Understanding protein function with a multimodal retrieval-augmented foundation model

Timothy F. Truong Jr
OpenProtein.AI
NY, USA
ttruong@openprotein.ai

Tristan Bepler
OpenProtein.AI
NY, USA
tbepler@openprotein.ai

Abstract

Protein language models (PLMs) learn probability distributions over natural protein sequences. By learning from hundreds of millions of natural protein sequences, protein understanding and design capabilities emerge. Recent works have shown that scaling these models improves structure prediction, but does not seem to improve mutation understanding and representation quality for protein function prediction. We introduce PoET-2, a multimodal, retrieval-augmented protein foundation model that incorporates in-context learning of family-specific evolutionary constraints with optional structure conditioning to learn generative distributions over protein sequences. PoET-2 uses a hierarchical transformer encoder that is equivariant to sequence context ordering and a dual decoder architecture with both causal and masked language modeling objectives, allowing PoET-2 to operate in both fully generative and bidirectional representation learning modes. PoET-2 achieves state-of-the-art performance on zero-shot variant effect prediction, excelling at scoring variants with multiple mutations and challenging indel mutations. In supervised settings, PoET-2 embeddings outperform previous methods for learning sequence-function relationships, especially with small datasets. This work highlights the benefits of combining retrieval augmentation with multimodal, family-centric modeling for advancing protein foundation models. [1]

1 Introduction

Proteins are polymer chains of amino acids that fold into 3-dimensional structures and carry out the vast majority of functions at the molecular level of life. Proteins catalyze chemical reactions, sense and respond to environmental signals, and defend against pathogens, among countless other functions. Their vast functional diversity arises from the astronomical space of possible amino acid sequences, which evolution has explored over billions of years through mutation and selection.

Accurate prediction of the effect of mutations on protein function is crucial for disease understanding, drug development, and protein engineering. Recent advances in protein language models (PLMs) have enabled more accurate zero-shot prediction of variant effects [1–4]. By learning to model probability distributions over natural protein sequences, PLMs output sequence likelihoods that capture relative fitness information and achieve remarkable correlation with functional and structural properties of proteins in deep mutational scanning and clinical mutation benchmarks [5]. However, several key challenges remain.

- Most PLMs use masked language model-based approaches that are limited to prediction of single substitution mutations. These approaches are unable to predict the effect of insertions and deletions (indels), as well as epistatic effects in higher order mutations.

[1] PoET-2 code and model weights available at: https://github.com/OpenProteinAI/PoET-2

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

- PLMs are usually evaluated in the zero-shot setting, which is fundamentally limited to evaluating the correlation between estimates of sequence fitness and actual functional properties of proteins. However, in protein engineering campaigns, practitioners seek to learn directly from limited mutagenesis data to optimize for specific functional properties. PLMs evaluated in this supervised few-shot setting are promising [6–8], but better data efficiency and generalization, particularly for sequence positions not observed during training, is still needed.
- Recent progress in PLMs have generally focused on scaling the number of parameters [9]. However, increasing model capacity primarily seems to benefit only structure prediction [9–11], while showing neutral or even negative impacts on fitness modeling and function prediction [1, 12]. This raises concerns of loss of generalizability due to model memorization, while also making these models increasingly expensive to train and deploy for inference. Recent PLMs have explored alternative approaches that incorporate additional information via multi-modal approaches [9, 13, 14], or retrieval augmentation [1], but not both.

To address these challenges, we propose PoET-2, a sequence-structure multimodal PLM that leverages retrieval-augmentation and dual training objectives to learn to generate new protein sequences conditioned on sequences and structures of homologs. PoET-2 combines three key ideas:

- **Multi-modality**: PoET-2 reasons over both sequence and structure. This enables conditioning on sequence and/or structural homologs, including structure-conditioned sequence generation from partially-observed backbone atomic structures
- **Retrieval-augmentation**: PoET-2 introduces a novel context-conditioning framework featuring a hierarchical attention architecture that is fully equivariant to the order of proteins in the context. This eliminates the need for multi-billion parameter models, while enabling in-context learning by allowing the model to be prompted with new sequences not present in the original training data.
- **Dual training objectives**: PoET-2 is trained using both a causal language modeling objective for sequence generation and likelihood calculation, as well as a masked language modeling objective for bidirectional understanding and sequence embedding.

PoET-2's novel architecture achieves remarkable performance on downstream protein understanding and design tasks. It is capable of solving problems not possible with existing PLMs, including zero-shot indel and higher-order variant effect prediction, improving on previous methods by as much 20%, on both deep mutational scanning and clinical datasets. Homology-augmented protein representations learned by PoET-2 also enable state-of-the-art accuracy in supervised few-shot function learning, reaching the performance of previous methods like Kermut [15] with substantially less data and outperforming other PLMs by a large margin in the contiguous and modulo dataset splits. In ablation experiments, we find that structure conditioning primarily contributes to zero-shot prediction of stability while offering little to no benefit on tasks like clinical variant effect prediction and supervised function prediction. PoET-2 offers fast inference with an efficient footprint of only 182M parameters and minimal GPU requirements.

2 Related Work

Zero-shot variant effect prediction has advanced significantly by integrating information across sequence, structure, and evolution (homologs). Early progress was driven by single-sequence PLMs such as ESM [2, 8, 10] and ProtT5 [16], which, when trained at evolutionary scale, demonstrated strong correlations between sequence likelihoods and protein fitness. Concurrently, family-specific models emerged, focusing on narrower evolutionary contexts [17, 18]. To capture broader evolutionary signals and enable knowledge transfer beyond single-family scopes, other models were trained across vast collections of distinct protein families. This was achieved through methods processing multiple sequence alignments (MSAs) e.g. MSA Transformer [19], and later via models utilizing unaligned homologs e.g. PoET-1 [1]. More recently, strategies for integrating structural information have been explored. These include using discrete structural tokens e.g. SaProt, ProSST, ESM3 [9, 13, 14], employing continuous geometric representations e.g. ProteinMPNN, ESM-IF [20–22], and explicitly leveraging protein surface information e.g. S3F, RSALOR [23, 24]. Lastly, ensemble methods e.g. VenusREM [25] combine different methods to further enhance performance.

Supervised variant effect prediction commonly utilizes likelihoods or embeddings from PLMs as features for downstream models, enabling fitness prediction from limited experimental data [6, 7, 15, 26–28]. For instance, ProteinNPT [7] integrated MSA Transformer embeddings and zero-shot scores within a non-parametric Transformer architecture. Kermut [15] further advanced this paradigm, achieving state-of-the-art results with a composite Gaussian Process (GP) kernel that incorporates features from multiple models, including ESM-2, ProteinMPNN, and AlphaFold2 [29].

3 PoET-2

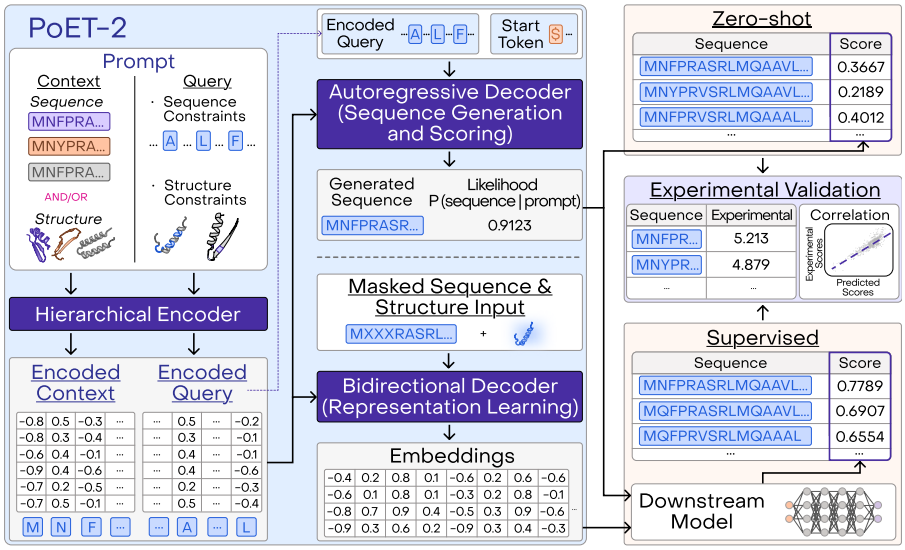

Figure 1: PoET-2 architecture and framework for zero-shot and supervised variant effect prediction. PoET-2 encodes a set of evolutionarily relevant proteins with an equivariant encoder, and decodes proteins with either of two decoders. Log-likelihoods from the autoregressive decoder are used for zero-shot prediction, and are combined with embeddings from the bidirectional decoder for supervised prediction.

3.1 Overview

PoET-2 is a multimodal generative model of protein families, designed for controllable protein generation and representation learning. By jointly modeling the sequences and structures of proteins within a protein family, PoET-2 infers—through in-context learning—the underlying evolutionary constraints that give rise to the family's characteristic sequence features, structural architectures, and/or functional properties. These inferred evolutionary constraints, coupled with a flexible grammar for specifying explicit sequence and structural constraints, enable the controlled generation of novel family members, including variants with enhanced characteristics relevant to their function.

PoET-2 is implemented as an encoder-decoder Transformer [30] with one encoder and two decoders (Figure 1). The encoder processes a user-provided **prompt** containing a *set of proteins* that guides the two decoders toward generating novel proteins with desired characteristics. The prompt is processed in a fully protein order equivariant manner, and consists of two optional components:

- **Context**: A protein family comprised of a set of proteins that the user believes are likely to exhibit at least one of the desired characteristics.
- **Query**: A single, partially specified protein that specifies the sequence and/or structure at only a subset of residues; when used, the query constrains the model to generate only proteins containing those sequence or structural elements. Common uses of the query include specifying the protein length, the presence of signal peptides, the inclusion of active sites, or the structure of the entire protein backbone (i.e. inverse folding).

Together, the context and query provide a flexible grammar that allows sequence generation to be controlled both (1) *implicitly*, via the context, which includes examples of proteins likely to exhibit desired characteristics, and (2) *explicitly*, via the query, which specifies explicit sequence and structure constraints. Clever prompt engineering via careful selection of the context allows PoET-2 to focus on the evolutionary, structural, or functional constraints of only the subspace of relevant proteins.

The decoders, conditioning on the encoder's output, generate new proteins aligned with the prompted protein family. PoET-2 employs two distinct decoders for complementary strengths:

- An **autoregressive decoder**, trained with a causal language modeling (CLM) objective, excels at generative tasks. By modeling the full joint probability distribution $P(\text{sequence}|\text{prompt})$, it allows for efficient, autoregressive generation of novel proteins and exact probabilistic scoring of sequence variants, including indels.
- A **bidirectional decoder**, trained with a masked language modeling (MLM) objective, specializes in representation learning. It produces powerful, context-aware embeddings where each residue's representation is informed by the entire sequence context (both preceding and succeeding residues). These rich representations capture deep structural and functional insights, crucial for tasks like structure and function prediction where understanding global dependencies is key.

Both decoders are also equivariant to the order of proteins in the prompt, meaning that PoET-2 as a whole is also equivariant. The complete loss function is thus composed of three components: $\mathcal{L} = \mathcal{L}_{\text{MLM encoder}} + \mathcal{L}_{\text{CLM decoder}} + \mathcal{L}_{\text{MLM decoder}}$.

The MLM encoder loss is the standard MLM loss applied to each protein in the prompt independently. The set of proteins is viewed as a sequence-of-sequences, where the order of proteins is arbitrary. Using the notation $x_j^{(i)}$ to denote the jth residue of ith protein in the prompt, the encoder loss is:

$$\mathcal{L}_{\text{MLM encoder}} = -\mathbb{E}_{x,m_x}\left[\frac{1}{|m_x|}\sum_{i,j\in m_x}\log p(x_j^{(i)}|x_{\setminus m_x})\right] \tag{1}$$

where m_x is the set of masked positions in the sequence-of-sequences x.

The CLM decoder loss is the standard CLM loss additionally conditioned on (1) the prompt, $x_{\setminus m_x}$, and (2) an optional index, q, indicating the index of the sequence in the prompt to use as the query:

$$\mathcal{L}_{\text{CLM decoder}} = -\mathbb{E}_{y,x,m_x,q}\left[\frac{1}{L_y}\sum_{i=1}^{L_y}\log p(y_i|y_{<i}, x_{\setminus m_x}, q)\right] \tag{2}$$

Here, y is a single sequence of length L_y; the notation y_i refers to the ith residue of y.

Likewise, the MLM decoder loss is the standard MLM loss additionally conditioned on $x_{\setminus m_x}$ and q:

$$\mathcal{L}_{\text{MLM decoder}} = -\mathbb{E}_{y,m_y,x,m_x,q}\left[\frac{1}{|m_y|}\sum_{i\in m_y}\log p(y_i|y_{\setminus m_y}, x_{\setminus m_x}, g)\right] \tag{3}$$

3.2 Architectural Details

In this section, we introduce the architectural details of PoET-2 that are relevant to the capabilities demonstrated in this paper; see Appendix B.5 for details relating to additional capabilities.

3.2.1 Model Inputs

Sequence Protein sequences (x_{seq}) are tokenized with a single token per amino acid, a start token (`$`) indicating the start of a sequence, a stop token (`*`) indicating the end of a sequence, a mask token (`X`) indicating a single residue with unknown identity, and a "gap" token (`-`) indicating zero or more residues with unknown identity.

Structure Protein structures backbones (N, Cα, and C atoms) are encoded in a roto-translation invariant way using two representations:

- Pairwise Cα distances (D): all pairwise Cα-Cα distances, discretized into 128 bins: 125 equal-width bins (2.5Å-48Å), one bin for distances > 48Å, one for low-confidence pairs (pLDDT < 70), and one for missing or masked pairs.
- Local structure backbone distances (x_{atomb}): the 36 pairwise distances between the backbone atoms of the residue being encoded and the residues to its left and right along the sequence. These distances capture backbone information not fully recoverable from just D.

Predicted structure confidence (pLDDT) As PoET-2 is trained only on predicted structures in AlphaFoldDB (AFDB) [31, 32], it also takes as input the predicted structure confidence, pLDDT, at each residue (x_{plddt}).

3.2.2 Input Embedding

The encoder and decoders share a common input embedding space that fuses the sequence (x_{seq}), local structure backbone (x_{atomb}), and pLDDT (x_{plddt}) into a single continuous latent space via summation (Appendix Algorithm 1).

3.2.3 Structure-based Attention Bias

To enhance structural information integration, PoET-2 employs a structure-based attention bias for all attention operations within individual protein sequences (but not between sequences) in both its encoder and decoders. This mechanism modifies attention scores by adding a learned bias term corresponding to the discretized pairwise Cα-Cα distance bin (D) for each residue pair (Appendix Figure 3). This approach is analogous to the relative positional bias in the T5 Transformer [33]; however, in PoET-2, the bias is determined by 3D structural proximity rather than linear sequence position.

3.2.4 Encoder Architecture

PoET-2's encoder layers (Appendix Algorithm 2) adopt the architecture of standard Transformer encoder layers with modern tweaks (rotary positional encodings, SwiGLU over MLP, and RMSNorm over LayerNorm) [30, 34–36], and modifications to (1) ensure protein order equivariance, and (2) improve handling of structural inputs. To process sets of input proteins in an order equivariant way, it replaces the standard one stage attention mechanism with the two stage, hierarchical attention mechanism introduced by PoET-1 [1]. Summarizing briefly,

- In the first stage, attention is applied only between residues of individual input proteins. Structure-based attention bias (§3.2.3) is used in this stage.
- In the second stage, attention is applied between residues of all input proteins. Additionally, relative positional encodings between residues reflect the absolute positions within each protein, rather than the absolute position in the sequence-of-sequences.

Attention in PoET-2's encoder is fully bidirectional, enabling the entire encoder to be protein order equivariant. This is in contrast to PoET-1, whose decoder-only, autoregressive architecture permits equivariance only in each individual decoder layer, and not the entire decoder.

3.2.5 Decoder Architecture

Similarly, PoET-2's decoder layers (Appendix Algorithm 4) adopt the architecture of the standard Transformer decoder layer with modifications. The modifications to the attention operations are as follows:

- In the first, self-attention stage, structure-based attention bias (§3.2.3) is used.
- In the second, cross-attention stage, protein order equivariance is maintained by using the same relative position scheme as in the second attention stage of the encoder.

Additionally, when the prompt includes a query, the input embedding of the decoder is modified to encode which protein in the prompt should be used as the query. The modified input embedding of each residue is simply the average of the unmodified input embedding, and the embedding, produced by the encoder, of corresponding residue of the query in the prompt.

3.3 Hyperparameters

PoET-2 is 182 million parameter model, structured with 12 layers and a 1024 hidden dimension in its encoder and decoders. Weights are tied between these modules to enhance parameter efficiency and promote shared representation learning. This approach leverages the ideal representational capabilities of bidirectional modeling, while also benefiting from the strong representations achievable with autoregressive architectures [7, 37].

3.4 Training Data

PoET-2 is trained on 62 million sets of homologous sequences. Each set corresponds to a sequence in UniRef50 Version 2304 [38], and contains all of its homologs in UniRef50 found using Diamond [39]. Each sequence may optionally be associated with a predicted structure from AFDB by matching on the UniRef100 identifier. To ensure that PoET-2 sees a variety of prompts during training, and to reduce the risk of overfitting, sequence and structure are randomly masked (Appendix C).

4 Experiments

Variant effect prediction is the task of predicting the effect of mutations on the ability of a protein to perform its function. To evaluate PoET-2's zero-shot and supervised variant effect prediction capabilities, we utilize the ProteinGym benchmark [5]. ProteinGym assesses the performance of a model by measuring its ability to predict variant effect in two types of datasets: (1) deep mutational scanning (DMS) datasets, which encompass over 200 distinct assays measuring the effect of mutations on a wide variety of proteins and protein functions spanning the tree of life, and (2) clinical datasets measuring the pathogenicity of mutations on >2,500 human genes.

Following ProteinGym conventions, we use Spearman's rank correlation coefficient (ρ) between experimental measurements and predicted fitness as the primary metric for continuous variables, and area under the receiver operating curve (AUROC) for binary variables.

As baselines, we use a subset of the evaluated methods recorded in ProteinGym as of May 12 2025, including the top methods for each benchmark subset, and methods covering a wide variety of approaches e.g. using structure information or not, using sequence homologs or not, etc.

4.1 PoET-2 achieves state-of-the-art zero-shot performance for challenging mutations and clinical variants

In zero-shot variant effect prediction, models must predict the effect mutations on protein fitness without training on experimental data for the specific protein or function of interest. To score the fitness of a mutated variant sequence relative to its wild-type (WT) using PoET-2, we use the log likelihood ratio (LLR) under PoET-2's CLM decoder: $\log P(\text{variant}|\text{prompt}) - \log P(\text{WT}|\text{prompt})$. To optimize predictions, we employ several prompt engineering and scoring adjustment strategies:

- **Ensembling Prompts**: Following PoET-1 [1], we average LLRs obtained from multiple prompts. Each prompt contains a different subsample of WT homologs identified by searching UniRef100 using the ColabFold MSA protocol [40]. Sampling parameters are detailed in Appendix D.1.
- **Structure Conditioning**: We employ two ways of incorporating structure in the prompt. The utility of structural information varies by task, as discussed in §4.3.
- **Length Adjustment for Indels**: To correct for potential biases in autoregressive models towards shorter sequences, we score indel variants using a length-adjusted log likelihood ratio (Appendix D.2).

Table 1: Performance (Spearman ρ) on zero-shot DMS substitutions and indels benchmarks. N/A indicates that the model cannot score indels.

Model	# Param	Model Inputs		Substitutions By MSA Depth				Indels
		Struct.	MSA	Low	Medium	High	All	
ESM-2	650M			0.340	0.410	0.513	0.415	N/A
ESM C	300M	×	×	0.338	0.401	0.519	0.407	N/A
ProGen2 M	764M			0.305	0.390	0.422	0.379	0.466
ProGen2 XL	6.4B			0.322	0.411	0.442	0.390	0.430
SaProt	650M			0.397	0.446	0.546	0.457	N/A
ESM-3 Open	1.4B	✓	×	0.402	0.465	0.575	0.467	N/A
ProSST	110M			0.468	0.506	0.581	0.508	N/A
MSA Transformer	100M			0.375	0.456	0.480	0.431	N/A
TranceptEVE L	700M	×	✓	0.434	0.473	0.491	0.456	0.414
GEMME	N/A			0.445	0.474	0.494	0.455	N/A
PoET-1	200M			0.479	0.477	0.511	0.470	0.517
S3F-MSA	910M			0.470	0.509	0.547	0.496	N/A
VenusREM	110M	✓	✓	0.498	0.524	0.578	0.519	N/A
PoET-2	182M			0.488	0.507	0.555	0.500	**0.566**
PoET-2 + VenusREM	292M			**0.528**	**0.550**	**0.593**	**0.543**	N/A

4.1.1 PoET-2 significantly advances our ability to predict the effects of indels and higher-order mutations

Predicting the effects of complex mutations such as insertions/deletions (indels) and higher-order substitutions is a challenge for many PLMs. State-of-the-art structure-aware predictors based on masked language modeling (e.g. VenusREM [25], S3F-MSA [23]) operate on fixed-length sequences, which prevents them from directly scoring length-altering indels. Furthermore, for higher-order mutations, they assume additive effects across mutated positions and therefore cannot fully model epistatic interactions. In contrast, PoET-2's autoregressive decoder conditions on both sequence and structure to model the full joint probability of a sequence, $P(\text{sequence}|\text{prompt})$. This approach naturally handles variable sequence lengths and epistatic effects.

DMS indels On the ProteinGym DMS indels benchmark (Indels column, Table 1), PoET-2 significantly outperforms all existing models. It achieves a substantial improvement of $\Delta\rho \approx 0.05$ ($p < 1e-5$) over PoET-1, the previous best. Compared to the top-performing non-PoET model, PoET-2 demonstrates an even larger lead of $\Delta\rho \approx 0.10$ ($p < 1e-5$), an improvement of over 20%.

DMS higher-order substitutions For higher-order substitution mutations (Table 2), PoET-2 demonstrates exceptional performance, particularly for variants with three or more mutations. When compared to VenusREM, the state-of-the-art model on the overall DMS substitutions benchmark, PoET-2 achieves substantial gains on these more complex variants ($\Delta\rho \approx 0.09$ for 3 mutations, $\Delta\rho \approx 0.10$ for 4 mutations, and $\Delta\rho \approx 0.075$ for 5+ mutations).

4.1.2 PoET-2 complements existing methods for substitution mutations

DMS substitutions On the DMS substitutions benchmark (Table 1), PoET-2 achieves performance comparable to VenusREM [25], the current state-of-the-art. VenusREM is an ensemble model combining ProSST [13], a structure-aware protein language model (PLM), with a Position-Specific Scoring Matrix (PSSM) derived from evolutionary alignments [41]. While PoET-2 slightly trails VenusREM on the primary Spearman correlation metric ($p < 1e-3$), it demonstrates superior or comparable performance on metrics emphasizing the prediction of beneficial mutations, such as normalized discounted cumulative gain (NDCG, 0.786 vs 0.766 for VenusREM, $p < 1e-5$; Appendix D.4). NDCG scores whether a model gives its highest scores to the sequences with highest fitness, indicating that PoET-2 is slightly better at identifying the most efficacious mutations versus ranking middling and deleterious ones as accurately.

Table 2: Performance (Spearman ρ) on zero-shot DMS substitutions benchmark, by mutation depth (i.e. by number of substitutions).

	Substitutions by Mutation Depth				
Model	**1**	**2**	**3**	**4**	**5+**
ESM-2	0.422	0.245	0.203	0.160	0.220
ESM C	0.417	0.255	0.189	0.150	0.217
ProGen2 M	0.372	0.126	0.149	0.131	0.178
ProGen2 XL	0.370	0.138	0.219	0.200	0.261
SaProt	0.460	0.310	0.271	0.268	0.337
ESM-3 Open	0.493	0.335	0.303	0.284	0.365
ProSST	0.523	0.391	0.316	0.274	0.334
MSA Transformer	0.427	0.216	0.358	0.365	0.401
TranceptEVE L	0.446	0.274	0.349	0.327	0.385
GEMME	0.449	0.273	0.329	0.338	0.419
PoET-1	0.467	0.295	0.412	0.393	0.421
S3F-MSA	0.501	0.330	0.377	0.343	0.387
VenusREM	0.536	0.394	0.352	0.320	0.372
PoET-2	0.508	0.355	**0.444**	**0.419**	**0.447**
PoET-2 + VenusREM	**0.558**	**0.400**	0.442	0.411	0.441

Table 3: Performance (AUROC) on zero-shot clinical substitutions and indels benchmarks. N/A indicates not applicable, whereas a dash (–) indicates applicable, but not computed.

Model	**Subs.**	**Indels**
Progen2 M	–	0.846
Progen2 L	–	0.851
Progen2 XL	–	0.842
RITA M	–	0.892
RITA L	–	0.922
RITA XL	–	0.916
PROVEAN	0.886	0.927
ESM-1b	0.892	N/A
TranceptEVE L	0.920	0.857
PoET-1	0.920	0.934
PoET-2	**0.928**	**0.952**

A simple ensemble combining PoET-2 and VenusREM (Appendix D.3) consistently outperforms both individual models and all other existing methods across all metrics ($p < 1e-5$; Table 1, Appendix D.4). This suggests PoET-2 and VenusREM capture distinct, complementary fitness signals. The ensemble shows robust performance across diverse protein subgroups, including those with varying MSA depths (Table 1) and assay functions (Appendix D.4). However, on higher-order substitutions (3 or more mutations), PoET-2 alone surpasses the ensemble's performance (Table 2), underscoring its strong intrinsic capability to model these complex mutational effects.

4.1.3 PoET-2 improves prediction of clinical variant pathogenicity

PoET-2 significantly improves our ability to distinguish between pathogenic and benign human protein mutations on the ProteinGym clinical benchmark (Table 3). Compared to PoET-1, the next best model, PoET-2 improves AUROC by 0.008 on the substitutions benchmark ($p < 9e-5$) and by a substantial 0.018 on the indels benchmark ($p < 3e-5$), establishing a new state-of-the-art for both.

4.2 PoET-2 embeddings and likelihoods enhance supervised learning of sequence-function relationships

While zero-shot prediction is valuable, protein engineering often involves learning from limited experimental data. We evaluate PoET-2's utility in this supervised setting on ProteinGym's primary supervised DMS benchmark, which focuses on single-site substitutions across all 217 DMS assays. This benchmark assesses generalization ability across three cross-validation (CV) schemes, varying in difficulty based on the relationship between training and test set mutation locations. In the random fold, mutations are distributed randomly across five CV folds. In the modulo fold, protein positions are assigned to one of five CV folds using a modulo-based strategy i.e. every fifth position belongs to the same fold. In the contiguous fold, the protein sequence is divided into five contiguous, equal-length segments, each constituting a CV fold.

We employ a Gaussian Process (GP) regression model to predict fitness. The GP uses a product kernel combining two Matérn 5/2 kernels: one operating on protein embeddings from PoET-2's MLM decoder, and the other on LLRs from the CLM decoder (as used in zero-shot prediction). This kernel was chosen for its relative simplicity and minimal hyperparameter tuning requirements, but it may not be optimal for all scenarios, such as predicting the effects of multi-mutation variants (Appendix E.2). Predictions are ensembled across GPs trained on features from different sequence-only prompts; structure was omitted from prompts because they did not improve supervised results (§4.3).

4.2.1 PoET-2 improves supervised variant effect prediction across diverse generalization regimes

Our PoET-2 based GP model (PoET-2 GP) substantially outperforms the previous state-of-the-art, Kermut [15], in all cross-validation folds on both Spearman correlation and mean squared error metrics (Table 4, $p < 1e-5$). For example, PoET-2 GP achieves an average Spearman ρ of 0.693, compared to 0.664 for Kermut. This improved performance is consistent across various protein and assay subgroups (Appendix E.3).

Table 4: Performance on supervised DMS substitutions benchmark.

	Spearman ρ (↑)				**Mean square error (↓)**			
Model	**Rand.**	**Mod.**	**Contig.**	**Avg.**	**Rand.**	**Mod.**	**Contig.**	**Avg.**
ProteinNPT	0.741	0.588	0.529	0.619	0.441	0.765	0.856	0.687
Kermut	0.746	0.635	0.613	0.664	0.413	0.649	0.697	0.586
ESM-2 (650 M) GP	0.749	0.573	0.549	0.624	0.404	0.720	0.768	0.630
ESM C GP	0.747	0.605	0.573	0.642	0.398	0.660	0.716	0.592
PoET-2 GP	**0.773**	**0.661**	**0.645**	**0.693**	**0.370**	**0.602**	**0.647**	**0.540**

4.2.2 PoET-2 has exceptional data efficiency for few-shot function learning

A critical aspect of practical protein engineering is a model's ability to learn effectively from limited experimental data. To assess data efficiency and compare the utility of different foundation models, we benchmark GP models using identical kernel functions but features derived from different foundation models, including PoET-2, PoET-1 [1], ESM-2 [10], and ESM C [42]. To simulate smaller training set sizes, we systematically vary the training data for each assay by subsampling its available training points, targeting specific sizes from as few as 10 points up to the maximum available.

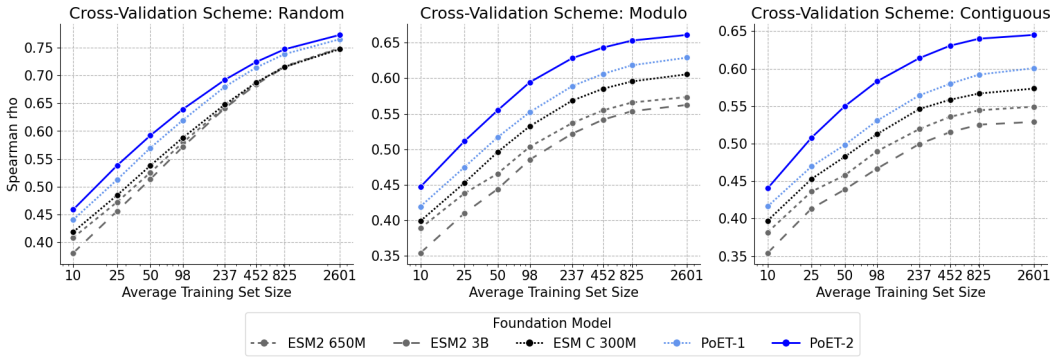

Figure 2: Impact of training set size on the performance of Gaussian Process (GP) models leveraging various foundation models, evaluated on the supervised DMS substitutions benchmark.

PoET-2 GP consistently outperforms GPs based on other foundation models across all evaluated training set sizes and cross-validation schemes (Figure 2). PoET-2's advantage is most pronounced in the challenging contiguous cross-validation split, where PoET-2 GP trained with at most 100 data points matches the performance of ESM C GP (the strongest non-PoET based GP) trained with the maximum training set size (averaging ~2600 data points across all assays). Moreover, PoET-2 GP trained with at most 250 data points achieves performance comparable to Kermut trained with the maximum training set size, demonstrating exceptional data efficiency for practical protein engineering applications compared to the existing state-of-the-art.

4.3 Structure conditioning improves zero-shot prediction, but has limited impact on supervised prediction

To leverage PoET-2's multimodality, we explore two methods for incorporating structure in the prompt. First, we include the predicted structures of homologous proteins in the context. Second,

following prior work showing that inverse-folding likelihood correlates with fitness [5], we use a query consisting of only the WT structure to score a variant's likelihood of adopting the same fold.

In the zero-shot setting, both strategies improve performance on DMS substitutions over using sequence alone (Table 5). Including predicted structures in the context and using the inverse-folding query each individually improve performance, with the best results coming from ensembling these different strategies. Consistent with prior work, this highlights that structural priors are highly informative for zero-shot prediction. This is particularly evident for stability-related assays, where PoET-2 achieves its largest performance gains over its sequence-only predecessor, PoET-1 (Table 1). For other tasks like clinical variant prediction, however, the benefit is less clear, with the inverse-folding query being slightly detrimental (Appendix Table 6). A detailed analysis of these strategies on both DMS and clinical benchmarks is in Appendix D.1.

Table 5: Performance (Spearman's ρ) of different strategies for including structure in the prompt on the zero-shot DMS benchmarks.

Strategy	Prompt		Substitutions	Indels
	Context Modalities	Query		
A	Sequence	None	0.47534	0.55589
B	Sequence and Structure	None	0.48374	**0.56666**
C	Sequence	Structure of WT	**0.49128**	N/A
D	Sequence and Structure	Structure of WT	0.48927	N/A
E	Ensemble of A and B (Different contexts, No query)		0.48260	**0.56606**
F	Ensemble of C and D (Different contexts, With query)		0.49256	N/A
G	Ensemble of A and C (Sequence only context, Different queries)		0.49632	N/A
H	Ensemble of B and D (Sequence and structure context, Different queries)		0.49987	N/A
I	Ensemble of A, B, C, and D		**0.49989**	N/A

In contrast, for supervised learning, including structural information offers little to no benefit (Appendix E.1). This suggests that PoET-2's embeddings already implicitly encode critical structural information, and as a result, our current supervised model does not gain additional predictive power from explicit structure conditioning. Unlocking further improvements may require more sophisticated methods capable of leveraging this explicit structural data to refine an already strong baseline.

5 Conclusion and Limitations

PoET-2 is a multimodal, retrieval-augmented protein language model. PoET-2 can learn from unaligned sequences and structure in-context and directly condition on atomic backbone structure for protein sequence generation and representation learning. As a result, PoET-2 achieves state-of-the-art performance for zero-shot indel and higher order mutation effect prediction, clinical variant effect prediction, and supervised variant effect prediction. However, it lags slightly behind recent structure-based masked language models on single mutant effects on DMS datasets. This seems to be largely driven by these models' superior performance on stability datasets. However, ensembling PoET-2 with VenusREM produces a predictor that outperforms all previous models on ProteinGym's DMS benchmark, suggesting there is still orthogonal information being learned by these methods. Structure-based methods have increasingly been adopting discrete structure tokenizations whereas PoET-2 operates directly on backbone atoms. We also find that structure conditioning is only helpful for some problems, in particular stability prediction in the ProteinGym DMS datasets. For clinical variant effect prediction and supervised learning, structure conditioning offers little to no benefit. In principle, predicted structure information should already be encoded in protein language model-based representations. PoET-2 offers state-of-the-art performance in a compact 182M parameter footprint. We expect PoET-2 to become a core part of protein machine learning and engineering workflows.

References

[1] Timothy Truong Jr and Tristan Bepler. Poet: A generative model of protein families as sequences-of-sequences. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 77379–77415. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/f4366126eba252699b280e8f93c0ab2f-Paper-Conference.pdf.

[2] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29287–29303. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/f51338d736f95dd42427296047067694-Paper.pdf.

[3] Pascal Notin, Lood Van Niekerk, Aaron W Kollasch, Daniel Ritter, Yarin Gal, and Debora Susan Marks. TranceptEVE: Combining family-specific and family-agnostic models of protein sequences for improved fitness prediction. In *NeurIPS 2022 Workshop on Learning Meaningful Representations of Life*, 2022. URL https://openreview.net/forum?id=l7Oo9DcLmR1.

[4] Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena Hurtado, Aidan N Gomez, Debora Marks, and Yarin Gal. Tranception: Protein fitness prediction with autoregressive transformers and inference-time retrieval. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16990–17017. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/notin22a.html.

[5] Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Yarin Gal, and Debora Marks. Proteingym: Large-scale benchmarks for protein fitness prediction and design. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 64331–64379. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/cac723e5ff29f65e3fcbb0739ae91bee-Paper-Datasets_and_Benchmarks.pdf.

[6] Tristan Bepler and Bonnie Berger. Learning the protein language: Evolution, structure, and function. *Cell Systems*, 12(6):654–669.e3, 2021. ISSN 2405-4712. doi: https://doi.org/10.1016/j.cels.2021.05.017. URL https://www.sciencedirect.com/science/article/pii/S2405471221002039.

[7] Pascal Notin, Ruben Weitzman, Debora Marks, and Yarin Gal. Proteinnpt: Improving protein property prediction and design with non-parametric transformers. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 33529–33563. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6a4d5d85f7a52f062d23d98d544a5578-Paper-Conference.pdf.

[8] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, April 2021. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2016239118. URL https://pnas.org/doi/full/10.1073/pnas.2016239118.

[9] Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language

model. July 2024. doi: 10.1101/2024.07.01.600583. URL https://www.biorxiv.org/content/10.1101/2024.07.01.600583v1.

[10] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574.

[11] Bo Chen, Xingyi Cheng, Yangli ao Geng, Shen Li, Xin Zeng, Boyan Wang, Jing Gong, Chiming Liu, Aohan Zeng, Yuxiao Dong, Jie Tang, and Le Song. xtrimopglm: Unified 100b-scale pre-trained transformer for deciphering the language of protein. *bioRxiv*, 2023. doi: 10.1101/2023.07.05.547496. URL https://www.biorxiv.org/content/early/2023/07/14/2023.07.05.547496.

[12] Eli N Weinstein, Alan Nawzad Amin, Jonathan Frazer, and Debora Susan Marks. Non-identifiability and the blessings of misspecification in models of molecular fitness. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=CwG-o0ind6t.

[13] Mingchen Li, Yang Tan, Xinzhu Ma, Bozitao Zhong, Ziyi Zhou, Huiqun Yu, Wanli Ouyang, Liang Hong, Bingxin Zhou, and Pan Tan. Prosst: Protein language modeling with quantized structure and disentangled attention. *bioRxiv*, 2024. doi: 10.1101/2024.04.15.589672. URL https://www.biorxiv.org/content/early/2024/05/17/2024.04.15.589672.1.

[14] Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. *bioRxiv*, 2023.

[15] Peter Mø rch Groth, Mads Herbert Kerrn, Lars Olsen, Jesper Salomon, and Wouter Boomsma. Kermut: Composite kernel regression for protein variant effects. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 29514–29565. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/34547650b2ca69d91f3b3c3ae8b21962-Paper-Conference.pdf.

[16] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44 (10):7112–7127, October 2022. ISSN 1939-3539. doi: 10.1109/TPAMI.2021.3095381. URL https://ieeexplore.ieee.org/document/9477085.

[17] Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K. Min, Kelly Brock, Yarin Gal, and Debora S. Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, Nov 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-04043-8. URL https://doi.org/10.1038/s41586-021-04043-8.

[18] Adam J. Riesselman, John B. Ingraham, and Debora S. Marks. Deep generative models of genetic variation capture the effects of mutations. *Nature Methods*, 15(10):816–822, Oct 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0138-4. URL https://doi.org/10.1038/s41592-018-0138-4.

[19] Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8844–8856. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/rao21a.html.

[20] Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael J. L. Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons, 2021. URL https://arxiv.org/abs/2009.01411.

[21] J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022. doi: 10.1126/science.add2187. URL https://www.science.org/doi/abs/10.1126/science.add2187.

[22] Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. *ICML*, 2022. doi: 10.1101/2022.04.10.487779. URL https://www.biorxiv.org/content/early/2022/04/10/2022.04.10.487779.

[23] Zuobai Zhang, Pascal Notin, Yining Huang, Aurélie Lozano, Vijil Chenthamarakshan, Debora Marks, Payel Das, and Jian Tang. Multi-scale representation learning for protein fitness prediction, 2024. URL https://arxiv.org/abs/2412.01108.

[24] Matsvei Tsishyn, Pauline Hermans, Fabrizio Pucci, and Marianne Rooman. Residue conservation and solvent accessibility are (almost) all you need for predicting mutational effects in proteins. *bioRxiv*, 2025. doi: 10.1101/2025.02.03.636212. URL https://www.biorxiv.org/content/early/2025/02/04/2025.02.03.636212.

[25] Yang Tan, Ruilin Wang, Banghao Wu, Liang Hong, and Bingxin Zhou. Retrieval-enhanced mutation mastery: Augmenting zero-shot prediction of protein language model. *arXiv:2410.21127*, 2024.

[26] Chloe Hsu, Hunter Nisonoff, Clara Fannjiang, and Jennifer Listgarten. Learning protein fitness models from evolutionary and assay-labeled data. *Nature Biotechnology*, 40(7):1114–1122, Jul 2022. ISSN 1546-1696. doi: 10.1038/s41587-021-01146-5. URL https://doi.org/10.1038/s41587-021-01146-5.

[27] Kevin K. Yang, Zachary Wu, and Frances H. Arnold. Machine-learning-guided directed evolution for protein engineering. *Nature Methods*, 16(8):687–694, Aug 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0496-6. URL https://doi.org/10.1038/s41592-019-0496-6.

[28] Tristan Bepler and Bonnie Berger. Learning protein sequence embeddings using information from structure. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SygLehCqtm.

[29] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, Aug 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://doi.org/10.1038/s41586-021-03819-2.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[31] Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, Augustin Žídek, Tim Green, Kathryn Tunyasuvunakool, Stig Petersen, John Jumper, Ellen Clancy, Richard Green, Ankur Vora, Mira Lutfi, Michael Figurnov, Andrew Cowie, Nicole Hobbs, Pushmeet Kohli, Gerard Kleywegt, Ewan Birney, Demis Hassabis, and Sameer Velankar. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, 11 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab1061. URL https://doi.org/10.1093/nar/gkab1061.

[32] Mihaly Varadi, Damian Bertoni, Paulyna Magana, Urmila Paramval, Ivanna Pidruchna, Malarvizhi Radhakrishnan, Maxim Tsenkov, Sreenath Nair, Milot Mirdita, Jingi Yeo, Oleg Kovalevskiy, Kathryn Tunyasuvunakool, Agata Laydon, Augustin Žídek, Hamish Tomlinson, Dhavanthi Hariharan, Josh Abrahamson, Tim Green, John Jumper, Ewan Birney, Martin Steinegger, Demis Hassabis, and Sameer Velankar. Alphafold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Research*, 52(D1):D368–D375, 11 2023. ISSN 0305-1048. doi: 10.1093/nar/gkad1011. URL https://doi.org/10.1093/nar/gkad1011.

[33] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

[34] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2021. URL https://arxiv.org/abs/2104.09864.

[35] Noam Shazeer. Glu variants improve transformer, 2020. URL https://arxiv.org/abs/2002.05202.

[36] Biao Zhang and Rico Sennrich. Root mean square layer normalization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/1e8a19426224ca89e83cef47f1e7f53b-Paper.pdf.

[37] Timothy Fei Truong Jr. and Tristan Bepler. Poet: A foundation model for high accuracy protein property prediction, Sep 2024. URL https://www.openprotein.ai/poet-foundation-model-for-high-accuracy-protein-property-prediction.

[38] The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 11 2022. ISSN 0305-1048. doi: 10.1093/nar/gkac1052. URL https://doi.org/10.1093/nar/gkac1052.

[39] Benjamin Buchfink, Klaus Reuter, and Hajk-Georg Drost. Sensitive protein alignments at tree-of-life scale using diamond. *Nature Methods*, 18(4):366–368, Apr 2021. ISSN 1548-7105. doi: 10.1038/s41592-021-01101-x. URL https://doi.org/10.1038/s41592-021-01101-x.

[40] Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature Methods*, 19(6):679–682, Jun 2022. ISSN 1548-7105. doi: 10.1038/s41592-022-01488-1. URL https://doi.org/10.1038/s41592-022-01488-1.

[41] Thomas A. Hopf, John B. Ingraham, Frank J. Poelwijk, Charlotta P. I. Schärfe, Michael Springer, Chris Sander, and Debora S. Marks. Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, 35(2):128–135, Feb 2017. ISSN 1546-1696. doi: 10.1038/nbt.3769. URL https://doi.org/10.1038/nbt.3769.

[42] ESM Team. ESM Cambrian: Revealing the mysteries of proteins with unsupervised learning. https://evolutionaryscale.ai/blog/esm-cambrian, 2024. EvolutionaryScale Website, December 4, 2024.

[43] Michel van Kempen, Stephanie S. Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L. M. Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature Biotechnology*, 42(2):243–246, Feb 2024. ISSN 1546-1696. doi: 10.1038/s41587-023-01773-0. URL https://doi.org/10.1038/s41587-023-01773-0.

[44] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4596–4604. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/shazeer18a.html.

[45] OpenProtein.AI. [new model] poet. URL https://github.com/OATML-Markslab/ProteinGym/pull/28. Accessed: 2025-5-15.

[46] Aadyot Bhatnagar, Sarthak Jain, Joel Beazer, Samuel C. Curran, Alexander M. Hoffnagle, Kyle Ching, Michael Martyn, Stephen Nayfach, Jeffrey A. Ruffolo, and Ali Madani. Scaling unlocks broader generation and deeper functional understanding of proteins. *bioRxiv*, 2025. doi: 10.1101/2025.04.15.649055. URL https://www.biorxiv.org/content/early/2025/04/16/2025.04.15.649055.

[47] Junming Zhao, Chao Zhang, and Yunan Luo. Contrastive fitness learning: Reprogramming protein language models for low-n learning of protein fitness landscape. *bioRxiv*, 2024. doi: 10.1101/2024.02.11.579859. URL https://www.biorxiv.org/content/early/2024/02/12/2024.02.11.579859.

[48] EvolutionaryScale. Cambrian open license agreement. URL https://www.evolutionaryscale.ai/policies/cambrian-open-license-agreement. Accessed: 2025-5-15.

[49] Marion Sourisseau, Daniel J. P. Lawrence, Megan C. Schwarz, Carina H. Storrs, Ethan C. Veit, Jesse D. Bloom, and Matthew J. Evans. Deep Mutational Scanning Comprehensively Maps How Zika Envelope Protein Mutations Affect Viral Growth and Antibody Escape. *Journal of Virology*, 93(23):e01291–19, December 2019. ISSN 0022-538X, 1098-5514. doi: 10.1128/JVI.01291-19. URL https://journals.asm.org/doi/10.1128/JVI.01291-19.

[50] Zhifeng Deng, Wanzhi Huang, Erol Bakkalbasi, Nicholas G. Brown, Carolyn J. Adamski, Kacie Rice, Donna Muzny, Richard A. Gibbs, and Timothy Palzkill. Deep Sequencing of Systematic Combinatorial Libraries Reveals beta-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.

[51] Clara J. Amorosi, Melissa A. Chiasson, Matthew G. McDonald, Lai Hong Wong, Katherine A. Sitko, Gabriel Boyle, John P. Kowalski, Allan E. Rettie, Douglas M. Fowler, and Maitreya J. Dunham. 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.

[52] Melissa A Chiasson, Nathan J Rollins, Jason J Stephany, Katherine A Sitko, Kenneth A Matreyek, Marta Verby, Song Sun, Frederick P Roth, Daniel DeSloover, Debora S Marks, Allan E Rettie, and Douglas M Fowler. 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.

[53] Emily E. Wrenbeck, Laura R. Azouz, and Timothy A. Whitehead. 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.

[54] Lisa Brenan, Aleksandr Andreev, Ofir Cohen, Sasha Pantel, Atanas Kamburov, Davide Cacchiarelli, Nicole S. Persky, Cong Zhu, Mukta Bagul, Eva M. Goetz, Alex B. Burgin, Levi A. Garraway, Gad Getz, Tarjei S. Mikkelsen, Federica Piccioni, David E. Root, and Cory M. Johannessen. 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.

[55] Krystian A. Kozek, Andrew M. Glazer, Chai-Ann Ng, Daniel Blackwell, Christian L. Egly, Loren R. Vanags, Marcia Blair, Devyn Mitchell, Kenneth A. Matreyek, Douglas M. Fowler, Bjorn C. Knollmann, Jamie I. Vandenberg, Dan M. Roden, and Brett M. Kroncke.

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.

[56] Dan Davidi, Melina Shamshoum, Zhijun Guo, Yinon M Bar-On, Noam Prywes, Aia Oz, Jagoda Jablonska, Avi Flamholz, David G Wernick, Niv Antonovsky, Benoit Pins, Lior Shachar, Dina Hochhauser, Yoav Peleg, Shira Albeck, Itai Sharon, Oliver Mueller-Cajar, and Ron Milo. Highly active rubiscos discovered by systematic interrogation of natural sequence diversity. *The EMBO Journal*, 39(18), September 2020. ISSN 0261-4189, 1460-2075. doi: 10.15252/embj.2019104081. URL https://onlinelibrary.wiley.com/doi/10.15252/embj.2019104081.

[57] Karen S. Sarkisyan, Dmitry A. Bolotin, Margarita V. Meer, Dinara R. Usmanova, Alexander S. Mishin, George V. Sharonov, Dmitry N. Ivankov, Nina G. Bozhanova, Mikhail S. Baranov, Onuralp Soylemez, Natalya S. Bogatyreva, Peter K. Vlasov, Evgeny S. Egorov, Maria D. Logacheva, Alexey S. Kondrashov, Dmitry M. Chudakov, Ekaterina V. Putintseva, Ilgar Z. Mamedov, Dan S. Tawfik, Konstantin A. Lukyanov, and Fyodor A. Kondrashov. Local fitness landscape of the green fluorescent protein. *Nature*, 533(7603):397–401, May 2016. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature17995. URL http://www.nature.com/articles/nature17995.

[58] C. Anders Olson, Nicholas C. Wu, and Ren Sun. 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.

[59] Maria Duenas-Decamp, Li Jiang, Daniel Bolon, and Paul R. Clapham. Saturation Mutagenesis of the HIV-1 Envelope CD4 Binding Loop Reveals Residues Controlling Distinct Trimer Conformations. *PLOS Pathogens*, 12(11):e1005988, November 2016. ISSN 1553-7374. doi: 10.1371/journal.ppat.1005988. URL https://dx.plos.org/10.1371/journal.ppat.1005988.

[60] Samuel Thompson, Yang Zhang, Christine Ingle, Kimberly A Reynolds, and Tanja Kortemme. 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.

[61] Lea M. Starita, Jonathan N. Pruneda, Russell S. Lo, Douglas M. Fowler, Helen J. Kim, Joseph B. Hiatt, Jay Shendure, Peter S. Brzovic, Stanley Fields, and Rachel E. Klevit. Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proceedings of the National Academy of Sciences*, 110(14), April 2013. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1303309110. URL https://pnas.org/doi/full/10.1073/pnas.1303309110.

[62] Carlos L. Araya, Douglas M. Fowler, Wentao Chen, Ike Muniez, Jeffery W. Kelly, and Stanley Fields. 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 https://pnas.org/doi/full/10.1073/pnas.1209751109.

[63] Nicholas C. Wu, Arthur P. Young, Laith Q. Al-Mawsawi, C. Anders Olson, Jun Feng, Hangfei Qi, Shu-Hwa Chen, I.-Hsuan Lu, Chung-Yen Lin, Robert G. Chin, Harding H. Luan, Nguyen Nguyen, Stanley F. Nelson, Xinmin Li, Ting-Ting Wu, and Ren Sun. High-throughput profiling of influenza A virus hemagglutinin gene at single-nucleotide resolution. *Scientific Reports*, 4(1):4942, December 2014. ISSN 2045-2322. doi: 10.1038/srep04942. URL https://www.nature.com/articles/srep04942.

[64] Heather J. Young, Matthew Chan, Balaji Selvam, Steven K. Szymanski, Diwakar Shukla, and Erik Procko. 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.

[65] Li Jiang, Ping Liu, Claudia Bank, Nicholas Renzette, Kristina Prachanronarong, Lutfu S. Yilmaz, Daniel R. Caffrey, Konstantin B. Zeldovich, Celia A. Schiffer, Timothy F. Kowalik, Jeffrey D. Jensen, Robert W. Finberg, Jennifer P. Wang, and Daniel N.A. Bolon. A Balance between Inhibitor Binding and Substrate Processing Confers Influenza Drug Resistance. *Journal of Molecular Biology*, 428(3):538–553, February 2016. ISSN 00222836. doi: 10.1016/j.jmb.2015.11.027. URL https://linkinghub.elsevier.com/retrieve/pii/S0022283615006907.

[66] Daniel Melamed, David L. Young, Caitlin E. Gamble, Christina R. Miller, and Stanley Fields. 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.

[67] Xiaoyan Jia, Bala Bharathi Burugula, Victor Chen, Rosemary M. Lemons, Sajini Jayakody, Mariam Maksutova, and Jacob O. Kitzman. 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.

[68] David Mavor, Kyle Barlow, Samuel Thompson, Benjamin A Barad, Alain R Bonny, Clinton L Cario, Garrett Gaskins, Zairan Liu, Laura Deming, Seth D Axen, Elena Caceres, Weilin Chen, Adolfo Cuesta, Rachel E Gate, Evan M Green, Kaitlin R Hulce, Weiyue Ji, Lillian R Kenner, Bruk Mensa, Leanna S Morinishi, Steven M Moss, Marco Mravic, Ryan K Muir, Stefan Niekamp, Chimno I Nnadi, Eugene Palovcak, Erin M Poss, Tyler D Ross, Eugenia C Salcedo, Stephanie K See, Meena Subramaniam, Allison W Wong, Jennifer Li, Kurt S Thorn, Shane Ó Conchúir, Benjamin P Roscoe, Eric D Chow, Joseph L DeRisi, Tanja Kortemme, Daniel N Bolon, and James S Fraser. 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.

[69] Andrew M. Glazer, Brett M. Kroncke, Kenneth A. Matreyek, Tao Yang, Yuko Wada, Tiffany Shields, Joe-Elie Salem, Douglas M. Fowler, and Dan M. Roden. Deep Mutational Scan of an *SCN5A* Voltage Sensor. *Circulation: Genomic and Precision Medicine*, 13(1):e002786, February 2020. ISSN 2574-8300. doi: 10.1161/CIRCGEN.119.002786. URL https://www.ahajournals.org/doi/10.1161/CIRCGEN.119.002786.

[70] Eric D. Kelsic, Hattie Chung, Niv Cohen, Jimin Park, Harris H. Wang, and Roy Kishony. 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.

[71] Liat Rockah-Shmuel, Ágnes Tóth-Petróczy, and Dan S. Tawfik. 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.

[72] Jacob O Kitzman, Lea M Starita, Russell S Lo, Stanley Fields, and Jay Shendure. 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.

[73] Christopher D. Aakre, Julien Herrou, Tuyen N. Phung, Barrett S. Perchuk, Sean Crosson, and Michael T. Laub. 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.

[74] Benedetta Bolognesi, Andre J. Faure, Mireia Seuma, Jörn M. Schmiedel, Gian Gaetano Tartaglia, and Ben Lehner. The mutational landscape of a prion-like domain. *Nature Communications*, 10(1):4162, December 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-12101-z. URL http://www.nature.com/articles/s41467-019-12101-z.

[75] Julia M. Flynn, Neha Samant, Gily Schneider-Nachum, David T. Barkan, Nese Kurt Yilmaz, Celia A. Schiffer, Stephanie A. Moquin, Dustin Dovala, and Daniel N.A. Bolon. Comprehensive fitness landscape of SARS-CoV-2 M^{pro} reveals insights into viral resistance mechanisms. preprint, Molecular Biology, January 2022. URL http://biorxiv.org/lookup/doi/10.1101/2022.01.26.477860.

[76] Hugh K Haddox, Adam S Dingens, Sarah K Hilton, Julie Overbaugh, and Jesse D Bloom. 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.

[77] Michael A. Stiffler, Doeke R. Hekstra, and Rama Ranganathan. Evolvability as a Function of Purifying Selection in TEM-1 beta-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.

[78] Arti Tripathi, Kritika Gupta, Shruti Khare, Pankaj C. Jain, Siddharth Patel, Prasanth Kumar, Ajai J. Pulianmackal, Nilesh Aghera, and Raghavan Varadarajan. 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.

[79] Gregory M. Findlay, Riza M. Daza, Beth Martin, Melissa D. Zhang, Anh P. Leith, Molly Gasperini, Joseph D. Janizek, Xingfan Huang, Lea M. Starita, and Jay Shendure. 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.

[80] Juhye M. Lee, John Huddleston, Michael B. Doud, Kathryn A. Hooper, Nicholas C. Wu, Trevor Bedford, and Jesse D. Bloom. Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants. *Proceedings of the National Academy of Sciences*, 115(35), August 2018. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1806133115. URL https://pnas.org/doi/full/10.1073/pnas.1806133115.

[81] Jochen Weile, Song Sun, Atina G Cote, Jennifer Knapp, Marta Verby, Joseph C Mellor, Yingzhou Wu, Carles Pons, Cassandra Wong, Natascha Lieshout, Fan Yang, Murat Tasan, Guihong Tan, Shan Yang, Douglas M Fowler, Robert Nussbaum, Jesse D Bloom, Marc Vidal, David E Hill, Patrick Aloy, and Frederick P Roth. 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.

[82] Hangfei Qi, C. Anders Olson, Nicholas C. Wu, Ruian Ke, Claude Loverdo, Virginia Chu, Shawna Truong, Roland Remenyi, Zugen Chen, Yushen Du, Sheng-Yao Su, Laith Q. Al-Mawsawi, Ting-Ting Wu, Shu-Hua Chen, Chung-Yen Lin, Weidong Zhong, James O. Lloyd-Smith, and Ren Sun. 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.

[83] Yvonne H. Chan, Sergey V. Venev, Konstantin B. Zeldovich, and C. Robert Matthews. 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.

[84] Alexandre Melnikov, Peter Rogov, Li Wang, Andreas Gnirke, and Tarjei S. Mikkelsen. 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.

[85] Christina Nutschel, Alexander Fulton, Olav Zimmermann, Ulrich Schwaneberg, Karl-Erich Jaeger, and Holger Gohlke. 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.

[86] Hervé Jacquier, André Birgy, Hervé Le Nagard, Yves Mechulam, Emmanuelle Schmitt, Jérémy Glodt, Beatrice Bercot, Emmanuelle Petit, Julie Poulain, Guilène Barnaud, Pierre-Alexis Gros, and Olivier Tenaillon. 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 https://pnas.org/doi/full/10.1073/pnas.1215206110.

[87] Parul Mishra, Julia M. Flynn, Tyler N. Starr, and Daniel N.A. Bolon. 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.

[88] Victoria O. Pokusaeva, Dinara R. Usmanova, Ekaterina V. Putintseva, Lorena Espinar, Karen S. Sarkisyan, Alexander S. Mishin, Natalya S. Bogatyreva, Dmitry N. Ivankov, Arseniy V. Akopyan, Sergey Ya. Avvakumov, Inna S. Povolotskaya, Guillaume J. Filion, Lucas B. Carey, and Fyodor A. Kondrashov. 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.

[89] Jason D. Fernandes, Tyler B. Faust, Nicolas B. Strauli, Cynthia Smith, David C. Crosby, Robert L. Nakamura, Ryan D. Hernandez, and Alan D. Frankel. 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.

[90] Sam Sinai, Nina Jain, George M Church, and Eric D Kelsic. Generative AAV capsid diversification by latent interpolation. preprint, Synthetic Biology, April 2021. URL http://biorxiv.org/lookup/doi/10.1101/2021.04.16.440236.

[91] Nicholas C. Wu, C. Anders Olson, Yushen Du, Shuai Le, Kevin Tran, Roland Remenyi, Danyang Gong, Laith Q. Al-Mawsawi, Hangfei Qi, Ting-Ting Wu, and Ren Sun. 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.

[92] Kenneth A. Matreyek, Jason J. Stephany, Ethan Ahler, and Douglas M. Fowler. Integrating thousands of PTEN variant activity and abundance measurements reveals variant subgroups and new dominant negatives in cancers. *Genome Medicine*, 13(1):165, December 2021. ISSN 1756-994X. doi: 10.1186/s13073-021-00984-x. URL https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-021-00984-x.

[93] Taylor L. Mighell, Sara Evans-Dutson, and Brian J. O'Roak. 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.

[94] William P. Russ, Matteo Figliuzzi, Christian Stocker, Pierre Barrat-Charlaix, Michael Socolich, Peter Kast, Donald Hilvert, Remi Monasson, Simona Cocco, Martin Weigt, and Rama Ranganathan. An evolution-based model for designing chorismate mutase enzymes. *Science*, 369(6502):440–445, July 2020. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aba3304. URL https://www.sciencemag.org/lookup/doi/10.1126/science.aba3304.

[95] Courtney E. Gonzalez, Paul Roberts, and Marc Ostermeier. Fitness Effects of Single Amino Acid Insertions and Deletions in TEM-1 beta-Lactamase. *Journal of Molecular Biology*, 431(12):2320–2330, May 2019. ISSN 00222836. doi: 10.1016/j.jmb.2019.04.030. URL https://linkinghub.elsevier.com/retrieve/pii/S0022283619302372.

[96] Christian B. Macdonald, David Nedrud, Patrick Rockefeller Grimes, Donovan Trinidad, James S. Fraser, and Willow Coyote-Maestas. DIMPLE: deep insertion, deletion, and missense mutation libraries for exploring protein variation in evolution, disease, and biology. *Genome Biology*, 24(1):36, February 2023. ISSN 1474-760X. doi: 10.1186/s13059-023-02880-6. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-023-02880-6.

[97] Eric M Jones, Nathan B Lubock, Aj Venkatakrishnan, Jeffrey Wang, Alex M Tseng, Joseph M Paggi, Naomi R Latorraca, Daniel Cancilla, Megan Satyadi, Jessica E Davis, M Madan Babu, Ron O Dror, and Sriram Kosuri. 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.

[98] John Z Chen, Douglas M Fowler, and Nobuhiko Tokuriki. 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.

[99] Paul Kennouche, Arthur Charles-Orszag, Daiki Nishiguchi, Sylvie Goussard, Anne-Flore Imhaus, Mathieu Dupré, Julia Chamot-Rooke, and Guillaume Duménil. Deep mutational scanning of the *Neisseria meningitidis* major pilin reveals the importance of pilus tip-mediated adhesion. *The EMBO Journal*, 38(22):e102145, November 2019. ISSN 0261-4189, 1460-2075. doi: 10.15252/embj.2019102145. URL https://www.embopress.org/doi/10.15252/embj.2019102145.

[100] Hugh K. Haddox, Adam S. Dingens, and Jesse D. Bloom. 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.

[101] Benjamin P. Roscoe and Daniel N.A. Bolon. 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.

[102] Justin R. Klesmith, John-Paul Bacik, Ryszard Michalczyk, and Timothy A. Whitehead. 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.

[103] Louisa Gonzalez Somermeyer, Aubin Fleiss, Alexander S Mishin, Nina G Bozhanova, Anna A Igolkina, Jens Meiler, Maria-Elisenda Alaball Pujol, Ekaterina V Putintseva, Karen S Sarkisyan, and Fyodor A Kondrashov. Heterogeneity of the GFP fitness landscape and data-driven protein design. *eLife*, 11:e75842, May 2022. ISSN 2050-084X. doi: 10.7554/eLife.75842. URL https://elifesciences.org/articles/75842.

[104] Michael Doud and Jesse Bloom. 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.

[105] Yq Shirleen Soh, Louise H Moncla, Rachel Eguia, Trevor Bedford, and Jesse D Bloom. 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.

[106] Mireia Seuma, Andre J Faure, Marta Badia, Ben Lehner, and Benedetta Bolognesi. 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.

[107] Rohan Dandage, Rajesh Pandey, Gopal Jayaraj, Manish Rai, David Berger, and Kausik Chakraborty. 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.

[108] Elad Firnberg, Jason W. Labonte, Jeffrey J. Gray, and Marc Ostermeier. 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.

[109] Bharat V. Adkar, Arti Tripathi, Anusmita Sahoo, Kanika Bajaj, Devrishi Goswami, Purbani Chakrabarti, Mohit K. Swarnkar, Rajesh S. Gokhale, and Raghavan Varadarajan. 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.

[110] Andrew O. Giacomelli, Xiaoping Yang, Robert E. Lintner, James M. McFarland, Marc Duby, Jaegil Kim, Thomas P. Howard, David Y. Takeda, Seav Huong Ly, Eejung Kim, Hugh S. Gannon, Brian Hurhula, Ted Sharpe, Amy Goodale, Briana Fritchman, Scott Steelman, Francisca Vazquez, Aviad Tsherniak, Andrew J. Aguirre, John G. Doench, Federica Piccioni, Charles W. M. Roberts, Matthew Meyerson, Gad Getz, Cory M. Johannessen, David E. Root, and William C. Hahn. 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 https://www.nature.com/articles/s41588-018-0204-y.

[111] Chase C. Suiter, Takaya Moriyama, Kenneth A. Matreyek, Wentao Yang, Emma Rose Scaletti, Rina Nishii, Wenjian Yang, Keito Hoshitsuki, Minu Singh, Amita Trehan, Chris Parish, Colton Smith, Lie Li, Deepa Bhojwani, Liz Y. P. Yuen, Chi-kong Li, Chak-ho Li, Yung-li Yang, Gareth J. Walker, James R. Goodhand, Nicholas A. Kennedy, Federico Antillon Klussmann, Smita Bhatia, Mary V. Relling, Motohiro Kato, Hiroki Hori, Prateek Bhatia, Tariq Ahmad, Allen E. J. Yeoh, Pål Stenmark, Douglas M. Fowler, and Jun J. Yang. 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 https://pnas.org/doi/full/10.1073/pnas.1915680117.

[112] Helen T. Hobbs, Neel H. Shah, Sophie R. Shoemaker, Jeanine F. Amacher, Susan Marqusee, and John Kuriyan. Saturation mutagenesis of a predicted ancestral Syk-family kinase. *Protein Science*, 31(10), October 2022. ISSN 0961-8368, 1469-896X. doi: 10.1002/pro.4411. URL https://onlinelibrary.wiley.com/doi/10.1002/pro.4411.

[113] Sarah Gersing, Matteo Cagiada, Marinella Gebbia, Anette P. Gjesing, Atina G. Coté, Gireesh Seesankar, Roujia Li, Daniel Tabet, Amelie Stein, Anna L. Gloyn, Torben Hansen, Frederick P. Roth, Kresten Lindorff-Larsen, and Rasmus Hartmann-Petersen. A comprehensive map of human glucokinase variant activity. preprint, Genetics, May 2022. URL http://biorxiv.org/lookup/doi/10.1101/2022.05.04.490571.

[114] Max V. Staller, Alex S. Holehouse, Devjanee Swain-Lenz, Rahul K. Das, Rohit V. Pappu, and Barak A. Cohen. A High-Throughput Mutational Scan of an Intrinsically Disordered Acidic Transcriptional Activation Domain. *Cell Systems*, 6(4):444–455.e6, April 2018. ISSN 24054712. doi: 10.1016/j.cels.2018.01.015. URL https://linkinghub.elsevier.com/retrieve/pii/S2405471218300528.

[115] Pradeep Bandaru, Neel H Shah, Moitrayee Bhattacharyya, John P Barton, Yasushi Kondo, Joshua C Cofsky, Christine L Gee, Arup K Chakraborty, Tanja Kortemme, Rama Ranganathan, and John Kuriyan. 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.

[116] Jessica L. Bridgford, Su Min Lee, Christine M. M. Lee, Paola Guglielmelli, Elisa Rumi, Daniela Pietra, Stephen Wilcox, Yash Chhabra, Alan F. Rubin, Mario Cazzola, Alessandro M. Vannucchi, Andrew J. Brooks, Matthew E. Call, and Melissa J. Call. 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.

[117] Jeffrey M. Spencer and Xiaoliu Zhang. Deep mutational scanning of S. pyogenes Cas9 reveals important functional domains. *Scientific Reports*, 7(1):16836, December 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-17081-y. URL https://www.nature.com/articles/s41598-017-17081-y.

[118] Richard N. McLaughlin, Jr., Frank J. Poelwijk, Arjun Raman, Walraj S. Gosal, and Rama Ranganathan. 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 https://www.nature.com/articles/nature11500.

[119] Michael B. Doud, Orr Ashenberg, and Jesse D. Bloom. 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.

[120] Florian Mattenberger, Victor Latorre, Omer Tirosh, Adi Stern, and Ron Geller. 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.

[121] Kenneth A. Matreyek, Lea M. Starita, Jason J. Stephany, Beth Martin, Melissa A. Chiasson, Vanessa E. Gray, Martin Kircher, Arineh Khechaduri, Jennifer N. Dines, Ronald J. Hause, Smita Bhatia, William E. Evans, Mary V. Relling, Wenjian Yang, Jay Shendure, and Douglas M. Fowler. 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 https://www.nature.com/articles/s41588-018-0122-z.

[122] Julia M Flynn, Ammeret Rossouw, Pamela Cote-Hammarlof, Inês Fragata, David Mavor, Carl Hollins, Claudia Bank, and Daniel Na Bolon. 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.

[123] Ethan Ahler, Ames C. Register, Sujata Chakraborty, Linglan Fang, Emily M. Dieter, Katherine A. Sitko, Rama Subba Rao Vidadala, Bridget M. Trevillian, Martin Golkowski, Hannah Gelman, Jason J. Stephany, Alan F. Rubin, Ethan A. Merritt, Douglas M. Fowler, and Dustin J. Maly. 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.

[124] Philip A. Romero, Tuan M. Tran, and Adam R. Abate. 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 https://pnas.org/doi/full/10.1073/pnas.1422285112.

[125] Andre J. Faure, Júlia Domingo, Jörn M. Schmiedel, Cristina Hidalgo-Carcedo, Guillaume Diss, and Ben Lehner. Mapping the energetic and allosteric landscapes of protein binding domains. *Nature*, 604(7904):175–183, April 2022. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-022-04586-4. URL https://www.nature.com/articles/s41586-022-04586-4.

[126] Benjamin P. Roscoe, Kelly M. Thayer, Konstantin B. Zeldovich, David Fushman, and Daniel N.A. Bolon. 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.

[127] Robert W. Newberry, Taylor Arhar, Jean Costello, George C. Hartoularos, Alison M. Maxwell, Zun Zar Chi Naing, Maureen Pittman, Nishith R. Reddy, Daniel M. C. Schwarz, Douglas R. Wassarman, Taia S. Wu, Daniel Barrero, Christa Caggiano, Adam Catching, Taylor B. Cavazos, Laurel S. Estes, Bryan Faust, Elissa A. Fink, Miriam A. Goldman, Yessica K. Gomez, M. Grace Gordon, Laura M. Gunsalus, Nick Hoppe, Maru Jaime-Garza, Matthew C. Johnson, Matthew G. Jones, Andrew F. Kung, Kyle E. Lopez, Jared Lumpe, Calla Martyn, Elizabeth E. McCarthy, Lakshmi E. Miller-Vedam, Erik J. Navarro, Aji Palar, Jenna Pellegrino, Wren Saylor, Christina A. Stephens, Jack Strickland, Hayarpi Torosyan, Stephanie A. Wankowicz, Daniel R. Wong, Garrett Wong, Sy Redding, Eric D. Chow, William F. DeGrado, and Martin Kampmann. Robust Sequence Determinants of alpha-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.

[128] Nicholas C Wu, Lei Dai, C Anders Olson, James O Lloyd-Smith, and Ren Sun. Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife*, 5:e16965, July 2016. ISSN 2050-084X. doi: 10.7554/eLife.16965. URL https://elifesciences.org/articles/16965.

[129] Tyler N. Starr, Allison J. Greaney, Sarah K. Hilton, Daniel Ellis, Katharine H.D. Crawford, Adam S. Dingens, Mary Jane Navarro, John E. Bowen, M. Alejandra Tortorici, Alexandra C. Walls, Neil P. King, David Veesler, and Jesse D. Bloom. 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.

[130] Daniel D. Brauer, Celine B. Santiago, Zoe N. Merz, Esther McCarthy, Danielle Tullman-Ercek, and Matthew B. Francis. Comprehensive Fitness Landscape of a Multi-Geometry Protein Capsid Informs Machine Learning Models of Assembly. preprint, Bioengineering, December 2021. URL http://biorxiv.org/lookup/doi/10.1101/2021.12.21.473721.

[131] Thomas W. Linsky, Renan Vergara, Nuria Codina, Jorgen W. Nelson, Matthew J. Walker, Wen Su, Christopher O. Barnes, Tien-Ying Hsiang, Katharina Esser-Nobis, Kevin Yu, Z. Beau Reneer, Yixuan J. Hou, Tanu Priya, Masaya Mitsumoto, Avery Pong, Uland Y. Lau, Marsha L. Mason, Jerry Chen, Alex Chen, Tania Berrocal, Hong Peng, Nicole S. Clairmont, Javier Castellanos, Yu-Ru Lin, Anna Josephson-Day, Ralph S. Baric, Deborah H. Fuller, Carl D. Walkey, Ted M. Ross, Ryan Swanson, Pamela J. Bjorkman, Michael Gale, Luis M. Blancas-Mejia, Hui-Ling Yen, and Daniel-Adriano Silva. De novo design of potent and resilient hACE2 decoys to neutralize SARS-CoV-2. *Science*, 370(6521):1208–1214, December 2020. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.abe0075. URL https://www.science.org/doi/10.1126/science.abe0075.

[132] Tobias Stadelmann, Daniel Heid, Michael Jendrusch, Jan Mathony, Stéphane Rosset, Bruno E. Correia, and Dominik Niopek. A deep mutational scanning platform to characterize the fitness landscape of anti-CRISPR proteins. preprint, Synthetic Biology, August 2021. URL http://biorxiv.org/lookup/doi/10.1101/2021.08.21.457204.

[133] Philipp Koch, Steven Schmitt, Alexander Heynisch, Anja Gumpinger, Irene Wüthrich, Marina Gysin, Dimitri Shcherbakov, Sven N. Hobbie, Sven Panke, and Martin Held. Optimization of the antimicrobial peptide Bac7 by deep mutational scanning. *BMC Biology*, 20(1):114, December 2022. ISSN 1741-7007. doi: 10.1186/s12915-022-01304-4. URL https://bmcbiol.biomedcentral.com/articles/10.1186/s12915-022-01304-4.

[134] Eran Kotler, Odem Shani, Guy Goldfeld, Maya Lotan-Pompan, Ohad Tarcic, Anat Gershoni, Thomas A. Hopf, Debora S. Marks, Moshe Oren, and Eran Segal. 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.

[135] Ian J. Campbell, Joshua T. Atkinson, Matthew D. Carpenter, Dru Myerscough, Lin Su, Caroline M. Ajo-Franklin, and Jonathan J. Silberg. Determinants of Multiheme Cytochrome Extracellular Electron Transfer Uncovered by Systematic Peptide Insertion. *Biochemistry*, 61 (13):1337–1350, July 2022. ISSN 0006-2960, 1520-4995. doi: 10.1021/acs.biochem.2c00148. URL https://pubs.acs.org/doi/10.1021/acs.biochem.2c00148.

[136] David Ding, Anna G. Green, Boyuan Wang, Thuy-Lan Vo Lite, Eli N. Weinstein, Debora S. Marks, and Michael T. Laub. Co-evolution of interacting proteins through non-contacting and non-specific mutations. *Nature Ecology & Evolution*, 6(5):590–603, March 2022. ISSN 2397-334X. doi: 10.1038/s41559-022-01688-0. URL https://www.nature.com/articles/s41559-022-01688-0.

[137] Kotaro Tsuboyama, Justas Dauparas, Jonathan Chen, Elodie Laine, Yasser Mohseni Behbahani, Jonathan J. Weinstein, Niall M. Mangan, Sergey Ovchinnikov, and Gabriel J. Rocklin. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature*, 620(7973):434–444, August 2023. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-023-06328-6. URL https://www.nature.com/articles/s41586-023-06328-6.

[138] Amporn Suphatrakul, Pratsaneeyaporn Posiri, Nittaya Srisuk, Rapirat Nantachokchawapan, Suppachoke Onnome, Juthathip Mongkolsapaya, and Bunpote Siridechadilok. Functional analysis of flavivirus replicase by deep mutational scanning of dengue NS5. March 2023.

[139] Jochen Weile, Nishka Kishore, Song Sun, Ranim Maaieh, Marta Verby, Roujia Li, Iosifina Fotiadou, Julia Kitaygorodsky, Yingzhou Wu, Alexander Holenstein, Céline Bürer, Linnea Blomgren, Shan Yang, Robert Nussbaum, Rima Rozen, David Watkins, Marinella Gebbia, Viktor Kozich, Michael Garton, D Sean Froese, and Frederick P Roth. Shifting landscapes of human MTHFR missense-variant effects. *American Journal of Human Genetics*, 108(7):1283–1300, July 2021.

[140] Michael J Xie, Gareth A Cromie, Katherine Owens, Martin S Timour, Michelle Tang, J Nathan Kutz, Ayman W El-Hattab, Richard N McLaughlin, and Aimée M Dudley. Predicting the functional effect of compound heterozygous genotypes from large scale variant effect maps. *bioRxiv*, January 2023.

[141] Hridindu Roychowdhury and Philip A Romero. Microfluidic deep mutational scanning of the human executioner caspases reveals differences in structure and regulation. *Cell Death Discovery*, 8(1):7, January 2022.

[142] Bryan Andrews and Stanley Fields. Distinct patterns of mutational sensitivity for λ resistance and maltodextrin transport in escherichia coli LamB. *Microbial Genomics*, 6(4), April 2020.

[143] Russell S Lo, Gareth A Cromie, Michelle Tang, Kevin Teng, Katherine Owens, Amy Sirr, J Nathan Kutz, Hiroki Morizono, Ljubica Caldovic, Nicholas Ah Mew, Andrea Gropman, and Aimée M Dudley. The functional impact of 1,570 individual amino acid substitutions in human OTC. *American Journal of Human Genetics*, 110(5):863–879, May 2023.

[144] Kiran S Gajula, Peter J Huwe, Charlie Y Mo, Daniel J Crawford, James T Stivers, Ravi Radhakrishnan, and Rahul M Kohli. High-throughput mutagenesis reveals functional determinants for DNA targeting by activation-induced deaminase. *Nucleic Acids Research*, 42(15):9964–9975, September 2014.

[145] Sujata Chakraborty, Ethan Ahler, Jessica J Simon, Linglan Fang, Zachary E Potter, Katherine A Sitko, Jason J Stephany, Miklos Guttman, Douglas M Fowler, and Dustin J Maly. Profiling of the drug resistance of thousands of src tyrosine kinase mutants uncovers a regulatory network that couples autoinhibition to catalytic domain dynamics. December 2021.

[146] Chenchun Weng, Andre J Faure, and Ben Lehner. The energetic and allosteric landscape for KRAS inhibition. December 2022.

[147] David Ding, Ada Shaw, Sam Sinai, Nathan Rollins, Noam Prywes, David F Savage, Michael T Laub, and Debora S Marks. Protein design using structure-based residue preferences. June 2023.

[148] Vanessa Nguyen, Ethan Ahler, Katherine A Sitko, Jason J Stephany, Dustin J Maly, and Douglas M Fowler. Molecular determinants of hsp90 dependence of src kinase revealed by deep mutational scanning. *Protein Science*, 32(7):e4656, July 2023.

[149] Mireia Seuma, Ben Lehner, and Benedetta Bolognesi. An atlas of amyloid aggregation: the impact of substitutions, insertions, deletions and truncations on amyloid beta fibril nucleation. *Nature Communications*, 13(1):7084, November 2022.

[150] Oana Ursu, James T Neal, Emily Shea, Pratiksha I Thakore, Livnat Jerby-Arnon, Lan Nguyen, Danielle Dionne, Celeste Diaz, Julia Bauman, Mariam Mounir Mosaad, Christian Fagre, April Lo, Maria McSharry, Andrew O Giacomelli, Seav Huong Ly, Orit Rozenblatt-Rosen, William C Hahn, Andrew J Aguirre, Alice H Berger, Aviv Regev, and Jesse S Boehm. Massively parallel phenotyping of coding variants in cancer with perturb-seq. *Nature Biotechnology*, 40(6): 896–905, June 2022.

[151] Emily E Wrenbeck, Matthew A Bedewitz, Justin R Klesmith, Syeda Noshin, Cornelius S Barry, and Timothy A Whitehead. An automated Data-Driven pipeline for improving heterologous enzyme expression. *ACS Synthetic Biology*, 8(3):474–481, March 2019.

[152] Nancy Hom, Lauren Gentles, Jesse D Bloom, and Kelly K Lee. Deep mutational scan of the highly conserved influenza a virus M1 matrix protein reveals substantial intrinsic mutational tolerance. *Journal of Virology*, 93(13), July 2019.

[153] Vanessa E Gray, Katherine Sitko, Floriane Z Ngako Kameni, Miriam Williamson, Jason J Stephany, Nicholas Hasle, and Douglas M Fowler. Elucidating the molecular determinants of Aβ aggregation with deep mutational scanning. *G3*, 9(11):3683–3689, November 2019.

[154] Veeramohan Veerapandian, Jan Ole Ackermann, Yogesh Srivastava, Vikas Malik, Mingxi Weng, Xiaoxiao Yang, and Ralf Jauch. Directed evolution of reprogramming factors by cell selection and sequencing. *Stem Cell Reports*, 11(2):593–606, August 2018.

[155] Bargavi Thyagarajan and Jesse D Bloom. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *Elife*, 3, July 2014.

[156] Martin K Ostermaier, Christian Peterhans, Rolf Jaussi, Xavier Deupi, and Jörg Standfuss. Functional map of arrestin-1 at single amino acid resolution. *Proceedings of the National Academy of Sciences*, 111(5):1825–1830, February 2014.

[157] Jesse D Bloom. An experimentally determined evolutionary model dramatically improves phylogenetic fit. *Molecular Biology and Evolution*, 31(8):1956–1978, August 2014.

[158] Ryan T Hietpas, Jeffrey D Jensen, and Daniel N A Bolon. Experimental illumination of a fitness landscape. *Proceedings of the National Academy of Sciences of the United States of America*, 108(19):7896–7901, May 2011.

[159] Kevin S Gill, Kritika Mehta, Jeremiah D Heredia, Vishnu V Krishnamurthy, Kai Zhang, and Erik Procko. Multiple mechanisms of self-association of chemokine receptors CXCR4 and CCR5 demonstrated by deep mutagenesis. *bioRxiv*, March 2023.

[160] Gianmarco Meier, Sujani Thavarasah, Kai Ehrenbolger, Cedric A J Hutter, Lea M Hürlimann, Jonas Barandun, and Markus A Seeger. Deep mutational scan of a drug efflux pump reveals its structure-function landscape. *Nature Chemical Biology*, 19(4):440–450, April 2023.

[161] Timothy J C Tan, Zongjun Mou, Ruipeng Lei, Wenhao O Ouyang, Meng Yuan, Ge Song, Raiees Andrabi, Ian A Wilson, Collin Kieffer, Xinghong Dai, Kenneth A Matreyek, and Nicholas C Wu. High-throughput identification of prefusion-stabilizing mutations in SARS-CoV-2 spike. *Nature Communications*, 14(1):2003, April 2023.

[162] Mark R MacRae, Dhenesh Puvanendran, Max A B Haase, Nicolas Coudray, Ljuvica Kolich, Cherry Lam, Minkyung Baek, Gira Bhabha, and Damian C Ekiert. Protein-protein interactions in the mla lipid transport system probed by computational structure prediction and deep mutational scanning. *Journal of Biological Chemistry*, 299(6):104744, June 2023.

[163] Dia A Ghose, Kaitlyn E Przydzial, Emily M Mahoney, Amy E Keating, and Michael T Laub. Marginal specificity in protein interactions constrains evolution of a paralogous family. *Proceedings of the National Academy of Sciences of the United States of America*, 120(18): e2221163120, May 2023.

[164] Thuy N Nguyen, Christine Ingle, Samuel Thompson, and Kimberly A Reynolds. The genetic landscape of a metabolic interaction. May 2023.

[165] Lene Clausen, Vasileios Voutsinos, Matteo Cagiada, Kristoffer E Johansson, Martin Grønbæk-Thygesen, Snehal Nariya, Rachel L Powell, Magnus K N Have, Vibe H Oestergaard, Amelie Stein, Douglas M Fowler, Kresten Lindorff-Larsen, and Rasmus Hartmann-Petersen. A mutational atlas for parkin proteostasis. June 2023.

[166] Rosario Vanella, Christoph Küng, Alexandre A Schoepfer, Vanni Doffini, Jin Ren, and Michael A Nash. Understanding Activity-Stability tradeoffs in biocatalysts by enzyme proximity sequencing. March 2023.

[167] Ruipeng Lei, Andrea Hernandez Garcia, Timothy J C Tan, Qi Wen Teo, Yiquan Wang, Xiwen Zhang, Shitong Luo, Satish K Nair, Jian Peng, and Nicholas C Wu. Mutational fitness landscape of human influenza H3N2 neuraminidase. *Cell Reports*, 42(1):111951, January 2023.

[168] Warren van Loggerenberg, Shahin Sowlati-Hashjin, Jochen Weile, Rayna Hamilton, Aditya Chawla, Marinella Gebbia, Nishka Kishore, Laure Frésard, Sami Mustajoki, Elena Pischik, Elena Di Pierro, Michela Barbaro, Ylva Floderus, Caroline Schmitt, Laurent Gouya, Alexandre Colavin, Robert Nussbaum, Edith C H Friesema, Raili Kauppinen, Jordi To-Figueras, Aasne K Aarsand, Robert J Desnick, Michael Garton, and Frederick P Roth. Systematically testing human HMBS missense variants to reveal mechanism and pathogenic variation. *bioRxiv*, February 2023.

[169] Ryan Weeks and Marc Ostermeier. Fitness and functional landscapes of the e. coli RNase III gene rnc. *Molecular Biology and Evolution*, 40(3), March 2023.

[170] Ayesha Muhammad, Maria E Calandranis, Bian Li, Tao Yang, Daniel J Blackwell, M Lorena Harvey, Jeremy E Smith, Ashli E Chew, John A Capra, Kenneth A Matreyek, Douglas M Fowler, Dan M Roden, and Andrew M Glazer. High-throughput functional mapping of variants in an arrhythmia gene, KCNE1 , reveals novel biology. *bioRxiv*, April 2023.

[171] Sook Wah Yee, Christian Macdonald, Darko Mitrovic, Xujia Zhou, Megan L Koleske, Jia Yang, Dina Buitrago Silva, Patrick Rockefeller Grimes, Donovan Trinidad, Swati S More, Linda Kachuri, John S Witte, Lucie Delemotte, Kathleen M Giacomini, and Willow Coyote-Maestas. The full spectrum of OCT1 (SLC22A1) mutations bridges transporter biophysics to drug pharmacogenomics. *bioRxiv*, June 2023.

[172] Yongcan Chen, Ruyun Hu, Keyi Li, Yating Zhang, Lihao Fu, Jianzhi Zhang, and Tong Si. Deep mutational scanning of an Oxygen-Independent fluorescent protein CreiLOV for comprehensive profiling of mutational and epistatic effects. *ACS Synthetic Biology*, 12(5): 1461–1473, May 2023.

[173] Sarah Gersing, Thea K Schulze, Matteo Cagiada, Amelie Stein, Frederick P Roth, Kresten Lindorff-Larsen, and Rasmus Hartmann-Petersen. Characterizing glucokinase variant mechanisms using a multiplexed abundance assay. *bioRxiv*, May 2023.

[174] Zachary M Huttinger, Laura M Haynes, Andrew Yee, Colin A Kretz, Matthew L Holding, David R Siemieniak, Daniel A Lawrence, and David Ginsburg. Deep mutational scanning of the plasminogen activator inhibitor-1 functional landscape. *Scientific Reports*, 11(1):18827, September 2021.

[175] Jason J. Kwon, Behnoush Hajian, Yuemin Bian, Lucy C. Young, Alvaro J. Amor, James R. Fuller, Cara V. Fraley, Abbey M. Sykes, Jonathan So, Joshua Pan, Laura Baker, Sun Joo Lee, Douglas B. Wheeler, David L. Mayhew, Nicole S. Persky, Xiaoping Yang, David E. Root, Anthony M. Barsotti, Andrew W. Stamford, Charles K. Perry, Alex Burgin, Frank McCormick, Christopher T. Lemke, William C. Hahn, and Andrew J. Aguirre. Structure–function analysis of the SHOC2–MRAS–PP1C holophosphatase complex. *Nature*, 609(7926):408–415, September 2022. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-022-04928-2. URL https://www.nature.com/articles/s41586-022-04928-2.

[176] Song Sun, Jochen Weile, Marta Verby, Yingzhou Wu, Yang Wang, Atina G. Cote, Iosifina Fotiadou, Julia Kitaygorodsky, Marc Vidal, Jasper Rine, Pavel Ješina, Viktor Kožich, and Frederick P. Roth. A proactive genotype-to-patient-phenotype map for cystathionine beta-synthase. *Genome Medicine*, 12(1):13, December 2020. ISSN 1756-994X. doi: 10.1186/s13073-020-0711-1. URL https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-020-0711-1.

[177] Aliete Wan, Emily Place, Eric A. Pierce, and Jason Comander. Characterizing variants of unknown significance in rhodopsin: A functional genomics approach. *Human Mutation*, 40 (8):1127–1144, August 2019. ISSN 1059-7794, 1098-1004. doi: 10.1002/humu.23762. URL https://onlinelibrary.wiley.com/doi/10.1002/humu.23762.

[178] Kui K. Chan, Danielle Dorosky, Preeti Sharma, Shawn A. Abbasi, John M. Dye, David M. Kranz, Andrew S. Herbert, and Erik Procko. Engineering human ACE2 to optimize binding to the spike protein of SARS coronavirus 2. *Science*, 369(6508):1261–1265, September 2020. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.abc0870. URL https://www.science.org/doi/10.1126/science.abc0870.

[179] Rachel A. Silverstein, Song Sun, Marta Verby, Jochen Weile, Yingzhou Wu, Marinella Gebbia, Iosifina Fotiadou, Julia Kitaygorodsky, and Frederick P. Roth. A systematic genotype-phenotype map for missense variants in the human intellectual disability-associated gene *GDI1*. preprint, Genetics, October 2021. URL http://biorxiv.org/lookup/doi/10.1101/2021.10.06.463360.

[180] Julia Zinkus-Boltz, Craig DeValk, and Bryan C. Dickinson. A Phage-Assisted Continuous Selection Approach for Deep Mutational Scanning of Protein–Protein Interactions. *ACS Chemical Biology*, 14(12):2757–2767, December 2019. ISSN 1554-8929, 1554-8937. doi: 10.1021/acschembio.9b00669. URL https://pubs.acs.org/doi/10.1021/acschembio.9b00669.

[181] Justin R. Klesmith, Lihe Su, Lan Wu, Ian A. Schrack, Fay J. Dufort, Alyssa Birt, Christine Ambrose, Benjamin J. Hackel, Roy R. Lobb, and Paul D. Rennert. Retargeting CD19 Chimeric Antigen Receptor T Cells via Engineered CD19-Fusion Proteins. *Molecular Pharmaceutics*, 16 (8):3544–3558, August 2019. ISSN 1543-8384, 1543-8392. doi: 10.1021/acs.molpharmaceut.9b00418. URL https://pubs.acs.org/doi/10.1021/acs.molpharmaceut.9b00418.

[182] Assaf Elazar, Jonathan Weinstein, Ido Biran, Yearit Fridman, Eitan Bibi, and Sarel Jacob Fleishman. Mutational scanning reveals the determinants of protein insertion and association energetics in the plasma membrane. *eLife*, 5:e12125, January 2016. ISSN 2050-084X. doi: 10.7554/eLife.12125. URL https://elifesciences.org/articles/12125.

[183] Willow Coyote-Maestas, David Nedrud, Yungui He, and Daniel Schmidt. Determinants of trafficking, conduction, and disease within a K+ channel revealed through multiparametric deep mutational scanning. *eLife*, 11:e76903, May 2022. ISSN 2050-084X. doi: 10.7554/eLife.76903. URL https://elifesciences.org/articles/76903.

[184] Yuan Li, Sarah Arcos, Kimberly R. Sabsay, Aartjan J.W. Te Velthuis, and Adam S. Lauring. Deep mutational scanning reveals the functional constraints and evolutionary potential of the influenza A virus PB1 protein. preprint, Microbiology, August 2023. URL http://biorxiv.org/lookup/doi/10.1101/2023.08.27.554986.

[185] Iana Meitlis, Eric J. Allenspach, Bradly M. Bauman, Isabelle Q. Phan, Gina Dabbah, Erica G. Schmitt, Nathan D. Camp, Troy R. Torgerson, Deborah A. Nickerson, Michael J. Bamshad,

David Hagin, Christopher R. Luthers, Jeffrey R. Stinson, Jessica Gray, Ingrid Lundgren, Joseph A. Church, Manish J. Butte, Mike B. Jordan, Seema S. Aceves, Daniella M. Schwartz, Joshua D. Milner, Susan Schuval, Suzanne Skoda-Smith, Megan A. Cooper, Lea M. Starita, David J. Rawlings, Andrew L. Snow, and Richard G. James. Multiplexed Functional Assessment of Genetic Variants in CARD11. *The American Journal of Human Genetics*, 107 (6):1029–1043, December 2020. ISSN 00029297. doi: 10.1016/j.ajhg.2020.10.015. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929720303736.

[186] UK Monogenic Diabetes Consortium, Myocardial Infarction Genetics Consortium, UK Congenital Lipodystrophy Consortium, Amit R Majithia, Ben Tsuda, Maura Agostini, Keerthana Gnanapradeepan, Robert Rice, Gina Peloso, Kashyap A Patel, Xiaolan Zhang, Marjoleine F Broekema, Nick Patterson, Marc Duby, Ted Sharpe, Eric Kalkhoven, Evan D Rosen, Inês Barroso, Sian Ellard, Sekar Kathiresan, Stephen O'Rahilly, Krishna Chatterjee, Jose C Florez, Tarjei Mikkelsen, David B Savage, and David Altshuler. Prospective functional classification of all possible missense variants in PPARG. *Nature Genetics*, 48(12): 1570–1575, December 2016. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.3700. URL https://www.nature.com/articles/ng.3700.

[187] Gabriella O. Estevam, Edmond M. Linossi, Christian B. Macdonald, Carla A. Espinoza, Jennifer M. Michaud, Willow Coyote-Maestas, Eric A. Collisson, Natalia Jura, and James S. Fraser. Conserved regulatory motifs in the juxtamembrane domain and kinase N-lobe revealed through deep mutational scanning of the MET receptor tyrosine kinase domain. preprint, Molecular Biology, August 2023. URL http://biorxiv.org/lookup/doi/10.1101/2023.08.03.551866.

[188] Peter G. Miller, Murugappan Sathappa, Jamie A. Moroco, Wei Jiang, Yue Qian, Sumaiya Iqbal, Qi Guo, Andrew O. Giacomelli, Subrata Shaw, Camille Vernier, Besnik Bajrami, Xiaoping Yang, Cerise Raffier, Adam S. Sperling, Christopher J. Gibson, Josephine Kahn, Cyrus Jin, Matthew Ranaghan, Alisha Caliman, Merissa Brousseau, Eric S. Fischer, Robert Lintner, Federica Piccioni, Arthur J. Campbell, David E. Root, Colin W. Garvie, and Benjamin L. Ebert. Allosteric inhibition of PPM1D serine/threonine phosphatase via an altered conformational state. *Nature Communications*, 13(1):3778, June 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-30463-9. URL https://www.nature.com/articles/s41467-022-30463-9.

[189] Steven Erwood, Teija M. I. Bily, Jason Lequyer, Joyce Yan, Nitya Gulati, Reid A. Brewer, Liangchi Zhou, Laurence Pelletier, Evgueni A. Ivakine, and Ronald D. Cohn. Saturation variant interpretation using CRISPR prime editing. *Nature Biotechnology*, 40(6):885–895, June 2022. ISSN 1087-0156, 1546-1696. doi: 10.1038/s41587-021-01201-1. URL https://www.nature.com/articles/s41587-021-01201-1.

[190] Rosanna Junchen Jiang. *Exhaustive Mapping of Missense Variation in Coronary Heart Disease-related Genes*. PhD thesis, University of Toronto, November 2019. URL https://hdl.handle.net/1807/98076.

[191] Tianlong Chen, Chengyue Gong, Daniel Jesus Diaz, Xuxi Chen, Jordan Tyler Wells, Qiang Liu, Zhangyang Wang, Andrew Ellington, Alex Dimakis, and Adam Klivans. HotProtein: A Novel Framework for Protein Thermostability Prediction and Editing. October 2022. URL https://openreview.net/forum?id=RtV_iEbWeGE.

A Broader Impact

PoET-2, a multimodal, retrieval-augmented protein model, aims to improve the prediction of mutational effects and enable controllable protein design by learning from sequence, structure, and evolutionary family context. This work can accelerate beneficial applications such as designing novel enzymes, therapeutics, and more stable proteins. PoET-2's enhanced data efficiency in supervised learning may also broaden access to advanced protein engineering, especially in data-limited settings. While PoET-2 is a foundational research tool, advanced capabilities in understanding and designing proteins could theoretically be misused, for example, in the development of dangerous drugs. We expect that future work will serve to address and mitigate such concerns.

B PoET-2 Architecture

This section elaborates on PoET-2's architecture, supplementing the description in the main text.

Notation To refer to a specific residue in a sequence-of-sequences x, we use the same notation as in the main text i.e. we use $x_j^{(i)}$ to denote the jth residue of the ith sequence in x. In general, the superscript $^{(i)}$ is used to refer to the ith sequence. For example, in the main text, we use D to refer to the matrix of pairwise discretized Cα distances of a protein. Thus, in the context of a sequence-of-sequences, the notation $D^{(i)}$ refers to the pairwise discretized Cα distances of the ith sequence in the sequence-of-sequences.

B.1 Input Embedding

The encoder and decoders share a common input embedding space. This space fuses representations derived from the input sequence (x_{seq}), the local structure backbone coordinates (x_{atomb}), and the pLDDT scores (x_{plddt}). Both x_{atomb} and x_{plddt} can contain entries that are masked, for instance, due to missing structural information, padding, or masking specified by a user. For each of these features, a corresponding binary mask (e.g., $x_{\text{atomb\_mask}}$, $x_{\text{plddt\_mask}}$) is provided, where a value of 1 indicates an observed or valid entry, and 0 indicates a masked or invalid entry. To process these potentially masked inputs, the algorithm first applies the respective binary mask to the feature data by setting the values at masked positions to zero. Subsequently, the binary mask itself is concatenated as an additional feature channel to this modified data. These augmented representations for x_{atomb} and x_{plddt} are then linearly projected into the target embedding dimension. The sequence x_{seq} is embedded directly. Finally, these three resulting embeddings are summed to form the single continuous latent representation (Algorithm 1).

Algorithm 1 `embed_inputs` – embeds a single sequence or a sequence-of-sequences

Require: $x_{\text{inputs}} = \{x_{\text{seq}} \in \{1..28\}^{L_x}, x_{\text{plddt}} \in [0,100]^{L_x}, x_{\text{plddt\_mask}} \in \{0,1\}^{L_x}, x_{\text{atomb}} \in \mathbb{R}^{L_x \times 36}, x_{\text{atomb\_mask}} \in \{0,1\}^{L_x \times 36}$

▷ Handle masks for x_{plddt} and x_{atomb} by applying the masks and concatenating along feature dimension

z_{plddt} = Concat(($x_{\text{plddt}} * x_{\text{plddt\_mask}}$, $x_{\text{plddt\_mask}}$), dim=1) ▷ $[0,100]^{L_x \times 2}$

z_{atomb} = Concat(($x_{\text{atomb}} * x_{\text{atomb\_mask}}$, $x_{\text{atomb\_mask}}$), dim=1) ▷ $\mathbb{R}^{L_x \times 72}$

▷ Embed inputs individually

$z_{\text{seq}} = \text{Embed}(x_{\text{seq}})$ ▷ $\mathbb{R}^{L_x \times d}$

$z_{\text{plddt}} = \text{Linear}(z_{\text{plddt}})$ ▷ $\mathbb{R}^{L_x \times d}$

$z_{\text{atomb}} = \text{Linear}(z_{\text{atomb}})$ ▷ $\mathbb{R}^{L_x \times d}$

$z_{\text{seqid}} = \text{Linear}(z_{\text{seqid}})$ ▷ $\mathbb{R}^{L_x \times d}$

return $z_{\text{seq}} + z_{\text{plddt}} + z_{\text{atomb}} + z_{\text{seqid}}$

B.2 Structure-based Attention Bias

See the main text (§3.2.3) for a description of the structure-based attention bias. Figure 3 visualizes the application of the structure-based attention bias.

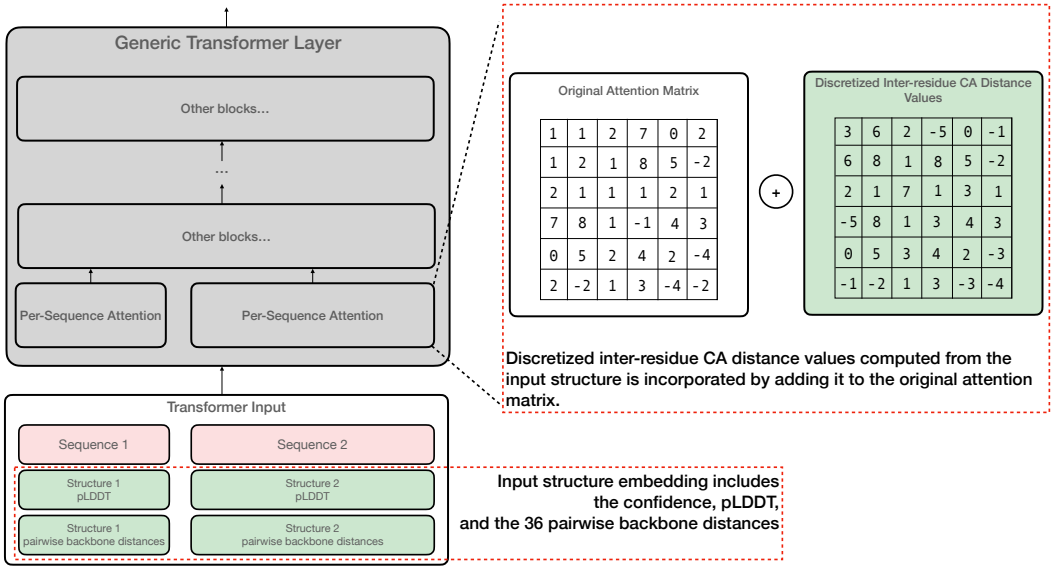

Figure 3: Structure-based attention bias

B.3 Encoder Architecture

After transforming the raw inputs x_{inputs} into embeddings, the encoder (Algorithm 3) further transforms the embeddings by applying n_{layers} of protein order equivariant encoder layers (Algorithm 2). The encoder has two outputs: per residue embeddings of the prompt, h_{encoder}, that are used in the decoders, and per residue sequence logits, z_{seq}, that are used to compute the encoder MLM loss $\mathcal{L}_{\text{MLM encoder}}$.

B.4 Decoder Architecture

The decoders each decode a single sequence y, conditioned on (1) the prompt embedding h_{encoder}, and (2) an optional query index q indicating which sequence in the prompt to use as the query, if any.

The transformations performed in each decoder are detailed in Algorithm 5; the CLM decoder and MLM decoder only differ in that the former uses a causal mask in the attention operations, and the latter does not. First, the input y is embedded. Next, if a query is present, the embedding of y and the embedding of the query (produced by the encoder) are averaged. Then, n_{layers} of decoder layers (Algorithm 4) that are equivariant to the protein order of the prompt are applied. The decoders each have a single output, z_{seq}, that is used to compute the corresponding loss ($\mathcal{L}_{\text{CLM decoder}}$ for the CLM decoder and $\mathcal{L}_{\text{MLM decoder}}$ for the MLM decoder).

Algorithm 2 `encoder_layer` (for brevity, this algorithm describes only a single attention head, but can be extended to multiple attention heads in the normal fashion)

Require: $\forall i \in \{1..\text{N}\}, j \in \{1..L_{x_i}\}$
- prompt embedding $h_j^{(i)} \in \mathbb{R}^d$
- pairwise discretized Cα distances $D_{m,n}^{(i)} \in \{1..128\}, m \in \{1..L_{x_i}\}, n \in \{1..L_{x_i}\}$

▷ First, apply self-attention with structure-based attention bias to each sequence individually
$f = \text{RMSNorm}(h)$ ▷ $\mathbb{R}^{L_x \times d}$
$q_j^{(i)} = \text{RoPE}(\text{Linear}(f_j^{(i)}), j)\ \forall i, j$ ▷ $\mathbb{R}^d$
$k_j^{(i)} = \text{RoPE}(\text{Linear}(f_j^{(i)}), j)\ \forall i, j$ ▷ $\mathbb{R}^d$
$v = \text{Linear}(f)$ ▷ $\mathbb{R}^{L_x \times d}$
▷ Compute attention score with structure-based bias
$A_{m,n}^{(i)} = {q_m^{(i)}}^T k_n^{(i)} + \texttt{structure\_bias}(D_{m,n}^{(i)})$ ▷ $\mathbb{R}$
$f^{(i)} = f^{(i)} + \text{softmax}(\frac{A^{(i)}}{\sqrt{d}})v^{(i)}\ \forall i$ ▷ $\mathbb{R}^{L_{x^{(i)}} \times d}$
▷ Next, apply self-attention to all sequences together
$g = \text{RMSNorm}(f)$ ▷ $\mathbb{R}^{L_x \times d}$
$q_j^{(i)} = \text{RoPE}(\text{Linear}(g_j^{(i)}), j)\ \forall i, j$ ▷ $\mathbb{R}^d$
$k_j^{(i)} = \text{RoPE}(\text{Linear}(g_j^{(i)}), j)\ \forall i, j$ ▷ $\mathbb{R}^d$
$v = \text{Linear}(g)$ ▷ $\mathbb{R}^{L_x \times d}$
$g = g + \text{Attention}(q, k, v)$ ▷ $\mathbb{R}^{L_x \times d}$
▷ Finally, the feedforward layer
$g' = \text{RMSNorm}(g)$ ▷ $\mathbb{R}^{L_x \times d}$
$h' = \text{SwiGLU}(g')$ ▷ $\mathbb{R}^{L_x \times \frac{8}{3}d}$
$g' = g + \text{Linear}(h')$ ▷ $\mathbb{R}^{L_x \times d}$
return g'

Algorithm 3 `encoder` – encodes a prompt x composed of a sequence-of-sequences

Require: $\forall i \in \{1..\text{N}\}$
- $x_{\text{inputs}} = \{x_{\text{seq}}, x_{\text{plddt}}, x_{\text{atomb}}\}$
- pairwise discretized Cα distances $D_{m,n}^{(i)} \in \{1..128\}, m \in \{1..L_{x_i}\}, n \in \{1..L_{x_i}\}$

$h_{\text{encoder}} = \texttt{embed\_inputs}(x_{\text{inputs}})$ ▷ $\mathbb{R}^{L_x \times d}$
for $l \in 1..n_{\text{layers}}$ **do**
    $h_{\text{encoder}} = \texttt{encoder\_layer}(h_{\text{encoder}}, D)$
end for
$z_{\text{seq}} = \text{Linear}(h)$ ▷ $\mathbb{R}^{L_x \times 28}$
return prompt embedding h_{encoder}, sequence logits z_{seq}

Algorithm 4 `decoder_layer` (for brevity, this algorithm describes only a single attention head, but can be extended to multiple attention heads in the normal fashion)

Require:
- decoder type $T \in \{\text{CLM}, \text{MLM}\}$
- encoder prompt embeddings $h^{(i)}_{\text{encoder},j} \in \mathbb{R}^d, i \in \{1..\text{N}\}, j \in \{1..L_{x_i}\}$
- decoder sequence embedding $h_{\text{decoder},i} \in \mathbb{R}^d, i \in \{1..L_y\}$
- pairwise discretized Cα distances $D_{m,n} \in \{1..128\}, m \in \{1..L_y\}, n \in \{1..L_y\}$

▷ First, apply self-attention with structure-based attention bias
$f = \text{RMSNorm}(h_{\text{decoder}})$ ▷ $\mathbb{R}^{L_y \times d}$
$q_i = \text{RoPE}(\text{Linear}(f_i), i)\ \forall i$ ▷ $\mathbb{R}^d$
$k_i = \text{RoPE}(\text{Linear}(f_i), i)\ \forall i$ ▷ $\mathbb{R}^d$
$v = \text{Linear}(f)$ ▷ $\mathbb{R}^{L_y \times d}$
▷ Compute attention score with structure-based bias
$A_{m,n} = {q_m}^T k_n + \texttt{structure\_bias}(D_{m,n})$ ▷ $\mathbb{R}$
if $T == \text{CLM}$ **then** $A = A + \text{CausalMask}(L_y)$ ▷ $\mathbb{R}^{L_y \times L_y}$
end if
$f = f + \text{softmax}(\frac{A}{\sqrt{d}})v$ ▷ $\mathbb{R}^{L_y \times d}$
▷ Next, apply cross-attention to prompt embeddings
$g = \text{RMSNorm}(f)$ ▷ $\mathbb{R}^{L_y \times d}$
$q_i = \text{RoPE}(\text{Linear}(g_i), i)\ \forall i$ ▷ $\mathbb{R}^d$
$k^{(i)}_j = \text{RoPE}(\text{Linear}(h^{(i)}_{\text{encoder},j}), j)\ \forall i, j$ ▷ $\mathbb{R}^d$
$v = \text{Linear}(h_{\text{encoder}})$ ▷ $\mathbb{R}^{L_y \times d}$
$g = g + \text{Attention}(q, k, v)$ ▷ $\mathbb{R}^{L_y \times d}$
▷ Finally, the feedforward layer
$g' = \text{RMSNorm}(g)$ ▷ $\mathbb{R}^{L_y \times d}$
$h'_{\text{decoder}} = \text{SwiGLU}(g')$ ▷ $\mathbb{R}^{L_y \times \frac{8}{3}d}$
$g' = g + \text{Linear}(h'_{\text{decoder}})$ ▷ $\mathbb{R}^{L_y \times d}$
return g'

Algorithm 5 `decoder` – decodes a single sequence y conditioned on prompt embeddings h_{encoder}

Require:
- decoder type $T \in \{\text{CLM}, \text{MLM}\}$
- optional query index $q \in \{0...\text{N}\}$
- encoder prompt embeddings $h^{(i)}_{\text{encoder},j} \in \mathbb{R}^d, i \in \{1..\text{N}\}, j \in \{1..L_{x_i}\}$
- decoder inputs $y_{\text{inputs}} = \{y_{\text{seq}}, y_{\text{plddt}}, y_{\text{atomb}}\}$
- pairwise discretized Cα distances $D_{m,n} \in \{1..128\}, m \in \{1..L_y\}, n \in \{1..L_y\}$

$h_{\text{decoder}} = \texttt{embed\_inputs}(y_{\text{inputs}})$ ▷ $\mathbb{R}^{L_y \times d}$
▷ Embed the query if there is one
if $q \neq 0$ **then**
    $h_{\text{decoder},i} = \frac{1}{2}(h_{\text{decoder},i} + h^{(q)}_{\text{encoder},i})\forall i$ ▷ $\mathbb{R}^d$
end if
for $l \in 1..n_{\text{layers}}$ **do**
    $h_{\text{decoder}} = \texttt{decoder\_layer}(T, h_{\text{encoder}}, h_{\text{decoder}}, D)$ ▷ $\mathbb{R}^{L_y \times d}$
end for
$z_{\text{seq}} = \text{Linear}(h_{\text{decoder}})$ ▷ $\mathbb{R}^{L_y \times 28}$
return sequence logits z_{seq}

B.5 Miscellaneous

This section details additional aspects of PoET-2's architecture that are related to capabilities that are *not* utilized in this paper's experiments.

B.5.1 3Di Token Prediction

In addition to predicting each protein's amino acid sequence, PoET-2 is also trained to predict each protein's 3Di structure token sequence [43] using the cross entropy loss. The 3Di structure tokens are only predicted when the predicted pLDDT of the residue is at least 70. The amino acid and 3Di structure token losses have equal weight i.e. the total loss is simply the sum of the two losses.

B.5.2 Conditioning on target homology

PoET-2 is also trained to generate sequences that must be within a specified sequence identity range of a protein in the prompt, referred to as the query protein. The query can be any protein in the prompt whose sequence is completely known (i.e. contains no unknown or masked amino acids). This generation mode is called "target homology". When using this generation mode, the structure of the generated protein does not have to contain any structural elements specified in the query protein.

The target homology generation mode is implemented with two modifications to the architecture discussed so far:

1. The input embedding is augmented with another feature, x_{seqid}. This feature represents the sequence identity range as two values $\in [0, 1]$ indicating the range's lower and upper bounds, and is repeated across all residues. Thus, it has shape $L \times 2$. If the target homology generation mode is inactive (e.g. as in the encoder) or not being used in the decoder, these two sequence identity values are both set to 0. To prepare x_{seqid} for input embedding, it is processed similarly to other features like x_{plddt} and x_{atomb}: any values at positions intended to be masked are set to zero, and then a corresponding binary mask (1 for observed/valid, 0 for masked/invalid) is concatenated as a third channel to these two sequence identity values. This augmented 3-channel tensor for x_{seqid} is then linearly projected to the model's hidden dimension and subsequently summed with the embeddings of other input features, as detailed in Algorithm 1.
2. Since the generated sequence does not necessarily have the same length as the query, we cannot simply combine the embeddings of the generated sequence and the embeddings of the query by summing the embeddings of all corresponding residues as in Line 3 of Algorithm 5, since the correspondence is unknown. Instead, we sum the embedding of each residue of the generated sequence with the embedding of the first residue of the query sequence. That is, when the target homology generation mode is used, we replace Line 3 of Algorithm 5 with $h_{\text{decoder},i} = \frac{1}{2}(h_{\text{decoder},i} + h^{(q)}_{\text{encoder},1})\forall i$.

B.5.3 Decoding queries of unknown length

Lastly, PoET-2 is trained to decode query sequences with contiguous segments of unknown length. These segments are represented with the gap token (`-`) mentioned in §3.2.1; the gap token indicates zero or more residues with unknown identity. For example, the query sequence `$MK-IP*` indicates that PoET-2 must generate a sequence that starts with the two amino acids M and K, then has 0 of more amino acids, and then ends with the amino acids I and P.

In order to decode such sequences, PoET-2 uses a special decoding scheme called the "insertion decoding scheme". The purpose of this decoding scheme is to align the residues of the query sequence and the generated sequence so that the embeddings of their corresponding residues can be summed in Line 3 of Algorithm 5.

In the standard decoding scheme, the alignment between the query sequence and the sequence being generated by the model may be ambiguous when gap tokens are present in the query. For example, suppose that the query sequence is `$-A`. In a normal decoding scheme, if the model predicts that the token following the start token is the token A, it is ambiguous if that token should be aligned with the gap token to indicate an insertion, or aligned with the token A to indicate that there are no insertions. In order to address this ambiguity, we train the decoders to instead output the gap

token when generating a token that is unmasked in the query sequence. Continuing the example, if the model wants to generate the token A as part of an insertion aligned with the gap token in the query sequence, then the model should simply output the token A as usual. However, if the model wants to generate the token A and have it be aligned with the token A at the third position in the query sequence, then the model should output the gap token. In this case, the gap token in the query sequence represents no insertions.

The insertion decoding scheme is visualized in Figure 4. Using the alignment provided by the insertion decoding scheme, Line 3 of Algorithm 5 can then be modified to be $h_{\text{decoder},i} = \frac{1}{2}(h_{\text{decoder},i} + h^{(q)}_{\text{encoder,alignment}_i})\forall i$, where alignment_i indicates the index of the residue of the query that the ith residue of the generated sequence is aligned to.

Although the insertion decoding scheme is primarily required for the CLM decoder, we apply the insertion decoding scheme to the MLM decoder as well by adjusting the outputs tokens in a similar manner i.e. since the MLM decoder is trained to unmask the token at the current position (as opposed to next token prediction for the CLM decoder), if the token at the current position is unmasked, the MLM decoder is trained to output the gap token.

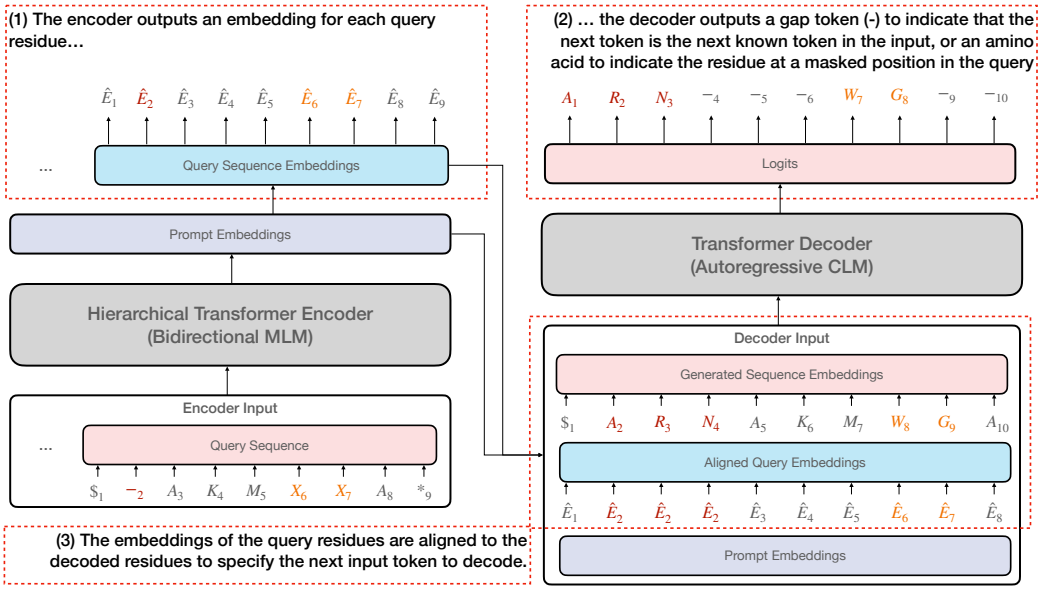

Figure 4: Visualization of insertion decoding scheme.

C PoET-2 Training Details

C.1 Training data

Sequence data PoET-2 is trained on sets of homologous sequences. Sets of homologous sequences are found and sampled using the same procedure used by PoET-1 [1]. Summarizing briefly, the sets of homologous sequences are found by using Diamond [39] to search UniRef50 in an all against all search using the following command:

```
diamond blastp -q uniref50.fasta -d diamond/uniref50 -f 6 -header -k 200000
-max-hsps 1 -e 0.001 -p 96 -o output.tab
```

Sets of homologous sequences are sampled with weight proportional to the inverse of the size of the set.

The main differences between the training data for PoET-2 and PoET-1 are as follows:

- UniRef Version 2304 is used instead of UniRef Version 2103.
- All sets of homologous sequences are used, rather than only sets with at least 10 members.

As a result of these differences, PoET-2 is trained on 62 million sets, as opposed to 29 million sets for PoET-1.

Structure data PoET-2 is trained only on predicted structures from AFDB [31, 32], and no experimentally solved structures. Sequences in the training data are associated with structures in AFDB using the UniRef100 sequence identifier. When the structure of a protein is used as input to PoET-2, if there is a conflict between the sequence in UniRef and AFDB (e.g. due to changes in UniRef), the sequence in AFDB is used. On the order of approximately half of sequences in the training data can be associated with a predicted structure using this methodology.

C.2 Noise schedule

For sequence inputs,

- Context sequence tokens are masked with a random masking rate chosen uniformly from 0%-30%.
- Query sequence tokens are randomly masked with a random masking rate chosen uniformly from 0%-100%.
- Decoder sequence tokens are masked with a random masking rate chosen uniformly from 0%-30%.

For structure inputs, the pLDDT and atomic backbone coordinates of the N, Cα, and C atoms are masked with a random masking rate chosen uniformly from 0%-100%.

For both sequence and structure inputs, with probability

- 50%, masking is performed randomly per residue.
- 25%, random contiguous spans of length L are masked, where L is drawn from the distribution `Poisson`(3) + 1.
- 25%, N random contiguous spans are masked, where N is drawn randomly from `Poisson`(2.5) half the time and from `Poisson`(13) the other half of the time.

C.3 Optimizer and learning rate schedule

PoET-2 is trained with the same optimizer and learning rate schedule as PoET-1 [1]. Namely, the optimizer is Adafactor [44], and the learning rate schedule consists of a linear warmup over the first 4000 steps to a peak learning rate of $1e-2$, and then a square root decay over the remaining training steps.

C.4 Compute requirements

PoET-2 is trained for 3 million steps on 8 x A100 GPUs with 40GB VRAM each. A batch size of 45056 tokens is used per GPU with gradient accumulation over two steps, for an effective batch size of 90112 tokens per GPU. The total training time on this hardware is approximately 2.5 months.

D Zero-shot variant effect prediction

D.1 Prompt engineering

Recall that the prompt consists of two optional components, a context containing proteins from the protein family of interest, and a query containing explicit sequence and/or structure constraints. Our goal in prompt engineering is to design prompts that enable us to accurately predict the properties of variants of a WT sequence.

To determine the set of proteins to use in the context, we use the same method as PoET-1 [1]. We first identify proteins in the protein family by searching for sequence homologs of WT in UniRef100 [38] using the ColabFold MSA protocol [40]. We then select the proteins to include in the context of a prompt by sampling a representative subset of the sequence homologs using the method from Hopf et. al [41].

Building off of this approach, we employ two prompt engineering strategies to improve predictions:

- Following the approach of PoET-1 [1], we ensemble over different prompts, where the context of each prompt contains a different subsample of sequence homologs. The ensemble prediction is simply the average of the individual predictions. Furthermore, for each context, we use different values for the context length (i.e. number of tokens or amino acids in the context) and maximum similarity of a sequence in the context to the WT sequence. The exact values of these parameters used in the ensemble for DMS and clinical datasets are specified in the sections below.
- Utilizing PoET-2's multimodal capabilities, we explore two methods for incorporating structure in the prompt. The first method incorporates structure in the context by associating sequences in the context with their predicted structure in AFDB [31, 32], if the sequence exists in AFDB. The second method incorporates structure by adding a query to the prompt that contains the structure (but not the sequence) of WT. The use of this "inverse-folding" query instructs PoET-2 to score the likelihood that a variant sequence will fold into the same structure as WT. Although it is not necessarily desirable for a variant to adopt the same structure as WT, the inverse-folding likelihood has been shown to be predictive of protein fitness, particularly for stability related properties [5].

  Not all methods of incorporating structure in the prompt are always helpful; the best method for doing so on the ProteinGym DMS and clinical substitutions benchmarks are ablated and identified in the following sections. Also, note that the query-based approach is not used for indel variants because indel variants have different lengths from WT and thus cannot adopt the same tertiary structure as WT.

D.1.1 Deep mutational scanning datasets

Ensembling over context length and maximum similarity Following PoET-1 [1], we ensemble over all combinations of values for context length $\in \{6144, 12288, 24576\}$ and maximum similarity $\in \{1.0, 0.95, 0.90, 0.70, 0.50\}$, resulting in 15 combinations in total.

Incorporating structure in the prompt Table 5 shows the performance of various strategies for incorporating or not incorporating structure in the prompt, and the performance of ensembling different strategies. First, we analyze the effect of different strategies on the substitutions benchmark. When not incorporating structure at all (Strategy A), thus using the same prompting strategy as PoET-1, PoET-2 performs marginally better than PoET-1 ($\Delta\rho = 0.005$; PoET-1 reported in Table 1). Both including the structure in the context (Strategy B), and in the query (Strategy C) improves performance, with the latter strategy offering a larger improvement. Interestingly, combining the two approaches (Strategy D) performs only about the same or slightly worse than Strategy C ($\Delta\rho = -0.002$).

Strategy I, which ensembles all of the above strategies (A-D), improves performance further by $\Delta\rho = 0.009$ vs Strategy C, the best individual strategy. However, we find that Strategy H, which only ensembles Strategies B and D and excludes the strategies that include only sequence in the context, performs similarly to Strategy I, with negligible performance loss. Therefore, we recommend the use of Strategy H, and use this strategy in performance comparisons with other models.

For indel variants, we observe a small improvement in performance by incorporating structure in the context (Strategy B) versus not (Strategy A). There is little or no benefit to ensembling these two strategies (Strategy E). Since a query cannot be used, the other strategies are not applicable. Therefore, we employ Strategy B when comparing PoET-2 to other models.

D.1.2 Clinical datasets

Ensembling over context length and maximum similarity Following PoET-1 [45], we use a context length of 49152 and ensemble over different values for maximum similarity $\in \{1.0, 0.95, 0.90, 0.70, 0.50\}$, resulting in 5 combinations in total. Note that the ensembling parameters for clinical datasets is less well studied than for DMS datasets, and there are likely better parameters for ensembling.

Table 6: Performance (AUROC) of different strategies for including structure in the prompt on the zero-shot clinical benchmarks.

Strategy	Prompt		Substitutions	Indels
	Context Modalities	Query		
A	Sequence	None	0.92544	0.94826
B	Sequence and Structure	None	**0.92624**	**0.95278**
C	Sequence	Structure of WT	0.91505	N/A
D	Sequence and Structure	Structure of WT	0.91292	N/A
E	Ensemble of A and B (Different contexts, No query)		**0.92789**	**0.95179**
F	Ensemble of C and D (Different contexts, With query)		0.91599	N/A
G	Ensemble of A and C (Sequence only context, Different queries)		0.92481	N/A
H	Ensemble of B and D (Sequence and structure context, Different queries)		0.92481	N/A
I	Ensemble of A, B, C, and D		0.92572	N/A

Incorporating structure in the prompt Table 6 shows the performance of various strategies for incorporating or not incorporating structure in the prompt, and the performance of ensembling different strategies. First, we analyze the effect of different strategies on the substitutions benchmark. Incorporating structure in the context (Strategy B) offers a very minor and not statistically significant improvement versus not incorporating structure in the context (Strategy A). Incorporating structure via the query (Strategy C), however, has a negative effect. Therefore, we do not consider strategies that use a query further.

Ensembling Strategies A and B (Strategy E) has a small positive effect over not ensembling, although due to the small effect size, it is unclear if the effect is simply due to ensembling more prompts, or due to ensembling different prompt strategies. Nevertheless, since Strategy E performs best, we use it in comparisons with other models.

For indel variants, we observe similar trends among applicable strategies (those not using a query), with some improvement observed for incorporating structure in the context, but variations in performance being fairly minor. Although Strategy E slightly underperforms Strategy B by a non-statistically significant amount, we employ Strategy E in comparisons with other models for consistency with the strategy used for the substitutions benchmark.

D.2 Length adjusted log likelihood (ratio)

Sequence likelihoods from autoregressive models trained with next-token prediction losses and teacher forcing can exhibit miscalibrated stop token probabilities that bias them towards shorter sequences. This likely arises because the loss function only operates at the token level – it does

not strongly penalize an early stop token as long as the early stop token is predicted to be relatively unlikely compared to other tokens.

This bias towards shorter sequences can be problematic when scoring indel variants with likelihoods, as indel variants can differ in length from WT and each other. To compensate for this, when scoring indel variants, we apply an adjustment to the log likelihood that favors longer sequences over shorter sequences. We find that on a sample of random protein families from UniRef50, the log likelihood decreases roughly linearly with sequence length, with a slope of -1.96 (Figure 5).

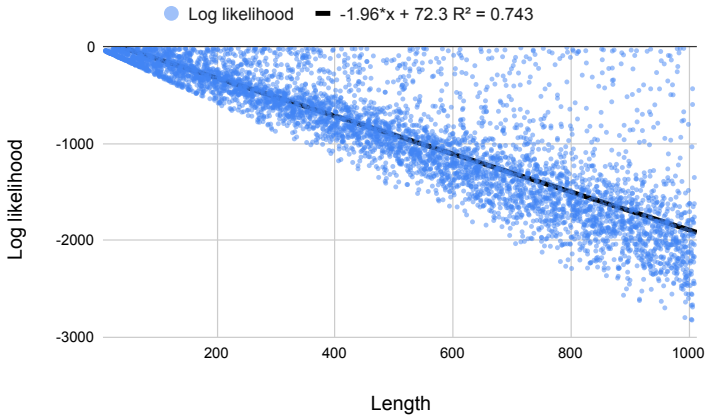

Figure 5: Plot of log likelihood vs length for random UniRef50 protein families.

Therefore, we compute the length adjusted log likelihood as follows:

$$\text{adjusted log likelihood} = \text{log likelihood} + \alpha \times \text{sequence length} \tag{4}$$

where $\alpha = 1.96$ is the length adjustment factor. To compute the adjusted log likelihood *ratio* for scoring variants, we simply use the adjusted log likelihood instead of the regular log likelihood.

On the DMS indels benchmark, we find that using the length adjusted likelihood ratio with $\alpha = 1.96$ improves performance (Figure 6). The length adjustment factor $\alpha = 1.96$ is near optimal, with the empirical best adjustment factor being 2.1.

On the other hand, on the clinical indels benchmark, we find that using the length adjusted likelihood ratio with $\alpha = 1.96$ slightly harms performance (Figure 7). However, the optimal adjustment factor is non-zero (between 0.90 and 1.20 depending on the benchmark version), indicating that some length adjustment is generally ideal for zero-shot fitness prediction, regardless of the specific task. Given that the relation between log likelihood and length is not completely linear, and that the length only explains ~75% of the variance in the log likelihood (Figure 5), there may be better ways to adjust the log likelihood using factors other than the length e.g. factors that may be dependent on the specific protein family of interest. We leave exploration of this to future work.

As our experiments show that adjusting log likelihoods for length is generally useful, even if $\alpha = 1.96$ is not necessarily optimal, we always use the length adjusted log likelihood when comparing the performance of PoET-2 to other models.

D.3 Model ensembles

We compute the zero-shot score for the ensemble model combining PoET-2 and VenusREM by computing a weighted average of the score from PoET-2 and the score from VenusREM:

$$\text{Ensemble Score} = w \times (\text{PoET-2 score}) + (1 - w) \times (\text{VenusREM score}) \tag{5}$$

where $w \in [0, 1]$. To select the weight w, we evaluate the performance of 21 values of w regularly spaced in the interval $[0, 1]$ (inclusive; increments of 0.05) on ProteinNPT's [7] validation of set 8 datasets:

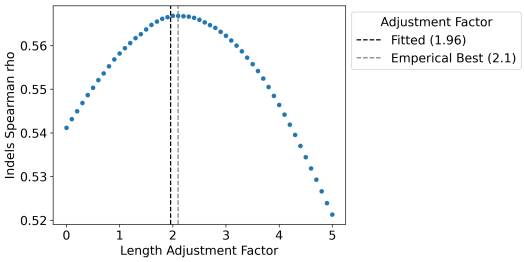

Figure 6: Plot of DMS indels performance (ρ) vs length adjustment factor.

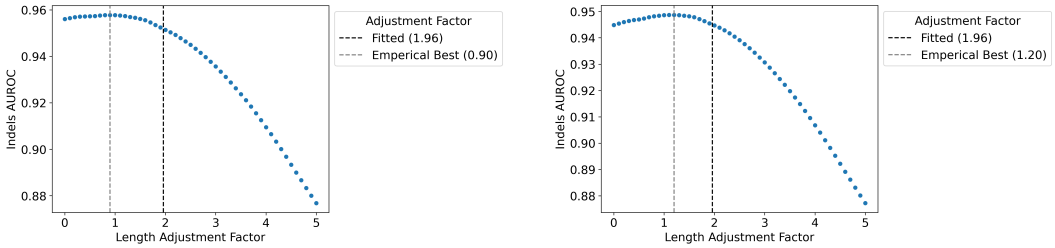

Figure 7: Plot of clinical indels performance (AUROC) versus length adjustment factor; results for ProteinGymV1 (left) and ProteinGymV1.3 (right).

`BLAT_ECOLX_Jacquier_2013, CALM1_HUMAN_Weile_2017, DYR_ECOLI_Thompson_2019, DLG4_RAT_McLaughlin_2012, P53_HUMAN_Giacomelli_2018_WT_Nutlin, REV_HV1H2_Fernandes_2016, RL40A_YEAST_Roscoe_2013, TAT_HV1BR_Fernandes_2016`

On this validation set, we find that the optimal value of w is simply 0.5, corresponding to a simple average.

D.4 Detailed results

The following tables detail results on the performance and standard error of models on the DMS zero-shot substitutions and indels benchmarks, broken by different metrics, and assay and protein subgroups.

- **Table 7**: Overall performance (Spearman, AUC, MCC, NDCG, Recall) on substitutions benchmark, with standard error of difference to PoET-2.
- **Table 8**: Overall performance (Spearman, AUC, MCC, NDCG, Recall) on substitutions benchmark, with standard error of difference to PoET-2 + VenusREM.
- **Table 9**: Overall performance (Spearman, AUC, MCC, NDCG, Recall) on indels benchmark, with standard error of difference to PoET-2.
- **Table 10**: Performance (Spearman) on substitutions benchmark broken down by assay type, with standard error of difference to PoET-2.
- **Table 11**: Performance (Spearman) on substitutions benchmark broken down by MSA depth, with standard error of difference to PoET-2.
- **Table 12**: Performance (Spearman) on substitutions benchmark broken down by taxonomy, with standard error of difference to PoET-2.
- **Table 13**: Performance (Spearman) on substitutions benchmark broken down by mutation depth, with standard error of difference to PoET-2.

D.5 Compute requirements

Inference with PoET-2 is performed on g5.xlarge instances from Amazon Web Services. The instances are equipped with A10G Nvidia GPUs with 24GB VRAM. For scoring sequences of average length (~350 amino acids), the inference throughput per prompt is approximately 125 sequences per second.

Table 7: Performance on zero-shot DMS substitutions benchmark. Standard error of difference to PoET-2 in parentheses.

Model	Metric				
	Spearman	**AUC**	**MCC**	**NDCG**	**Recall**
ESM-2	0.415 (0.013)	0.729 (0.007)	0.328 (0.010)	0.742 (0.007)	0.217 (0.007)
ESM C	0.407 (0.014)	0.727 (0.007)	0.323 (0.011)	0.742 (0.007)	0.213 (0.006)
ProGen2 M	0.379 (0.009)	0.711 (0.005)	0.299 (0.007)	0.747 (0.006)	0.203 (0.006)
ProGen2 XL	0.390 (0.009)	0.717 (0.005)	0.306 (0.008)	0.764 (0.004)	0.198 (0.005)
SaProt	0.457 (0.007)	0.751 (0.004)	0.358 (0.007)	0.764 (0.005)	0.232 (0.006)
ESM-3 Open	0.467 (0.005)	0.756 (0.003)	0.368 (0.005)	0.773 (0.004)	0.241 (0.006)
ProSST	0.508 (0.008)	0.777 (0.004)	0.399 (0.007)	0.752 (0.007)	0.235 (0.009)
MSA Transformer	0.431 (0.009)	0.736 (0.005)	0.338 (0.007)	0.774 (0.004)	0.225 (0.005)
TranceptEVE L	0.456 (0.006)	0.751 (0.003)	0.356 (0.005)	0.783 (0.003)	0.231 (0.005)
GEMME	0.455 (0.009)	0.749 (0.005)	0.352 (0.007)	0.773 (0.003)	0.212 (0.004)
PoET-1	0.470 (0.004)	0.759 (0.002)	0.367 (0.004)	0.779 (0.002)	0.226 (0.003)
S3F-MSA	0.496 (0.006)	0.771 (0.003)	0.388 (0.005)	0.788 (0.002)	0.244 (0.004)
VenusREM	0.519 (0.007)	0.783 (0.003)	0.405 (0.006)	0.766 (0.006)	0.243 (0.009)
PoET-2	0.500 (0.000)	0.773 (0.000)	0.391 (0.000)	0.786 (0.000)	0.238 (0.000)
PoET-2 + VenusREM	**0.543 (0.005)**	**0.796 (0.002)**	**0.423 (0.004)**	**0.791 (0.004)**	**0.254 (0.006)**

Table 8: Performance on zero-shot DMS substitutions benchmark. Standard error of difference to PoET-2 + VenusREM in parentheses.

Model	Metric				
	Spearman	**AUC**	**MCC**	**NDCG**	**Recall**
ESM-2	0.415 (0.014)	0.729 (0.007)	0.328 (0.011)	0.742 (0.007)	0.217 (0.007)
ESM C	0.407 (0.014)	0.727 (0.008)	0.323 (0.012)	0.742 (0.007)	0.213 (0.007)
ProGen2 M	0.379 (0.010)	0.711 (0.005)	0.299 (0.008)	0.747 (0.007)	0.203 (0.008)
ProGen2 XL	0.390 (0.010)	0.717 (0.005)	0.306 (0.009)	0.764 (0.006)	0.198 (0.007)
SaProt	0.457 (0.008)	0.751 (0.005)	0.358 (0.007)	0.764 (0.005)	0.232 (0.006)
ESM-3 Open	0.467 (0.007)	0.756 (0.004)	0.368 (0.006)	0.773 (0.004)	0.241 (0.006)
ProSST	0.508 (0.005)	0.777 (0.003)	0.399 (0.004)	0.752 (0.006)	0.235 (0.006)
MSA Transformer	0.431 (0.010)	0.736 (0.005)	0.338 (0.008)	0.774 (0.006)	0.225 (0.007)
TranceptEVE L	0.456 (0.007)	0.751 (0.004)	0.356 (0.006)	0.783 (0.004)	0.231 (0.006)
GEMME	0.455 (0.009)	0.749 (0.005)	0.352 (0.010)	0.773 (0.005)	0.212 (0.007)
PoET-1	0.470 (0.006)	0.759 (0.003)	0.367 (0.006)	0.779 (0.004)	0.226 (0.005)
S3F-MSA	0.496 (0.006)	0.771 (0.003)	0.388 (0.007)	0.788 (0.003)	0.244 (0.006)
VenusREM	0.519 (0.003)	0.783 (0.002)	0.405 (0.003)	0.766 (0.004)	0.243 (0.004)
PoET-2	0.500 (0.004)	0.773 (0.002)	0.391 (0.004)	0.786 (0.004)	0.238 (0.006)
PoET-2 + VenusREM	**0.543 (0.000)**	**0.796 (0.000)**	**0.423 (0.000)**	**0.791 (0.000)**	**0.254 (0.000)**

Table 9: Performance on zero-shot DMS indels benchmark. Standard error of difference to PoET-2 in parentheses.

Model	Metric				
	Spearman	**AUC**	**MCC**	**NDCG**	**Recall**
ProGen2 M	0.463 (0.037)	0.770 (0.021)	0.370 (0.031)	0.757 (0.018)	0.305 (0.017)
ProGen2 XL	0.427 (0.022)	0.747 (0.010)	0.323 (0.019)	0.749 (0.015)	0.297 (0.012)
TranceptEVE L	0.410 (0.020)	0.749 (0.011)	0.348 (0.020)	0.725 (0.013)	0.258 (0.014)
PoET-1	0.515 (0.006)	0.803 (0.005)	0.434 (0.011)	0.763 (0.006)	0.310 (0.010)
PoET-2	**0.567 (0.000)**	**0.831 (0.000)**	**0.478 (0.000)**	**0.795 (0.000)**	**0.340 (0.000)**

Table 10: Performance (Spearman ρ) on zero-shot DMS substitutions benchmark broken down by assay type. Standard error of difference to PoET-2 in parentheses.

Model	Substitutions By Assay Type				
	Activity	**Binding**	**Expression**	**Organismal Fitness**	**Stability**
ESM-2	0.429 (0.028)	0.336 (0.043)	0.417 (0.030)	0.368 (0.024)	0.523 (0.016)
ESM C	0.426 (0.028)	0.313 (0.044)	0.408 (0.032)	0.360 (0.027)	0.526 (0.017)
ProGen2 M	0.396 (0.028)	0.291 (0.018)	0.434 (0.015)	0.379 (0.015)	0.396 (0.020)
ProGen2 XL	0.406 (0.021)	0.300 (0.028)	0.415 (0.023)	0.384 (0.014)	0.445 (0.012)
SaProt	0.461 (0.017)	0.380 (0.016)	0.488 (0.009)	0.366 (0.020)	0.592 (0.011)
ESM-3 Open	0.432 (0.013)	0.403 (0.013)	0.470 (0.008)	0.388 (0.012)	0.641 (0.011)
ProSST	0.480 (0.019)	0.444 (0.021)	0.532 (0.016)	0.430 (0.018)	**0.653 (0.011)**
MSA Transformer	0.477 (0.009)	0.324 (0.036)	0.447 (0.013)	0.416 (0.017)	0.492 (0.011)
TranceptEVE L	0.490 (0.014)	0.371 (0.018)	0.459 (0.012)	0.458 (0.007)	0.501 (0.011)
GEMME	0.485 (0.008)	0.380 (0.037)	0.440 (0.015)	0.450 (0.008)	0.519 (0.012)
PoET-1	0.498 (0.006)	0.391 (0.019)	0.466 (0.006)	0.474 (0.006)	0.519 (0.005)
S3F-MSA	0.506 (0.008)	0.437 (0.025)	0.480 (0.013)	0.476 (0.007)	0.582 (0.008)
VenusREM	0.499 (0.016)	0.452 (0.018)	0.535 (0.014)	0.459 (0.016)	0.651 (0.011)
PoET-2	0.508 (0.000)	0.423 (0.000)	0.503 (0.000)	0.482 (0.000)	0.582 (0.000)
PoET-2 + VenusREM	**0.538 (0.010)**	**0.475 (0.013)**	**0.552 (0.010)**	**0.505 (0.010)**	0.644 (0.007)

Table 11: Performance (Spearman ρ) on zero-shot DMS substitutions benchmark broken down by MSA depth. Standard error of difference to PoET-2 in parentheses.

Model	Substitutions By MSA Depth		
	Low	**Medium**	**High**
ESM-2	0.340 (0.038)	0.410 (0.018)	0.513 (0.013)
ESM C	0.338 (0.040)	0.401 (0.020)	0.519 (0.011)
ProGen2 M	0.305 (0.031)	0.390 (0.016)	0.422 (0.016)
ProGen2 XL	0.322 (0.024)	0.411 (0.011)	0.442 (0.013)
SaProt	0.397 (0.027)	0.446 (0.014)	0.546 (0.011)
ESM-3 Open	0.402 (0.017)	0.465 (0.011)	0.575 (0.011)
ProSST	0.468 (0.029)	0.506 (0.013)	0.581 (0.013)
MSA Transformer	0.375 (0.024)	0.456 (0.011)	0.480 (0.012)
TranceptEVE L	0.434 (0.015)	0.473 (0.008)	0.491 (0.009)
GEMME	0.445 (0.017)	0.474 (0.008)	0.494 (0.009)
PoET-1	0.479 (0.008)	0.477 (0.006)	0.511 (0.005)
S3F-MSA	0.470 (0.017)	0.509 (0.005)	0.547 (0.007)
VenusREM	0.498 (0.023)	0.524 (0.011)	0.578 (0.013)
PoET-2	0.488 (0.000)	0.507 (0.000)	0.555 (0.000)
PoET-2 + VenusREM	**0.528 (0.016)**	**0.550 (0.007)**	**0.593 (0.008)**

Table 12: Performance (Spearman ρ) on zero-shot DMS substitutions benchmark broken down by taxonomy. Standard error of difference to PoET-2 in parentheses.

Model	Substitutions By Taxonomy			
	Human	**Other Eukaryote**	**Prokaryote**	**Virus**
ESM-2	0.457 (0.011)	0.488 (0.031)	0.459 (0.019)	0.262 (0.043)
ESM C	0.467 (0.010)	0.482 (0.030)	0.442 (0.022)	0.245 (0.051)
ProGen2 M	0.412 (0.011)	0.418 (0.027)	0.355 (0.027)	0.334 (0.035)
ProGen2 XL	0.385 (0.012)	0.459 (0.017)	0.417 (0.017)	0.401 (0.024)
SaProt	0.478 (0.010)	0.530 (0.018)	0.515 (0.013)	0.323 (0.037)
ESM-3 Open	0.480 (0.008)	0.549 (0.015)	0.530 (0.016)	0.407 (0.028)
ProSST	0.518 (0.013)	0.577 (0.019)	0.550 (0.018)	0.449 (0.028)
MSA Transformer	0.439 (0.012)	0.517 (0.012)	0.445 (0.014)	0.419 (0.028)
TranceptEVE L	0.472 (0.007)	0.515 (0.014)	0.455 (0.013)	0.460 (0.020)
GEMME	0.468 (0.009)	0.519 (0.013)	0.467 (0.011)	0.471 (0.019)
PoET-1	0.481 (0.005)	0.543 (0.009)	0.464 (0.008)	0.491 (0.011)
S3F-MSA	0.501 (0.008)	0.561 (0.010)	0.521 (0.008)	0.502 (0.012)
VenusREM	0.530 (0.011)	0.586 (0.017)	0.550 (0.016)	0.489 (0.023)
PoET-2	0.506 (0.000)	0.569 (0.000)	0.507 (0.000)	0.528 (0.000)
PoET-2 + VenusREM	**0.548 (0.008)**	**0.604 (0.011)**	**0.562 (0.009)**	**0.551 (0.013)**

Table 13: Performance (Spearman ρ) on zero-shot DMS substitutions benchmark broken down by mutation depth. Standard error of difference to PoET-2 in parentheses.

Model	Substitutions By Mutation Depth				
	1	**2**	**3**	**4**	**5+**
ESM-2	0.423 (0.011)	0.248 (0.021)	0.203 (0.077)	0.160 (0.077)	0.220 (0.073)
ESM C	0.416 (0.012)	0.257 (0.022)	0.189 (0.073)	0.150 (0.073)	0.217 (0.074)
ProGen2 M	0.372 (0.010)	0.131 (0.025)	0.149 (0.059)	0.131 (0.066)	0.178 (0.062)
ProGen2 XL	0.384 (0.008)	0.181 (0.023)	0.267 (0.046)	0.229 (0.046)	0.283 (0.047)
SaProt	0.459 (0.010)	0.312 (0.018)	0.271 (0.048)	0.268 (0.051)	0.337 (0.056)
ESM-3 Open	0.487 (0.009)	0.336 (0.019)	0.303 (0.047)	0.284 (0.046)	0.365 (0.050)
ProSST	0.520 (0.009)	0.393 (0.025)	0.316 (0.046)	0.274 (0.049)	0.334 (0.060)
MSA Transformer	0.427 (0.008)	0.220 (0.019)	0.358 (0.026)	0.365 (0.017)	0.401 (0.022)
TranceptEVE L	0.446 (0.006)	0.277 (0.014)	0.349 (0.041)	0.327 (0.039)	0.385 (0.046)
GEMME	0.447 (0.006)	0.275 (0.017)	0.329 (0.044)	0.338 (0.028)	0.419 (0.022)
PoET-1	0.466 (0.004)	0.298 (0.010)	0.412 (0.019)	0.393 (0.014)	0.421 (0.011)
S3F-MSA	0.499 (0.005)	0.332 (0.011)	0.377 (0.017)	0.343 (0.017)	0.387 (0.033)
VenusREM	0.534 (0.008)	0.395 (0.023)	0.352 (0.046)	0.320 (0.045)	0.372 (0.050)
PoET-2	0.506 (0.000)	0.357 (0.000)	**0.444 (0.000)**	**0.419 (0.000)**	**0.447 (0.000)**
PoET-2 + VenusREM	**0.556 (0.005)**	**0.402 (0.014)**	0.442 (0.023)	0.411 (0.020)	0.441 (0.028)

E Supervised variant effect prediction

Overview As described in the main text, our supervised variant effect prediction methodology employs a Gaussian Process (GP) regression model to predict fitness scores. The GP is configured with a constant mean function and a product kernel. This kernel integrates information from two Matérn 5/2 sub-kernels, each operating on distinct features derived from PoET-2.

One sub-kernel operates on protein embeddings derived from the last layer of PoET-2's MLM decoder. Given the high dimensionality of the full per-residue embeddings produced by PoET-2, a dimensionality reduction step is applied prior to their use in the GP. Specifically, we utilize Singular Value Decomposition (SVD) to project the full embeddings into a 1024-dimensional space. This dimensionality reduction strategy is conceptually similar to the PCA-based approach used by Bepler et al. [6] to improve the runtime of their learning algorithm, which is also a GP for protein fitness prediction. For each wild-type (WT) protein in an assay, the SVD transformation is fitted on a set of 1536 variants: this set comprises the WT sequence itself and a random sample of 1535 single and double substitution mutants of that WT. The number of variants used for fitting SVDs (1536) was chosen to be approximately 50% larger than the number of SVD components (1024). This was deemed a practical trade-off to provide a reasonable basis for the decomposition while managing the computational requirements of fitting SVD transformations for each of the numerous proteins in the ProteinGym benchmark; fitting the SVD on a larger or more diverse set of variants may improve performance.

The second Matérn 5/2 sub-kernel in the product utilizes the log likelihood ratios (LLRs) obtained from PoET-2's CLM decoder, as used in zero-shot prediction (§4.1). The final GP model thus learns from both the reduced-dimensionality MLM embeddings and the CLM-derived LLRs.

Gaussian Process Hyperparameter Priors To improve the stability and performance of GP training, particularly with small training datasets, we incorporate empirical priors on the GP hyperparameters. This process involves three steps: (1) We first fit individual GP models (with the architecture described above) to each of the 8 validation datasets specified by ProteinNPT [7] (listed in Appendix D.3). (2) From these 8 trained GPs, we extract the optimized hyperparameters and fit empirical distributions to them. Specifically, a Normal distribution is fitted to the learned constant mean values, while Gamma distributions are fitted to the other hyperparameters (i.e. the lengthscales of both Matérn kernels, the outputscale of the product kernel, and the likelihood noise term). (3) These fitted Normal and Gamma distributions then serve as priors for the respective hyperparameters when training GPs on the ProteinGym benchmark assays.

This prior-informed approach is particularly beneficial for assays with limited training data (e.g. fewer than ~50 data points), where the prior helps guide the optimization process. For larger training set sizes, the influence of the prior diminishes. When conducting ablation studies involving different foundation models to generate the input embeddings and LLRs, the procedure for deriving these priors (steps 1 and 2) is repeated to learn a separate, appropriate set of hyperparameter priors for each distinct foundation model.

E.1 Prompt engineering

Similar to our approach to prompt engineering for zero-shot variant effect prediction (Appendix D.1), for supervised prediction, we also explore two prompt engineering methods.

Ensembling over context length and maximum similarity We ensemble over different values of context length $\in \{6144, 12288, 24576, 49152, 98304\}$ and always use a maximum similarity value of 0.95, resulting in 5 combinations in total. We use a wide range of context lengths compared to zero-shot prediction because we observed in early experiments on ProteinNPT's validation set of 8 datasets (Appendix D.3) that longer contexts lengths generally had a small positive impact on supervised prediction performance (in contrast, for PoET-1, it was observed that long context lengths could have a negative effect on zero-shot prediction [1]). Figure 8 shows the performance of the GP model with different values for the context length, and the performance of the ensemble model. We use a fixed value of 0.95 for maximum similarity because 0.95 is typically the best value for zero-shot prediction, and did not explore other values in order to conserve compute. Therefore, it is most likely the case that there exists more optimal parameters for ensembling.

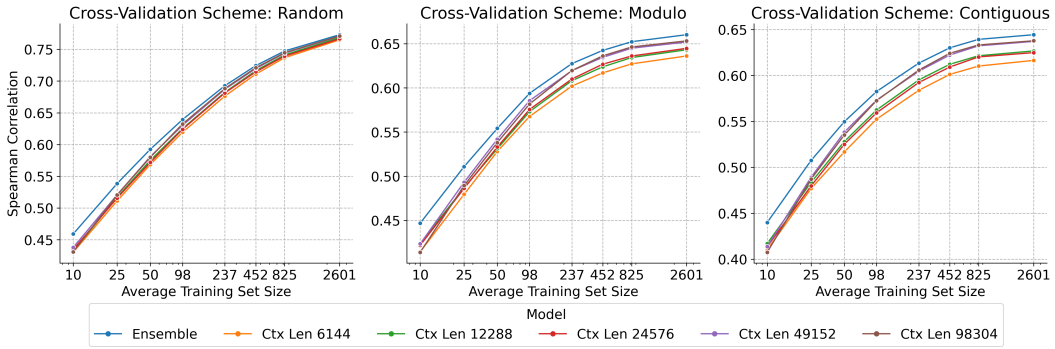

Figure 8: Performance on the supervised DMS substitutions benchmark as a function of training set size of (1) PoET-2 GP models using prompts with 5 different context lengths and (2) the ensemble GP model that ensembles the five models from (1).

Incorporating structure in the prompt Table 14 shows the performance of various strategies for incorporating structure in the prompt on an extended validation set of 30 assays from ProteinGym's DMS supervised substitutions benchmark. We find that all strategies, both those that do and don't incorporate structure in the prompt, perform about the same for supervised variant effect prediction, despite significant improvements for zero-shot variant effect prediction when using strategies that incorporate structure in the prompt. Therefore, for supervised variant effect prediction, we do not include structure in the prompt.

Table 14: Performance (Spearman's ρ) of different strategies for including structure in the prompt on a validation set of 30 datasets from the supervised DMS substitutions benchmark.

Prompt		**Supervised Cross-Validation Scheme**				**Zero-shot**
Context Modalities	**Query**	**Rand.**	**Mod.**	**Contig.**	**Avg.**	
Sequence	None	**0.70963**	**0.60398**	**0.49490**	**0.60284**	0.40711
Sequence and Structure	None	0.70689	0.59785	0.49312	0.59929	0.42554
Sequence	Structure of WT	0.70436	0.60179	0.49142	0.59919	0.42277
Sequence and Structure	Structure of WT	0.70103	0.58561	0.47449	0.58704	**0.43748**

The extended validation set consists of ProteinNPT's 8 validation datasets (Appendix D.3), plus 22 randomly selected datasets. The full list of extended validation datasets is as follows:

`BLAT_ECOLX_Jacquier_2013, CALM1_HUMAN_Weile_2017, DYR_ECOLI_Thompson_2019, DLG4_RAT_McLaughlin_2012, P53_HUMAN_Giacomelli_2018_WT_Nutlin, REV_HV1H2_Fernandes_2016, RL40A_YEAST_Roscoe_2013, TAT_HV1BR_Fernandes_2016, ACE2_HUMAN_Chan_2020, BCHB_CHLTE_Tsuboyama_2023_2KRU, CAR11_HUMAN_Meitlis_2020_gof, CP2C9_HUMAN_Amorosi_2021_abundance, ENVZ_ECOLI_Ghose_2023, F7YBW7_MESOW_Ding_2023, GCN4_YEAST_Staller_2018, GLPA_HUMAN_Elazar_2016, HCP_LAMBD_Tsuboyama_2023_2L6Q, KCNH2_HUMAN_Kozek_2020, LYAM1_HUMAN_Elazar_2016, MBD11_ARATH_Tsuboyama_2023_6ACV, MTHR_HUMAN_Weile_2021, OBSCN_HUMAN_Tsuboyama_2023_1V1C, OPSD_HUMAN_Wan_2019, PA_I34A1_Wu_2015, PSAE_SYNP2_Tsuboyama_2023_1PSE, PTEN_HUMAN_Matreyek_2021, Q53Z42_HUMAN_McShan_2019_binding-TAPBPR, RNC_ECOLI_Weeks_2023, SPG1_STRSG_Olson_2014, TADBP_HUMAN_Bolognesi_2019`

E.2 Gaussian Process kernels

Our primary Gaussian Process (GP) model for supervised variant effect prediction utilizes a product kernel combining two Matérn 5/2 sub-kernels: one operating on PoET-2 MLM embeddings and the other on PoET-2 CLM log likelihood ratios (LLRs). While this product kernel is effective for single-site substitutions, its performance can be suboptimal when predicting the effects of multi-mutation variants due to the behavior of the LLR-based sub-kernel under distributional shift.

Specifically, the distribution of LLRs for multi-mutation variants often differs significantly from that of single-site mutations, with multi-mutation LLRs tending towards much lower or higher values. Consequently, if a GP is trained predominantly on single-site variants and then applied to predict multi-mutation variants, the LLRs of the test set can be markedly out-of-distribution (OOD) relative to the training data.

This OOD characteristic poses a challenge for the Matérn 5/2 kernel operating on LLRs. Stationary kernels like the Matérn assume that covariance is a function of the distance between inputs; when test LLRs are far outside the training distribution's range, their covariance with the training data (as modeled by this kernel) diminishes significantly. Since our model employs a product kernel, if the LLR sub-kernel assigns low covariance to a test point, the overall covariance for that point will also be low, irrespective of the embedding-based sub-kernel. In such cases, the GP prediction tends to revert towards the prior mean, offering limited predictive power for these OOD multi-mutation variants.

We leave the exploration of more sophisticated methods for incorporating LLRs into the GP for multi-mutation contexts (e.g. using LLR transformations or non-stationary kernels) to future work. For the current work, a pragmatic approach to mitigate this issue when predicting the effects of multi-mutation variants is to utilize a GP model that relies solely on the embedding-based Matérn kernel, thereby omitting the LLR-based sub-kernel for these specific predictions. Even with this simplification, a PoET-2 based GP using only embeddings can achieve strong performance in predicting the effects of multi-mutation variants when trained on data from single-site or lower-order mutants, outperforming other state-of-the-art methods as discussed in Appendix E.6.

E.3 Detailed results

The following tables detail results on the performance and standard error of models on the DMS supervised benchmark, broken by different metrics, cross-validation schemes, and assay and protein subgroups.

- **Table 15** Performance (Spearman) broken down by cross-validation scheme, with standard error of difference to PoET-2.
- **Table 16** Performance (MSE) broken down by cross-validation scheme, with standard error of difference to PoET-2.
- **Table 17** Performance (average Spearman across cross-validation schemes) on substitutions benchmark broken down by assay type, with standard error of difference to PoET-2.
- **Table 18** Performance (average Spearman across cross-validation schemes) on substitutions benchmark broken down by MSA depth, with standard error of difference to PoET-2.
- **Table 19** Performance (average Spearman across cross-validation schemes) on substitutions benchmark broken down by taxonomy, with standard error of difference to PoET-2.
- **Table 20** Performance (average Spearman across cross-validation schemes) on substitutions benchmark for the smallest training set size ($n = 10$) with standard error of difference to PoET-2. This table demonstrates that PoET-2's performance advantage is statistically significant even in the extreme few-shot regime. Differences for larger dataset sizes ($n > 10$) are all statistically significant with $p < 1e - 5$ and are omitted for brevity.

Table 15: Performance (Spearman ρ) on supervised DMS substitutions benchmark. Standard error of difference to PoET-2 GP in parentheses.

	Spearman ρ (↑)			
Model	**Random**	**Modulo**	**Contiguous**	**Average**
ProteinNPT	0.741 (0.003)	0.588 (0.012)	0.529 (0.018)	0.619 (0.010)
Kermut	0.746 (0.004)	0.635 (0.008)	0.613 (0.009)	0.664 (0.006)
ESM-2 (650 M) GP	0.749 (0.002)	0.573 (0.009)	0.549 (0.010)	0.624 (0.007)
ESM C GP	0.747 (0.004)	0.605 (0.007)	0.573 (0.010)	0.642 (0.006)
PoET-2 GP	**0.773 (0.000)**	**0.661 (0.000)**	**0.645 (0.000)**	**0.693 (0.000)**

Table 16: Performance (MSE) on supervised DMS substitutions benchmark. Standard error of difference to PoET-2 GP in parentheses.

	MSE (↓)			
Model	**Random**	**Modulo**	**Contiguous**	**Average**
ProteinNPT	0.441 (0.012)	0.765 (0.023)	0.856 (0.025)	0.687 (0.015)
Kermut	0.413 (0.004)	0.649 (0.010)	0.697 (0.010)	0.586 (0.007)
ESM-2 (650 M) GP	0.404 (0.004)	0.720 (0.011)	0.768 (0.011)	0.630 (0.008)
ESM C GP	0.398 (0.004)	0.660 (0.008)	0.716 (0.009)	0.592 (0.007)
PoET-2 GP	**0.370 (0.000)**	**0.602 (0.000)**	**0.647 (0.000)**	**0.540 (0.000)**

Table 17: Performance (Spearman ρ) on supervised DMS substitutions benchmark broken down by assay type. Standard error of difference to PoET-2 GP in parentheses.

Model	**Substitutions By Assay Type**				
	Activity	**Binding**	**Expression**	**Organismal Fitness**	**Stability**
ProteinNPT	0.590 (0.007)	0.541 (0.045)	0.631 (0.011)	0.558 (0.010)	0.776 (0.005)
Kermut	0.606 (0.007)	0.627 (0.027)	0.680 (0.010)	0.584 (0.007)	0.825 (0.004)
ESM-2 (650 M) GP	0.569 (0.014)	0.577 (0.026)	0.633 (0.012)	0.545 (0.010)	0.795 (0.006)
ESM C GP	0.575 (0.015)	0.601 (0.021)	0.656 (0.013)	0.550 (0.013)	0.828 (0.006)
PoET-2 GP	**0.630 (0.000)**	**0.667 (0.000)**	**0.691 (0.000)**	**0.622 (0.000)**	**0.854 (0.000)**

Table 18: Performance (Spearman ρ) on supervised DMS substitutions benchmark broken down by MSA depth. Standard error of difference to PoET-2 GP in parentheses.

Model	**Substitutions By MSA Depth**		
	Low	**Medium**	**High**
ProteinNPT	0.576 (0.016)	0.621 (0.010)	0.705 (0.006)
Kermut	0.619 (0.012)	0.658 (0.005)	0.743 (0.005)
ESM-2 (650 M) GP	0.561 (0.019)	0.618 (0.007)	0.721 (0.006)
ESM C GP	0.581 (0.019)	0.627 (0.010)	0.749 (0.005)
PoET-2 GP	**0.667 (0.000)**	**0.689 (0.000)**	**0.769 (0.000)**

Table 19: Performance (Spearman ρ) on supervised DMS substitutions benchmark broken down by taxonomy. Standard error of difference to PoET-2 GP in parentheses.

Model	Substitutions By Taxonomy			
	Human	**Other Eukaryote**	**Prokaryote**	**Virus**
ProteinNPT	0.633 (0.007)	0.673 (0.006)	0.666 (0.012)	0.602 (0.019)
Kermut	0.671 (0.006)	0.712 (0.006)	0.707 (0.004)	0.628 (0.012)
ESM-2 (650 M) GP	0.649 (0.006)	0.668 (0.016)	0.673 (0.009)	0.552 (0.016)
ESM C GP	0.674 (0.005)	0.682 (0.017)	0.696 (0.007)	0.542 (0.023)
PoET-2 GP	**0.696 (0.000)**	**0.738 (0.000)**	**0.736 (0.000)**	**0.690 (0.000)**

Table 20: Mean Spearman correlation (ρ) and standard error on supervised DMS substitutions benchmark when training datasets are limited to no more than $n = 10$ data points. Values in parentheses are the standard error of the difference in mean performance relative to PoET-2 GP, computed via 10,000 bootstrap samples. All differences are statistically significant ($p < 4.5e-3$).

	Spearman ρ (↑)		
Model	**Random**	**Modulo**	**Contiguous**
ESM-2 (650M) GP	0.408 (0.008)	0.389 (0.010)	0.382 (0.012)
ESM-2 (3B) GP	0.381 (0.008)	0.354 (0.009)	0.354 (0.009)
ESM C (300M) GP	0.418 (0.008)	0.399 (0.009)	0.397 (0.011)
PoET-1 GP	0.440 (0.004)	0.420 (0.008)	0.416 (0.008)
PoET-2 GP	**0.459 (0.000)**	**0.447 (0.000)**	**0.440 (0.000)**

E.4 Comparison of MLM and CLM decoder embeddings

Table 21 compares the performance of GP models trained on embeddings from PoET-2's bidirectional (MLM) decoder versus its autoregressive (CLM) decoder. We find that the MLM decoder embeddings consistently outperform the CLM decoder embeddings.

Table 21: Performance of embeddings from PoET-2's MLM and CLM decoders on ProteinGym's supervised DMS substitutions benchmark.

	Spearman ρ (↑)				Mean Square Error (↓)			
Model	**Rand.**	**Mod.**	**Contig.**	**Avg.**	**Rand.**	**Mod.**	**Contig.**	**Avg.**
PoET-2 CLM GP	0.757	0.622	0.601	0.660	0.398	0.660	0.713	0.590
PoET-2 MLM GP	**0.771**	**0.652**	**0.637**	**0.687**	**0.374**	**0.616**	**0.657**	**0.549**

E.5 Compute requirements

- The computation of the SVD of embeddings from protein foundation models is performed on r6a.4xlarge instances from Amazon Web Services. These instances are equipped with 16 vCPUs and 128GB of RAM. The amount of RAM required to fit the SVD depends on the number of training samples, the length of the WT sequence, and the embedding dimension of the foundation model used. Generally we use 1536 training samples, as described at the start of this section. For some long sequences and models with very large embedding dimension, the number of samples may need to be decreased to fit within the available RAM.
- Embeddings and log likelihood ratios for supervised variant effect prediction are computed using the same computational resources as log likelihood ratios for zero-shot variant effect prediction (Appendix D.5)
- Gaussian process models are trained on g5.xlarge instances from Amazon Web Services. These instances are equipped with A10G Nvidia GPUs that have 24GB of VRAM.

E.6 Prediction of mutational effects of multi-mutation variants

As discussed in Appendix E.2, when predicting multi-mutation variant effects from data on single-site or lower-order mutants, we utilize a Gaussian Process (GP) model with a kernel based solely on PoET-2 embeddings, omitting log-likelihood ratios. To evaluate this embedding-only PoET-2 GP in such challenging generalization scenarios, we benchmark it on the multi-mutation dataset introduced in the ProGen3 publication [46]. The ProGen3 authors report that their model outperforms Kermut [15] and matches ConFit [47] on this specific benchmark. Our PoET-2 based GP, however, surpasses all three aforementioned models (Table 22).

The ProGen3 multi-mutation benchmark comprises 8 datasets selected from ProteinGym. The selection criteria, as stated by the ProGen3 authors, is as follows: "We identify all assays in ProteinGym with at least 3 mutations, and we train on all variants at most k mutations from the wild type, where k is the smallest number required for the train split to exceed 500 sequences. To ensure that the train and test splits contain proteins of similar fitness, we require that the total variation distance between the train and test distributions of fitness scores be less than 1. These filters yield 8 assays that measure diverse functional attributes for a wide range of proteins."

Table 22 details the performance of PoET-2 GP alongside ProGen3, Kermut, and ConFit on this multi-mutation benchmark; PoET-2 GP outperforms all other models.

Table 22: Performance comparison on multi-mutation variant effect prediction benchmark.

Model	**Spearman** ρ (↑)
Kermut	0.628
ConFit	0.679
ProGen3	0.673
PoET-2 GP	**0.708**

F Licenses

Existing assets used in this paper are licensed as follows:

- ProteinGym benchmark: MIT license
- UniRef protein database: CC BY 4.0 license
- AlphaFold database: CC BY 4.0 license
- ESM C [42] model: Cambrian Open License Agreement [48]

G Additional details

G.1 Statistical significance analysis

Statistical significance for all experiments is assessed by performing a non-parametric, two-sided bootstrap test with at least 10,000 samples. Bootstrap is performed using the same methodology as in ProteinGym [5].

H ProteinGym Assay Dataset Sources

We thank the authors of the original publications from which the ProteinGym assays were derived for making their experimental data publicly available [49–191].

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims of the paper are stated in the abstract and introduction. The claims are supported by experimental evidence as discussed in Section 4.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the paper are discussed in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Pre-training details are provided in Appendix C. Training and evaluation details of downstream models are described in Section 4, Appendix D, and Appendix E.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Code and model weights are planned for future public release.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental setting is described in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Statistical significance of results relating to claims are provided in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Information about compute resources required to reproduce experiments are provided in Appendix C, Appendix D.5, and Appendix E.5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conform with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The broader impacts are discussed in Appendix A.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Papers that produced existing assets used in this paper are cited. The licenses are specified in Appendix F.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLMs are not used for the development of core methods in this research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.