MSA-Conditioned Generative Protein Language Models for Fitness Landscape Modelling and Design

Alex Hawkins-Hooker [*]
University College London

David T. Jones
University College London

Brooks Paige
University College London

Abstract

Recently a number of works have demonstrated successful applications of a fully data-driven approach to protein design, based on learning generative models of the distribution of a family of evolutionarily related sequences. Language modelling techniques promise to generalise this design paradigm across protein space, however have for the most part neglected the rich evolutionary signal in multiple sequence alignments and relied on fine-tuning to adapt the learned distribution to a particular family. Inspired by the recent development of alignment-based language models, exemplified by the MSA Transformer, we propose a novel alignment-based generative model which combines an input MSA encoder with an autoregressive sequence decoder, yielding a generative sequence model which can be explicitly conditioned on evolutionary context. To test the benefits of this generative MSA-based approach in design-relevant settings we focus on the problem of unsupervised fitness landscape modelling. Across three unusually diverse fitness landscapes, we find evidence that directly modelling the distribution over full sequence space leads to improved unsupervised prediction of variant fitness compared to scores computed with non-generative masked language models. We believe that combining explicit encoding of evolutionary information with a generative decoder's representation of a distribution over sequence space provides a powerful framework generalising traditional family-based generative models.

1 Introduction

Generative models of protein sequences have played a central role in the computational study of protein evolution, structure and function. The dominant modelling paradigm relies on modelling the distribution of sequences within a single family to learn structural and functional constraints. In this way, profile HMMs capture the patterns of evolutionary conservation that are required to assign homology [6], Potts models capture the pairwise dependencies between residues that are the hallmarks of global structural and functional constraints within families [27, 3], and higher-order models such as VAEs have proved especially useful for predicting the fitness effects of mutations [19]. Building on these approaches a number of recent works have demonstrated the successful generation of functional proteins by models trained on single families [18, 21, 9, 24]. Many of these models are trained on a family-specific multiple sequence alignment (MSA). Alignments of sets of related sequences are a rich source of information due to the coevolutionary signal encoded in patterns of mutations in pairs (or larger subsets) of columns. However, a reliance on MSA inputs makes the construction of a suitably deep and diverse alignment a necessary precondition for using these models,

[*] alex.hawkins-hooker.20@ucl.ac.uk

Machine Learning for Structural Biology Workshop, NeurIPS 2021.

limiting their effectiveness on smaller families and *de novo* designed proteins for which evolutionary information is unavailable.

One of the promises of the language modelling techniques which have recently been imported from NLP is the possibility of addressing these limitations by training models which learn patterns of sequence variation which can be transferred across protein space [1, 20, 12]. However, it remains unclear how best to leverage (co)evolutionary information to yield models which successfully distil both global and local patterns of sequence variation.

1.1 Generalising family-based models

The organisation of protein sequence space into families based on evolutionary and structural relationships means that it is not obvious that a single, universal model of sequence variation is appropriate or easily learned. One popular solution is to adopt a two-stage procedure in which a language model is first pre-trained across all of protein space and subsequently fine-tuned on a relevant subset of sequences, such as a set of proteins with an identifiable evolutionary relationship to a design target [1, 14, 12]. This amounts to an implicit conditioning of the model on the local evolutionary context, with the pre-trained weights regularising the resulting family-specific model. An alternative strategy is to instead *explicitly* condition on the local evolutionary context by directly representing it. This is the route taken by the MSA Transformer [16], which is a masked language model which operates on sets of rows from a multiple sequence alignment rather than individual sequences, allowing a masked residue's evolutionary context to inform its prediction. Explicit conditioning removes the need for fine-tuning to adapt predictions to a target family, bringing test-time usage closer to pretraining. Moreover, directly using MSAs as inputs allows inference of alignment and homology to be handled in the input space, simplifying the representational demands placed on the neural network. In practice this leads to more structurally informative representations than those learned by purely sequence-based models despite the use of substantially fewer parameters [16]; notably a similar approach to MSA representation was used as a component of AlphaFold's end-to-end structure prediction pipeline [11].

We believe this explicit alignment-based conditioning represents a potentially powerful paradigm for generalising single-family models, and seek to apply it to the design setting. Direct application of the MSA Transformer to design problems is not straightforward because, like other masked language models, it models masked residues as conditionally independent given the unmasked context and does not naturally define a distribution over sequence space under which to sample or score new sequences. We therefore propose to augment the MSA Transformer with an autoregressive decoder, yielding an MSA-conditioned *generative* sequence model. To attempt to assess the benefits of learning a full distribution over sequences, we study the problem of unsupervised fitness landscape modelling. Variant scoring rules derived from masked language models including the MSA Transformer have proved highly effective at predicting the effects of small numbers of mutations [14], comparable to state-of-the-art family-based generative models. However, they rely on joint masking of sets of mutations to mimic pre-training, with the consequence that the resulting scoring rules treat the effects of multiple mutations as additive, neglecting any interactions between mutations in a fixed context.

To compare the merits of the generative and masked language modelling approaches in a more challenging and design-relevant setting, we focus on a set of three unusually diverse empirical fitness landscapes, selected to probe the ability of models to capture the sequence-function relationship across increasingly global subsets of sequence space. Briefly, they are: the local neighbourhood of a green fluorescent protein (GFP) [22], evolutionarily relevant combinations of sets of multiple mutations within the yeast His3 gene coding for imidazoleglycerol-phosphate dehydratase (IGPD) [15], and a set of highly diverse sequences designed by a Potts model trained to capture the global statistical properties of the AroQ family of chorismate mutase (CM) enzyme sequences [21]. Further details on each landscape are provided in Table 1 and Appendix A.4.

2 Model

2.1 MSA-conditioned generative model of protein sequences

Single-family generative modelling approaches such as Potts models involve specifying a highly simplified evolutionary model over multiple sequence alignments of related proteins, with rows in

Table 1: Empirical fitness landscapes

Protein	Diversity generation type	Reference	# variants	Avg distance to WT
GFP	random	[22]	51,715	3.9
IGPD	MSA substitutions	[15]	485,010	6.9
Chorismate mutase	designed	[21]	1618	71.6

the alignment treated as *i.i.d.* samples from an underlying distribution. We propose to *meta-learn* this form of evolutionary model: instead of learning a set of parameters that maximise the likelihood of a sample of sequences from a single family, we learn a parametric function which maps an input sample of aligned homologues to a distribution over all sequences in that MSA, parameterised via an autoregressive sequence model. The mapping is learned by training across a large number of MSAs. During training, each datapoint consists of a randomly sampled input alignment to be encoded and a disjoint randomly sampled set of rows to be decoded independently.

In practice, we implement the model as a conditional autoregressive Transformer model, with a similar basic structure to the standard sequence-to-sequence Transformer [26], except that the conditioning input is not a single sequence but a subsampled MSA, and the pre-alignment of encoded MSA and decoded sequences removes the need for encoder-decoder attention. We train the model on a set of almost 7000 alignments taken from [2], after excluding alignments showing homology with the target proteins.

2.2 Architecture details

The model has an encoder-decoder structure, with the encoder taking an input MSA and producing a set of embeddings, and the decoder autoregressively producing a sequence of logits conditioned on these embeddings. We use an MSA Transformer as our encoder, using the pretrained weights to warm-start the model. The decoder is a three-layer Transformer, with causal masking used to enforce an autoregressive factorisation of the model's likelihood. The outputs of the encoder are fed into the decoder through two conditioning paths: the final layer embeddings are averaged over the rows of the input MSA and used in place of positional embeddings in the decoder, while weighted combinations of the encoder's row attentions are used to bias the decoder's attention. Finally, drawing on the MSA Transformer's use of shared row attention and on connections between Transformers and Potts models [23, 4, 25], we enforce shared decoder attention across all sequences in a family by using a modified form of self-attention. A fuller description of the architecture is given in Appendix A.2 and details of training are provided in Appendix A.3.

3 Results

For each of the three fitness landscapes we compute the likelihood of all variants under our model and report Spearman's rank correlation coefficient between the likelihoods and the fitness values, as well as the AUC after binarising the fitness values via landscape-specific thresholds. We additionally compare to fitness predictions computed with the MSA Transformer and ESM-1v by using the scoring scheme of [14]. Scoring variants under this scheme relies on using a wild-type sequence as a reference and masking mutations with respect to this reference; here in each case the choice of reference is made for consistency with the functional assay. The presence of (aligned) indels relative to the wild-type sequence in many of the designed CM variants precluded scoring with ESM-1v. For each landscape we also compare to previous work for which variant fitness predictions are readily available [5, 19, 21]. Further details on baselines and variant scoring schemes are provided in Appendix A.1.

To produce inputs for the MSA Transformer and MSA-conditioned generative model we first generated MSAs following the protocol in [16]. We then randomly sub-sampled these alignments to a depth of 32 or 64 sequences depending on the landscape, and used the same set of sequences as input to both MSA-based models. Average prediction results across 5 randomly sub-sampled input MSAs are reported in Table 2. Average rather than ensembled results are also reported for ESM-1v and other baselines where appropriate. Results of ensembling are reported in Table 3.

The MSA-based models show relatively strong performance across the three landscapes, with the autoregressive version performing comparably or better than relevant baselines on all three, and the MSA Transformer on two of three. This is striking evidence of the benefits of explicit encoding

Table 2: Summary of fitness landscape modelling results

	GFP		IGPD		CM	
	Spearman	AUC	Spearman	AUC	Spearman	AUC
MSA-conditioned Transformer	**0.68**	**0.91**	0.48	**0.78**	**0.44**	**0.79**
MSA Transformer	**0.68**	**0.91**	0.43	0.75	0.35	0.75
ESM-1v	0.10	0.56	0.44	0.74	-	-
eUniRep	0.65	0.89	-	-	-	-
eUniRep (rnd init.)	0.57	0.84	-	-	-	-
DeepSequence (VAE)	-	-	0.45	0.75	-	-
EVMutation (Potts)	-	-	**0.51**	**0.78**	-	-
EVMutation profile model	-	-	0.46	0.75	-	-
bmDCA (Potts)	-	-	-	-	0.41	0.78

of evolutionary context, which yields models capable of summarising the fitness-relevant variation within a single family given just a small subset of the family's MSA, in contrast to the family-specific baselines which are trained on thousands of sequences in each case. GFP provides an example of the kind of small family for which such an approach might yield particular benefits: the full GFP alignment used here contains only 176 homologues, potentially posing difficulties for single-family models and fine-tuning approaches. To avoid these difficulties the eUniRep model was fine-tuned on a much broader set of putative homologues identified using a very permissive homology detection threshold [5], but the resulting model still underperforms the MSA-based approaches. ESM-1v also performs relatively poorly on GFP, providing an example of a case where 'zero-shot' prediction of mutation effects is comparably ineffective. This particular failure may be due to the fact that GFP's relatively few homologues will have limited representation in the pretraining set (UniRef90), or simply have to do with the specific way in which GFP function is encoded in the sequence.

The improved performance of the autoregressive MSA-conditioned model relative to the MSA Transformer on the IGPD and chorismate mutase landscapes may reflect the larger average number of mutations displayed by variants in these landscapes, which reduces the context available to the MSA Transformer and could increase the importance of modelling dependencies between mutations. A breakdown of performance by number of mutations for the IGPD landscape seems to indicate a decline in performance at increasing numbers of mutations for both ESM-1v and the MSA Transformer (Figure 1). Finally, after ensembling predictions across subsampled input MSAs, the MSA-conditioned model outperforms all baselines across the IGPD and CM landscapes, while the MSA Transformer narrowly outperforms it on GFP (Table 3).

4 Discussion

Alignment-based language models provide a natural framework for generalising traditional single-family modelling approaches. We proposed a novel generative extension of this framework, by augmenting the MSA Transformer with an autoregressive decoder. Across three diverse fitness landscapes, the resulting model's likelihoods better predicted the fitnesses of variants than scores computed using the MSA Transformer or the masked language model ESM-1v, both of which may be limited by their inability to represent a distribution over full sequence space. While these results are promising, it seems clear that further work comparing different modelling strategies is required. In particular, with the exception of UniRep, we do not compare to autoregressive alignment-free language models such as ProGen [13], nor the variety of pretraining approaches made available by ProtTrans [7]. Objective evaluation remains a challenge for generative approaches, and while well-characterised fitness landscapes of the sort studied here seem like a promising test-bed, they are limited in number, and may vary widely in experimental protocol and practical definition of fitness. Better characterising the properties of these fitness landscapes, for example by analysing patterns of epistasis as well as aggregate fitness, may help further elucidate differences between models. Finally, we note that efficient scoring and sampling are valuable benefits of the kind of generative approach we propose here; we leave exploration of the applications of these properties as well as further development of our proposed approach to future work.

References

[1] Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., and Church, G. M. “Unified rational protein engineering with sequence-based deep representation learning”. *Nature Methods* 16.12 (2019), pp. 1315–1322.

[2] Anishchenko, I., Ovchinnikov, S., Kamisetty, H., and Baker, D. “Origins of coevolution between residues distant in protein 3D structures”. *Proceedings of the National Academy of Sciences* 114.34 (2017), pp. 9122–9127.

[3] Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S.-I., and Langmead, C. J. “Learning generative models for protein fold families”. *Proteins: Structure, Function, and Bioinformatics* 79.4 (2011), pp. 1061–1078.

[4] Bhattacharya, N., Thomas, N., Rao, R., Daupras, J., Koo, P. K., Baker, D., Song, Y. S., and Ovchinnikov, S. “Single Layers of Attention Suffice to Predict Protein Contacts” (2020).

[5] 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 (2021), pp. 389–396.

[6] Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

[7] Elnaggar, A., Heinzinger, M., Dallago, C., Rihawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Bhowmik, D., and Rost, B. “ProtTrans: Towards Cracking the Language of Life’s Code Through Self-Supervised Deep Learning and High Performance Computing”. *bioRxiv* (2020), p. 2020.07.12.199554.

[8] Figliuzzi, M., Barrat-Charlaix, P., and Weigt, M. “How Pairwise Coevolutionary Models Capture the Collective Residue Variability in Proteins?” *Molecular Biology and Evolution* 35.4 (2018), pp. 1018–1027.

[9] Hawkins-Hooker, A., Depardieu, F., Baur, S., Couairon, G., Chen, A., and Bikard, D. “Generating functional protein variants with variational autoencoders”. *PLOS Computational Biology* 17.2 (2021), e1008736.

[10] Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Schärfe, C. P. I., Springer, M., Sander, C., and Marks, D. S. “Mutation effects predicted from sequence co-variation”. *Nature Biotechnology* 35.2 (2017), pp. 128–135.

[11] Jumper, J. et al. “Highly accurate protein structure prediction with AlphaFold”. *Nature* 596.7873 (2021), pp. 583–589.

[12] Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., Olmos, J. L., Xiong, C., Sun, Z. Z., Socher, R., Fraser, J. S., and Naik, N. “Deep neural language modeling enables functional protein generation across families”. *bioRxiv* (2021).

[13] 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”. *arXiv:2004.03497 [cs, q-bio, stat]* (2020).

[14] 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).

[15] Pokusaeva, V. O. et al. “An experimental assay of the interactions of amino acids from orthologous sequences shaping a complex fitness landscape”. *PLOS Genetics* 15.4 (2019), e1008079.

[16] Rao, R. M., Liu, J., Verkuil, R., Meier, J., Canny, J., Abbeel, P., Sercu, T., and Rives, A. “MSA Transformer”. *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021, pp. 8844–8856.

[17] Remmert, M., Biegert, A., Hauser, A., and Söding, J. “HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment”. *Nature Methods* 9.2 (2012), pp. 173–175.

[18] Repecka, D. et al. “Expanding functional protein sequence spaces using generative adversarial networks”. *Nature Machine Intelligence* 3.4 (2021), pp. 324–333.

[19] Riesselman, A. J., Ingraham, J. B., and Marks, D. S. “Deep generative models of genetic variation capture the effects of mutations”. *Nature Methods* (2018).

[20] Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., and Fergus, R. “Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences”. *Proceedings of the National Academy of Sciences of the United States of America* 118.15 (2021), e2016239118.

[21] Russ, W. P., Figliuzzi, M., Stocker, C., Barrat-Charlaix, P., Socolich, M., Kast, P., Hilvert, D., Monasson, R., Cocco, S., Weigt, M., and Ranganathan, R. “An evolution-based model for designing chorismate mutase enzymes”. *Science* 369.6502 (2020), pp. 440–445.

[22] Sarkisyan, K. S. et al. “Local fitness landscape of the green fluorescent protein”. *Nature* 533.7603 (2016), pp. 397–401.

[23] Sercu, T., Verkuil, R., Meier, J., Amos, B., Lin, Z., Chen, C., Liu, J., LeCun, Y., and Rives, A. “Neural Potts Model”. *bioRxiv* (2021).

[24] 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 (2021), p. 2403.

[25] Trinquier, J., Uguzzoni, G., Pagnani, A., Zamponi, F., and Weigt, M. “Efficient generative modeling of protein sequences using simple autoregressive models”. *Nature Communications* 12.1 (2021), p. 5800.

[26] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. “Attention is all you need”. *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS’17. Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 6000–6010.

[27] Weigt, M., White, R. A. H., Szurmant, H., Hoch, J. A., and Hwa, T. “Identification of direct residue contacts in protein-protein interaction by message passing.” *Proceedings of the National Academy of Sciences of the United States of America* (2009).

A Appendix

A.1 Details of baseline variant scoring computations

Potts models A Potts model is a pairwise Markov Random Field, which defines an energy over arbitrary aligned sequences. EVMutation [10] scores variants by the difference in energy between mutant and wild-type [10]:

$$s_{\text{EVMut}}(\mathbf{x}_{var}) = E(\mathbf{x}_{var}) - E(\mathbf{x}_{wt}) \tag{1}$$

For the chorismate mutase sequences, we simply use the reported energies under the bmDCA Potts model, which is equivalent to the EVMutation scoring scheme up to the constant wild-type energy. The difference between the bmDCA and EVMutation Potts models is the learning algorithm: while the EVMutation model uses the pseudolikelihood approximation to maximum likelihood, the bmDCA model uses a more accurate maximum likelihood estimation procedure based on Boltzmann machine learning [8].

ESM-1v and MSA Transformer Meier et al. [14] demonstrated the use of models trained via masked language modelling for unsupervised variant effect prediction. The best performing models were the MSA Transformer and ESM-1v, a Transformer-based masked language model trained on the UniRef90 representative sequences and otherwise similar to ESM-1b [20].

Masked language models define a conditional distribution over masked tokens given unmasked tokens $p(\mathbf{x}_{\mathbf{m}}|\mathbf{x}_{/\mathbf{m}}) = p(x_{m_1}, x_{m_2}...x_{m_M}|\mathbf{x}_{/\mathbf{m}})$. Meier et al. [14] exploit this to define a scoring rule which compares the probabilities of variant and wild-type amino acids at the mutated positions given the common amino acids at the unmutated positions:

$$s_{\text{ESM}}(\mathbf{x}_{var}) = \log\frac{p(\mathbf{x}_{mt}^{(var)}|\mathbf{x}_{/mt})}{p(\mathbf{x}_{mt}^{(wt)}|\mathbf{x}_{/mt})} \tag{2}$$

Here $\mathbf{x}_{mt}$ denotes the set of amino acids in positions in which wild-type and variant differ; $\mathbf{x}_{mt}^{(var)}$ and $\mathbf{x}_{mt}^{(wt)}$ are then the corresponding sets of amino acids in the variant and wild-type sequences respectively.

Computing these probabilities involves masking all mutated positions and running a forward pass of the model to obtain logits from which the probabilities can be derived. Similar to during pre-training, the mutated positions are as a result conditionally independent given the unmutated positions, which means that the scoring rule amounts to an assumption that the effects of mutations are *additive*:

$$s_{\text{ESM}}(\mathbf{x}_{var}) = \log \frac{p(\mathbf{x}_{mt}^{(var)}|\mathbf{x}_{/mt})}{p(\mathbf{x}_{mt}^{(wt)}|\mathbf{x}_{/mt})} = \sum_{i}^{N_{\text{muts}}} \log \frac{p(x_{mt_i}^{(var)}|\mathbf{x}_{/mt})}{p(x_{mt_i}^{(wt)}|\mathbf{x}_{/mt})} = \sum_{i}^{N_{\text{muts}}} s_{mt_i} \tag{3}$$

In the case of the MSA Transformer, only the first sequence in the input MSA is masked and scored in this way, with all the other sequences contributing to the conditioning context.

eUniRep For the GFP case study we also include likelihood scores from evotuned UniRep [1, 5]. UniRep is an mLSTM based autoregressive language model that was pretrained on the UniRef50 database. Notably evotuned UniRep (eUniRep) was specifically fine-tuned on a large set of putative relatives of GFP obtained via a non-stringent jackhmmer-based homology search. Biswas et al. [5] provide likelihoods from this model for all mutants in the local fitness landscape from [22], which we use to compute the reported metrics.

A.2 Details of MSA conditioned autoregressive transformer model

Encoder The encoder is an MSA Transformer [16], which produces a set of embeddings $\mathbf{H} \in \mathbb{R}^{M \times L \times D}$ representing an input MSA of M rows by L columns. In addition to these embeddings, we extract the row attention maps $\mathbf{A} \in \mathbb{R}^{H \times L \times L}$, where $H = n \times h$ is the total number of heads across the 12 layers of the model. Both final layer embeddings and row attention maps are used to condition the decoder.

Decoder and conditioning mechanism The decoder is a shallow autoregressive Transformer. Unlike standard encoder-decoder Transformers which employ encoder-decoder attention as a conditioning mechanism, we exploit the fact that the encoded and decoded sequences are pre-aligned in a single MSA to achieve attention-free conditioning of the decoder. The final layer embeddings from the MSA Transformer encoder are averaged over the row dimension to produce a sequence of L D-dimensional MSA column embeddings (Equation 4). These are then passed through a learned linear projection (Equation 5), and used in place of position embeddings in the decoder module.

$$\mathbf{c}_{l_d} = \tfrac{1}{M} \sum_{m} \mathbf{H}_{mld} \tag{4}$$

$$\mathbf{p}_l = W\mathbf{c}_l + \mathbf{b} \tag{5}$$

Additionally, since the shared row attention in the MSA Transformer directly encodes contacting residues, we use a learned weighted combination of the concatenated log-transformed post-softmax row attentions from all layers of the encoder to bias the decoder attentions in each layer (Equation 6).

$$a_{ij}^{(h)} = \text{softmax}(z_{ij}^{(h)} + \sum_{l} w_l^{(h)} \log(\mathbf{A}_{lij}) + b^{(h)}) \tag{6}$$

Where h indexes an attention head and $z_{ij}^{(h)}$ denotes the pre-softmax decoder attentions between positions i and j.

Inspired by the use of shared row attention in the MSA Transformer to enforce a common set of relationships between sequence positions across all members of a family (reflecting shared structural constraints), we employ a shared attention mechanism in the decoder. To achieve this, we require that the key and query embeddings in each layer are independent of the decoded sequence, obtaining them directly as learned layer-specific projections of the MSA Transformer-conditioned position embeddings (Equation 7). The values are learned projections of the sequence-dependent residue representations as in standard self-attention.

$$z_{ij}^{(h)} = (U_q^{(h)} \mathbf{p}_i)^T (U_k^{(h)} \mathbf{p}_j) \tag{7}$$

Hyperparameters The decoder consists of 3 layers with an embedding dimension of 256 and 8 attention heads in each layer. The details of the MSA Transformer encoder are provided in [16]. Our implementation is based on the ESM pytorch library [20].

A.3 Training details

The model is trained on a set of multiple sequence alignments. We make use of the set generated by Anishchenko et al. [2] of alignments for sequences represented in the PDB. These alignments were generated using hhblits [17] and processed to remove redundant and low-coverage sequences. We additionally exclude all alignments containing fewer than 100 effective sequences, to avoid overfitting to small families. We split the qualifying alignments randomly 90/5/5 into training, validation and test datasets, using performance on the validation set to manually tune hyperparameters. Additionally, after computing the original splits we used hhsearch to search for significant homology between alignments representing each of the three proteins whose fitness landscapes were studied and alignments in the training set. All identified alignments were removed from the training set, leaving a total of 6543 training set alignments.

Models were trained for 21 epochs using the Adam optimizer with a base learning rate of $1e^{-4}$, a batch size of 4, and a standard Transformer-style learning rate schedule consisting of a linear warmup period of 2000 updates and subsequent inverse square root decay. We leverage the pretrained MSA Transformer weights to initialise the parameters of our MSA encoder, and initialise the decoder weights randomly.

During training, each datapoint consists in a randomly subsampled input MSA together with 32 target sequences sampled randomly from the rest of the MSA on which the likelihood is computed. The number of input sequences is uniformly sampled between 8 and 104. The target sequences are decoded independently. To avoid memory issues when handling long sequences, we randomly crop input and target alignments exceeding 150 residues to a width of 150. At test time, the full-width alignments are used.

A.4 Fitness landscape task details

Following Rao et al. [16], we constructed a multiple sequence alignment for each protein using hhblits against UniClust30 (2017-10), using 3 rounds of searching and default parameters otherwise. Inputs for the MSA-based models (our MSA-conditioned generative model and the MSA Transformer) were subsampled from these alignments. We restrict the number of sequences in each input alignment to 64, and weight the probabilities of selection for each sequence by the density of the local sequence neighbourhood as in [14]. Only 32 inputs are used in the case of GFP due to the low diversity of the input alignment. In each case the wild-type sequence is always included amongst the 32/64 inputs, though its role is different in the two models: in the MSA Transformer it is masked, and the corresponding logits are used to derive the relevant probabilities of the mutant and wild-type amino acids, whereas in the MSA-conditioned autoregressive model the wild-type sequence has the same status as the other MSA inputs. A set of 5 input MSAs was sampled in this way for each protein.

GFP We downloaded quantitative measurements of fluorescence of variants generated by mutating the *Aequorea victoria* GFP-coding gene [22]. We removed all variants including a non-terminal stop codon, leaving 51,715 variants with associated fitness values. Following the authors, we binarised fitness by considering all variants with log-fluorescence less than 3 to be non functional.

For ESM-1v we used the full 238-residue avGFP UniProt sequence as input (UniProt accession P42212); for the MSA-based models the first residue was trimmed.

IGPD The experimental data published by [15] constitute the results of 12 indepedent mutational scans of the IGPD-coding His3 gene in *Saccharomyces cerevisiae*. Each of the 12 experiments assays combinations of mutations within a small local 'segment' containing between 11 and 20 mutable positions in two nearby variable regions separated by a short constant region. Within each segment the assayed mutations reflect different possible combinations of amino acid substitutions that are individually observed in a multiple sequence alignment of His3 orthologues. The 12 segments are interleaved and together cover the length of the His3 gene. We aggregated results across all segments. To facilitate direct comparison with fitness predictions reported by [19], we filtered out all variants

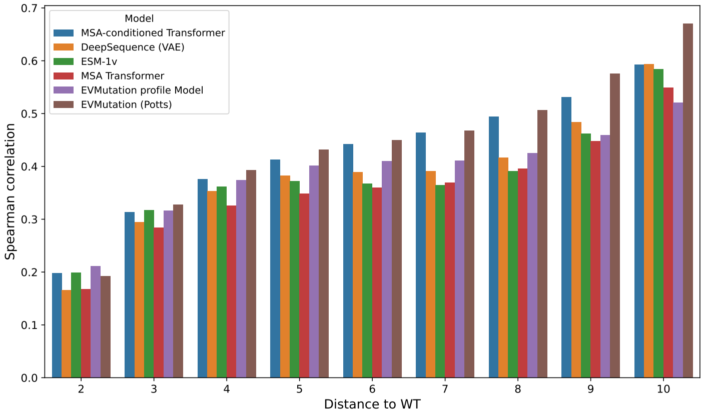

Figure 1: Breakdown of performance as measured by Spearman correlation by number of mutations for the IGPD landscape

for which DeepSequence scores were not provided, leaving 496,137 sequences. We additionally filtered out around 10,000 variants which include mutations in areas outside of the intended mutable region for each segment. The authors rescaled and thresholded fitness values so that a fitness value of 0 is assigned to 95% of all nonsense mutations and a fitness value of 1 is the maximum within each segment. We compute Spearman correlation with respect to these rescaled values, and binarise the fitness values by regarding any variant with a rescaled fitness value > 0 as functional.

The reference sequence for masked language models and for homology search was the 220-residue protein encoded by the His3 gene in yeast (UniProt accession P06633).

CM Russ et al. [21] reported fitness values for a set of 1618 designed chorismate mutase sequences, generated by sampling from Potts models trained on alignments of naturally occurring members of the AroQ family of chorismate mutases. The sequences were assayed for their ability to show evidence of chorismate mutase activity when expressed in *Escherichia Coli*. We used the authors' proposed value of 0.42 (normalized relative enrichment) to separate nonfunctional and functional variants. The alignment used to train the Potts models from which sequences were sampled contains two positions in which the wild-type *E. Coli* sequence (the first 95 residues of the CM domain of the CM-prephenate dehydratase from *E. Coli*, UniProt ID P0A9J8) contains gaps; because the Potts model generates sequences corresponding to rows in this alignment, many of the designed aligned sequences contain amino acids in these positions corresponding to insertions relative to the WT, as well as gaps in other positions corresponding to deletions relative to the WT. To score the resulting sequences with the MSA-based models we generated MSAs containing the same set of columns as the MSA used by the authors. The presence of indels precluded scoring with the ESM-1v model using the mask-based scheme described above. We use the authors' reported Potts model energies to compute baseline fitness predictions. We note that unlike in the case of the other landscapes, the diversity of the variants means that some variants are actually closer by Hamming distance to other natural sequences than they are to the *E. Coli* sequence used as a reference input to the MSA Transformer. We nonetheless stick to using a single reference for consistency with the functional assay and with the scoring proposed in [14].

A.5 Effect of number of mutations on scoring schemes

The IGPD landscape was constructed by mutating sets of positions within a series of 12 segments of around 10-20 amino acids in size, incorporating combinations of individual substitutions that were observed in homologues of the reference yeast sequence. The resulting variants are distributed relatively evenly across a range of distances to wild type, allowing an assessment of the effect of the number of mutations on performance (the other landscapes are comparatively skewed towards either very distant or very close variants). We report Spearman correlations for variants by distance

Table 3: Fitness landscape modelling results for prediction ensembles

	GFP		IGPD		CM	
	spearman	AUC	spearman	AUC	spearman	AUC
MSA-conditioned Transformer	0.69	0.91	**0.54**	**0.81**	**0.46**	**0.80**
MSA Transformer	**0.71**	**0.92**	0.51	0.79	0.36	0.76
ESM-1v	0.10	0.56	0.46	0.75	-	-

from wild type, excluding distances with either too few assayed variants (<1000), or for which the composition of variants is heavily skewed towards a single segment (we discard distances for which more than 50% of variants come from a single segment).

Notably, since the individual substitutions are all present in homologues of the wild-type yeast IGPD, individual mutations are generally better tolerated than the random mutations assayed in typical mutational scans, and epistasis - i.e. non-additivity of mutation effects - is more prominent [15]. Figure 1 shows some evidence for a decline in performance at increasing numbers of mutations for both ESM-1v and the MSA Transformer which may indicate a reduced accuracy of the assumption of additive mutation effects for larger numbers of individually well-tolerated mutations.

A.6 Ensembling predictions across input MSAs

Since the scores of the MSA-based models depend on the identities of the sequences selected in the input MSA, a natural and rapid way to form an ensemble is to generate predictions on multiple input MSAs and combine them, as proposed previously in [14]. We follow this strategy for the MSA-based models, using the mean of the predicted scores as the ensembled prediction. We also generate an ensemble for the ESM-1v predictions using the five sets of pretrained weights.