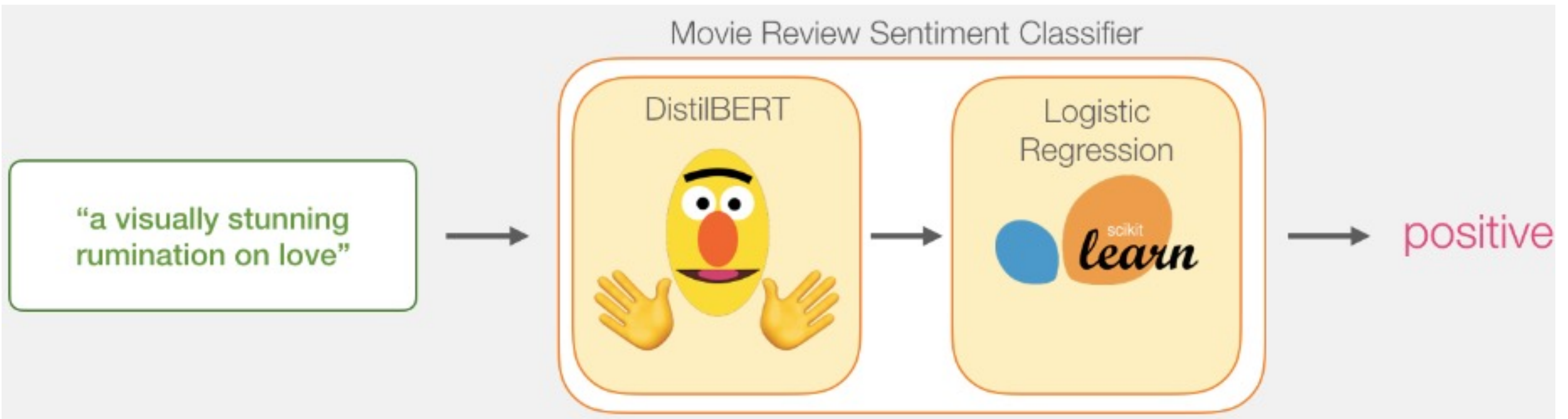
Transformers are good at Transfer Learning

Pre-Training improves downstream performance



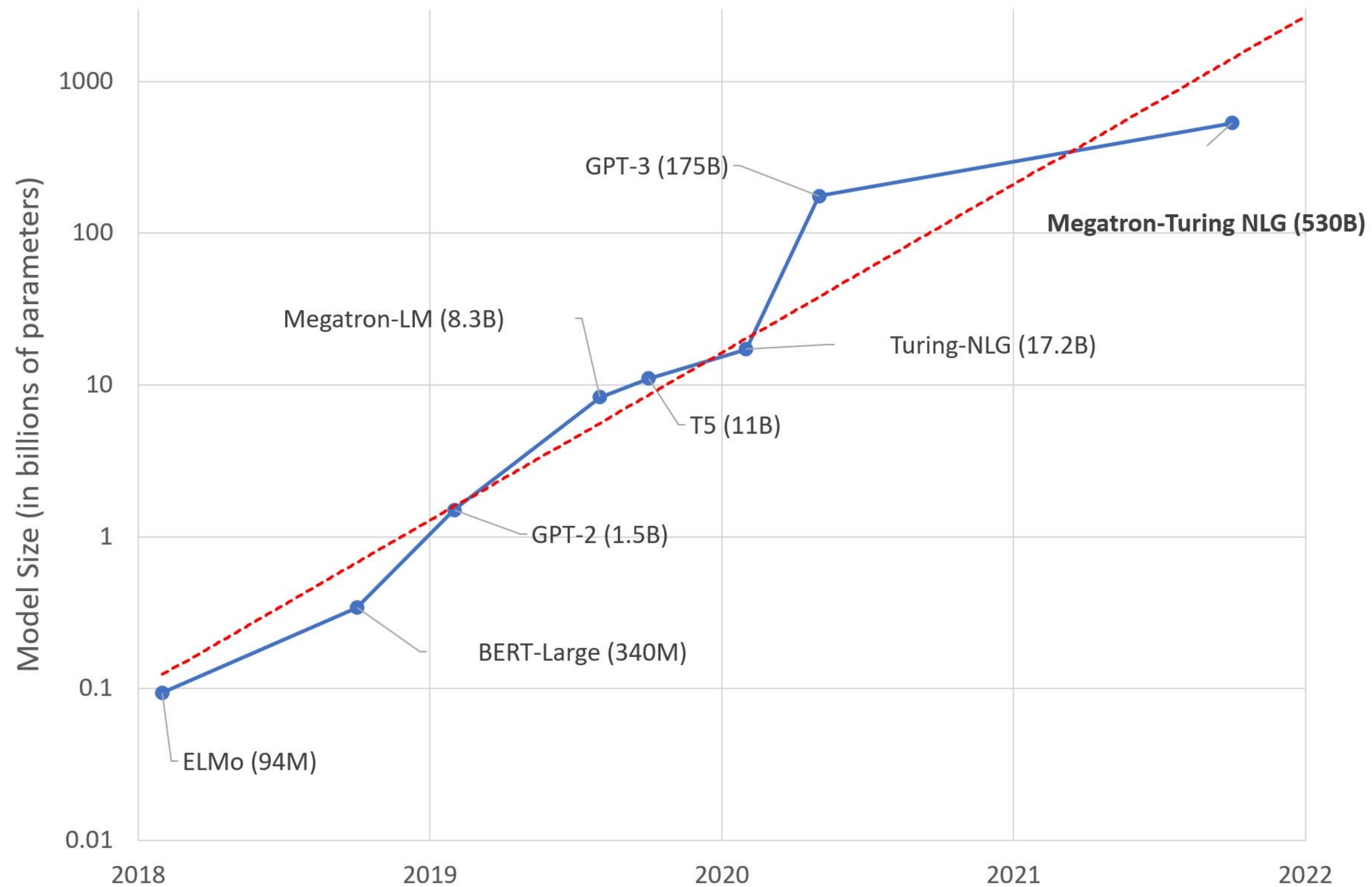
Transformers are good at Transfer Learning

Pre-Training improves downstream performance



Language Models

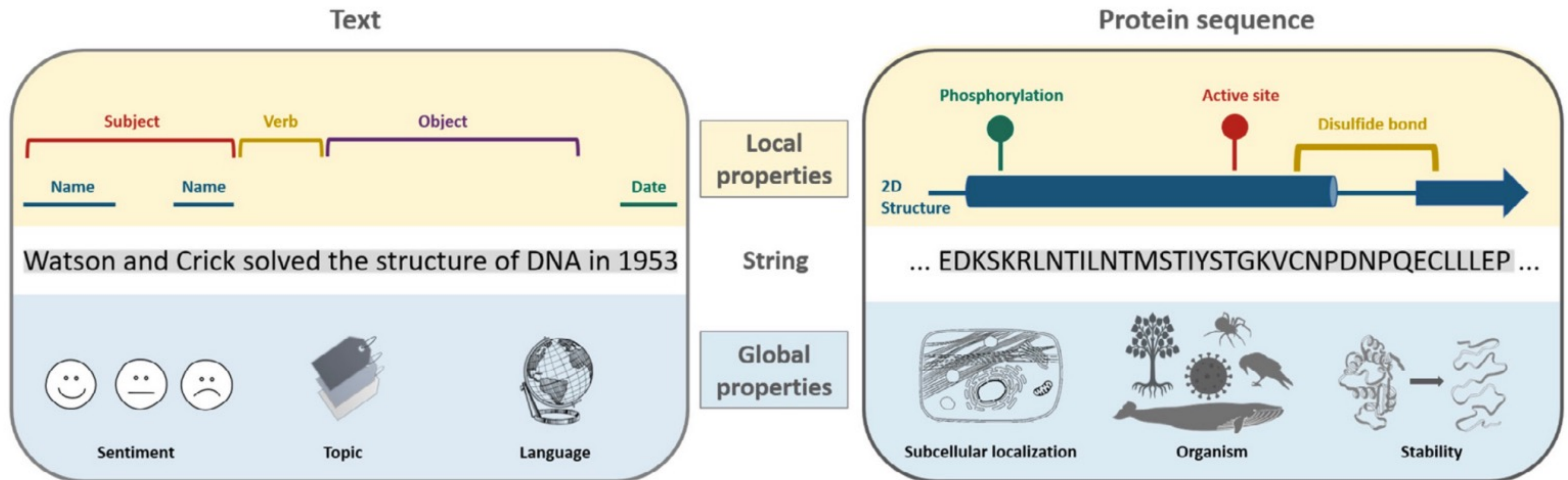
Bigger = Better ?



3. Protein Linguistics: Language Models in Biology

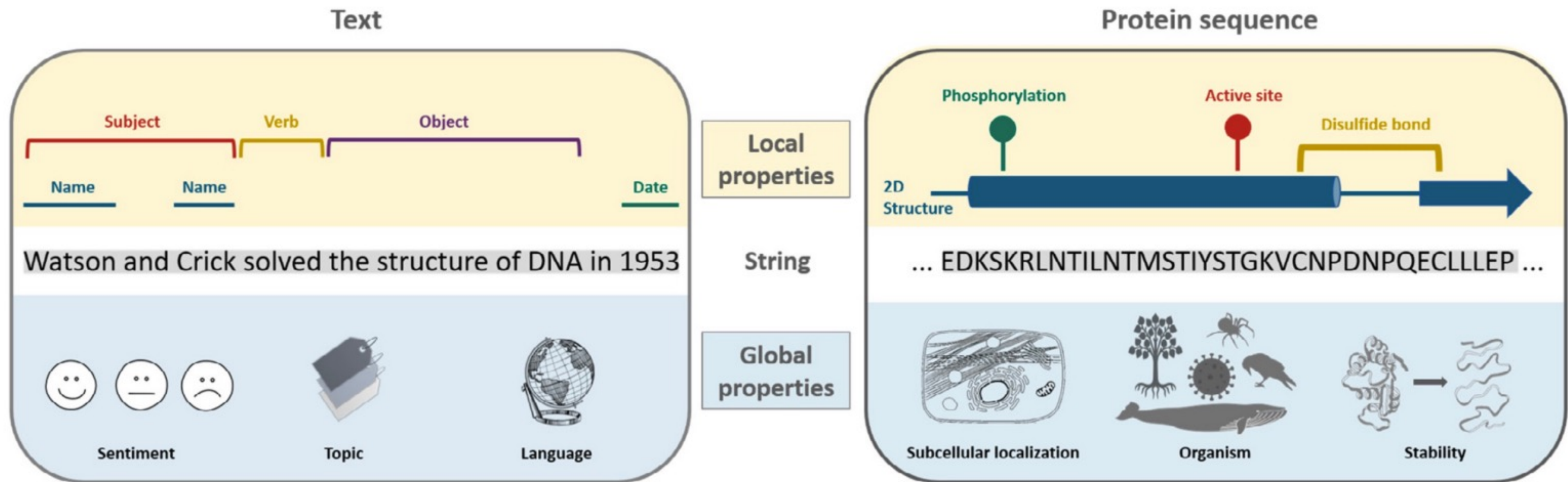
Proteins vs Sentences: The same?

Similar, but also important differences



Proteins vs Sentences: The same?

Similar, but also important differences



Can you *read* a protein?

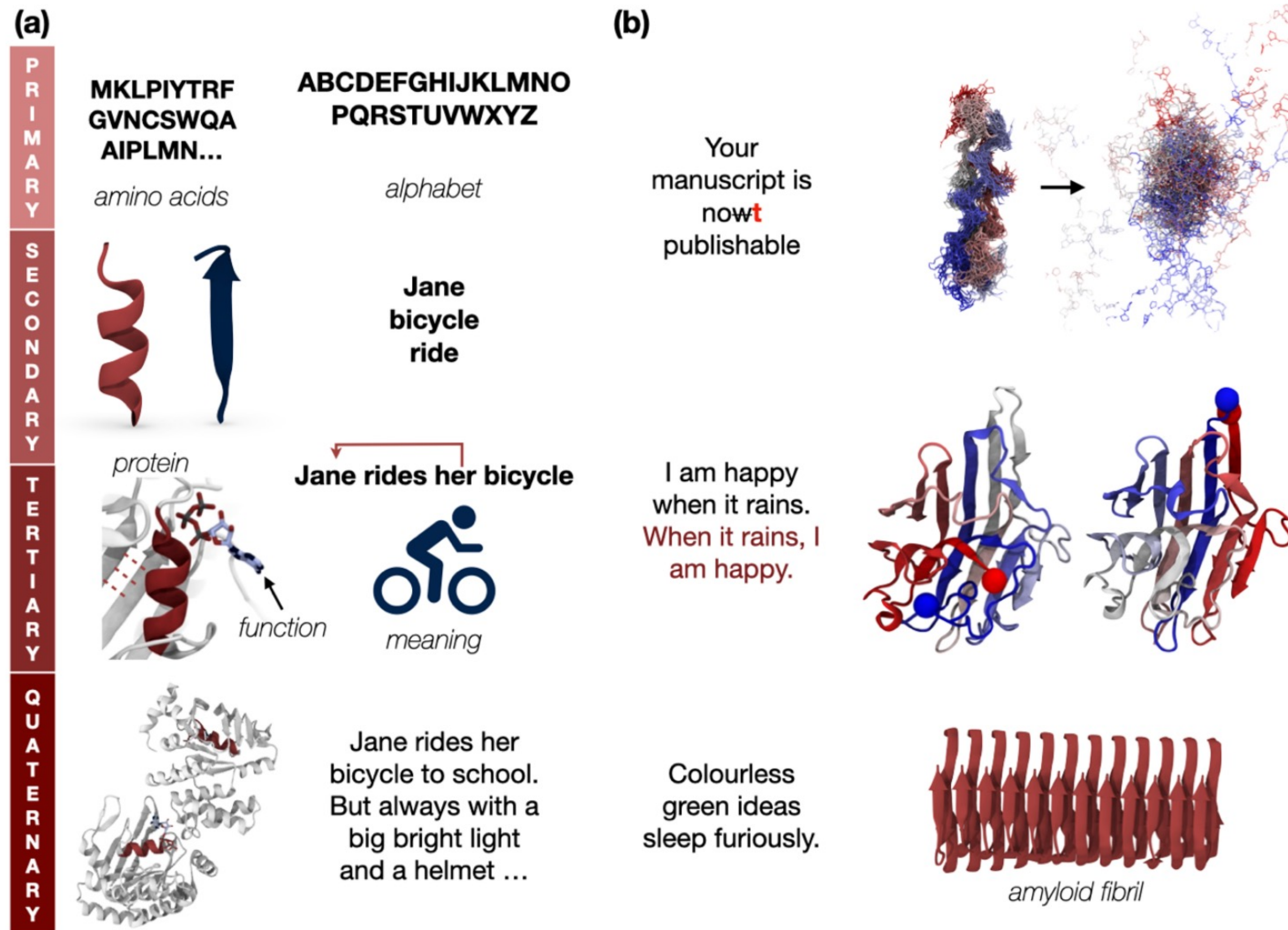
Distant interactions in protein structures

Past and Future tense?

Bias in Sequencing/Research

Proteins vs Sentences: The same?

Similar, but also important differences



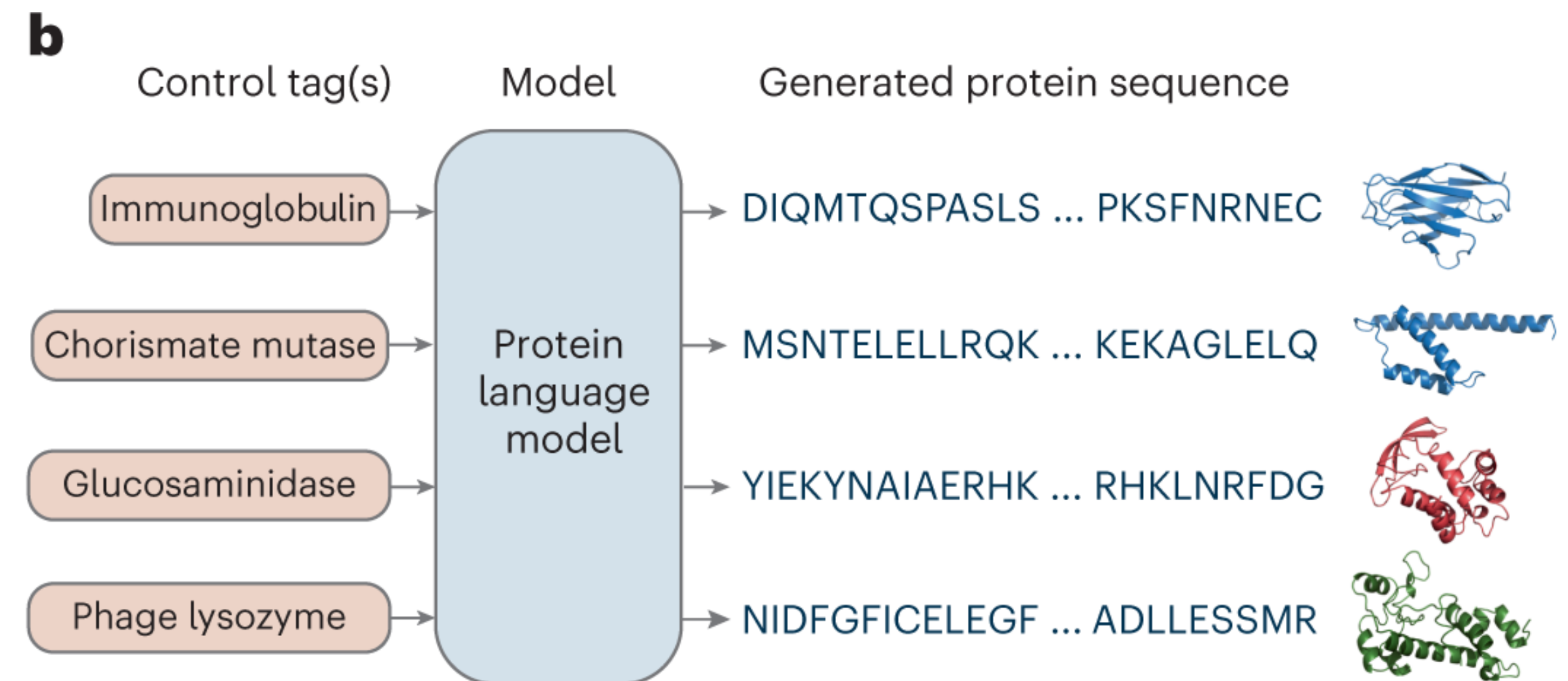
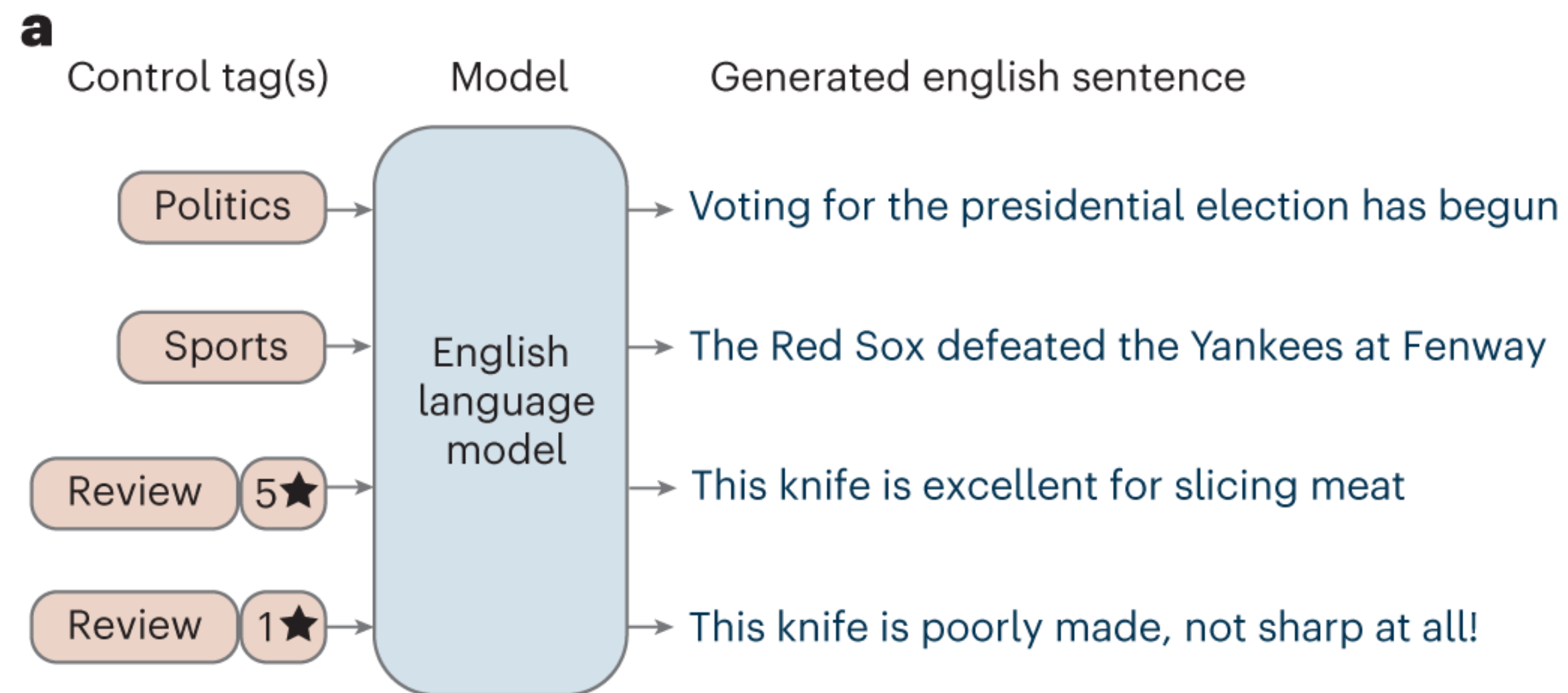
The linguistic hypothesis

Did evolution force proteins to develop a “language”?

- 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.

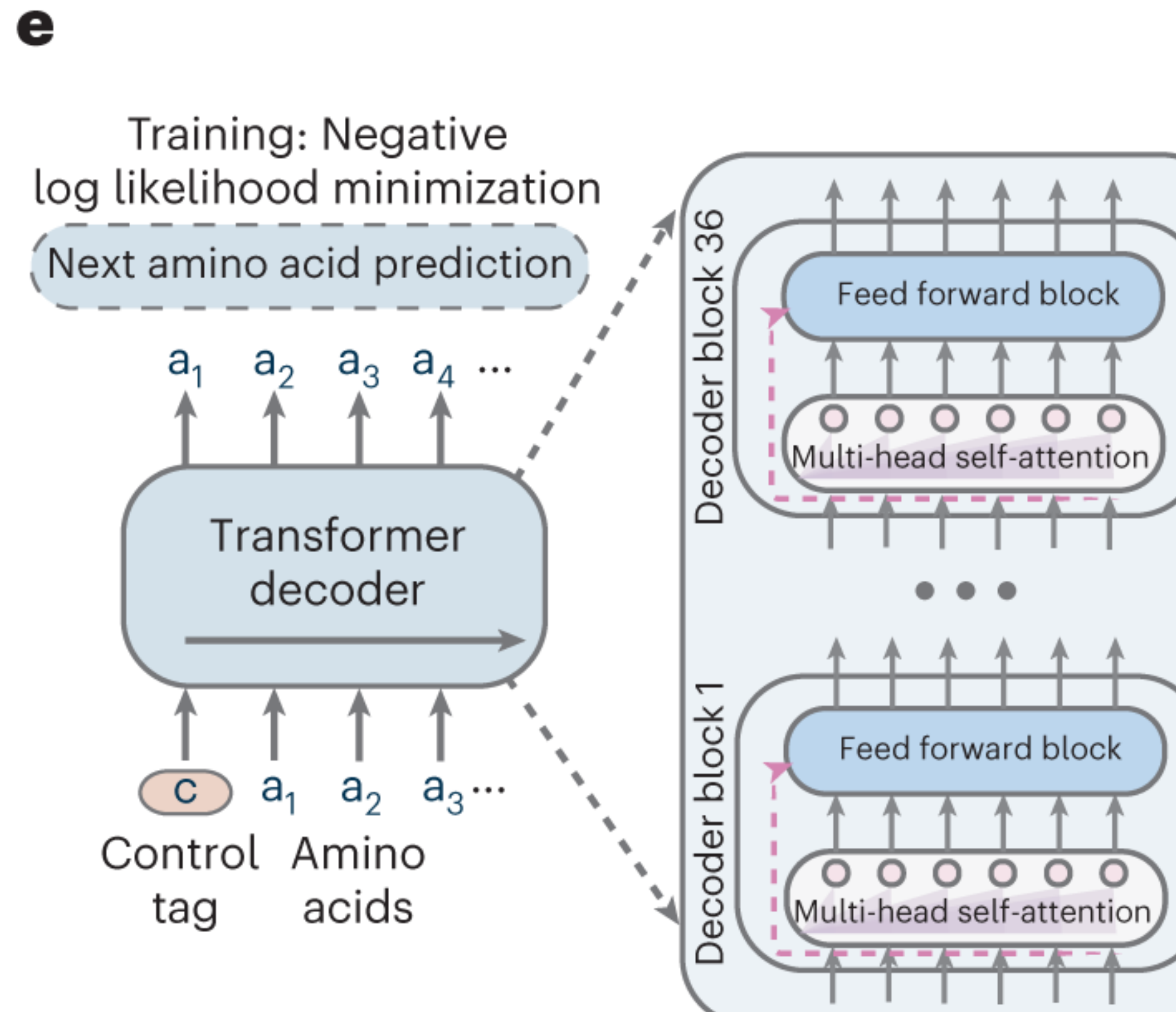
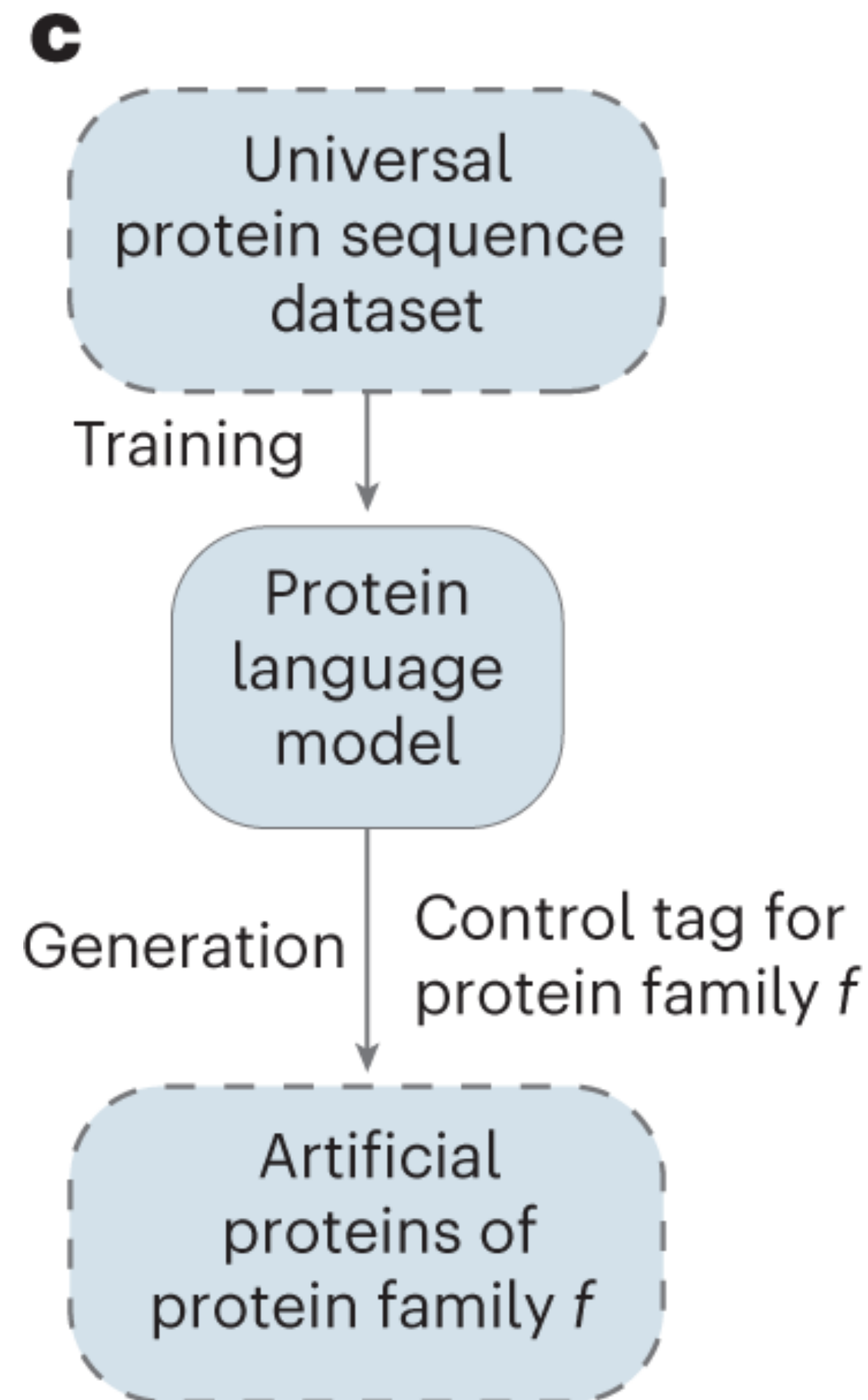
Protein Language Models

Train a model to understand the language of proteins



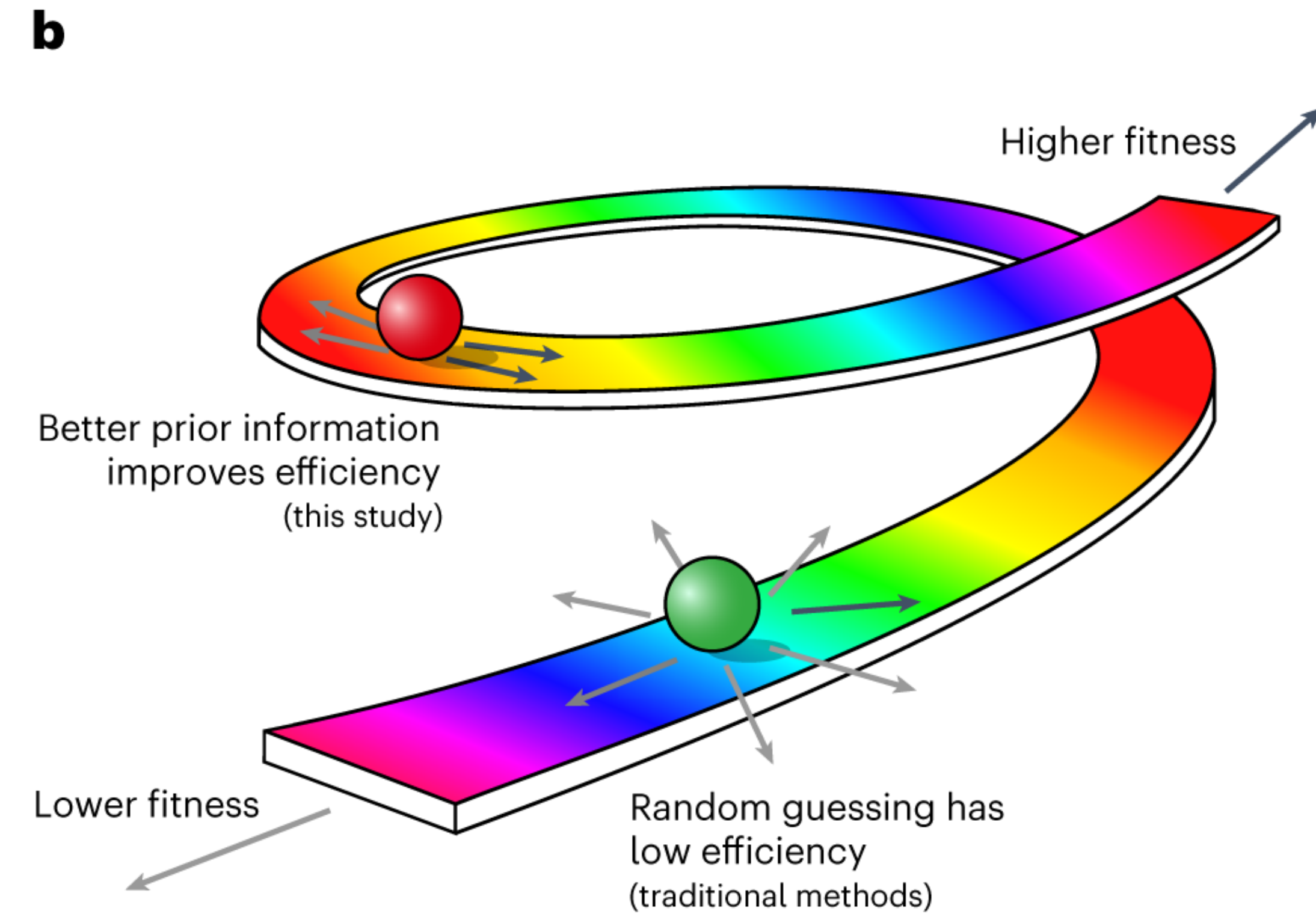
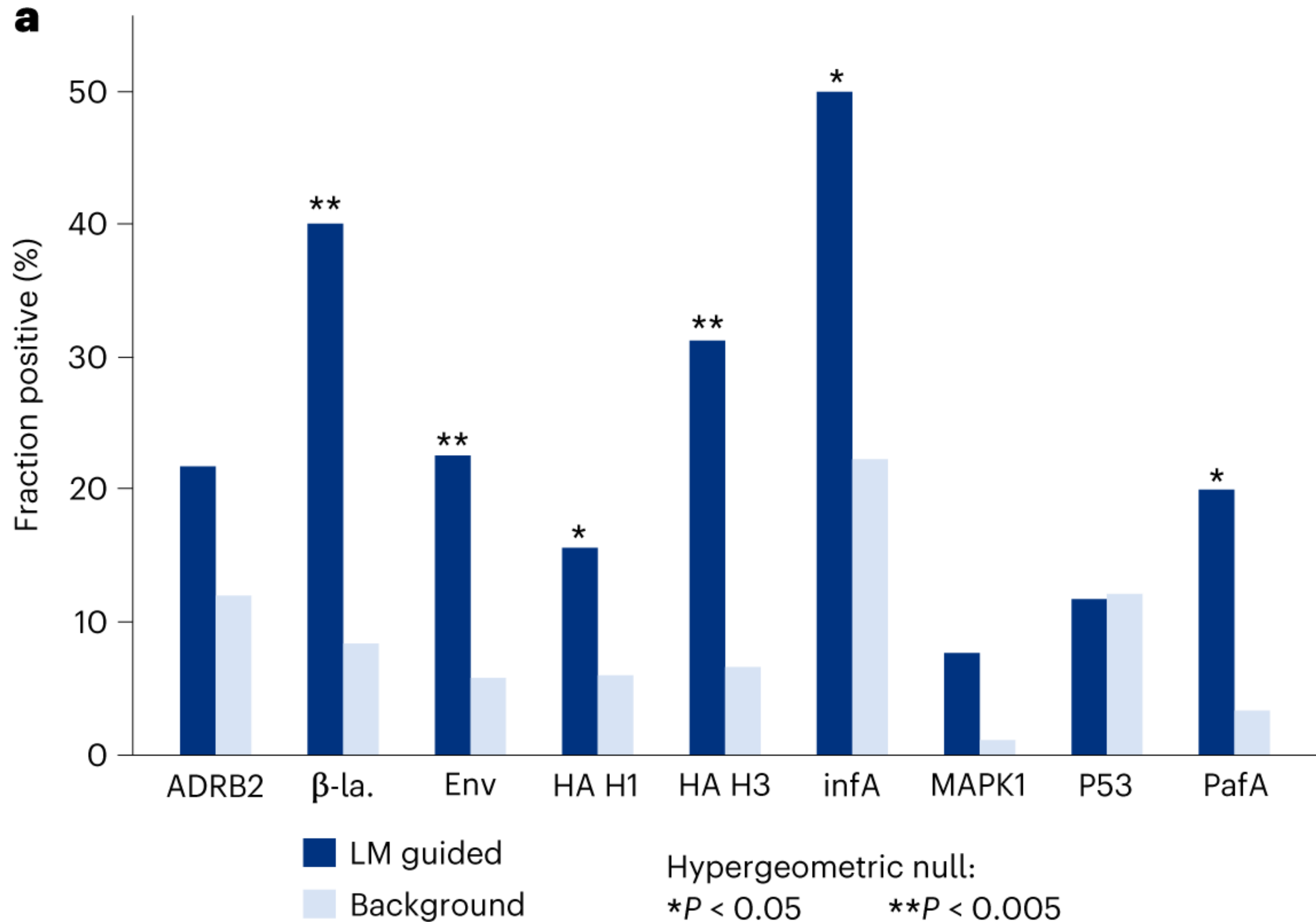
Protein Language Models

Train a model to understand the language of proteins



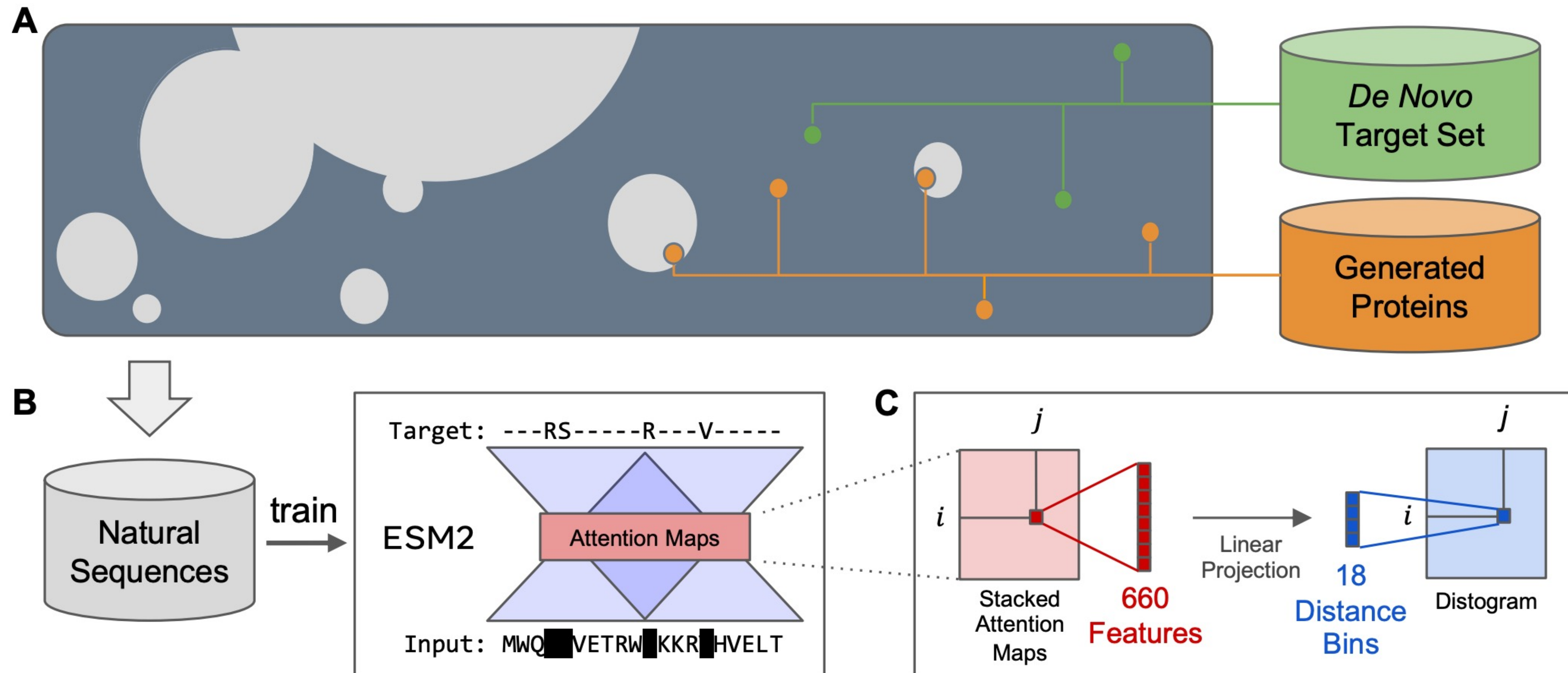
Applications of PLM

Efficient Evolution



Generalisation beyond natural proteins?

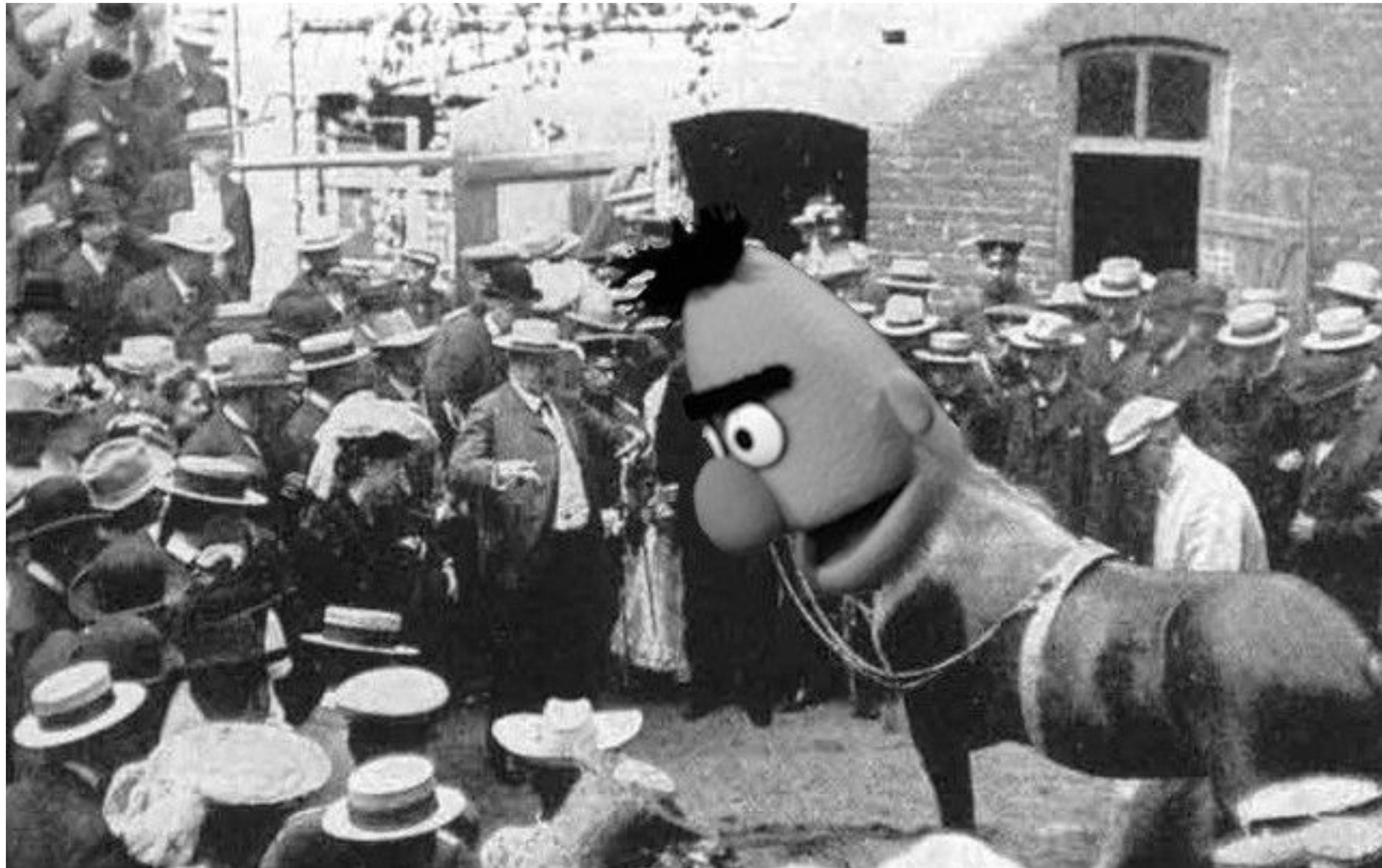
Promising signs, but no consensus yet among researchers



4. Practical Considerations

Clever Hans Moment in NLP

Do our models learn what we want them to learn?



Clever Hans Moment in NLP

Often they just learn heuristic shortcuts

Article: Super Bowl 50

Paragraph: *“Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver’s Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.”*

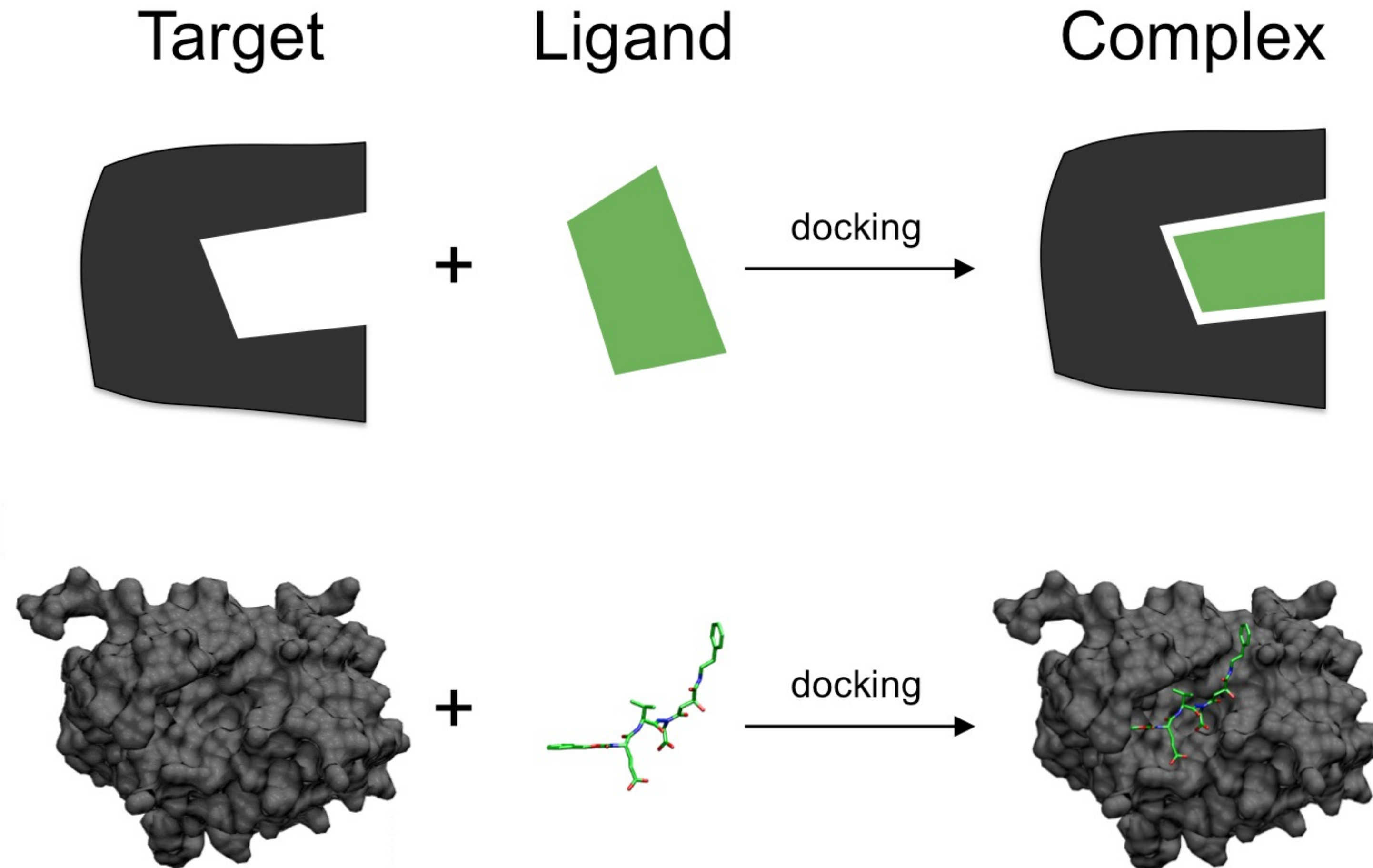
Question: *“What is the name of the quarterback who was 38 in Super Bowl XXXIII?”*

Original Prediction: John Elway

Prediction under adversary: Jeff Dean

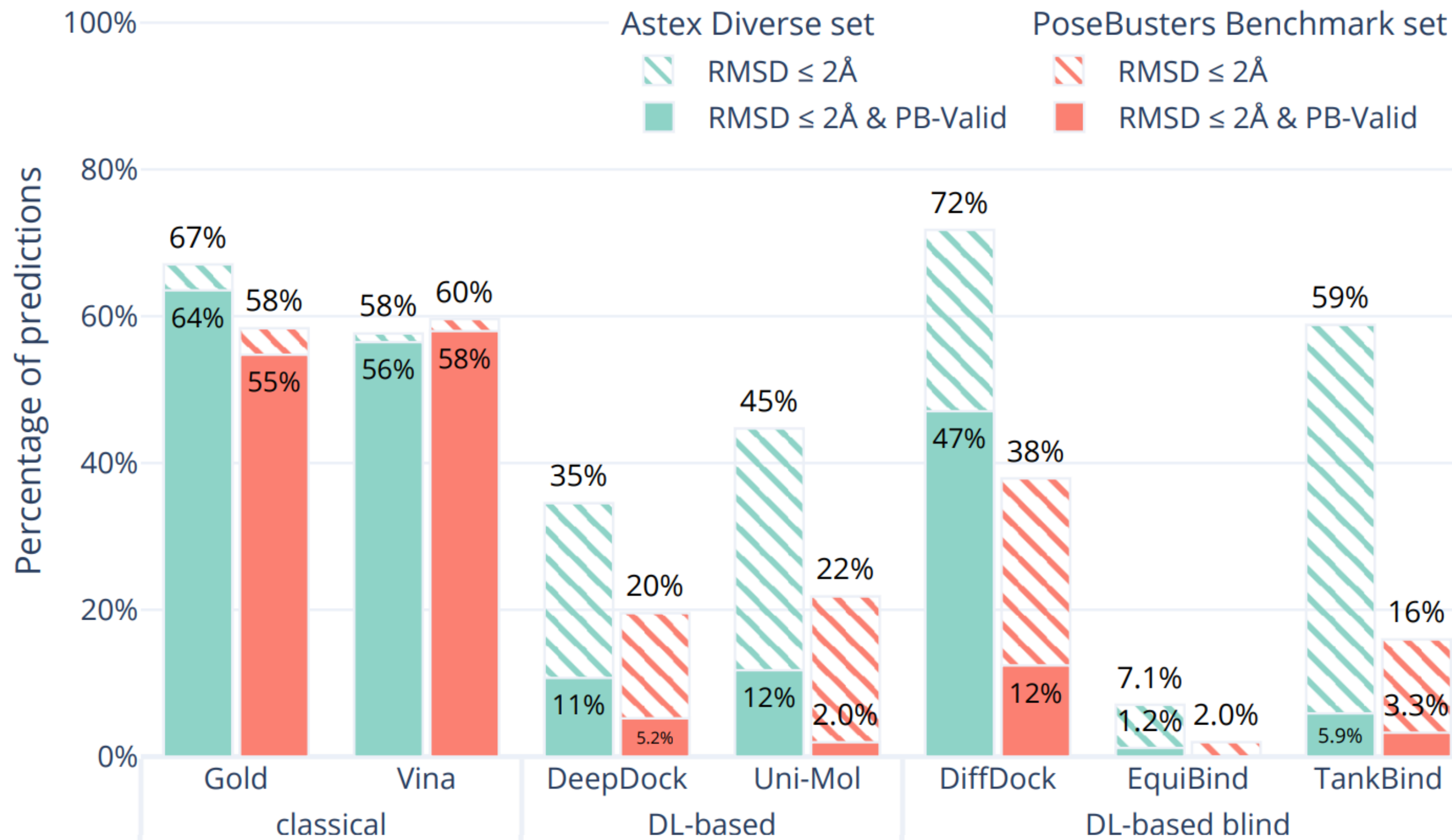
What can you do?

Spend significant time on good evaluations!



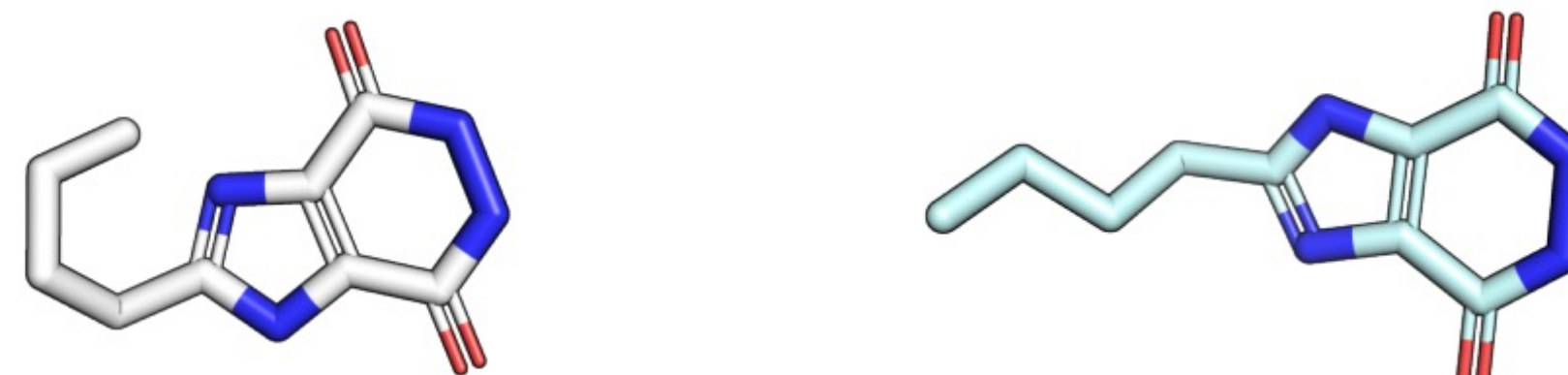
What can you do?

Spend significant time on good evaluations!



What can you do?

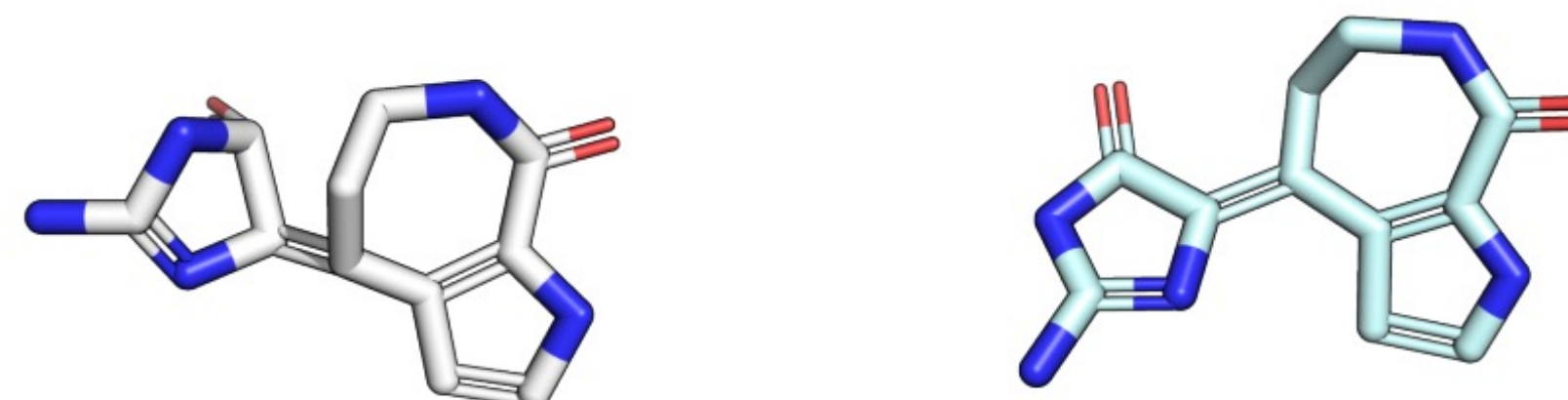
Spend significant time on good evaluations!



(d) Internal clash. DeepDock prediction for ligand BDI of protein-ligand complex 1N2V. RMSD 1.6 Å.



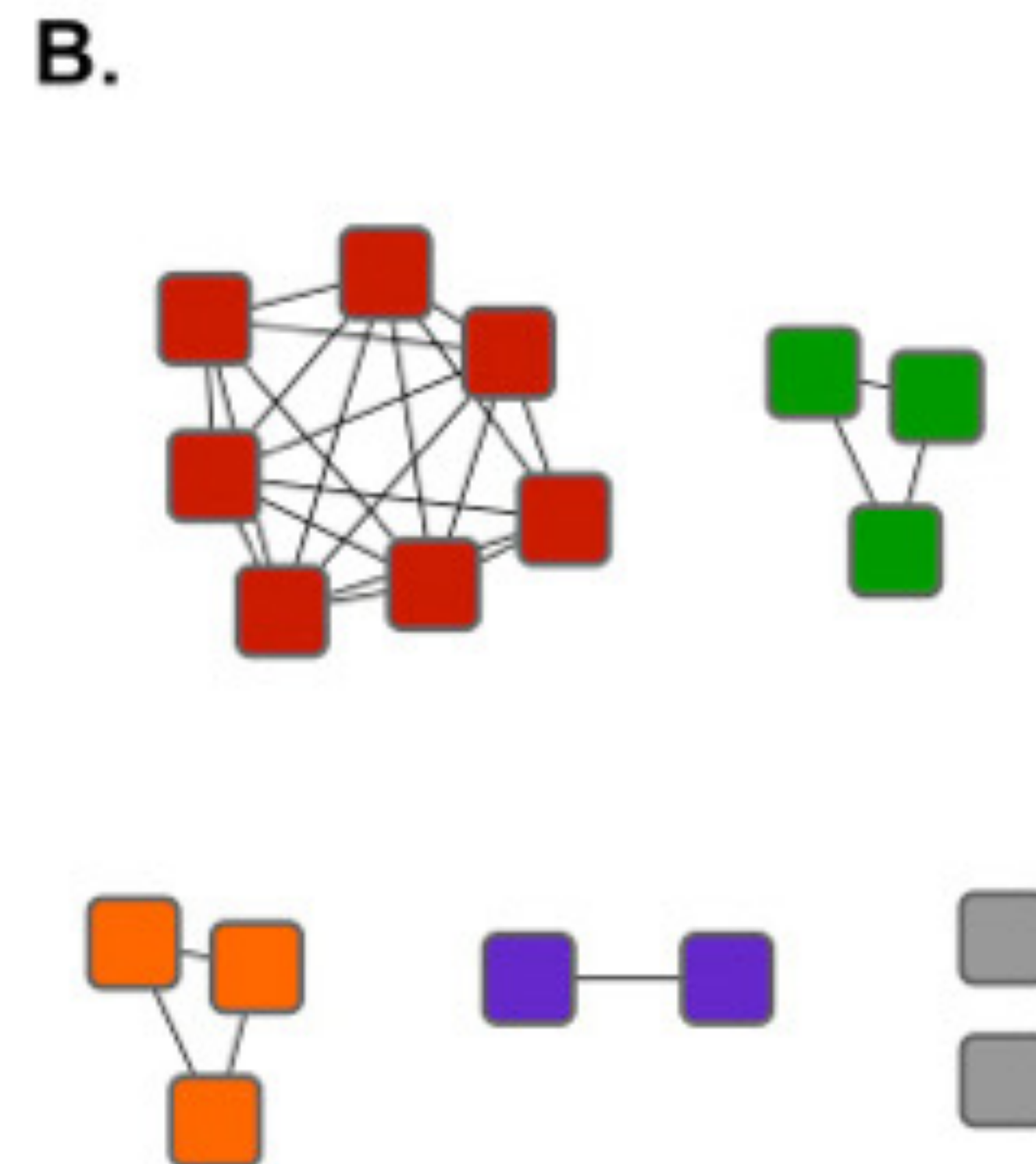
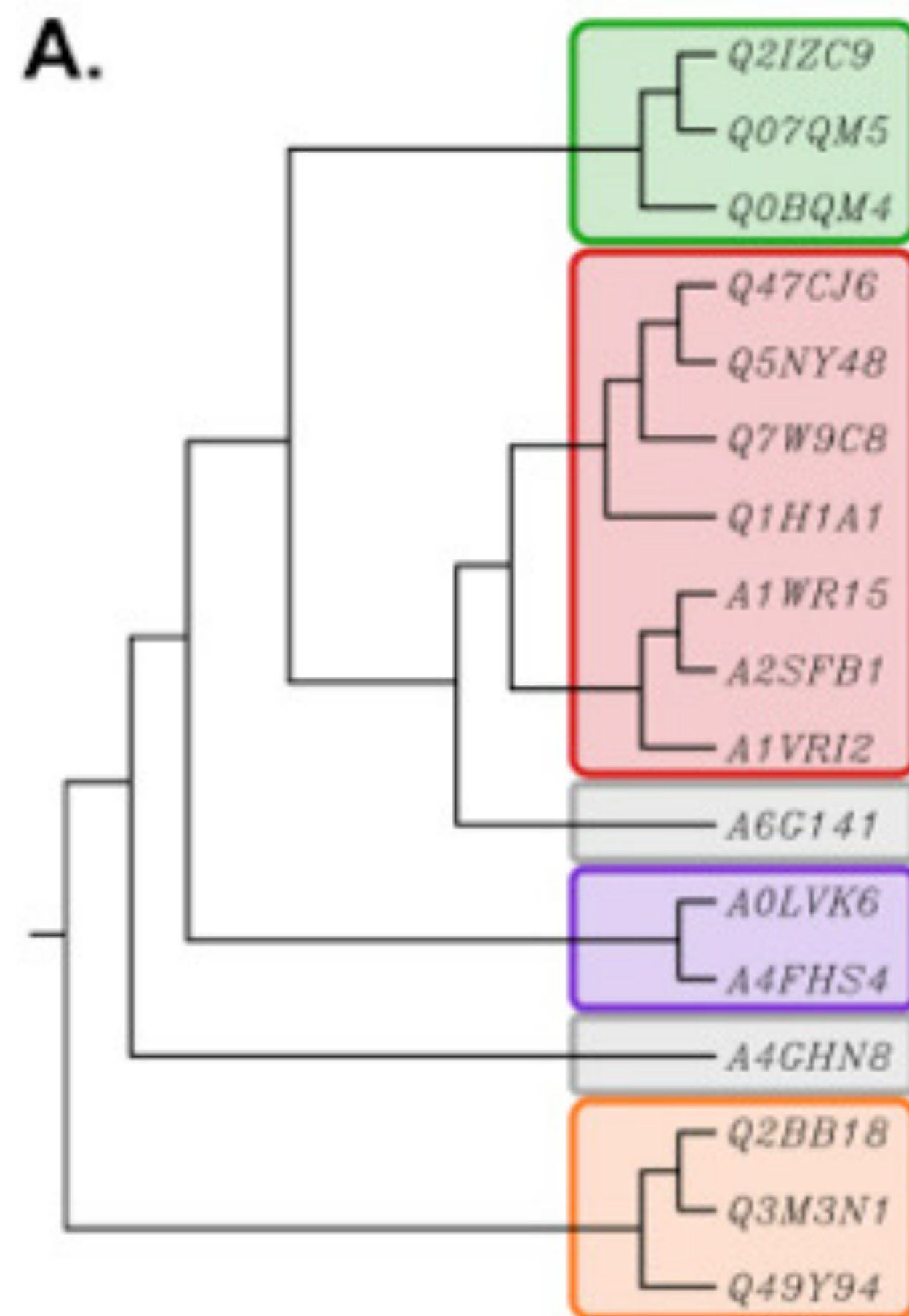
(e) Aromatic rings not flat. TankBind prediction for ligand CRZ of protein-ligand complex 1TOW. RMSD 2.2 Å.



(f) Double bond not flat. TankBind prediction for ligand DBQ of protein-ligand complex 1U4D. RMSD 1.7 Å.

What can you do?

Prepare your datasets to avoid heuristic short-cuts for models!



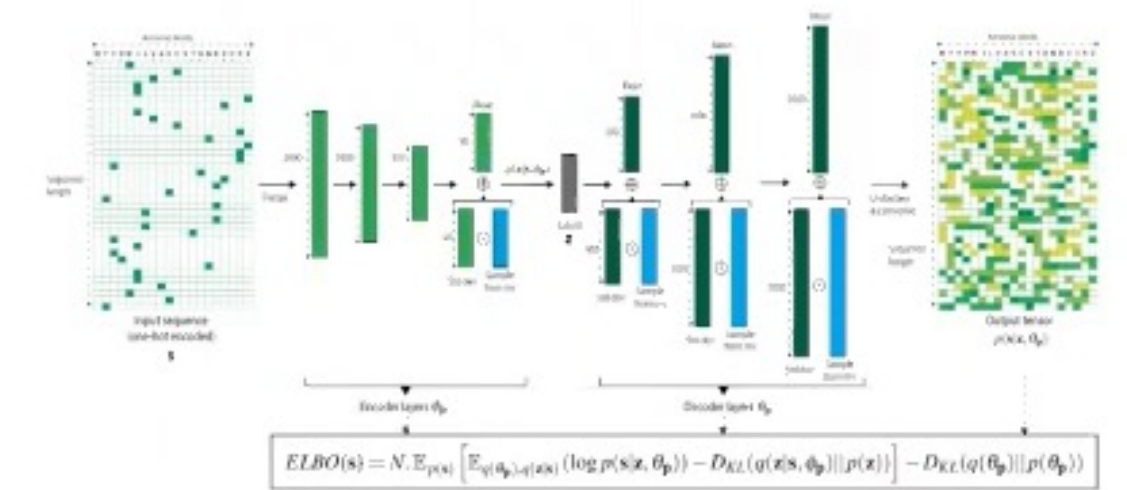
5. Current Research

Limitations of current approaches

Both alignments and PLMs have pros and cons

Alignment-based models

- Learn a distribution from sequences in a **Multiple-Sequence Alignment (MSA)** -- either at **position level** (e.g., Site independent¹), **pairs of positions** (eg., EVmutation¹) or **full sequence** (eg., DeepSequence², EVE³)
- Limitations:
 - **Unable to score insertions & deletions** ('indels')
 - **Need fairly deep alignments** to learn complex dependencies across positions (certain proteins are difficult to align eg., disordered proteins)
 - **Lack of information sharing** across families (each model is trained from scratch)



Protein language models

- Train a **(masked) language model** on large quantities of **aligned sequences** (eg., MSA Transformer⁴) or **non-aligned sequences** (eg., ESM-1v⁵) **across protein families**
- Since MLMs **do not learn a proba over full protein sequences**, fitness is approximated via the **masked-marginals heuristic**:

$$\sum_{t \in T} \log p(x_t = x_t^{mt} | x_{\setminus T}) - \log p(x_t = x_t^{wt} | x_{\setminus T})$$

- Limitations (MLMs):
 - **Unable to score insertions & deletions** ('indels')
 - **Approximation for multiple mutations**: ignore dependencies across mutations
 - **Mismatch between training Vs inference**: mask 15% tokens during training Vs 1+ token(s) at inference

How can we augment our language models?

Give them access to tools and resources in the outside world

Retrieval



Augment with a bigger corpus

Chains



Augment with more LLM calls

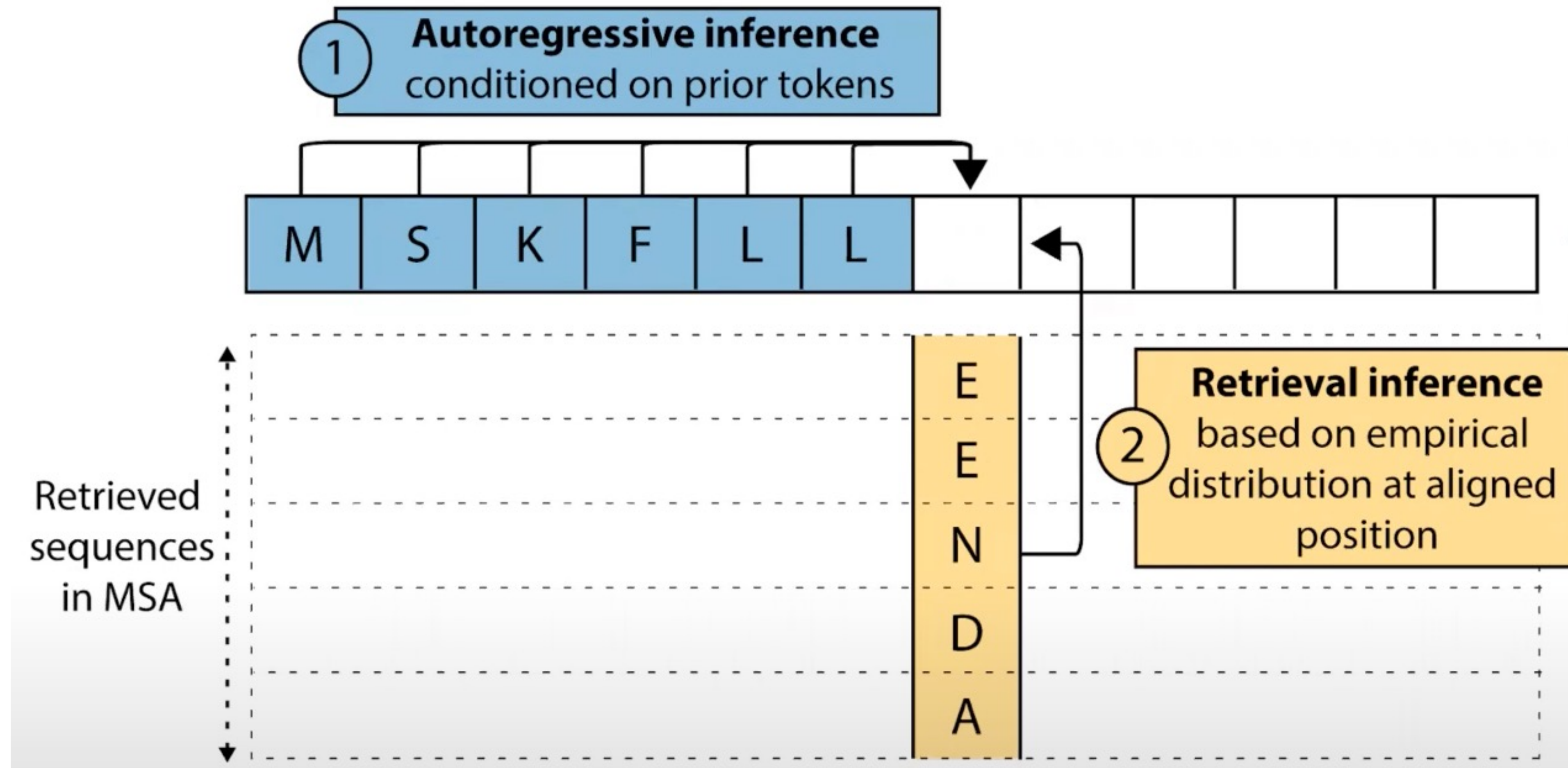
Tools



Augment with outside sources

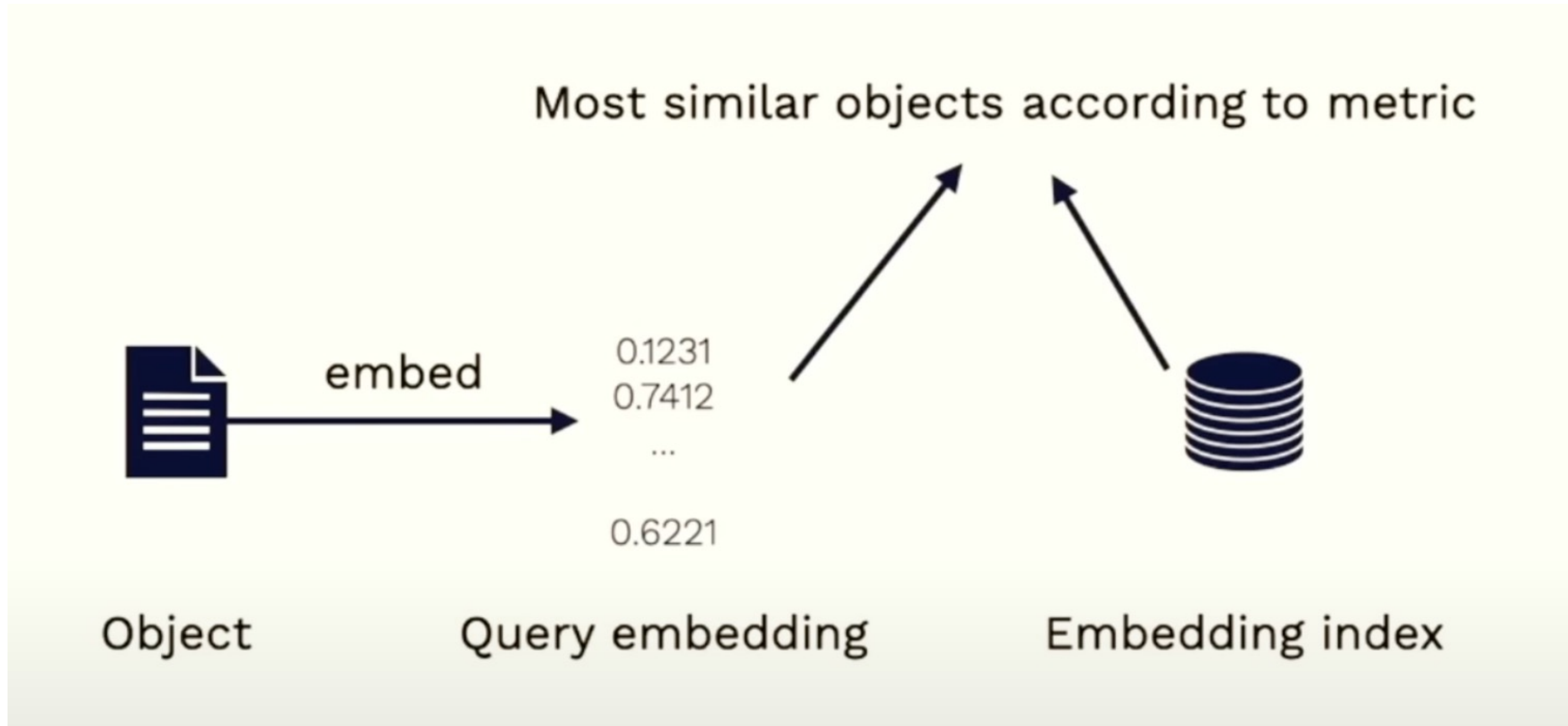
Tranception: a transformer with retrieval

Combine the best of both worlds



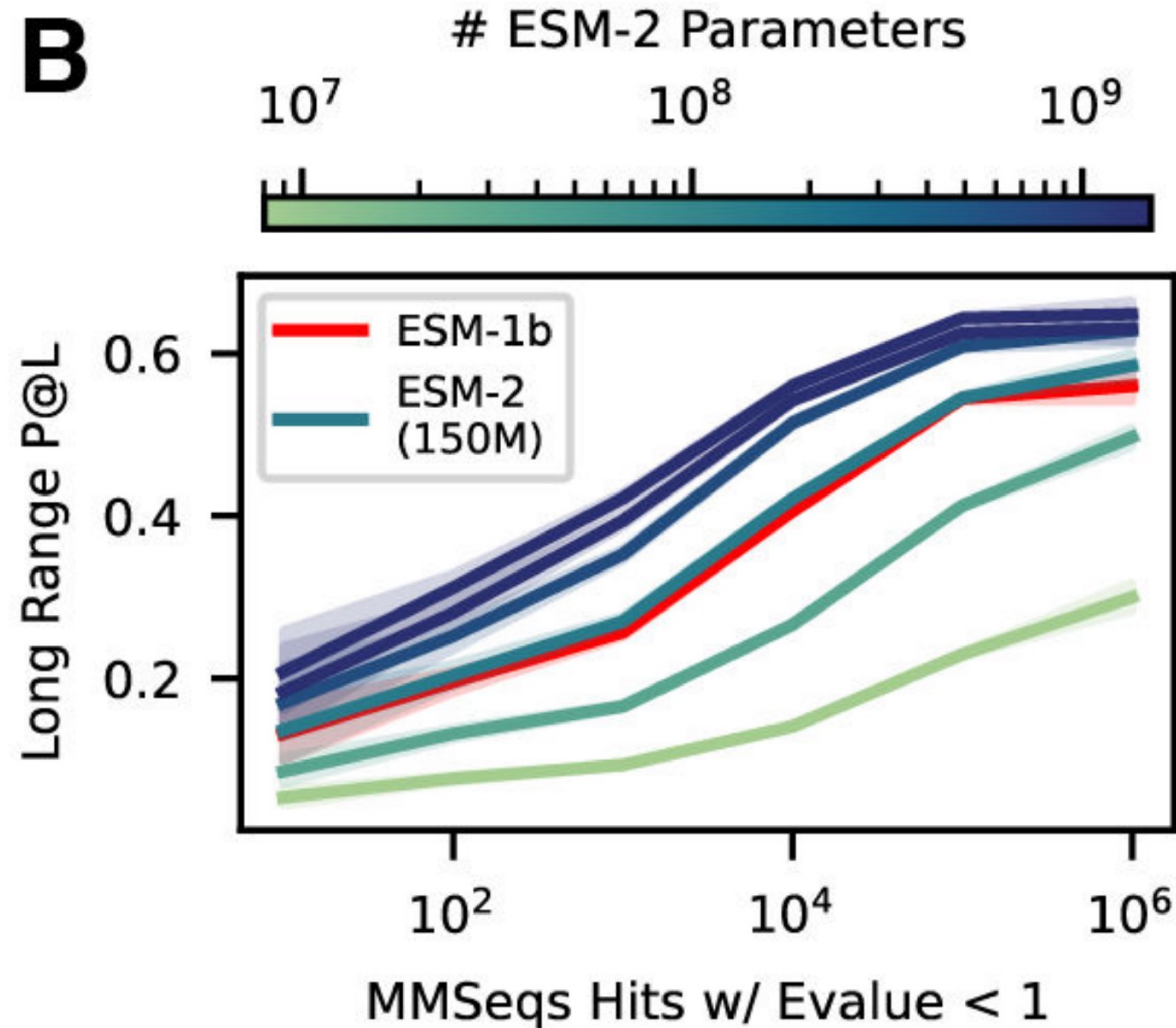
We can get clever with retrieval

We can use learned embeddings to compare similarities!



Scaling Laws: Bigger is better?

Are we at the end of scaling? A controversial topic





Takeaway



While **protein language models** show strong performance on a number of tasks, **relevant and meaningful evaluations** are still an active area of research.