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

Pascal Notin [1] **Mafalda Dias** [2] **Jonathan Frazer** [2] **Javier Marchena-Hurtado** [2] **Aidan Gomez** [1 3]
Debora S. Marks [† 2] **Yarin Gal** [† 1]

Abstract

The ability to accurately model the fitness landscape of protein sequences is critical to a wide range of applications, from quantifying the effects of human variants on disease likelihood, to predicting immune-escape mutations in viruses and designing novel biotherapeutic proteins. Deep generative models of protein sequences trained on multiple sequence alignments have been the most successful approaches so far to address these tasks. The performance of these methods is however contingent on the availability of sufficiently deep and diverse alignments for reliable training. Their potential scope is thus limited by the fact many protein families are hard, if not impossible, to align. Large language models trained on massive quantities of non-aligned protein sequences from diverse families address these problems and show potential to eventually bridge the performance gap. We introduce Tranception, a novel transformer architecture leveraging autoregressive predictions and retrieval of homologous sequences at inference to achieve state-of-the-art fitness prediction performance. Given its markedly higher performance on multiple mutants, robustness to shallow alignments and ability to score indels, our approach offers significant gain of scope over existing approaches. To enable more rigorous model testing across a broader range of protein families, we develop ProteinGym – an extensive set of multiplexed assays of variant effects, substantially increasing both the number and diversity of assays compared to existing benchmarks.

1. Introduction

Unsupervised models predicting the effects of mutations in protein sequences are emerging as central tools in drug design, pathogen forecasting, identification of disease causing variants and more. Several modeling approaches have been introduced in recent years, offering various trade-offs in terms of performance, diversity of proteins which can be modelled and types of sequence variation which can be scored. Current state-of-the-art methods for predicting the effect of single amino acid substitutions are trained on a multiple sequence alignment (MSA) for each protein sequence or domain of interest. In this context, MSAs serve two purposes. First, they act as a data acquisition tool by identifying sequences related to the target in a large protein database, in order to then train a model on a relevant set of sequences. Second, they align sequences by modelling insertions, deletions and substitutions, resulting in a coordinate system which enables the amino acid at a given position to be compared across the training set. While training models on protein-specific alignments has been shown to be an effective approach for thousands of proteins (Frazer et al., 2021) it nevertheless brings severe limitations. For instance, such models can not make predictions for sequences which are incompatible with the coordinate system of the MSA used in training (eg., insertions and deletions), thereby limiting scope. Additionally, a large fraction of the proteome corresponds to regions that can not be aligned such as so-called disordered regions – around half of all human proteins contain regions of at least 40 amino acids classified as disordered (Radivojac et al., 2004; Toth-Petroczy et al., 2016). Even when alignments are accessible, the protein function might be taxa specific, and the MSA algorithm may not retrieve a large enough set of homologous sequences for model training. Alignment-based models may be relatively sensitive to the characteristics of the MSAs they are trained on – including the choice of hyperparameters used to retrieve these MSAs. Lastly, there is a lack of information sharing across models that are independently trained on different data subsets. Recently, language models trained on large quantities of non-aligned (Meier et al., 2021) or aligned (Rao et al., 2021) protein sequences have made first steps towards addressing these issues. However, mod-

† Equal senior authorship [1]OATML Group, Department of Computer Science, University of Oxford, Oxford, UK [2]Marks Group, Department of Systems Biology, Harvard Medical School, Boston, MA, USA [3]Cohere, Toronto, Canada. Correspondence to: Pascal Notin <pascal.notin@cs.ox.ac.uk>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

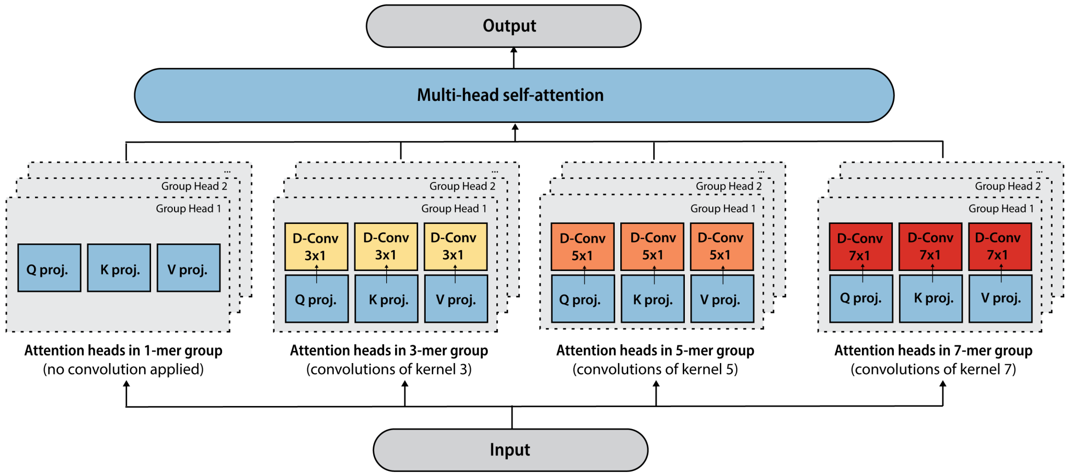
Output
Multi-head self-attention
Group Head 2
Group Head 1
Q proj.
K proj.
V proj.
D-Conv 3x1
D-Conv 5x1
D-Conv 7x1
Attention heads in 1-mer group
(no convolution applied)
Attention heads in 3-mer group
(convolutions of kernel 3)
Attention heads in 5-mer group
(convolutions of kernel 5)
Attention heads in 7-mer group
(convolutions of kernel 7)
Input

Figure 1. **Tranception attention mechanism.** Attention heads at each layer are split into 4 distinct groups. Except for the first group that does not mix information across tokens (ie., '1-mer' group), separate spatial depthwise convolutions are applied after the Query, Key and Value projections of the other 3 groups with kernel size of 3, 5 and 7 respectively. This incentivizes attention heads within each group to mix information at various ranges across the sequence length. We preserve autoregressiveness of the architecture by applying the right amount of left padding for each convolution.

els trained on non-aligned sequences still fail to match the performance of alignment-based methods without further fine-tuning on sequences obtained with a MSA (Meier et al., 2021). Furthermore, approaches based on masked language modeling objectives are unable to estimate the log likelihood of full sequences leading to heuristics when predicting mutation effects – in particular for multiple mutants, and do not support the scoring of 'indels' (Appendix D). To address these limitations, we introduce Tranception, an autoregressive transformer architecture which is pretrained on large quantities of non-aligned sequences and leverages retrieval at inference to achieve state-of-the-art fitness prediction performance. Our model outperforms all prior baselines, especially on proteins with shallow alignments. Since we do not train on MSAs, our performance is less sensitive to their characteristics (eg., their depth, § 7) and we have the flexibility to operate without them, should the number of sequences retrieved be too low. Lastly, the model has broader scope than any other model to date, while exceeding the performance of more specialized models in particular when extrapolating to mutated proteins that are further away in sequence space.

Our contributions are as follows:

- We introduce Tranception, a novel autoregressive transformer architecture that promotes specialization across attention heads for enhanced protein modeling (§3);
- We combine autoregressive predictions and homology from retrieved sequences at inference (§ 4) to reach state-of-the-art fitness prediction performance on both substitutions and indels (§ 6);
- We curate an extensive set of multiplexed assays of variant effects – the ProteinGym benchmarks – substantially increasing both the number and diversity of assays compared to existing benchmarks (Riesselman et al., 2018) (§5).

2. Background

2.1. Mutation effect prediction with aligned sequences

Predicting the effect of genetic variation using aligned protein sequences from diverse organisms, and in particular, predicting if a variant is likely to be disease-causing in humans, has a long history (Ng & Henikoff, 2001; Ramensky et al., 2002; Reva et al., 2011). While initial models focused on extracting position-specific information from alignments (Ng & Henikoff, 2001), subsequent work sought to capture more complex patterns. Hopf et al. (2017) proposed to model interactions between pairs of distinct positions with energy based models. Riesselman et al. (2018) later expanded on the concept with DeepSequence: Variational Autoencoders trained on protein-specific MSAs to learn a distribution of amino acid sequence which capture higher-order interactions. Focusing on predicting the pathogenicity of protein variants in human disease-related genes, EVE (Frazer et al., 2021) subsequently enhanced the DeepSequence architecture to reach higher fitness prediction performance.

2.2. Modeling proteins without alignments

While MSAs capture meaningful information about protein functions and structures, they also have certain limitations: not all proteins are alignable and, if they are, the depth of the corresponding alignments may not be enough to train models sufficiently large to learn the complex interactions between residues. This has led to a stream of research investigating alternative modeling approaches that do not rely on aligned sequences. Shin et al. (2021); Weinstein & Marks (2021) developed models that could be trained on non-aligned sequences, although they still relied on MSA routines to recover protein-specific sets of homologous sequences to serve as training data. Alley et al. (2019); Heinzinger et al. (2019) were the first to introduce models trained across protein families, relying on LSTM architectures (Hochreiter & Schmidhuber, 1997). Building on advances in the Natural Language Processing literature to train larger-scale language models, Madani et al. (2020); Rives et al. (2021); Nambiar et al. (2020) proposed the usage of transformer architectures to model protein sequences. Rao et al. (2020) introduced the MSA transformer, an architecture to learn a model of MSAs across thousands of protein families, while ESM-1v (Meier et al., 2021) and ProtTrans (Elnaggar et al., 2020) focused on learning patterns exclusively from non-aligned sequences from very large protein databases. Closely related models which combine unsupervised protein embedding with supervised data are seeing diverse applications, from supervised protein design tasks (Biswas et al., 2021), to task agnostic sequence representations (Bepler & Berger, 2019), to protein structure prediction (Jumper et al., 2021; Baek et al., 2021).

2.3. Deep Mutational Scanning benchmarks

Using a large number of Deep Mutational Scanning experiments (DMS) or Multiplex Assays of Variant Effects (MAVEs) to assess the performance of protein models was first proposed in Hopf et al. (2017), with a benchmark of $\sim$ 20 different assays. Riesselman et al. (2018) later doubled the size of this benchmark ($\sim$ 40 assays). The list of DMS assays curated to benchmark fitness predictors has seen only modest updates thereafter. More recent benchmarks for protein modeling (Rao et al., 2019; Dallago et al., 2021) have introduced additional assays focused on assessing model performance across a diverse set of downstream tasks. We include all assays related to fitness prediction from these prior benchmarks when building ProteinGym.

2.4. Retrieval

Retrieval aims at identifying objects related to a target one in a reference database to improve the processing or modeling of that object. In the Natural Language Processing literature, retrieval has been leveraged for open domain question answering (Robertson & Zaragoza, 2009; Wang et al., 2018; Karpukhin et al., 2020) or to augment pretrained language models to find relevant information in massive datasets at inference. Grave et al. (2017) and Khandelwal et al. (2020) both extended language models with a k-NN retrieval over pretrained embeddings at test time for question answering, while ORQA (Lee et al., 2019) and REALM (Guu et al., 2020) architectures jointly trained both the 'retriever' and 'reader' models end-to-end. RAG (Lewis et al., 2020) then applied similar concepts to the broader task of generative language modelling. More recently, RETRO (Borgeaud et al., 2021) demonstrated the benefits of retrieval at the scale of trillions of tokens. The bioinformatics literature has also heavily contributed to retrieval systems, in particular through Multiple Sequence Alignments (Thompson et al., 1994; Edgar, 2004; Sievers et al., 2011; Remmert et al., 2012) who may be used to both retrieve and align homologous sequences.

3. Tranception

Tranception is a novel autoregressive transformer architecture that was designed with two core principles in mind: 1) promoting specialization across attention heads 2) explicitly extracting patterns from contiguous subsequences.

3.1. Tranception attention

The concept of 'k-mers' is well-established in biological sequence analysis: k-mers are contiguous subsequences of k elements (typically nucleotides or amino acids) which have proved to be critically useful abstractions in several applications such as de novo assembly, read correction, repeat detection, comparison of genomes, metagenomes (Manekar & Sathe, 2018). The majority of protein language models to date (§ 2.2) have focused on extracting patterns (via sequence tokenization or attention mechanisms) at the amino acid level only. In this work we investigate the benefits from explicitly attending over contiguous subsequences of amino acid tokens via a novel attention mechanism – Tranception attention (Fig. 1) – which combines ideas from Primer (So et al., 2021) and Inception (Szegedy et al., 2014) networks. Similar to Primer, we leverage squared ReLU activations and depthwise convolutions after the different multi-head attention projections. Instead of using similar-sized kernels for each depthwise convolution, we split the attention heads at each layer in 4 groups and apply convolutions with different kernel sizes on each group, thereby combining information at different resolutions as in Inception. This incentivizes each attention head to specialize to pattern extractions at different k-mer sizes and leads to both more efficient training and downstream task performance compared with Primer and GPT2 (Radford et al., 2018) (Fig. 2 and Appendix B.1).

3.2. Grouped ALiBi position encoding

In order to further promote specialization across attention heads and enhance predictions for protein sequences that

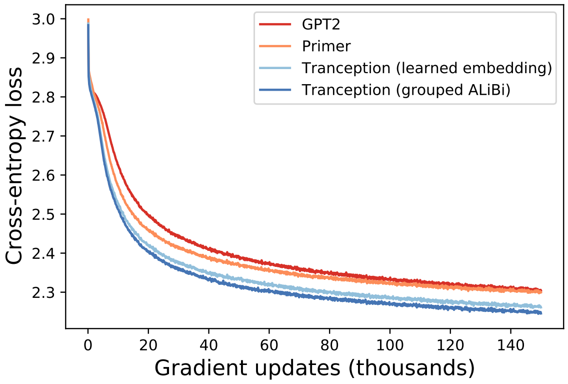

Figure 2. **Training loss comparison across transformer architectures for protein modeling.** We plot the training loss as a function of the number of gradient steps for GPT2 (Radford et al., 2018), Primer (So et al., 2021), Tranception with learned position embeddings and Tranception with grouped AliBi. All models have similar number of parameters and only differ by they attention mechanism, non-linear activations and position encodings. Tranception converges faster and to a lower loss compared with other architectures. This translates into higher downstream task performance (Appendix B.1).

are longer than the context length, we replace the learned or sinusoidal position encodings typically used in autoregressive transformer architectures (Vaswani et al., 2017; Brown et al., 2020), with a variant of ALiBi (Press et al., 2021) called 'Grouped ALiBi'. Similar to ALiBi, we remove the position encodings added to input token embeddings, and bias the query-key attention scores with a term proportional to their distance. We however apply the mechanism on each group of attention heads independently, in adequacy with the Tranception attention scheme, to enable each group to learn attention patterns at different distances. Compared to using learned position encodings, this helps reduce the number of parameters, converge faster during training and leads to better downstream task performance (Fig. 2 and Appendix B.1).

3.3. Data processing and augmentations

Our models are trained on UniRef (Suzek et al., 2014), a large scale protein sequence database. We perform thorough ablations when developing Tranception (Appendix B.1). Similar to Meier et al. (2021), we investigate the impact of training on protein sequences clustered at different levels of similarity. Unlike what was observed for masked-language model architectures, we find that keeping as much of the granularity available in the dataset is beneficial to downstream task performance. We therefore train our final model (700M parameters) on UniRef100 which, after preprocessing (Appendix B.2), leads to a training dataset of ~250 million protein sequences. Our vocabulary is comprised of the standard 20 amino acids (Kessel & Ben-Tal, 2018). We find that averaging the predictions obtained by scoring each sequence and its reverse at inference time leads to higher downstream performance (Appendix B.4), and therefore apply sequence mirroring at random during training to teach our model to score sequences from both directions. Our model has a maximum context size of 1024 tokens, which is wider than the length of 98% of protein sequences in UniRef100 (Table 7). At train time, if a protein is longer than the maximum context size of the model (after accounting for the special start and end of sequence tokens), we extract a randomly-selected contiguous slice of width equal to that maximum context size. Indeterminate amino acids are imputed at random during training and inference.

3.4. Scoring sequences for fitness prediction

The goal of fitness prediction is to assess the effects of mutations (eg., amino acid substitutions, insertions or deletions) on the ability of the corresponding mutated protein sequence to perform its function. A common approach to estimating mutation effects is to quantify the likelihood ratio between the mutated sequence and a naturally occurring reference sequence for that protein family, referred to as the 'wild-type' sequence (Riesselman et al., 2018). More formally, we represent each protein x as a sequence of amino acids $(x_1, x_2, ..., x_l)$. Our model is trained in a self-supervised fashion to predict the next token x_i in the sequence based on the context of the prior $i-1$ tokens, such that the probability of the full sequence factorizes as:

$$P(x) = \prod_{i=1}^{l} P(x_i|x_1, ..., x_{i-1}) = \prod_{i=1}^{l} P(x_i|x_{<i}) \qquad (1)$$

The fitness F_x of a given mutated protein x^{mut} is then measured via the log-likelihood ratio with the wild-type sequence x^{wt}:

$$F_x = \log \frac{P(x^{mut})}{P(x^{wt})} \qquad (2)$$

When assessing fitness in sequences longer than the context length, we select the sequence slice providing the widest left and right context for the set of mutations considered (Appendix 3.4). At inference time and building on our data augmentations, we take the arithmetic average of the log-likelihood ratios obtained by scoring each sequence and its reverse.

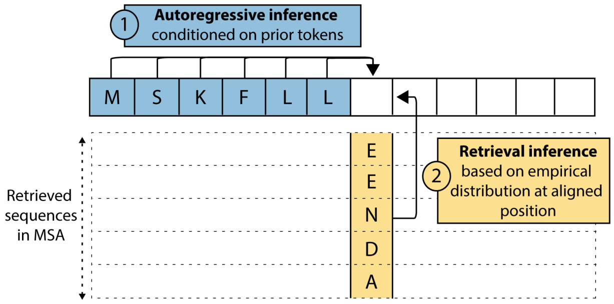
1 Autoregressive inference
conditioned on prior tokens
M
S
K
F
L
L
Retrieved sequences in MSA
E
E
N
D
A
2 Retrieval inference
based on empirical distribution at aligned position

Figure 3. **Combining autoregressive inference and retrieval inference.** Predictions in Tranception are based on two complementary modes of inference: autoregressive predictions based on the context of previously generated tokens and predictions based on the empirical distribution of amino acid at each position in the retrieved set of homologous sequences.

4. Inference-time retrieval

4.1. Multiple sequence alignments

Multiple sequence alignments retrieve neighboring proteins in sequence space and align them in the position coordinate system of the seed sequence. The overwhelming majority of fitness prediction models rely on MSAs as they inherently capture critical information about homology, phylogeny, and 3D structure of the corresponding protein family (Thompson et al., 1994; 1997). At a given position, the observed distribution of amino acids over sequences in the MSA recapitulates evolutionary constraints: the protein sequences that are part of the MSA are the variants that maintain fitness and that were not selected out by evolution.

4.2. Two modes of inference

We propose to augment the autoregressive inference mode from Tranception with a second inference mode – retrieval inference (Fig. 3) – that is based on the empirical distribution of amino acids observed across sequences in the MSA. To that end, the first step consists of retrieving a MSA at inference time for a wild-type sequence from the protein family of interest. When focusing on amino acid substitutions, the retrieved set of homologous sequences is common to both the wild-type and the mutated sequences: we do a retrieval step once per family, and amortize the cost over all mutated sequences to be scored. When dealing with insertions and deletions, we tailor the retrieved MSA to each mutated sequence by deleting columns in the MSA corresponding to deleted positions and adding zero-filled columns in the MSA at inserted positions in the mutated protein. At inference time, the inserted columns or fully non-covered positions are ignored and the model solely relies on its autoregressive mode to make predictions at these positions. The second step is to compute the empirical distribution of amino acids for each aligned position based on pseudocounts (ignoring gaps) and Laplace smoothing (Jurafsky & Martin, 2008) (Appendix B.5). Since the distribution of sequences found in protein databases is biased by human sampling (certain organisms are more studied than others) and evolutionary sampling (groups of species that arose from large radiations will have over-represented proteins), we re-weigh sequences in the MSA using the scheme described in Hopf et al. (2017). Finally, we estimate the log likelihood $\log P(x)$ for a protein sequence x by a weighted arithmetic average of the log likelihood $\log P_A(x)$ from the autoregressive inference mode and the log likelihood $\log P_R(x)$ obtained from the retrieval inference mode. This can be equivalently viewed as a weighted geometric average in probability space, and form a proper probability distribution up to a normalization constant:

$$\log P(x) = \frac{1}{C}[(1-\alpha)\log P_A(x) + \alpha \log P_R(x)] \quad (3)$$

The normalization constant C cancels out when computing the log likelihood ratio as per equation 2, and thus we do not need to estimate it in practice. Using equation 1, we further obtain:

$$\log P(x) \propto \sum_{i=1}^{l} [(1-\alpha)\log P_A(x_i|x_{<i}) + \alpha \log P_R(x_i)] \quad (4)$$

The above scoring decomposition as the sum of position level scores is advantageous in practice as it allows us to ignore one of the two inference modes as needed (eg., ignoring retrieval inference at positions with insertions when scoring indels). Following the scoring procedure described in § 3.4, we first traverse the sequence in the canonical order from left to right and compute the sequence log probability as per equation4 (in practice, all computations are performed in parallel across positions using masks in the self-attention layers to preserve autoregressiveness). We then perform the symmetric operation traversing the protein sequence from right to left, and finally average the two log likelihood ratios. Retrieving the MSA and computing the corresponding pseudocounts at inference is relatively cheap computationally. Since we only rely on aggregate statistics per position, our proposed approach is not very sensitive to the characteristics of the MSA for the protein family of interest (Fig. 4), nor to the hyperparameters chosen to retrieve that MSA. As Tranception is trained on a diverse set of non-aligned sequences it is not subject to the biases that may stem from solely training on proteins that can be aligned, and thus exhibits higher performance on difficult to align sequences such as disordered proteins (§ 7).

5. ProteinGym

ProteinGym is an extensive set of Deep Mutational Scanning (DMS) assays (Appendix A) curated to enable thorough comparisons of various mutation effect predictors in different regimes. ProteinGym is comprised of two benchmarks: 1) a *substitution benchmark* which consists of the experimental characterisation of $\sim$1.5M missense variants across 87 DMS assays 2) an *indel benchmark* that includes $\sim$300k mutants across 7 DMS assays.

The relationship between protein fitness measured experimentally and as reflected by the distribution of sequences selected by evolution is complex. The fitness landscape of naturally occurring proteins is the result of an intricate set of overlapping constraints that these proteins are subjected to in an organism. Consequently, it is often challenging to identify a single molecular property that is both easy to measure experimentally and that reflects that complexity. To build this benchmark we therefore prioritized assays where both the experimentally-measured property for each mutated protein is expected to reflect the role of the protein in organism fitness as well as - where available - their quality measured via experimental replicates. The resulting set of DMS assays covers a wide range of functional properties (eg., thermostability, ligand binding, aggregation, viral replication, drug resistance), spans a diverse protein families (eg., kinases, ion channel proteins, g-protein coupled receptors, polymerases, transcription factors, tumor suppressors) and different taxa (eg., humans, other eukaryotes, prokaryotes, viruses).

ProteinGym is the largest and most diverse set of DMS experiments specifically targeted at the task of variant effect prediction. It contains more than twice the number of assays and variants present in the DeepSequence benchmark (Riesselman et al., 2018), which had been created for the same purpose (Table 1). While most of the curated DMS assays (in our benchmark or others) probe the effect of single amino-acid substitutions, our collection also includes several multiple amino-acid variants which are critical to assess the ability of models to extrapolate further away in sequence space from the naturally occurring proteins they are trained on. Lastly, as most mutation effect predictors are not able to quantify the effect of insertions and deletions, indels have been absent from the majority of prior benchmarks. We expand on the set of DMS available in Shin et al. (2021) and Dallago et al. (2021) to address this gap.

The relationship between protein function and organism fitness has been shown to often be non-linear (Boucher et al., 2016). As such, we use Spearman's rank correlation coefficient between model scores and the experimental measurements as the standard measure of model performance (Riesselman et al., 2018; Meier et al., 2021). In certain instances, the DMS measurements are characterized by a bimodal profile for which rank correlations are not well suited. To that end, we provide additional measures of model performance: the area under the ROC curve (AUC) and Matthews correlation coefficient (MCC) between model scores and the experimental measurements (Appendix E.1).

6. Results

6.1. Baselines

We compare the ability of various models to predict the effects of mutations across the DMS assays in ProteinGym. We focus on the main approaches described in § 2, including a number of protein-specific alignment-based methods – Site independent model and EVmutation (Hopf et al., 2017), DeepSequence (Riesselman et al., 2018), EVE (Frazer et al., 2021) – and large-scale protein language models trained across protein families that leverage alignments during training, such as the MSA Transformer (Rao et al., 2021) or that are alignment-free, such as ESM-1v (Meier et al., 2021). Although technically trained on subsets of unaligned sequences, Wavenet (Shin et al., 2021) models are protein-specific and use MSAs to extract their training data (as such, we group them with other alignment-based models). Tranception falls in the category of large-scale protein language models trained across families and, thanks to its two modes of inference, can be seen as a hybrid between ESM-1v and the MSA transformer. As in ESM-1v, it is trained on a large set of unaligned sequences, which makes the training procedure scalable and removes biases that would result from training on alignable proteins only. Similar to the MSA transformer, it leverages the information in a retrieved MSA to enhance fitness predictions. The critical difference is that Tranception is never trained on MSAs and is therefore less sensitive to their characteristics and limitations (Fig. 4). To allow fair comparisons across models we focus on single seed scoring only, but provide additional results with model ensembles in Appendix E.5.

6.2. ProteinGym substitution benchmark

We compute the Spearman's rank correlation coefficient ρ, AUC and MCC between model scores and experimental measurements for all DMS assays in the ProteinGym substitution benchmark (Appendix E.2, Fig. 6), and draw very similar conclusions across metrics. Tranception with retrieval outperforms all other baselines on the overall benchmark, with markedly higher performance in the regime of low-depth MSAs (Table 2) and on multiple mutants (Table 3). When analyzing performance at the taxon level (Table 11), we observe consistently high performance from Tranception across categories, in particular on human proteins and other eukaryotes. This high performance on human proteins has immediate clinical applications, since Tranception outperforms EVE and directly extends to modelling the entire human proteome, while EVE models need to be trained for each new protein of interest and are available for only $\sim$3k proteins at the time of writing. Without retrieval, Trancep-

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**

Table 1. **Comparison of the ProteinGym and DeepSequence benchmarks**. ProteinGym contains a substantially higher number of assays and variants compared to DeepSequence. It addresses notable gaps such as the limited number of viral DMS assays, limited multiple substitutions assays and absence of indels benchmark.

Model type	Model name	Spearman's rank correlation by MSA depth ↑				AUC ↑
		Low	Medium	High	All	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**

Table 2. **Average AUC and Spearman's rank correlation between model scores and experimental measurements by MSA depth on the ProteinGym substitution benchmark.** Alignment depth for the various proteins is measured by the ratio of the effective number of sequences N_{eff} in the MSA, using the same weighting scheme as in (Hopf et al., 2017), by the length covered L: shallow alignments (ie., 'Low' group) typically have a low value of N_{eff}/L. Specifically, Low: N_{eff}/L <1; Medium: 1< N_{eff}/L <100; High: N_{eff}/L >100. Tranception outperforms all other baselines overall, with the largest performance gaps observed on the low depth proteins.

tion outperforms ESM-1v, the only baseline which also does not leverage alignments for inference. The performance lift is particularly significant on proteins with shallow alignments, on multiple mutations and viral proteins.

6.3. ProteinGym indel benchmark

We report the performance metrics (Spearman, AUC, MCC) on the ProteinGym indel benchmark (Table 4) and compare the performance of Tranception with Wavenet, the only baseline from the set described in § 6.1 that is able to quantify the effects of deletions or insertions. Other alignment-based models are constrained by the fixed coordinate system from the original MSA they have been trained on. ESM-1v and MSA transformer both rely on a scoring heuristic (AppendixD) that requires the mutated position to exist in the wild-type sequence. On the indel benchmark, Tranception outperforms Wavenet both with and without retrieval.

7. Discussion

Gain of scope for proteins with shallow alignments. Tranception with retrieval outperforms other mutation effects predictors on the ProteinGym substitutions and indels benchmarks. While retrieval is key to achieve such results, our method only relies on high level statistics of the retrieved alignments. It is thus fairly robust to the various hyperparameters selected to generate these alignments – even when they result in relatively shallow alignments. When progressively removing sequences in the MSA based on their minimum similarity to the seed sequence (Fig. 4), we observe that Tranception has consistently high performance, unlike EVE or MSA Transformer. Given that one can obtain shallow MSAs for far more proteins than deep MSAs, our lightweight inference-time retrieval can be effectively leveraged to enhance mutation effect predictions for the vast majority of proteins. Disordered proteins are examples of

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**

Table 3. **Average Spearman's rank correlation between model scores and experimental measurements by mutation depth.** Mutation depth is measured by the number of distinct substitutions compared with the wild-type sequence.

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

Table 4. **Average AUC and Spearman's rank correlation between model scores and experimental measurements on the ProteinGym indel benchmark**.

proteins that are notoriously difficult to align. The ProteinGym substitution benchmark includes two proteins with disordered regions, A4 and GCN4. Tranception markedly outperforms all baselines on these two assays (Fig. 6). Finally, our approach benefits from the additional flexibility to fully ignore the retrieval-based mode of inference when scoring positions that are not present in the MSAs (eg., indels) and proteins that are not alignable or for which the corresponding set of homologous sequences is extremely shallow (eg., a few sequences).

Gain of coverage. A major benefit from Tranception is gain of coverage. For example, BRCA1, the gene encoding Breast cancer type 1 susceptibility protein, has 1863 amino acids making it challenging to model with alignment based methods. Trying to obtain an alignment for the full protein results in poor diversity. Not only alignment-based models (eg., EVE) obtain relatively weak performance when trained on such alignments (Table 15), but they are unable to make predictions for all mutations impacting regions of the protein with insufficient coverage (see Appendix D for more details on these limitations). We know from previous work (Frazer et al., 2021) that it is possible to obtain higher quality alignments and model performance when dealing with the two RING and BRCT domains of BRCA1 separately. But this comes at the cost of not being able to make predictions for the majority of the protein (outside of these two domains), and ignore potential dependencies across domains. Tranception is not subject to these coverage limitations and is able to score all potential mutations of the protein. It obtains relatively high performance without retrieval, yet benefits from leveraging the retrieval inference mode based on full-protein alignment or domain-specific alignments, with slightly higher performance for the latter.

Extrapolating far away in sequence space Compared with other baselines, Tranception obtains much higher performance on multiple mutants, with the gap widening with mutation depth. While we sought to include as many assays with multiple mutants as possible when building ProteinGym, our benchmarks are still relatively biased towards single mutant assays: on the substitution benchmark, only 11 assays out of 87 included multiple mutants. Since the actual number of multiple mutants is exponentially higher than the number of single mutants for any protein family, it is likely that the reported gap between Tranception and other baselines would only increase on a benchmark with a higher proportion of multiple mutant assays. Performing well on multiple mutants has important ramifications in several applications, in particular in machine-learning guided protein design for which accurate extrapolation of sequence likelihood far away in sequence space is critical to uncover novel desirable candidate proteins.

8. Conclusion

Models of the effects of genetic variations are emerging as powerful tools with diverse applications – from quantifying genetic predispositions to certain pathologies to designing novel proteins. Yet our understanding of how to build these models is still in its infancy. To make progress, we need methods that can scale and perform well across the whole protein universe, including regions that are hard to align or that are recent in evolutionary terms. The model we present

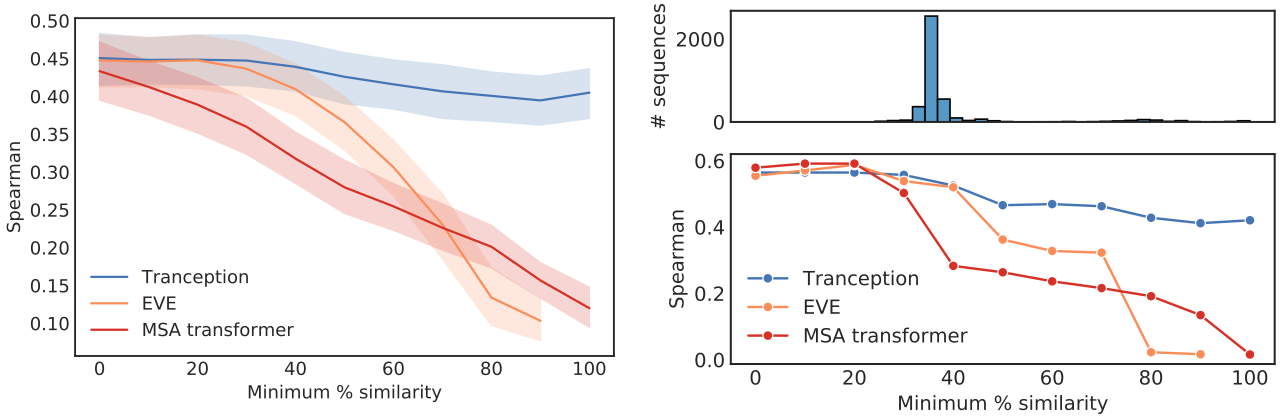

Figure 4. **Robustness to alignment depth.** We measure the average performance (Spearman's correlation with DMS measurements) of Tranception, EVE and MSA Transformer as we filter out an increasing proportion of MSA sequences based on their similarity to the seed sequence. The left figure aggregates results across all 87 substitution assays in ProteinGym, the right figure focuses on the tumor protein P53 (Kotler et al. (2018) assay; other examples are provided in Appendix E.3). The performance of Tranception is robust to MSA depth, while that of EVE and MSA transformer drops significantly as diversity in the MSA is reduced. Rightmost points for Tranception and MSA transformer correspond to performance with no retrieval and with a single-sequence input MSA (the seed) respectively. On the right figure, the histogram reports the number of sequences in the MSA per similarity to the seed sequence groupings.

in this work makes progress in this direction on five main aspects.

Firstly, the combination of the Tranception model architecture together with retrieval at inference delivers state-of-the-art fitness prediction performance, with a significantly stronger ability to extrapolate to multiple mutants. Scaling up the size of our transformer (Hesslow et al., 2022), together with training on a larger and more diverse set of protein sequences (Mitchell et al., 2020; Steinegger & Söding, 2018) will likely improve our performance further.

Secondly, our suggested retrieval at inference approach is fairly robust to alignment characteristics, and Tranception performs well with retrieval of just the nearest homologs. Our model can make use of deep alignments when they are available, and small, or no alignments when need be, resulting in both high-performance and broad scope. This is a significant advantage over alignment-based methods like EVE which require deep enough alignments to capture the complex relationships across residues in the protein sequence of interest.

Thirdly, unlike most existing mutation effect predictors, Tranception is able to handle insertions and deletions out-of-the-box, and outperforms prior baselines in that regime as well.

Fourthly, we find our approach to be robust across taxa and protein families, making it well suited to a broad range of tasks. This includes predicting the effect of mutations in viruses, a key component of forecasting outbreaks, and predicting disease causing variants in humans, of value to both diagnosis and preventative care.

Finally, our autoregressive model is naturally suited to sequence generation and hence has great potential for protein design.

Acknowledgements

We thank Lood Van Niekerk and Aaron Kollasch for their help running the DeepSequence and Wavenet baselines on the ProteinGym benchmark, the broader OATML group and Marks Lab for helpful discussions when writing this manuscript, and the Google Cloud Platform team for research credits and TRC access when developing and training our models. P.N. is supported by GSK and the UK Engineering and Physical Sciences Research Council (EPSRC ICASE award no. 18000077). M.D., J.F. and J.M.H. are supported by the Chan Zuckerberg Initiative CZI2018-191853. A.G. is a Clarendon Scholar and Open Philanthropy AI Fellow. D.S.M. holds a Ben Barres Early Career Award by the Chan Zuckerberg Initiative as part of the Neurodegeneration Challenge Network, CZI2018-191853. Y.G. holds a Turing AI Fellowship (Phase 1) at the Alan Turing Institute, which is supported by EPSRC grant reference V030302/1.

References

Aakre, C., Herrou, J., Phung, T., Perchuk, B., Crosson, S., and Laub, M. Evolving New Protein-Protein Interaction Specificity through Promiscuous Intermediates. *Cell*, 163(3):594–606, October 2015. ISSN 00928674. doi: 10.1016/j.cell.2015.

09.055. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867415012726.

Adkar, B., Tripathi, A., Sahoo, A., Bajaj, K., Goswami, D., Chakrabarti, P., Swarnkar, M., Gokhale, R., and Varadarajan, R. Protein Model Discrimination Using Mutational Sensitivity Derived from Deep Sequencing. *Structure*, 20(2):371–381, February 2012. ISSN 09692126. doi: 10.1016/j.str.2011.11.021. URL https://linkinghub.elsevier.com/retrieve/pii/S0969212612000068.

Ahler, E., Register, A. C., Chakraborty, S., Fang, L., Dieter, E. M., Sitko, K. A., Vidadala, R. S. R., Trevillian, B. M., Golkowski, M., Gelman, H., Stephany, J. J., Rubin, A. F., Merritt, E. A., Fowler, D. M., and Maly, D. J. A Combined Approach Reveals a Regulatory Mechanism Coupling Src's Kinase Activity, Localization, and Phosphotransferase-Independent Functions. *Molecular Cell*, 74(2):393–408.e20, April 2019. ISSN 10972765. doi: 10.1016/j.molcel.2019.02.003. URL https://linkinghub.elsevier.com/retrieve/pii/S1097276519300930.

Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., and Church, G. M. Unified rational protein engineering with sequence-only deep representation learning. *bioRxiv*, pp. 589333, 2019.

Amorosi, C. J., Chiasson, M. A., McDonald, M. G., Wong, L. H., Sitko, K. A., Boyle, G., Kowalski, J. P., Rettie, A. E., Fowler, D. M., and Dunham, M. J. Massively parallel characterization of CYP2C9 variant enzyme activity and abundance. *The American Journal of Human Genetics*, 108(9):1735–1751, September 2021. ISSN 00029297. doi: 10.1016/j.ajhg.2021.07.001. URL https://linkinghub.elsevier.com/retrieve/pii/S000292972100269X.

Araya, C. L., Fowler, D. M., Chen, W., Muniez, I., Kelly, J. W., and Fields, S. A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proceedings of the National Academy of Sciences*, 109(42):16858–16863, October 2012. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1209751109. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1209751109.

Baek, M., Dimaio, F., Anishchenko, I. V., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A. M., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., Opperman, D. J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M. K., Dalwadi, U., Yip, C. K., Burke, J. E., Garcia, K. C., Grishin, N. V., Adams, P. D., Read, R. J., and Baker, D. Accurate prediction of protein structures and interactions using a 3-track neural network. *Science (New York, N.Y.)*, 373:871 – 876, 2021.

Bandaru, P., Shah, N. H., Bhattacharyya, M., Barton, J. P., Kondo, Y., Cofsky, J. C., Gee, C. L., Chakraborty, A. K., Kortemme, T., Ranganathan, R., and Kuriyan, J. Deconstruction of the Ras switching cycle through saturation mutagenesis. *eLife*, 6:e27810, July 2017. ISSN 2050-084X. doi: 10.7554/eLife.27810. URL https://elifesciences.org/articles/27810.

Bepler, T. and Berger, B. Learning protein sequence embeddings using information from structure. *arXiv preprint arXiv:1902.08661*, 2019.

Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M., and Church, G. M. Low-n protein engineering with data-efficient deep learning. *Nature Methods*, 18(4):389–396, 2021.

Bolognesi, B., Faure, A. J., Seuma, M., Schmiedel, J. M., Tartaglia, G. G., and Lehner, B. The mutational landscape of a prion-like domain. *Nature communications*, 10(1):1–12, 2019.

Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G., Lespiau, J.-B., Damoc, B., Clark, A., de Las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan, T. W., Huang, S., Maggiore, L., Jones, C., Cassirer, A., Brock, A., Paganini, M., Irving, G., Vinyals, O., Osindero, S., Simonyan, K., Rae, J. W., Elsen, E., and Sifre, L. Improving language models by retrieving from trillions of tokens. *ArXiv*, abs/2112.04426, 2021.

Boucher, J. I., Bolon, D. N., and Tawfik, D. S. Quantifying and understanding the fitness effects of protein mutations: Laboratory versus nature. *Protein Science*, 25(7):1219–1226, 2016.

Brenan, L., Andreev, A., Cohen, O., Pantel, S., Kamburov, A., Cacchiarelli, D., Persky, N., Zhu, C., Bagul, M., Goetz, E., Burgin, A., Garraway, L., Getz, G., Mikkelsen, T., Piccioni, F., Root, D., and Johannessen, C. Phenotypic Characterization of a Comprehensive Set of MAPK1 /ERK2 Missense Mutants. *Cell Reports*, 17(4):1171–1183, October 2016. ISSN 22111247. doi: 10.1016/j.celrep.2016.09.061. URL https://linkinghub.elsevier.com/retrieve/pii/S2211124716313171.

Bridgford, J. L., Lee, S. M., Lee, C. M. M., Guglielmelli, P., Rumi, E., Pietra, D., Wilcox, S., Chhabra, Y., Rubin, A. F., Cazzola, M., Vannucchi, A. M., Brooks, A. J., Call, M. E., and Call, M. J. Novel drivers and modifiers of MPL-dependent oncogenic transformation identified by deep mutational scanning. *Blood*, 135(4):287–292, January 2020. ISSN 0006-4971, 1528-0020. doi: 10.1182/blood.2019002561. URL https://ashpublications.org/blood/article/135/4/287/381157/Novel-drivers-and-modifiers-of-MPLdependent.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners, 2020.

Chan, Y. H., Venev, S. V., Zeldovich, K. B., and Matthews, C. R. Correlation of fitness landscapes from three orthologous TIM barrels originates from sequence and structure constraints. *Nature Communications*, 8(1): 14614, April 2017. ISSN 2041-1723. doi: 10.1038/ncomms14614. URL http://www.nature.com/articles/ncomms14614.

Chen, J. Z., Fowler, D. M., and Tokuriki, N. Comprehensive exploration of the translocation, stability and substrate recognition requirements in VIM-2 lactamase. *eLife*, 9:e56707, June 2020. ISSN 2050-084X. doi: 10.7554/eLife.56707. URL https://elifesciences.org/articles/56707.

Chiasson, M. A., Rollins, N. J., Stephany, J. J., Sitko, K. A., Matreyek, K. A., Verby, M., Sun, S., Roth, F. P., DeSloover, D., Marks, D. S., Rettie, A. E., and Fowler, D. M. Multiplexed measurement of variant abundance and activity reveals VKOR topology, active site and human variant impact. *eLife*, 9: e58026, September 2020. ISSN 2050-084X. doi: 10.7554/eLife.58026. URL https://elifesciences.org/articles/58026.

Dallago, C., Mou, J., Johnston, K. E., Wittmann, B. J., Bhattacharya, N., Goldman, S., Madani, A., and Yang, K. K. Flip: Benchmark tasks in fitness landscape inference for proteins. 2021.

Dandage, R., Pandey, R., Jayaraj, G., Rai, M., Berger, D., and Chakraborty, K. Differential strengths of molecular determinants guide environment specific mutational fates. *PLOS Genetics*, 14(5):e1007419, May 2018. ISSN 1553-7404. doi: 10.1371/journal.pgen.1007419. URL https://dx.plos.org/10.1371/journal.pgen.1007419.

Davidi, D., Shamshoum, M., Guo, Z., Bar-On, Y. M., Prywes, N., Oz, A., Jablonska, J., Flamholz, A., Wernick, D. G., Antonovsky, N., et al. Highly active rubiscos discovered by systematic interrogation of natural sequence diversity. *The EMBO journal*, 39(18):e104081, 2020.

Deng, Z., Huang, W., Bakkalbasi, E., Brown, N. G., Adamski, C. J., Rice, K., Muzny, D., Gibbs, R. A., and Palzkill, T. Deep Sequencing of Systematic Combinatorial Libraries Reveals β-Lactamase Sequence Constraints at High Resolution. *Journal of Molecular Biology*, 424(3-4):150–167, December 2012. ISSN 00222836. doi: 10.1016/j.jmb.2012.09.014. URL https://linkinghub.elsevier.com/retrieve/pii/S0022283612007711.

Doud, M. and Bloom, J. Accurate Measurement of the Effects of All Amino-Acid Mutations on Influenza Hemagglutinin. *Viruses*, 8(6):155, June 2016. ISSN 1999-4915. doi: 10.3390/v8060155. URL http://www.mdpi.com/1999-4915/8/6/155.

Doud, M. B., Ashenberg, O., and Bloom, J. D. Site-Specific Amino Acid Preferences Are Mostly Conserved in Two Closely Related Protein Homologs. *Molecular Biology and Evolution*, 32(11):2944–2960, November 2015. ISSN 0737-4038, 1537-1719. doi: 10.1093/molbev/msv167. URL https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msv167.

Duenas-Decamp, M., Jiang, L., Bolon, D., and Clapham, P. R. Saturation mutagenesis of the hiv-1 envelope cd4 binding loop reveals residues controlling distinct trimer conformations. *PLoS pathogens*, 12(11):e1005988, 2016.

Eddy, S. R. Accelerated profile hmm searches. *PLoS computational biology*, 7(10):e1002195, 2011.

Edgar, R. C. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32 5:1792–7, 2004.

Elnaggar, A., Heinzinger, M., Dallago, C., Rihawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al. Prottrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing. *arXiv preprint arXiv:2007.06225*, 2020.

Faure, A. J., Domingo, J., Schmiedel, J. M., Hidalgo-Carcedo, C., Diss, G., and Lehner, B. Mapping the energetic and allosteric landscapes of protein binding domains. *Nature*, 604(7904):175–183, 2022.

Fernandes, J. D., Faust, T. B., Strauli, N. B., Smith, C., Crosby, D. C., Nakamura, R. L., Hernandez, R. D., and Frankel, A. D. Functional Segregation of Overlapping Genes in HIV. *Cell*, 167(7):1762–1773.e12, December 2016. ISSN 00928674. doi: 10.1016/j.cell.2016.

11.031. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867416316038.

Findlay, G. M., Daza, R. M., Martin, B., Zhang, M. D., Leith, A. P., Gasperini, M., Janizek, J. D., Huang, X., Starita, L. M., and Shendure, J. Accurate classification of BRCA1 variants with saturation genome editing. *Nature*, 562(7726):217–222, October 2018. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-018-0461-z. URL http://www.nature.com/articles/s41586-018-0461-z.

Firnberg, E., Labonte, J. W., Gray, J. J., and Ostermeier, M. A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape. *Molecular Biology and Evolution*, 31 (6):1581–1592, June 2014. ISSN 1537-1719, 0737-4038. doi: 10.1093/molbev/msu081. URL https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msu081.

Flynn, J. M., Rossouw, A., Cote-Hammarlof, P., Fragata, I., Mavor, D., Hollins, C., Bank, C., and Bolon, D. N. Comprehensive fitness maps of Hsp90 show widespread environmental dependence. *eLife*, 9:e53810, March 2020. ISSN 2050-084X. doi: 10.7554/eLife.53810. URL https://elifesciences.org/articles/53810.

Fowler, D. M. and Fields, S. Deep mutational scanning: a new style of protein science. *Nature methods*, 11(8): 801–807, 2014.

Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J. K., Brock, K. P., Gal, Y., and Marks, D. S. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 2021.

Giacomelli, A. O., Yang, X., Lintner, R. E., McFarland, J. M., Duby, M., Kim, J., Howard, T. P., Takeda, D. Y., Ly, S. H., Kim, E., Gannon, H. S., Hurhula, B., Sharpe, T., Goodale, A., Fritchman, B., Steelman, S., Vazquez, F., Tsherniak, A., Aguirre, A. J., Doench, J. G., Piccioni, F., Roberts, C. W. M., Meyerson, M., Getz, G., Johannessen, C. M., Root, D. E., and Hahn, W. C. Mutational processes shape the landscape of TP53 mutations in human cancer. *Nature Genetics*, 50(10):1381–1387, October 2018. ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-018-0204-y. URL http://www.nature.com/articles/s41588-018-0204-y.

Glazer, A. M., Kroncke, B. M., Matreyek, K. A., Yang, T., Wada, Y., Shields, T., Salem, J.-E., Fowler, D. M., and Roden, D. M. Deep Mutational Scan of an *SCN5A* Voltage Sensor. *Circulation: Genomic and Precision Medicine*, 13(1), February 2020. ISSN 2574-8300. doi: 10.1161/CIRCGEN.119.002786. URL https://www.ahajournals.org/doi/10.1161/CIRCGEN.119.002786.

Gonzalez, C. E., Roberts, P., and Ostermeier, M. Fitness effects of single amino acid insertions and deletions in tem-1 β-lactamase. *Journal of molecular biology*, 431 (12):2320–2330, 2019.

Grave, E., Cissé, M., and Joulin, A. Unbounded cache model for online language modeling with open vocabulary. In *NIPS*, 2017.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M.-W. Realm: Retrieval-augmented language model pretraining. *ArXiv*, abs/2002.08909, 2020.

Haddox, H. K., Dingens, A. S., and Bloom, J. D. Experimental Estimation of the Effects of All Amino-Acid Mutations to HIV's Envelope Protein on Viral Replication in Cell Culture. *PLOS Pathogens*, 12(12):e1006114, December 2016. ISSN 1553-7374. doi: 10.1371/journal.ppat.1006114. URL https://dx.plos.org/10.1371/journal.ppat.1006114.

Haddox, H. K., Dingens, A. S., Hilton, S. K., Overbaugh, J., and Bloom, J. D. Mapping mutational effects along the evolutionary landscape of HIV envelope. *eLife*, 7:e34420, March 2018. ISSN 2050-084X. doi: 10.7554/eLife.34420. URL https://elifesciences.org/articles/34420.

Heinzinger, M., Elnaggar, A., Wang, Y., Dallago, C., Nechaev, D., Matthes, F., and Rost, B. Modeling the language of life–deep learning protein sequences. *bioRxiv*, pp. 614313, 2019.

Hesslow, D., ed. Zanichelli, N., Notin, P., Poli, I., and Marks, D. S. Rita: a study on scaling up generative protein sequence models. *ArXiv*, abs/2205.05789, 2022.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Computation*, 9:1735–1780, 1997.

Hopf, T. A., Schärfe, C. P., Rodrigues, J. P., Green, A. G., Kohlbacher, O., Sander, C., Bonvin, A. M., and Marks, D. S. Sequence co-evolution gives 3d contacts and structures of protein complexes. *Elife*, 3:e03430, 2014.

Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Schärfe, C. P., Springer, M., Sander, C., and Marks, D. S. Mutation effects predicted from sequence co-variation. *Nature biotechnology*, 35(2):128–135, 2017.

Jacquier, H., Birgy, A., Le Nagard, H., Mechulam, Y., Schmitt, E., Glodt, J., Bercot, B., Petit, E., Poulain, J., Barnaud, G., Gros, P.-A., and Tenaillon, O. Capturing the mutational landscape of the beta-lactamase TEM-1. *Proceedings of the National*

Academy of Sciences, 110(32):13067–13072, August 2013. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1215206110. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1215206110.

Jia, X., Burugula, B. B., Chen, V., Lemons, R. M., Jayakody, S., Maksutova, M., and Kitzman, J. O. Massively parallel functional testing of MSH2 missense variants conferring Lynch syndrome risk. *The American Journal of Human Genetics*, 108(1):163–175, January 2021. ISSN 00029297. doi: 10.1016/j.ajhg.2020.12.003. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929720304390.

Jiang, L., Liu, P., Bank, C., Renzette, N., Prachanronarong, K., Yilmaz, L. S., Caffrey, D. R., Zeldovich, K. B., Schiffer, C. A., Kowalik, T. F., et al. A balance between inhibitor binding and substrate processing confers influenza drug resistance. *Journal of molecular biology*, 428(3): 538–553, 2016.

Jones, E. M., Lubock, N. B., Venkatakrishnan, A., Wang, J., Tseng, A. M., Paggi, J. M., Latorraca, N. R., Cancilla, D., Satyadi, M., Davis, J. E., Babu, M. M., Dror, R. O., and Kosuri, S. Structural and functional characterization of G protein–coupled receptors with deep mutational scanning. *eLife*, 9:e54895, October 2020. ISSN 2050-084X. doi: 10.7554/eLife.54895. URL https://elifesciences.org/articles/54895.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Jurafsky, D. and Martin, J. H. Speech and language processing, 2nd edition. 2008.

Kaplan, J., McCandlish, S., Henighan, T. J., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *ArXiv*, abs/2001.08361, 2020.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L. Y., Edunov, S., Chen, D., and tau Yih, W. Dense passage retrieval for open-domain question answering. *ArXiv*, abs/2004.04906, 2020.

Kelsic, E. D., Chung, H., Cohen, N., Park, J., Wang, H. H., and Kishony, R. RNA Structural Determinants of Optimal Codons Revealed by MAGE-Seq. *Cell Systems*, 3(6):563–571.e6, December 2016. ISSN 24054712. doi: 10.1016/j.cels.2016.11.004. URL https://linkinghub.elsevier.com/retrieve/pii/S2405471216303684.

Kennouche, P., Charles-Orszag, A., Nishiguchi, D., Goussard, S., Imhaus, A., Dupré, M., Chamot-Rooke, J., and Duménil, G. Deep mutational scanning of the *Neisseria meningitidis* major pilin reveals the importance of pilus tip-mediated adhesion. *The EMBO Journal*, 38(22), November 2019. ISSN 0261-4189, 1460-2075. doi: 10.15252/embj.2019102145. URL https://onlinelibrary.wiley.com/doi/10.15252/embj.2019102145.

Kessel, A. and Ben-Tal, N. *Introduction to Proteins: Structure, Function, and Motion, SECOND EDITION (Chapman & Hall/CRC Mathematical and Computational Biology)*. 03 2018. ISBN 9781498747172. doi: 10.1201/9781315113876.

Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L., and Lewis, M. Generalization through memorization: Nearest neighbor language models. *ArXiv*, abs/1911.00172, 2020.

Kitzman, J. O., Starita, L. M., Lo, R. S., Fields, S., and Shendure, J. Massively parallel single-amino-acid mutagenesis. *Nature Methods*, 12(3):203–206, March 2015. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.3223. URL http://www.nature.com/articles/nmeth.3223.

Klesmith, J. R., Bacik, J.-P., Michalczyk, R., and Whitehead, T. A. Comprehensive Sequence-Flux Mapping of a Levoglucosan Utilization Pathway in *E. coli*. *ACS Synthetic Biology*, 4(11):1235–1243, November 2015. ISSN 2161-5063, 2161-5063. doi: 10.1021/acssynbio.5b00131. URL https://pubs.acs.org/doi/10.1021/acssynbio.5b00131.

Kotler, E., Shani, O., Goldfeld, G., Lotan-Pompan, M., Tarcic, O., Gershoni, A., Hopf, T. A., Marks, D. S., Oren, M., and Segal, E. A Systematic p53 Mutation Library Links Differential Functional Impact to Cancer Mutation Pattern and Evolutionary Conservation. *Molecular Cell*, 71(1):178–190.e8, July 2018. ISSN 10972765. doi: 10.1016/j.molcel.2018.06.012. URL https://linkinghub.elsevier.com/retrieve/pii/S1097276518304544.

Kozek, K. A., Glazer, A. M., Ng, C.-A., Blackwell, D., Egly, C. L., Vanags, L. R., Blair, M., Mitchell, D., Matreyek, K. A., Fowler, D. M., Knollmann, B. C., Vandenberg, J. I., Roden, D. M., and Kroncke, B. M. High-throughput discovery of trafficking-deficient variants in the cardiac potassium channel KV11.1. *Heart Rhythm*, 17(12):2180–2189, December 2020. ISSN 15475271. doi: 10.1016/j.hrthm.2020.05.041. URL https://linkinghub.elsevier.com/retrieve/pii/S1547527120305427.

Lee, J. M., Huddleston, J., Doud, M. B., Hooper, K. A., Wu, N. C., Bedford, T., and Bloom, J. D. Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants. *Proceedings of the National Academy of Sciences*, 115(35):E8276–E8285, August 2018. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1806133115. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1806133115.

Lee, K., Chang, M.-W., and Toutanova, K. Latent retrieval for weakly supervised open domain question answering. *ArXiv*, abs/1906.00300, 2019.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-augmented generation for knowledge-intensive nlp tasks. *ArXiv*, abs/2005.11401, 2020.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *ICLR*, 2019.

Madani, A., McCann, B., Naik, N., Keskar, N. S., Anand, N., Eguchi, R. R., Huang, P.-S., and Socher, R. Progen: Language modeling for protein generation, 2020.

Manekar, S. C. and Sathe, S. R. A benchmark study of k-mer counting methods for high-throughput sequencing. *GigaScience*, 7(12):giy125, 2018.

Matreyek, K. A., Starita, L. M., Stephany, J. J., Martin, B., Chiasson, M. A., Gray, V. E., Kircher, M., Khechaduri, A., Dines, J. N., Hause, R. J., Bhatia, S., Evans, W. E., Relling, M. V., Yang, W., Shendure, J., and Fowler, D. M. Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nature Genetics*, 50(6):874–882, June 2018. ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-018-0122-z. URL http://www.nature.com/articles/s41588-018-0122-z.

Matreyek, K. A., Stephany, J. J., Ahler, E., and Fowler, D. M. Integrating thousands of pten variant activity and abundance measurements reveals variant subgroups and new dominant negatives in cancers. *Genome medicine*, 13(1):1–17, 2021.

Mattenberger, F., Latorre, V., Tirosh, O., Stern, A., and Geller, R. Globally defining the effects of mutations in a picornavirus capsid. *eLife*, 10:e64256, January 2021. ISSN 2050-084X. doi: 10.7554/eLife.64256. URL https://elifesciences.org/articles/64256.

Mavor, D., Barlow, K., Thompson, S., Barad, B. A., Bonny, A. R., Cario, C. L., Gaskins, G., Liu, Z., Deming, L., Axen, S. D., Caceres, E., Chen, W., Cuesta, A., Gate, R. E., Green, E. M., Hulce, K. R., Ji, W., Kenner, L. R., Mensa, B., Morinishi, L. S., Moss, S. M., Mravic, M., Muir, R. K., Niekamp, S., Nnadi, C. I., Palovcak, E., Poss, E. M., Ross, T. D., Salcedo, E. C., See, S. K., Subramaniam, M., Wong, A. W., Li, J., Thorn, K. S., Conchúir, S. O., Roscoe, B. P., Chow, E. D., DeRisi, J. L., Kortemme, T., Bolon, D. N., and Fraser, J. S. Determination of ubiquitin fitness landscapes under different chemical stresses in a classroom setting. *eLife*, 5:e15802, April 2016. ISSN 2050-084X. doi: 10.7554/eLife.15802. URL https://elifesciences.org/articles/15802.

McLaughlin Jr, R. N., Poelwijk, F. J., Raman, A., Gosal, W. S., and Ranganathan, R. The spatial architecture of protein function and adaptation. *Nature*, 491(7422):138–142, November 2012. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature11500. URL http://www.nature.com/articles/nature11500.

Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., and Rives, A. Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv*, 2021. doi: 10.1101/2021.07.09.450648. URL https://www.biorxiv.org/content/early/2021/07/10/2021.07.09.450648.

Melamed, D., Young, D. L., Gamble, C. E., Miller, C. R., and Fields, S. Deep mutational scanning of an RRM domain of the *Saccharomyces cerevisiae* poly(A)-binding protein. *RNA*, 19(11):1537–1551, November 2013. ISSN 1355-8382, 1469-9001. doi: 10.1261/rna.040709.113. URL http://rnajournal.cshlp.org/lookup/doi/10.1261/rna.040709.113.

Melnikov, A., Rogov, P., Wang, L., Gnirke, A., and Mikkelsen, T. S. Comprehensive mutational scanning of a kinase *in vivo* reveals substrate-dependent fitness landscapes. *Nucleic Acids Research*, 42(14):e112–e112, August 2014. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gku511. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gku511.

Mighell, T. L., Evans-Dutson, S., and O'Roak, B. J. A Saturation Mutagenesis Approach to Understanding PTEN Lipid Phosphatase Activity and Genotype-Phenotype Relationships. *The American Journal of Human Genetics*, 102(5):943–955, May 2018. ISSN 00029297. doi: 10.1016/j.ajhg.2018.03.018. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929718301071.

Mishra, P., Flynn, J., Starr, T., and Bolon, D. Systematic Mutant Analyses Elucidate General and Client-Specific

Aspects of Hsp90 Function. *Cell Reports*, 15(3):588–598, April 2016. ISSN 22111247. doi: 10.1016/j.celrep.2016.03.046. URL https://linkinghub.elsevier.com/retrieve/pii/S2211124716303175.

Mitchell, A. L., Almeida, A., Beracochea, M., Boland, M. A., Burgin, J., Cochrane, G., Crusoe, M. R., Kale, V., Potter, S. C., Richardson, L. J., Sakharova, E. A., Scheremetjew, M., Korobeynikov, A. I., Shlemov, A., Kunyavskaya, O., Lapidus, A. L., and Finn, R. D. Mgnify: the microbiome analysis resource in 2020. *Nucleic Acids Research*, 48:D570 – D578, 2020.

Nambiar, A., Heflin, M., Liu, S., Maslov, S., Hopkins, M., and Ritz, A. Transforming the language of life: transformer neural networks for protein prediction tasks. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pp. 1–8, 2020.

Newberry, R. W., Arhar, T., Costello, J., Hartoularos, G. C., Maxwell, A. M., Naing, Z. Z. C., Pittman, M., Reddy, N. R., Schwarz, D. M. C., Wassarman, D. R., Wu, T. S., Barrero, D., Caggiano, C., Catching, A., Cavazos, T. B., Estes, L. S., Faust, B., Fink, E. A., Goldman, M. A., Gomez, Y. K., Gordon, M. G., Gunsalus, L. M., Hoppe, N., Jaime-Garza, M., Johnson, M. C., Jones, M. G., Kung, A. F., Lopez, K. E., Lumpe, J., Martyn, C., McCarthy, E. E., Miller-Vedam, L. E., Navarro, E. J., Palar, A., Pellegrino, J., Saylor, W., Stephens, C. A., Strickland, J., Torosyan, H., Wankowicz, S. A., Wong, D. R., Wong, G., Redding, S., Chow, E. D., DeGrado, W. F., and Kampmann, M. Robust Sequence Determinants of α-Synuclein Toxicity in Yeast Implicate Membrane Binding. *ACS Chemical Biology*, 15(8):2137–2153, August 2020. ISSN 1554-8929, 1554-8937. doi: 10.1021/acschembio.0c00339. URL https://pubs.acs.org/doi/10.1021/acschembio.0c00339.

Ng, P. C. and Henikoff, S. Predicting deleterious amino acid substitutions. *Genome research*, 11(5):863–874, 2001.

Nutschel, C., Fulton, A., Zimmermann, O., Schwaneberg, U., Jaeger, K.-E., and Gohlke, H. Systematically Scrutinizing the Impact of Substitution Sites on Thermostability and Detergent Tolerance for *Bacillus subtilis* Lipase A. *Journal of Chemical Information and Modeling*, 60(3):1568–1584, March 2020. ISSN 1549-9596, 1549-960X. doi: 10.1021/acs.jcim.9b00954. URL https://pubs.acs.org/doi/10.1021/acs.jcim.9b00954.

Olson, C., Wu, N., and Sun, R. A Comprehensive Biophysical Description of Pairwise Epistasis throughout an Entire Protein Domain. *Current Biology*, 24(22):2643–2651, November 2014. ISSN 09609822. doi: 10.1016/j.cub.2014.09.072. URL https://linkinghub.elsevier.com/retrieve/pii/S0960982214012688.

Pokusaeva, V. O., Usmanova, D. R., Putintseva, E. V., Espinar, L., Sarkisyan, K. S., Mishin, A. S., Bogatyreva, N. S., Ivankov, D. N., Akopyan, A. V., Avvakumov, S. Y., Povolotskaya, I. S., Filion, G. J., Carey, L. B., and Kondrashov, F. A. An experimental assay of the interactions of amino acids from orthologous sequences shaping a complex fitness landscape. *PLOS Genetics*, 15(4):e1008079, April 2019. ISSN 1553-7404. doi: 10.1371/journal.pgen.1008079. URL https://dx.plos.org/10.1371/journal.pgen.1008079.

Press, O., Smith, N. A., and Lewis, M. Train short, test long: Attention with linear biases enables input length extrapolation, 2021.

Qi, H., Olson, C. A., Wu, N. C., Ke, R., Loverdo, C., Chu, V., Truong, S., Remenyi, R., Chen, Z., Du, Y., Su, S.-Y., Al-Mawsawi, L. Q., Wu, T.-T., Chen, S.-H., Lin, C.-Y., Zhong, W., Lloyd-Smith, J. O., and Sun, R. A Quantitative High-Resolution Genetic Profile Rapidly Identifies Sequence Determinants of Hepatitis C Viral Fitness and Drug Sensitivity. *PLoS Pathogens*, 10(4):e1004064, April 2014. ISSN 1553-7374. doi: 10.1371/journal.ppat.1004064. URL https://dx.plos.org/10.1371/journal.ppat.1004064.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2018. URL https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

Radivojac, P., Obradovic, Z., Smith, D. K., Zhu, G., Vucetic, S., Brown, C. J., Lawson, J. D., and Dunker, A. K. Protein flexibility and intrinsic disorder. *Protein Science*, 13(1):71–80, 2004.

Ramensky, V., Bork, P., and Sunyaev, S. Human nonsynonymous snps: server and survey. *Nucleic acids research*, 30(17):3894–3900, 2002.

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J. F., Abbeel, P., and Song, Y. S. Evaluating protein transfer learning with TAPE. *CoRR*, abs/1906.08230, 2019. URL http://arxiv.org/abs/1906.08230.

Rao, R., Meier, J., Sercu, T., Ovchinnikov, S., and Rives, A. Transformer protein language models are unsupervised structure learners. In *International Conference on Learning Representations*, 2020.

Rao, R., Liu, J., Verkuil, R., Meier, J., Canny, J. F., Abbeel, P., Sercu, T., and Rives, A. Msa transformer. *bioRxiv*, 2021. doi: 10.1101/2021.02.12.430858. URL https://www.biorxiv.org/content/early/2021/02/13/2021.02.12.430858.

Remmert, M., Biegert, A., Hauser, A., and Söding, J. Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. *Nature Methods*, 9:173–175, 2012.

Reva, B., Antipin, Y., and Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research*, 39(17):e118–e118, 2011.

Riesselman, A. J., Ingraham, J. B., and Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nature methods*, 15(10):816–822, 2018.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15), 2021.

Robertson, S. E. and Zaragoza, H. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3:333–389, 2009.

Rockah-Shmuel, L., Tóth-Petróczy, A., and Tawfik, D. S. Systematic Mapping of Protein Mutational Space by Prolonged Drift Reveals the Deleterious Effects of Seemingly Neutral Mutations. *PLOS Computational Biology*, 11(8):e1004421, August 2015. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004421. URL https://dx.plos.org/10.1371/journal.pcbi.1004421.

Romero, P. A., Tran, T. M., and Abate, A. R. Dissecting enzyme function with microfluidic-based deep mutational scanning. *Proceedings of the National Academy of Sciences*, 112(23):7159–7164, June 2015. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1422285112. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1422285112.

Roscoe, B. P. and Bolon, D. N. Systematic Exploration of Ubiquitin Sequence, E1 Activation Efficiency, and Experimental Fitness in Yeast. *Journal of Molecular Biology*, 426(15):2854–2870, July 2014. ISSN 00222836. doi: 10.1016/j.jmb.2014.05.019. URL https://linkinghub.elsevier.com/retrieve/pii/S0022283614002587.

Roscoe, B. P., Thayer, K. M., Zeldovich, K. B., Fushman, D., and Bolon, D. N. Analyses of the Effects of All Ubiquitin Point Mutants on Yeast Growth Rate. *Journal of Molecular Biology*, 425(8):1363–1377, April 2013. ISSN 00222836. doi: 10.1016/j.jmb.2013.01.032. URL https://linkinghub.elsevier.com/retrieve/pii/S0022283613000636.

Russ, W. P., Figliuzzi, M., Stocker, C., Barrat-Charlaix, P., Socolich, M., Kast, P., Hilvert, D., Monasson, R., Cocco, S., Weigt, M., et al. An evolution-based model for designing chorismate mutase enzymes. *Science*, 369 (6502):440–445, 2020.

Sarkisyan, K. S., Bolotin, D. A., Meer, M. V., Usmanova, D. R., Mishin, A. S., Sharonov, G. V., Ivankov, D. N., Bozhanova, N. G., Baranov, M. S., Soylemez, O., et al. Local fitness landscape of the green fluorescent protein. *Nature*, 533(7603):397–401, 2016.

Seuma, M., Faure, A. J., Badia, M., Lehner, B., and Bolognesi, B. The genetic landscape for amyloid beta fibril nucleation accurately discriminates familial Alzheimer's disease mutations. *eLife*, 10:e63364, February 2021. ISSN 2050-084X. doi: 10.7554/eLife.63364. URL https://elifesciences.org/articles/63364.

Shin, J.-E., Riesselman, A. J., Kollasch, A. W., McMahon, C., Simon, E., Sander, C., Manglik, A., Kruse, A. C., and Marks, D. S. Protein design and variant prediction using autoregressive generative models. *Nature communications*, 12(1):1–11, 2021.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., and Higgins, D. G. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular Systems Biology*, 7:539 – 539, 2011.

Sinai, S., Jain, N., Church, G. M., and Kelsic, E. D. Generative aav capsid diversification by latent interpolation. *bioRxiv*, 2021.

So, D. R., Mańke, W., Liu, H., Dai, Z., Shazeer, N., and Le, Q. V. Primer: Searching for efficient transformers for language modeling, 2021.

Soh, Y. S., Moncla, L. H., Eguia, R., Bedford, T., and Bloom, J. D. Comprehensive mapping of adaptation of the avian influenza polymerase protein PB2 to humans. *eLife*, 8:e45079, April 2019. ISSN 2050-084X. doi: 10.7554/eLife.45079. URL https://elifesciences.org/articles/45079.

Sourisseau, M., Lawrence, D. J. P., Schwarz, M. C., Storrs, C. H., Veit, E. C., Bloom, J. D., and Evans, M. J. Deep Mutational Scanning Comprehensively Maps How Zika Envelope Protein Mutations Affect Viral Growth and Antibody Escape. *Journal of Virology*, 93(23), December 2019. ISSN 0022-538X, 1098-5514. doi: 10.1128/JVI.01291-19. URL https://journals.asm.org/doi/10.1128/JVI.01291-19.

Staller, M. V., Holehouse, A. S., Swain-Lenz, D., Das, R. K., Pappu, R. V., and Cohen, B. A. A high-throughput mutational scan of an intrinsically disordered acidic transcriptional activation domain. *Cell systems*, 6(4):444–455, 2018.

Starita, L. M., Pruneda, J. N., Lo, R. S., Fowler, D. M., Kim, H. J., Hiatt, J. B., Shendure, J., Brzovic, P. S., Fields, S., and Klevit, R. E. Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proceedings of the National Academy of Sciences*, 110(14):E1263–E1272, April 2013. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1303309110. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1303309110.

Starr, T. N., Greaney, A. J., Hilton, S. K., Ellis, D., Crawford, K. H., Dingens, A. S., Navarro, M. J., Bowen, J. E., Tortorici, M. A., Walls, A. C., King, N. P., Veesler, D., and Bloom, J. D. Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell*, 182(5):1295–1310.e20, September 2020. ISSN 00928674. doi: 10.1016/j.cell.2020.08.012. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867420310035.

Steinegger, M. and Söding, J. Clustering huge protein sequence sets in linear time. *Nature Communications*, 9, 2018.

Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S. J., and Söding, J. Hh-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*, 20, 2019.

Stiffler, M., Hekstra, D., and Ranganathan, R. Evolvability as a Function of Purifying Selection in TEM-1 β-Lactamase. *Cell*, 160(5):882–892, February 2015. ISSN 00928674. doi: 10.1016/j.cell.2015.01.035. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867415000781.

Suiter, C. C., Moriyama, T., Matreyek, K. A., Yang, W., Scaletti, E. R., Nishii, R., Yang, W., Hoshitsuki, K., Singh, M., Trehan, A., Parish, C., Smith, C., Li, L., Bhojwani, D., Yuen, L. Y. P., Li, C.-k., Li, C.-h., Yang, Y.-l., Walker, G. J., Goodhand, J. R., Kennedy, N. A., Klussmann, F. A., Bhatia, S., Relling, M. V., Kato, M., Hori, H., Bhatia, P., Ahmad, T., Yeoh, A. E. J., Stenmark, P., Fowler, D. M., and Yang, J. J. Massively parallel variant characterization identifies *NUDT15* alleles associated with thiopurine toxicity. *Proceedings of the National Academy of Sciences*, 117(10):5394–5401, March 2020. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1915680117. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1915680117.

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., and the UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6): 926–932, 11 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu739. URL https://doi.org/10.1093/bioinformatics/btu739.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions, 2014.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, 22 22:4673–80, 1994.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. The clustal_x windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic acids research*, 25 24: 4876–82, 1997.

Thompson, S., Zhang, Y., Ingle, C., Reynolds, K. A., and Kortemme, T. Altered expression of a quality control protease in E. coli reshapes the in vivo mutational landscape of a model enzyme. *eLife*, 9:e53476, July 2020. ISSN 2050-084X. doi: 10.7554/eLife.53476. URL https://elifesciences.org/articles/53476.

Toth-Petroczy, A., Palmedo, P., Ingraham, J., Hopf, T. A., Berger, B., Sander, C., and Marks, D. S. Structured states of disordered proteins from genomic sequences. *Cell*, 167 (1):158–170, 2016.

Tripathi, A., Gupta, K., Khare, S., Jain, P. C., Patel, S., Kumar, P., Pulianmackal, A. J., Aghera, N., and Varadarajan, R. Molecular Determinants of Mutant Phenotypes, Inferred from Saturation Mutagenesis Data. *Molecular Biology and Evolution*, 33(11):2960–2975, November 2016. ISSN 0737-4038, 1537-1719. doi: 10.1093/molbev/msw182. URL https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msw182.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2017.

Wang, S., Yu, M., Guo, X., Wang, Z., Klinger, T., Zhang, W., Chang, S., Tesauro, G., Zhou, B., and Jiang, J. R3:

Reinforced ranker-reader for open-domain question answering. In *AAAI*, 2018.

Weile, J., Sun, S., Cote, A. G., Knapp, J., Verby, M., Mellor, J. C., Wu, Y., Pons, C., Wong, C., Lieshout, N., Yang, F., Tasan, M., Tan, G., Yang, S., Fowler, D. M., Nussbaum, R., Bloom, J. D., Vidal, M., Hill, D. E., Aloy, P., and Roth, F. P. A framework for exhaustively mapping functional missense variants. *Molecular Systems Biology*, 13(12):957, December 2017. ISSN 1744-4292, 1744-4292. doi: 10.15252/msb.20177908. URL https://onlinelibrary.wiley.com/doi/10.15252/msb.20177908.

Weinstein, E. N. and Marks, D. S. A structured observation distribution for generative biological sequence prediction and forecasting. *bioRxiv*, pp. 2020–07, 2021.

Wrenbeck, E. E., Azouz, L. R., and Whitehead, T. A. Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded. *Nature Communications*, 8(1):15695, August 2017. ISSN 2041-1723. doi: 10.1038/ncomms15695. URL http://www.nature.com/articles/ncomms15695.

Wu, N. C., Young, A. P., Al-Mawsawi, L. Q., Olson, C. A., Feng, J., Qi, H., Chen, S.-H., Lu, I.-H., Lin, C.-Y., Chin, R. G., Luan, H. H., Nguyen, N., Nelson, S. F., Li, X., Wu, T.-T., and Sun, R. High-throughput profiling of influenza A virus hemagglutinin gene at single-nucleotide resolution. *Scientific Reports*, 4(1):4942, May 2014. ISSN 2045-2322. doi: 10.1038/srep04942. URL http://www.nature.com/articles/srep04942.

Wu, N. C., Olson, C. A., Du, Y., Le, S., Tran, K., Remenyi, R., Gong, D., Al-Mawsawi, L. Q., Qi, H., Wu, T.-T., and Sun, R. Functional Constraint Profiling of a Viral Protein Reveals Discordance of Evolutionary Conservation and Functionality. *PLOS Genetics*, 11(7):e1005310, July 2015. ISSN 1553-7404. doi: 10.1371/journal.pgen.1005310. URL https://dx.plos.org/10.1371/journal.pgen.1005310.

Young, H. J., Chan, M., Selvam, B., Szymanski, S. K., Shukla, D., and Procko, E. Deep Mutagenesis of a Transporter for Uptake of a Non-Native Substrate Identifies Conformationally Dynamic Regions. preprint, Biochemistry, April 2021. URL http://biorxiv.org/lookup/doi/10.1101/2021.04.19.440442.

Appendix

A. Glossary

Multiple Sequence Alignment (MSA): a set of homologous protein sequences that are aligned in the position coordinate system of the seed sequence. They have been shown to capture information about the homology, phylogeny, and structure of the corresponding protein family (Thompson et al., 1994; 1997). The depth of the MSA refers to the number of sequences it contains.

Deep Mutational Scanning (DMS): a technique which combines saturation mutagenesis with high-throughput functional screening to obtain a thorough description of the fitness landscape of the corresponding proteins (Fowler & Fields, 2014).

Gain of scope: increase in the number of distinct protein families that can be modeled. Proteins that have been historically difficult to model are the ones with shallow MSA or that are difficult to align (e.g., disordered proteins).

B. Tranception model architecture and training details

B.1. Ablation studies

Tranception is an autoregressive transformer architecture designed to explicitly promote head specialization and extraction of contiguous protein subsequence patterns, building on ideas introduced in Primer (So et al., 2021) and Inception (Szegedy et al., 2014). We performed thorough ablations when developing Tranception and summarize the main variants tested in Table 5. Our largest transformer model, Tranception L, has 700M parameters and is trained on UniRef100 (Suzek et al., 2014). In early iterations we also experimented training our architecture on UniRef90 and UniRef50, clustered versions of UniRef100 at 90% and 50% similarity levels respectively, but observed superior performance from training on UniRef100 (see Appendix B.3 and Table 6).

Hyperparameter	GPT2 S	Primer S	Tranception LS	Tranception S	Tranception M	Tranception L
Parameters	85M	85M	85M	85M	300M	700M
Attention heads	12	12	12	12	16	20
Layers	12	12	12	12	24	36
Embedding size	768	768	768	768	1,024	1,280
Activation function	GELU	Squared ReLU	Squared ReLU	Squared ReLU	Squared ReLU	Squared ReLU
Position encoding	Learned embedding	Learned embedding	Learned embedding	Grouped ALiBi	Grouped ALiBi	Grouped ALiBi

Table 5. **Characteristics of different model variants used in ablations.** All models had a max context length of 1024 tokens, and we use the default dropout value of 0.1 in all variants. Tranception LS differs from Tranception S by the use of learned position embeddings instead of Grouped ALiBi.

To decide between different architecture options while not overfitting these decisions to our benchmark, we selected a small yet representative subset of DMS assays in the ProteinGym substitution benchmark (10 out of 87 substitution DMS assays):

- BLAT ECOLX (Jacquier et al., 2013)
- CALM1 HUMAN (Weile et al., 2017)
- CCDB ECOLI (Tripathi et al., 2016)
- DLG4 RAT (McLaughlin Jr et al., 2012)
- PA I34A1 (Wu et al., 2015)
- Q2N0S5 9HIV1 (Haddox et al., 2018)
- RL401 YEAST (Roscoe et al., 2013)
- SPG1 STRSG (Olson et al., 2014)
- SPIKE SARS2 (Starr et al., 2020)
- TPOR HUMAN (Bridgford et al., 2020)

Together, these 10 assays cover the different taxa (3 viral proteins, 4 human and other eukaryote proteins, 3 prokaryote proteins), mutation depths (3 low, 4 medium and 3 high as per the classification described in Table2) and include one assay with multiple mutants (which matches the overall proportion of assays with multiple mutants within ProteinGym). Downstream performance of the different ablations on this validation set, and the overall substitution set are reported in Table 6.

Model variant	Training data	Position encoding	Spearman validation set	Spearman full set
GPT2 S	Uniref100	Learned embedding	0.324	0.320
Primer S	Uniref100	Learned embedding	0.314	0.315
Tranception LS	Uniref100	Learned embedding	0.330	0.333
Tranception S	Uniref100	Grouped ALiBi	0.344	0.335
Tranception S	Uniref90	Grouped ALiBi	0.264	0.275
Tranception S	Uniref50	Grouped ALiBi	0.248	0.247
Tranception M	Uniref100	Grouped ALiBi	0.358	0.376
Tranception L	Uniref100	Grouped ALiBi	**0.399**	**0.404**

Table 6. **Performance of the different model variants in ablation studies.** Performance is measured via Spearman's rank correlation ρ between model scores and experimental measurements, following the approach discussed in E.1. Retrieval inference is excluded from this analysis. Model selection is performed on the validation set described in Appendix B.1.

Metric	Value
Number of sequences	249M
Max sequence length	40,921
95^{th} percentile of length	939
75^{th} percentile of length	470
Median sequence length	314
25^{th} percentile of length	198
5^{th} percentile of length	92
Min sequence length	12

Table 7. **High level statistics of protein sequences in UniRef100 after preprocessing.** About 98% of sequences in UniRef100 have length lower than 1,024.

Hyperparameter	Value
Training steps	150k
Batch size	1,024
Peak learning rate	$3 * 10^{-4}$
Weight decay	10^{-4}
Optimizer	AdamW

Table 8. **Model training hyperparameters.**

B.2. Data processing

Except for the two ablations focusing on UniRef50 and UniRef90, all models are trained on UniRef100. We perform very mild filtering steps of the data to remove fragments and low quality sequences, and preserve as much sequence diversity as possible. For each UniRef100 sequence cluster, we map the corresponding UniRef50 cluster which pools together sequences within 50% similarity from one another [1]. We use 99% of the data ($\sim$ 249 million sequences) for training and set aside 1% of the data for validation ($\sim$ 2.5 million sequences). All singletons at the UniRef50 cluster level are removed (eg., isolated fragments). We further exclude from the training and validation datasets all sequences that contained the infrequent Pyrrolysine (O) or Selenocysteine (U) amino acids, or with two or more consecutive indeterminate amino acids X to remove lower quality sequences. The remaining indeterminate amino acids (X, B, J, Z) are kept at train time and randomly imputed as follows: X is imputed to any of the 20 standard amino acids, B to either D (Aspartic acid) or N (Asparagine), J to either I (Isoleucine) or L (Leucine), Z to either E (Glutamic acid) or Q (Glutamine). All sequences with indeterminates are excluded from the validation set. Table 7 recapitulates key statistics of sequences in UniRef100 after applying these filtering criteria. The observed distribution of sequences guided our choice for the maximum context length of 1,024 tokens for our transformer models, as it allows to handle 98% of protein sequences in UniRef100 without truncation.

B.3. Model training

All model variants are trained for 150k steps, with a batch size of 1,024 sequences. During training, we reverse sequences at random and truncate sequences if longer than the maximum context size as per the scoring scheme details described in

[1] More precisely, UniRef100 is first clustered at the 90% identity to generate UniRef90. Cluster representatives in UniRef90 are then clustered at 50% identity to yield UniRef50.

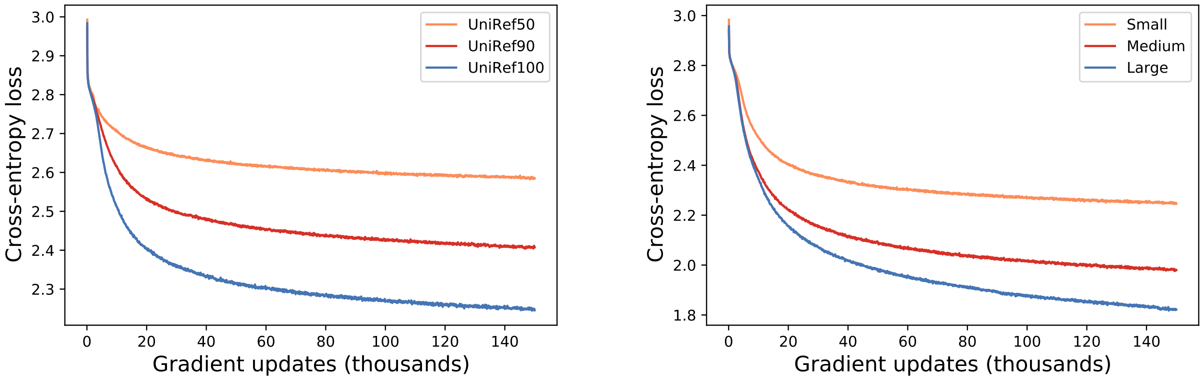

Figure 5. **Cross-entropy loss Vs number of gradient steps** Left: Tranception S architecture trained on different UniRef datasets (UniRef50, UniRef90, UniRef100). Right: Small (S), Medium (M) or Large (L) Tranception architectures trained on UniRef100.

DMS set	Model	Unidirectional scoring	Bidirectional scoring
Validation set	Tranception (w/o retrieval)	0.376	0.399
	Tranception (w/ retrieval)	0.447	0.452
Full set	Tranception (w/o retrieval)	0.376	0.401
	Tranception (w/ retrieval)	0.432	0.445

Table 9. **Comparison of unidirectional Vs bidirectional scoring** Performance is measured via Spearman's rank correlation ρ between model scores and experimental measurements. Analysis is performed with Tranception L (with and without retrieval), and scores are reported on both the validation DMS set and the full DMS set. The bidirectional scoring is the average of log likelihood scores obtained by traversing the sequence in the canonical direction (left to right) and the reverse direction (right to left).

Appendix B.4. We train with the AdamW optimizer (Loshchilov & Hutter, 2019), with a learning rate schedule annealed over the first 10k steps up to the maximum value ($3 * 10^{-4}$), and then linearly decreased until the end of training. Other training hyperparameters are summarized in Table 8. In terms of computing resources, small architectures are trained on 8 V100 GPUs for $\sim$1 week, medium architectures with 32 V100 GPUs for $\sim$1 week, and our largest model, Tranception L, is trained on 64 A100 GPUs for $\sim$2 weeks.

We provide the training loss curves of the different architecture variants in Fig. 2 and Fig. 5. While trained beyond the optimal number of steps for transformer models used in NLP tasks (Kaplan et al., 2020), our larger networks still appear to be undertrained. We note that cross-entropy is not necessarily a good indicator of downstream performance when comparing datasets with very different characteristics. For instance, UniRef50 is by design less redundant than UniRef100, thus we expect the loss to be typically lower on the latter (as can be seen in Fig.5). It so happens that, for the particular task we are interested in (fitness prediction), the more granular UniRef100 both leads to lower cross-entropy loss and higher downstream task performance (Table 6), hence we trained our larger models on that dataset.

B.4. Scoring protein sequences

Mirrored sequences While protein sequences are not strictly invariant to mirroring (ie., the same sequence of amino acids assembled in the left to right order may not lead to the same 3D structure if assembled from right to left), prior autoregressive transformer architectures (eg., (Madani et al., 2020)) have proposed to augment their training dataset by including all sequences and their reverse. In this work we further investigate the benefits of using each sequence and its reverse *at inference* when predicting fitness (§ 3.4). To remove potential discrepancies between training and inference, we apply a data augmentation similar to that of Madani et al. (2020) at training time and randomly reverse a subset of sequences in each batch (each sequence in the batch is flipped with a probability 0.5). We find that averaging the log-likelihood ratios obtained by scoring each sequence and its mirror image leads to superior downstream task performance, with or without retrieval (Table 9).

Scoring window When scoring protein sequences that are longer than the maximum context length of the model (1,024 amino acids in our model), we have the choice to either truncate the sequence or combine predictions from slided (overlapping or non-overlapping) windows of the sequence. We observe very little performance difference between the different approaches in terms of average Spearman's rank correlation with validation DMS assays measurements, and therefore leveraged the former for simplicity. For single mutations, the optimal scoring window is selected so as to maximize the context available around that mutation for prediction when scoring the sequence from left to right and right to left. When scoring multiple mutants, we applied a very similar approach, maximizing context around the barycenter of the various mutant positions.

B.5. Retrieval

Augmenting the autoregressive predictions at each position via retrieval inference only occurs at the positions covered (even partially) by the MSA – outside of theses positions, we fully rely on autoregressive predictions. We compute the pseudocounts at each position of the alignment via weighted Laplace smoothing (Jurafsky & Martin, 2008), with a small smoothing parameter (10^{-5}). As discussed in § 4, sequence are weighted as per the procedure described in (Hopf et al., 2017) and we fully ignore gaps in the MSA when computing the pseudocounts.

We optimize the retrieval inference weight α from equation 4 via linearly-spaced grid search between 0.0 (no retrieval) and 1.0 (full retrieval on covered positions, autoregressive only otherwise). As per the results in table 10, the optimal α value on the validation DMS set is 0.6. Except for the analysis discussed in this section, whenever retrieval is used in this paper,it is with this 0.6 retrieval inference weight.

DMS set	**Retrieval inference weight α**										
	0.0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
Validation set	0.399	0.412	0.424	0.435	0.444	0.450	**0.452**	0.449	0.443	0.431	0.397
Full set	0.401	0.410	0.419	0.428	0.435	0.441	0.445	**0.446**	0.442	0.432	0.404

Table 10. **Retrieval inference weight optimization.** We perform a linearly-spaced grid search for α on our validation DMS set and obtain an optimal rate of 0.6.

C. Multiple Sequence Alignments

For every assay in ProteinGym, we build MSAs of the corresponding protein by performing five search iterations of the profile HMM homology search tool Jackhmmer (Eddy, 2011) on the UniRef100 database of non-redundant proteins (Suzek et al., 2014), downloaded on November 24 2021. We explore a range of 9 bit score thresholds, from 0.1 to 0.9. We subsequently select the alignment with the highest number of significant Evolutionary Couplings (ECs) (Hopf et al., 2014). To be consistent in our approach across all assays, we build a single MSA for each protein, and do not investigate domain-specific alignments where relevant. As noted in Appendix E.4, crafting domain-specific alignments, when these domains are known, may help further increase performance of the different models. Characteristics of the different MSAs (eg., number of sequences, selected bit score) we obtained with the above process are provided in the ProteinGym reference file available in our GitHub repository.

D. Baselines

Except for the enhancements we discuss in this section, we use the official codebases for all baselines included in § 6. For the Site independent and EVmutation models we use the EVcouplings library. For Wavenet, we use the code made available in the SeqDesign repository. All alignment-based models are trained on the MSAs we obtain via the process discussed in Appendix C, and the same MSAs are used at inference time with MSA transformer or for retrieval in Tranception.

D.1. ESM-1v and MSA Transformer

We start from the official ESM codebase, and extend it as follows to support:

- the scoring of multiple mutants – as discussed in Meier et al. (2021) this is achieved by independently summing the effects of each single mutations that comprise the multiple mutant;

- the scoring of sequences that are longer than the context size of the ESM-1v and MSA Transformer models. We leverage the same routine we developed for Tranception to select the optimal scoring window (see Appendix B.4);
- the weighted sampling of sequences of an input MSA in the MSA Transformer. We compute sequence weights with the same procedure used for retrieval inference in Tranception, ie. based on the procedure described in Hopf et al. (2017).

When scoring with ESM-1v or MSA Transformer, we use the masked-marginals approach introduced in Meier et al. (2021). This scoring heuristics has two main limitations: 1) as discussed above, scores for multiple mutants are computed as the sum of the effects from the individual single mutants comprising the multiple mutation, which leads to ignoring potentially important epistatic effects 2) masked scoring implies a fixed frame of reference in which the mutated position exists in the original wild-type sequence. Consequently these two models are unable to score indels with the masked-marginals heuristic.

As discussed in § 6, we compute scores reported in the main results tables are obtained with a single model seed for fair comparison across models, but we also report ensemble performance in Appendix E.5. We use the most recent checkpoints made publicly available for each model (resp. esm1v_t33_650M_UR90S_1 and esm_msa1b_t12_100M_UR50S at the time of writing). For MSA transformer, we follow (Meier et al., 2021) and first filter sequences in the input MSA with HHFilter (Steinegger et al., 2019) to ensure minimum coverage of 75% and maximum sequence identity of 90%, and then sample 384 sequences (with the weights discussed above).

D.2. EVE and DeepSequence

For both DeepSequence and EVE, we use the Pytorch implementation available in the official EVE github repository (we use the optimal parameters for EVE as per Frazer et al. (2021), and as per Riesselman et al. (2018) for DeepSequence).

EVE and DeepSequence models are trained on protein-specific MSAs. Sequences with more than 50% of gaps in the MSA are removed from training. More importantly, positions in the MSA with more than 30% of gaps in the MSA are removed, such that EVE and DeepSequence models do not provide scores for mutations impacting these particular positions. When the full protein is difficult to cover, this sometimes results in a full domain of the protein being dropped (see the BRCA1 protein example in Appendix E.4). For that reason, when reporting performance results in § 6 we compare all models on the exact same set of mutants, excluding the ones that are not scored by EVE and DeepSequence. We nonetheless provide a full performance comparison on the entire set of mutants in the files available in our GitHub repository for the models that provide scores for all mutants (see § E).

Except for Wavenet, all alignment-based models are unable to score indels since they rely on the fixed coordinate system from the original MSA they have been trained on.

E. Detailed performance analysis

E.1. Performance reporting methodology

We report performance based on the 3 metrics described in § 6:

- **Spearman's rank correlation ρ between model scores and DMS experimental measurements.** Since certain DMS assays are relatively difficult to model resulting in very low (and sometimes negative) ρ values, we report the signed ρ value instead of the absolute ρ values reported in prior works (Riesselman et al., 2018; Meier et al., 2021). To that end, we adjust the signs of measured phenotypes where needed, to ensure consistency in the directionality across assays. In ProteinGym, a higher DMS score is always associated with higher fitness;
- **AUC between model scores and DMS experimental measurements.** This metric is particularly relevant when focusing on assays with a bimodal distribution of the measured phenotype. We binarize DMS measurements by setting the threshold manually when the assay is clearly bimodal and there is no ambiguity about the correct threshold value to select. We use a median cutoff in all other instances. We report the numerical value of the cutoff and the binarization method (median or manual) in the ProteinGym reference file (available in our GitHub repository);
- **Matthew's correlation coefficient (MCC) between model scores and DMS experimental measurements.** This complements the analysis obtained with AUC. DMS measurements are binarized with the same thresholds as for AUC. Predictions are binarized by using their median value as threshold.

In order to standardize measured outcomes as much as possible across assays, we further preprocess the raw DMS measurements as follows:

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**

Table 11. **Average Spearman's rank correlation ρ between model scores and experimental measurements by taxon.**

- **Silent mutations**: certain assays include nucleotide substitutions resulting in silent mutations. We remove these from our benchmark;
- **Duplicate mutations**: certain assays include duplicate mutants – either nucleotide substitutions resulting in mutant repeats in the experiments, or indels resulting in identical protein sequences. We remove duplicates by averaging all DMS measurements across duplicate mutants;
- **Missing measurements**: mutants with missing assay measurement are dropped.

Finally, ProteinGym contains several DMS assays for the same protein, for a handful of proteins (eg., we include 4 assays for BLAT ECOLX, 4 assays for P53). While all different and important in their own right, these different assays have experimental measurements that are strongly correlated. Consequently, models that tend to do well for one of these assays relative to others, tend to perform well across all of the assays for the same protein. In order to remove the potential biases resulting from these correlated assays, we first aggregate all performance metrics at the Uniprot ID level and then compute the different average results in § 6 (ie., when there 4 DMS assays for the same protein, each of these assays carries a weight of 0.25 in the aggregate results).

E.2. Detailed results

We report below aggregated performance results by taxon for the different metrics described in Appendix E.1 on the ProteinGym substitution benchmark: Spearman's rank in Table 11, AUC in Table 12 and MCC in Table 13. Overall model ranking is the same across all performance metrics. The Spearman's rank correlation coefficients between model scores and DMS experimental measurements for Tranception, EVE, ESM-1v and MSA Transformer for each DMS assay in the substitution benchmark are shown in Fig. 6. Performance for the same metric for Tranception and Wavenet on the ProteinGym indel benchmark is provided in Table 14.

Detailed performance tables for all models at the Uniprot ID level and at the DMS level for both the Protein substitutions and indels benchmarks are made available in the Tranception GitHub repository.

E.3. MSA Filtering analyses

We report additional DMS level results for the MSA filtering analysis described in § 7 in Fig. 7. We observe that the performance of Tranception is less sensitive to MSA depth compared to EVE or MSA Transformer.

E.4. Full protein Vs domain-specific models

For all DMS assays we focused on full protein alignments. It may be possible to further increase the performance (and the coverage) of models leveraging MSAs by tailoring the creation of the MSA to known protein domains. For instance, we can obtain more diverse alignments for the RING and BRCT domains of BRCA1, thereby increasing the predictive performance of alignment-based models trained specifically on these alignments (Table 15). This also helps improve predictions made by Tranception with retrieval.

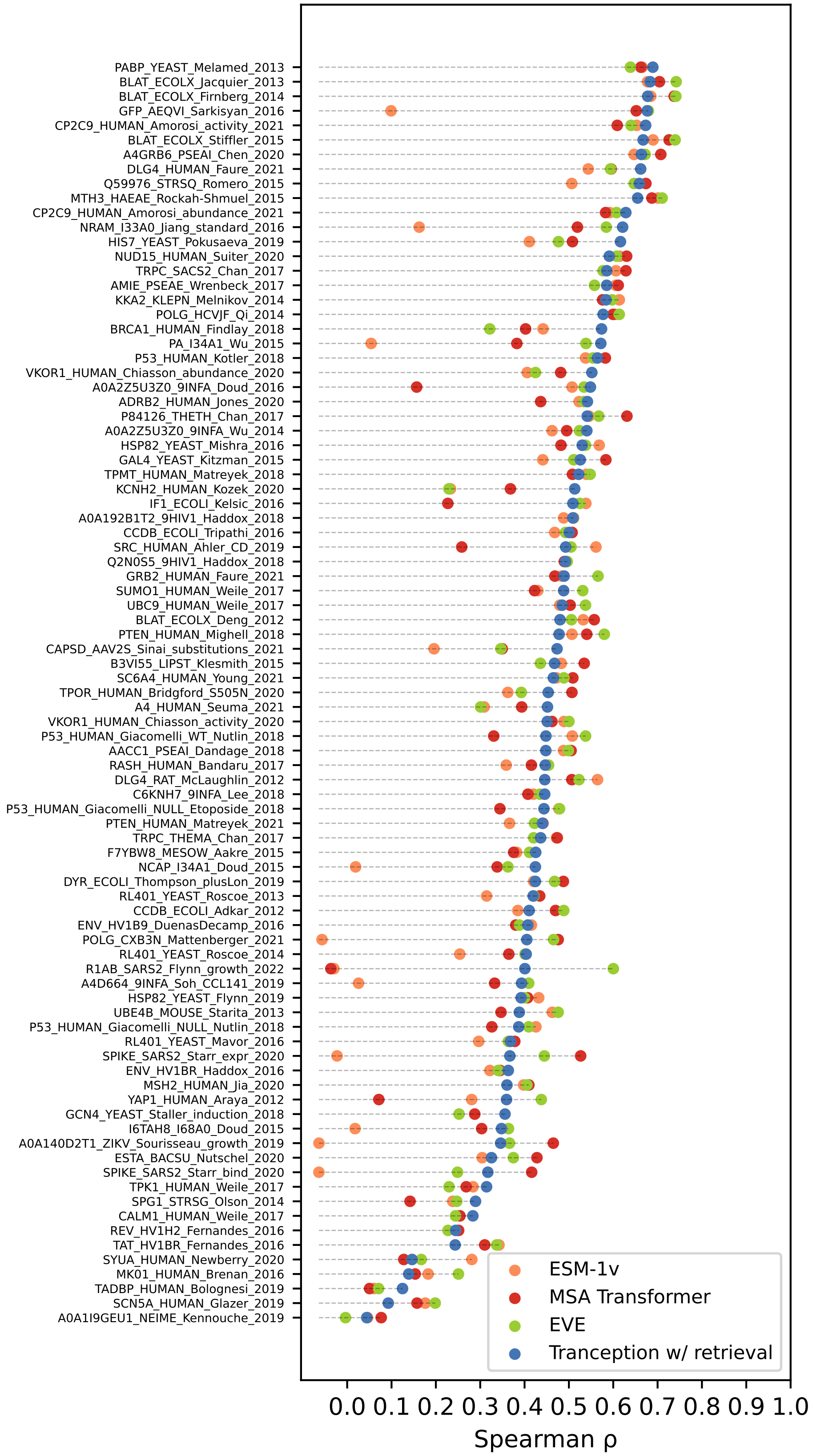

Figure 6. **Model performance on the ProteinGym substitution benchmark.** We report the DMS-level performance (measured by the Spearman's rank correlation ρ between model scores and experimental measurements) of Tranception with retrieval, ESM-1v (Meier et al., 2021), MSA Transformer (Rao et al., 2021) and EVE (Frazer et al., 2021) for each DMS in the ProteinGym substitution benchmark.

Model type	Model name	AUC by taxa category ↑ Human	Other Eukaryote	Prokaryote	Virus	All
Alignment-based models	Site indep	0.732	0.768	0.696	0.720	0.725
	Wavenet	0.730	0.766	0.763	0.665	0.725
	Deepsequence	0.729	0.779	0.766	0.685	0.733
	EVmutation	0.735	0.775	0.760	0.702	0.738
	EVE	0.742	0.782	**0.769**	**0.732**	0.751
Protein language models	ESM-1v	0.734	0.749	0.762	0.620	0.713
	MSA Transformer	0.723	0.784	0.768	0.702	0.737
	Tranception (w/o retrieval)	0.716	0.762	0.745	0.709	0.728
	Tranception (w/ retrieval)	**0.749**	**0.795**	0.766	0.727	**0.754**

Table 12. **Average AUC between model scores and experimental measurements by taxon.**

Model type	Model name	MCC by taxa category ↑ Human	Other Eukaryote	Prokaryote	Virus	All
Alignment-based models	Site indep	0.323	0.344	0.279	0.311	0.312
	Wavenet	0.322	0.351	0.369	0.233	0.314
	Deepsequence	0.327	0.369	0.386	0.261	0.330
	EVmutation	0.333	0.364	0.370	0.288	0.334
	EVE	0.338	0.366	0.387	**0.333**	0.352
Protein language models	ESM-1v	0.321	0.325	0.377	0.174	0.296
	MSA Transformer	0.311	0.388	**0.389**	0.290	0.334
	Tranception (w/o retrieval)	0.302	0.349	0.350	0.298	0.319
	Tranception (w/ retrieval)	**0.348**	**0.390**	0.379	0.330	**0.356**

Table 13. **Average Matthew's correlation coefficient (MCC) between model scores and experimental measurements by taxon.**

E.5. Model ensembling

Prior work from Riesselman et al. (2018); Meier et al. (2021) noted that additional performance gains on the fitness prediction task may be achieved in practice by ensembling several independently trained versions of the same model (eg., trained with different random seeds). We report the performance from several ensemble versions in Table 16. Additionally, we revisit how ensembling has been approached so far for fitness prediction – we propose that instead of ensembling several times the *same* model architecture, we may obtain substantially higher performance in practice by ensembling *different* yet complementary model architectures. We ensemble different pairs of models together and obtain the strongest results by combining Tranception with EVE (Table 17).

F. ProteinGym curation

To build the ProteinGym benchmark, we initially collected a set of 137 deep mutational scanning assays. We then filtered out 43 of these assays based on the following criteria: Data had not been made publicly available (9), non-protein assays (UTR, tRNA, promoter, etc.; 7), synthetic proteins (3), insufficient number of measurements (3), outdated (ie., a more recent improved assay on the same protein and same property was found; 4), majority of data hitting experimental floor (6), low dynamic range (6), assay not relevant to fitness prediction (5).

The final version of the ProteinGym consists of the experimental measurements of 94 deep mutational scanning experiments (87 substitutions assays and 7 indels assays) from the following 77 publications: (Adkar et al., 2012; Jones et al., 2020; Kozek et al., 2020; Firnberg et al., 2014; Jia et al., 2021; Chan et al., 2017; Wu et al., 2015; Mavor et al., 2016; Ahler et al., 2019; Newberry et al., 2020; Fernandes et al., 2016; Mighell et al., 2018; Roscoe & Bolon, 2014; Kotler et al., 2018; Rockah-Shmuel et al., 2015; Giacomelli et al., 2018; Melamed et al., 2013; Melnikov et al., 2014; Brenan et al., 2016; Suiter et al., 2020; Chen et al., 2020; Tripathi et al., 2016; Pokusaeva et al., 2019; Aakre et al., 2015; Haddox et al., 2016; Soh

DMS assay	**Wavenet**	**Tranception w/o retrieval**	**Tranception w/ retrieval**
A0A1J4YT16 9PROT (Davidi et al., 2020)	0.117	0.178	**0.191**
B1LPA6 ECOSM (Russ et al., 2020)	0.385	0.321	**0.415**
BLAT ECOLX (Gonzalez et al., 2019)	**0.546**	0.296	0.357
PTEN HUMAN (Mighell et al., 2018)	**0.699**	0.563	0.598
CAPSD AAV2S (Sinai et al., 2021)	0.457	0.549	**0.586**
HIS7 YEAST (Pokusaeva et al., 2019)	0.680	**0.707**	0.692
P53 HUMAN (Kotler et al., 2018)	0.001	0.395	**0.401**
Average	0.412	0.430	**0.463**

Table 14. **Average Spearman's rank correlation ρ between model scores and experimental measurements on the ProteinGym indel benchmark.**

Domain	Tranception (w/o retrieval)	Tranception (retrieval full MSA)	Tranception (retrieval domain MSA)	EVE (full MSA)	EVE (domain MSA)
RING	0.567	0.588	**0.607**	0.320	0.573
BRCT	0.354	0.490	0.504	N/A	**0.593**

Table 15. **BRCA1 model performance summary by domain, as measured by Spearman's ρ between model scores and experimental measurements** When using a full-protein alignment, EVE is unable to score mutations in the BRCT domain due to insufficient coverage in that region, as per the limitations discussed in Appendix D

et al., 2019; Deng et al., 2012; Weile et al., 2017; Klesmith et al., 2015; Doud & Bloom, 2016; Wu et al., 2014; Findlay et al., 2018; Stiffler et al., 2015; Faure et al., 2022; Jacquier et al., 2013; Kennouche et al., 2019; Sourisseau et al., 2019; Wrenbeck et al., 2017; Dandage et al., 2018; Seuma et al., 2021; Haddox et al., 2018; Doud et al., 2015; Mishra et al., 2016; Flynn et al., 2020; Amorosi et al., 2021; McLaughlin Jr et al., 2012; Nutschel et al., 2020; Kitzman et al., 2015; Kelsic et al., 2016; Lee et al., 2018; Mattenberger et al., 2021; Matreyek et al., 2018; Thompson et al., 2020; Romero et al., 2015; Qi et al., 2014; Roscoe et al., 2013; Bandaru et al., 2017; Young et al., 2021; Glazer et al., 2020; Chiasson et al., 2020; Olson et al., 2014; Bridgford et al., 2020; Starita et al., 2013; Starr et al., 2020; Araya et al., 2012; Davidi et al., 2020; Russ et al., 2020; Gonzalez et al., 2019; Sinai et al., 2021; Bolognesi et al., 2019; Duenas-Decamp et al., 2016; Jiang et al., 2016; Matreyek et al., 2021; Sarkisyan et al., 2016; Staller et al., 2018).

We preprocessed the raw DMS assays as indicated in § E.1. The ProteinGym benchmarks are made publicly available (both raw and processed assay data) on our GitHub repository.

Model type	**Single model**	**Ensemble of 5 models**
Tranception w/o retrieval	0.406	-
Tranception w/ retrieval	**0.451**	-
ESM-1v	0.371	0.401
MSA Transformer	0.422	0.434
EVE	0.448	**0.452**

Table 16. **Single-architecture ensemble analysis.** Tranception without retrieval (single seed) achieves higher average performance than the ESM-1v ensemble. Tranception with retrieval (single seed) achieves higher performance than the MSA Transformer ensemble. The ensemble of 5 EVE models does marginally better than a single Tranception model with retrieval. Performance is measured via Spearman's rank correlation between model scores and DMS measurements.

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

Table 17. **Paired architecture ensemble analysis** In this analysis we always use Tranception without retrieval. We note however that ensembling Tranception with retrieval and EVE provides only marginally higher overall performance (0.475) suggesting that the retrieval inference may capture information analogous to that of alignment-based models like EVE, while the autoregressive inference provides different but complementary information.

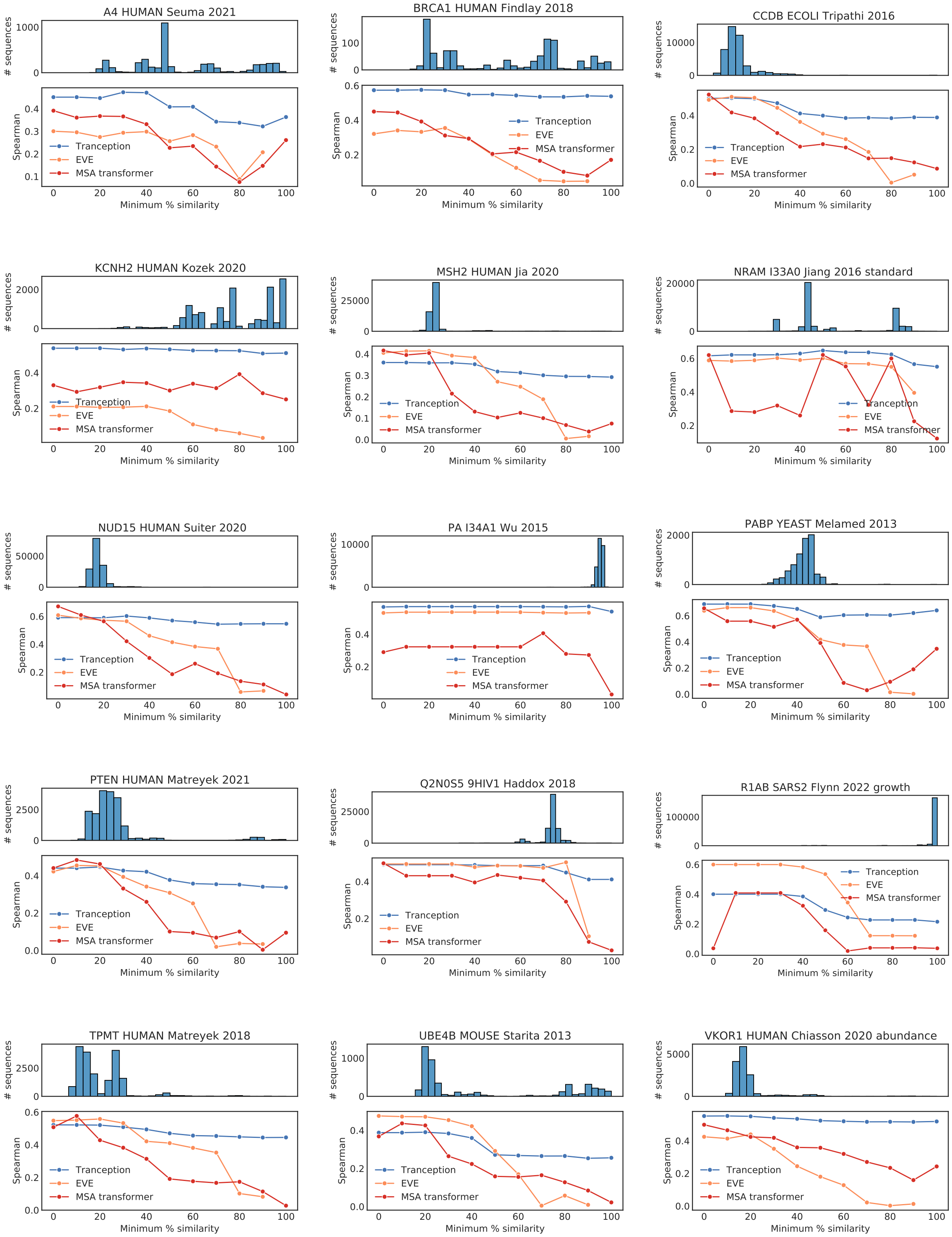

Figure 7. **Robustness to alignment depth for several DMS assays in the ProteinGym substitution benchmark**. We measure the Spearman's rank correlation between model score and experimental measurement for Tranception, EVE and MSA Transformer as we progressively exclude sequences in the corresponding MSAs based on their similarity to the sees sequence used to create the alignment.