International Conference on Machine Learning
July 2022

Tranception: Protein Fitness Prediction with Autoregressive Transformers and Inference-time Retrieval

Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena-Hurtado, Aidan Gomez, Debora Marks & Yarin Gal

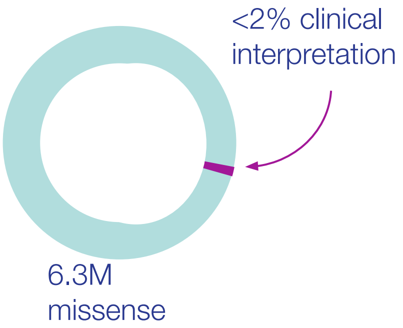
DEBORA MARKS LAB
DEPARTMENT OF SYSTEMS BIOLOGY, HARVARD MEDICAL SCHOOL
HARVARD MEDICAL SCHOOL
UNIVERSITY OF OXFORD DEPARTMENT OF COMPUTER SCIENCE
OATML

Motivations

Accurately modeling the fitness landscape of protein sequences is critical to:

Human variant annotation

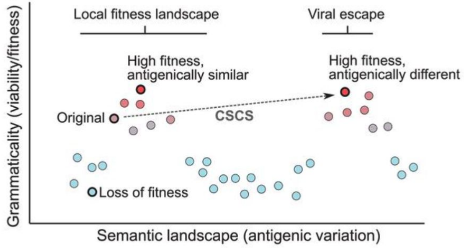
- The large majority of human variants[1] have no known interpretation



- Example: **EVE**[2], protein-specific alignment-based generative models for mutation effects prediction

Viral escape prediction

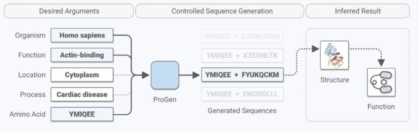
- Viral escape mutations are the ones that both **maintain fitness** while **disrupting Ab binding**



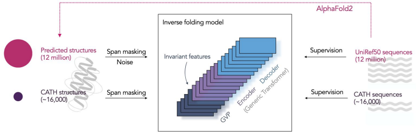
- Example: **Hie et al.**[3], use a single LLM to decompose escape in terms **semantic & grammaticality changes**

Protein design

- Generating **novel** yet **fit** sequences, conditioning on:
  - **Labels**: Madani et al., Progen[4]



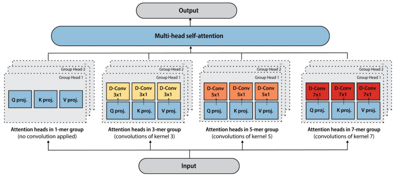
  - **Structure** (Inverse folding): Ingraham et al[5], Hsu et al[6].



1. Landrum & Kattman. ClinVar at five years: Delivering on the promise. Hum Mutat 39, 1623-1630.
2. Frazer, Notin, Dias et al. Disease variant prediction with deep generative models of evolutionary data. Nature, 2021.
3. Hie et al. Learning the language of viral evolution and escape. Science, 2021.
4. Madani et al. ProGen: Language Modeling for Protein Generation. 2020.
5. Ingraham et al. Generative Models for Graph-Based Protein Design. NeurIPS, 2019.
6. Hsu et al. Learning inverse folding from millions of predicted structures. 2022.

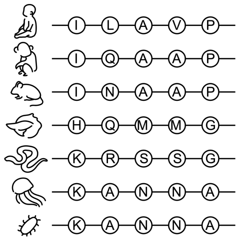
Overview

1 Tranception inference (autoregressive transformer)

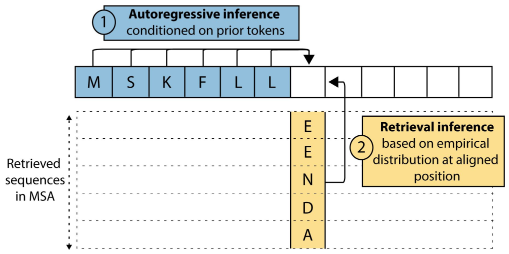


- Tranception attention
- Grouped ALiBi

2 Retrieval inference (via MSA)



3 Combining the 2 modes of inference



4 ProteinGym benchmarks

Measure	Category	DeepSequence	ProteinGym
Number of assays by taxon	Human	9	33
	Other eukaryotes	10	14
	Prokaryotes	13	24
	Virus	5	22
	All taxa	**37**	**93**
Number of variants by type	Single substitutions	0.12M	0.36M
	Multiple substitutions	0.55M	1.26M
	Indels	0	0.27M
	All variants	**0.67M**	**1.89M**

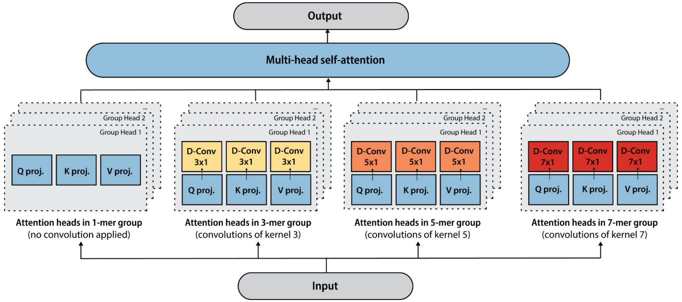
- Substitution benchmark
- Indel benchmark

5 Detailed performance analysis

Model type	Model name	Spearman's rank correlation by MSA depth ↑ Low	Medium	High	All	AUC ↑ All
Alignment-based models	Site indep	0.428	0.403	0.350	0.397	0.725
	Wavenet	0.319	0.398	0.469	0.398	0.725
	DeepSequence	0.375	0.397	0.506	0.415	0.733
	EVmutation	0.401	0.421	0.468	0.427	0.738
	EVE	0.408	**0.440**	**0.507**	0.448	0.751
Protein language models	ESM-1v	0.321	0.348	0.484	0.371	0.713
	MSA Transformer	0.373	0.418	0.482	0.422	0.737
	Tranception (w/o retrieval)	0.394	0.398	0.439	0.406	0.728
	Tranception (w/ retrieval)	**0.453**	0.438	0.488	**0.451**	**0.754**

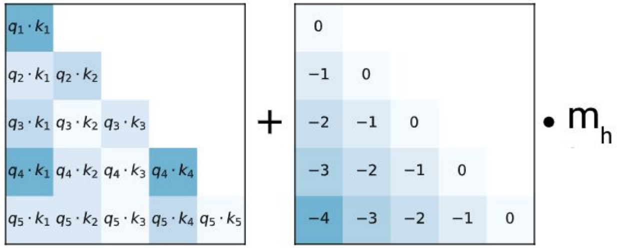
1 The two key components of the Tranception autoregressive transformer: Tranception attention and Grouped ALiBi

Tranception attention



- Our scheme differs from the standard autoregressive architecture (eg. GPT-2[1]) by promoting:
  - **extraction of sequence patterns of different lengths** (ie., k-mers)
  - **head specialization**
- Combines ideas from **Primer**[2] (D-conv after attention linear projections) and **Inception**[3] (split attention heads into 4 groups and apply a convolution w/ different kernel size to each group)

Grouped ALiBi



- **ALiBi**[4] is a relative position embedding method (used in lieu of learned / sinusoidal position embeddings)
- m_h is an attention **head-specific constant.** For a transformer with n attention heads:

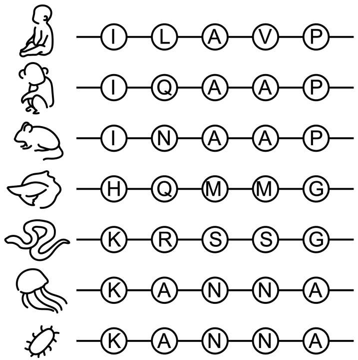
$$m_h = 2^{\frac{8.h}{n}}, with\ h \in [1, n]$$

- Leads to **faster training convergence** & **memory savings**
- We introduce **Grouped ALiBi,** in which we split attention heads in 4 groups and apply ALiBi to each group

1. Radford, Wu et al. Language Models are Unsupervised Multitask Learners. 2019
2. So et al. Primer: Searching for Efficient Transformers for Language Modeling. 2021
3. Szegedy et al. Going deeper with convolutions. CVPR, 2015
4. Press et al. Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation. 2021

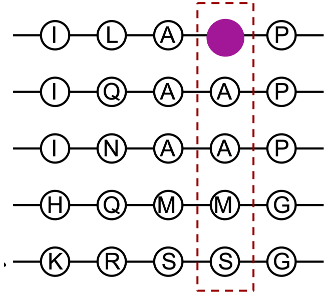
2 Inference-time retrieval

We retrieve a **Multiple Sequence Alignment (MSA)** for each protein sequence to be scored …



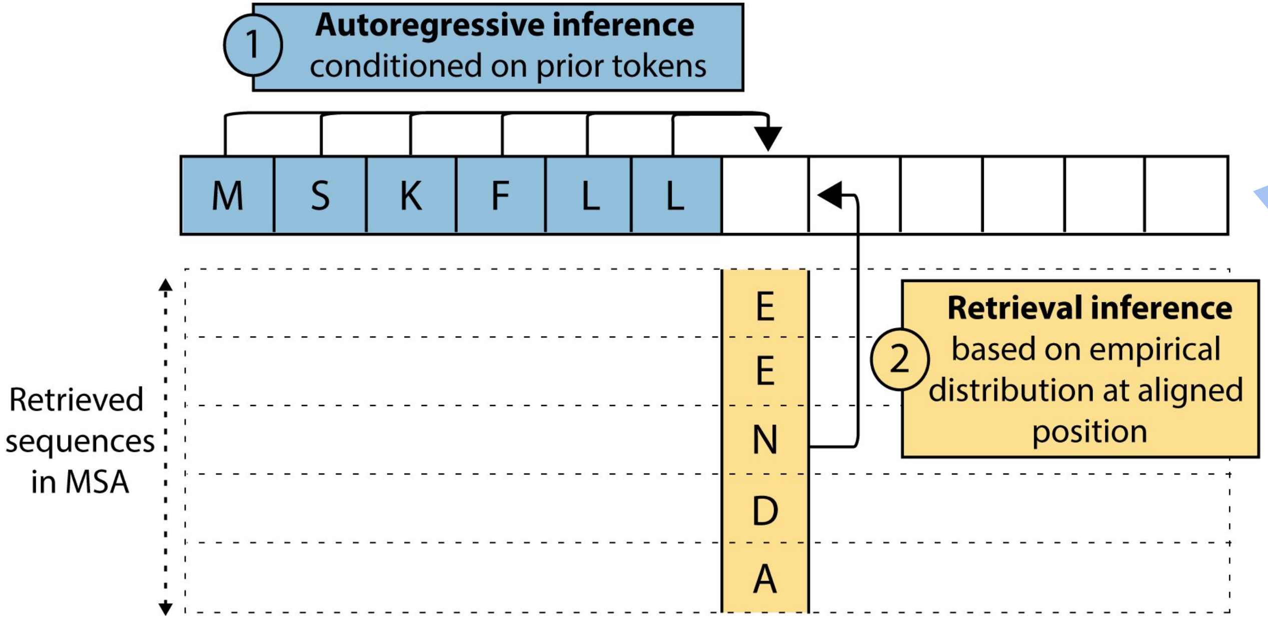
- **Substitution scoring:** one MSA retrieval amortized across all substitutions (singles and multiples)
- **Indel scoring:** we tailor the retrieved MSA to each mutated sequence by a) deleting columns in the MSA corresponding to deleted positions and b) adding zero-filled columns in the MSA at inserted positions in the mutated protein

… and compute **weighted pseudocounts** at each position to infer a distribution over AA at that position



- Pseudocounts at each position of the alignment computed via **weighted Laplace smoothing** (Jurafsky & Martin, 2008), with a small smoothing parameter (10^{-5})
- We fully **ignore gaps** in the MSA when computing the pseudocounts
- Sequence are **weighted** as per the procedure described in Hopf et al., 2017

3 At test time, we combine the autoregressive inference with retrieval inference



- During training of Tranception we apply **random sequence mirroring** as a data augmentation
- That allows us at inference to **score the sequence from both directions** (left to right and from right to left) and **average the two scores**

$$\log P(x) \propto \sum_{i=1}^{l} [(1-\alpha)\log P_A(x_i|x_{<i}) + \alpha \log P_R(x_i)]$$

4 ProteinGym benchmarks

- **ProteinGym** is a set of DMS-based benchmarks for fitness prediction
- **Two benchmarks:** substitutions and indels
- **Significant increase** in terms of **numbers of assays, number of mutants, diversity of assays** (more balanced share of human & viral proteins, more multiple assays) compared with prior benchmarks (eg., DeepSequence)

Measure	Category	DeepSequence	ProteinGym	Fold increase
Number of assays by taxon	Human	9	33	3.7
	Other eukaryotes	10	14	1.4
	Prokaryotes	13	24	1.8
	Virus	5	22	4.4
	All taxa	**37**	**93**	**2.5**
Number of variants by type	Single substitutions	0.12M	0.36M	2.9
	Multiple substitutions	0.55M	1.26M	2.3
	Indels	0	0.27M	-
	All variants	**0.67M**	**1.89M**	**2.8**

Comparison of the ProteinGym and DeepSequence benchmarks

5 Performance analysis: Robustness to MSA depth and gain of scope (1/3)

Performance by MSA depth

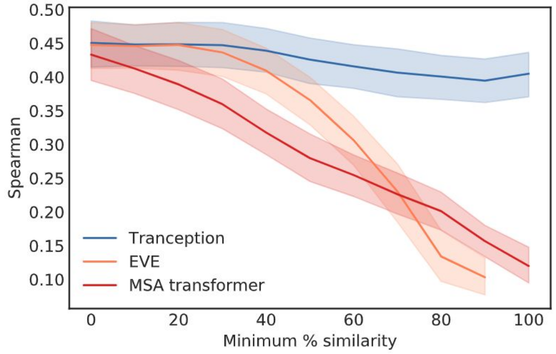
Avg. Spearman's rank correlation w/ experimental measurements

Model type	Model name	Spearman's rank correlation by MSA depth ↑ Low	Medium	High	All
Alignment-based models	Site indep	0.428	0.403	0.350	0.397
	Wavenet	0.319	0.398	0.469	0.398
	DeepSequence	0.375	0.397	0.506	0.415
	EVmutation	0.401	0.421	0.468	0.427
	EVE	0.408	**0.440**	**0.507**	0.448
Protein language models	ESM-1v	0.321	0.348	0.484	0.371
	MSA Transformer	0.373	0.418	0.482	0.422
	Tranception (w/o retrieval)	0.394	0.398	0.439	0.406
	Tranception (w/ retrieval)	**0.453**	0.438	0.488	**0.451**

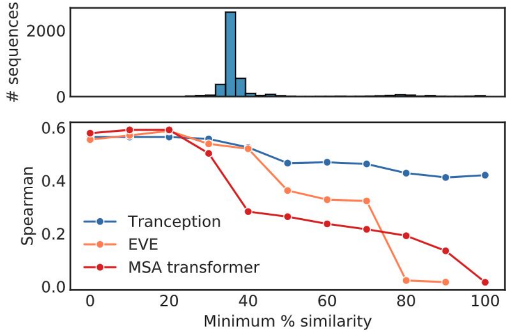
Robustness to MSA depth analysis

Avg. Spearman's rank correlation w/ experimental measurements when progressively filtering the MSA (based on min similarity to the wild type sequence)

All proteins in ProteinGym



Example for the P53 protein



5 Performance analysis: Versatility of usage (2/3)

ProteinGym substitution benchmark
Avg. Spearman's rank correlation w/ experimental measurements

By mutation depth

Model type	Model name	Spearman's rank correlation by mutation depth ↑ 1	2	3	4	5+	All
Alignment-based models	Site indep	0.396	0.325	0.286	0.319	0.421	0.397
	Wavenet	0.394	0.344	0.329	0.281	0.396	0.398
	DeepSequence	0.415	0.394	0.372	0.304	0.418	0.415
	EVmutation	0.427	0.392	0.379	0.319	0.433	0.427
	EVE	**0.448**	0.392	0.375	0.334	0.420	0.448
Protein language models	ESM-1v	0.372	0.291	0.190	0.160	0.245	0.371
	MSA Transformer	0.423	0.359	0.390	0.327	0.431	0.422
	Tranception (w/o retrieval)	0.397	0.412	0.425	0.335	0.479	0.406
	Tranception (w/ retrieval)	**0.448**	**0.435**	**0.443**	**0.368**	**0.499**	**0.451**

By taxon

Model type	Model name	Spearman correlation by taxa category ↑ Human	Other Eukaryote	Prokaryote	Virus	All
Alignment-based models	Site indep	0.398	0.446	0.350	0.410	0.397
	Wavenet	0.388	0.453	0.480	0.308	0.398
	Deepsequence	0.391	0.482	0.487	0.350	0.415
	EVmutation	0.405	0.475	0.484	0.380	0.427
	EVE	0.411	0.485	**0.497**	**0.435**	0.448
Protein language models	ESM-1v	0.394	0.420	0.482	0.216	0.371
	MSA Transformer	0.379	0.491	0.494	0.380	0.422
	Tranception (w/o retrieval)	0.369	0.441	0.453	0.396	0.406
	Tranception (w/ retrieval)	**0.426**	**0.502**	0.485	0.429	**0.451**

ProteinGym indel benchmark
Avg. AUC & Spearman's rank correlation w/ experimental measurements

Model name	Spearman ↑	AUC ↑
Wavenet	0.412	0.724
Tranception (w/o retrieval)	0.430	0.740
Tranception (w/ retrieval)	**0.463**	**0.759**

5 Performance analysis: Flexibility and modularity (3/3)

If we have additional knowledge about the protein, we may use it to create better MSA (eg., domain-level)

Avg. Spearman's rank correlation w/ experimental measurements; BRCA1 example

Domain	Tranception (w/o retrieval)	Tranception (retrieval full MSA)	Tranception (retrieval domain MSA)
RING	0.567	0.588	**0.607**
BRCT	0.354	0.490	0.504

- Since the Tranception autoregressive transformer and retrieval are **two modular components**, we have the flexibility to **not use retrieval**, for example if MSA depth is **too shallow**
- If we have additional knowledge about the protein (eg., separate domains), we can **manually craft better MSA** leading to **better performance**

We may combine Tranception with more complex models of the retrieved MSA at inference

Avg. Spearman's rank correlation w/ experimental measurements

Model pair ensembled	**Spearman**
Tranception w/o retrieval	0.406
Tranception + ESM-1v	0.427
Tranception + MSA Transformer	0.449
Tranception + EVE	**0.473**

- Ensembling **Tranception** (w/o retrieval) with an **EVE** model trained on the retrieved MSA at inference **yields even higher performance**
- **Trade-off** between **performance and compute budget** needed to train additional model
- **Flexibility** to train a complex model on MSA when its **depth is sufficient** Vs keep simpler retrieval mechanism otherwise

Conclusion

POSTER: Today 6:30-8:30pm; Hall E, #122

Paper: https://arxiv.org/abs/2205.13760
Code: https://github.com/OATML-Markslab/Tranception

Summary

- **State-of-the-art performance** on both substitutions and indels predictions
- Higher performance on **multiple mutants**, which increases with depth
- **One model for all proteins** -- performs well across taxa
- Performance **robust to MSA depth** / out performs other models in **shallow regime**
- **Flexibility to use or not MSAs**; to **curate MSAs** to particular application based on domain knowledge (eg., BRCA1) and to ensemble Tranception w/ more powerful alignment-based models to be trained on the retrieved MSA

Future directions

Model improvements

- **Scaling model size** (scaling laws for protein LLMs[1])
- **Training /w more data** (eg., MGnify, GISAID)
- Taking **phylogeny** into account[2]
- **Retrieval at train time** (eg., as in RETRO[3])
- Leverage **protein structure** more explicitly

Applications

- **Supporting clinical annotations in humans**, in particular for disordered proteins / regions
- Predicting **viral escape** mutants
- **Inverse folding** problem

1. Hesslow et al. RITA: a Study on Scaling Up Generative Protein Sequence Models. 2022
2. Weinstein, Amin et al. Non-identifiability and the Blessings of Misspecification in Models of Molecular Fitness and Phylogeny. 2022
3. Borgeaud, Mensch, Hoffmann et al. Improving language models by retrieving from trillions of tokens. 2021