Few-shot Protein Fitness Prediction via In-context Learning and Test-time Training

Felix Teufel[1,2,3] **Aaron W. Kollasch**[1] **Yining Huang**[1]
Ole Winther[2,5] **Kevin K. Yang**[4] **Pascal Notin**[1,*] **Debora S. Marks**[1,*]
[*]Corresponding Author
[1]Department of Systems Biology, Harvard Medical School [2]Department of Biology, University of Copenhagen
[3]Machine Intelligence, Novo Nordisk A/S [4]Microsoft Research, Cambridge, MA, USA
[5]Dept. of Applied Mathematics and Computer Science, Technical University of Denmark
pascal_notin@hms.harvard.edu, debbie@hms.harvard.edu

Abstract

Accurately predicting protein fitness with minimal experimental data is a persistent challenge in protein engineering. We introduce PRIMO (PRotein In-context Mutation Oracle), a transformer-based framework that leverages in-context learning and test-time training to adapt rapidly to new proteins and assays without large task-specific datasets. By encoding sequence information, auxiliary zero-shot predictions, and sparse experimental labels from many assays as a unified token set in a pre-training masked-language modeling paradigm, PRIMO learns to prioritize promising variants through a preference-based loss function. Across diverse protein families and properties—including both substitution and indel mutations—PRIMO outperforms zero-shot and fully supervised baselines. This work underscores the power of combining large-scale pre-training with efficient test-time adaptation to tackle challenging protein design tasks where data collection is expensive and label availability is limited.

1 Introduction

Protein engineering has rapidly advanced in recent years, driven by breakthroughs in both experimental and computational methods. High-throughput (HT) experimental methods, such as Deep Mutational Scanning (DMS) assays [1], enable large-scale exploration of sequence space by generating and testing thousands of variants for a desired function. The data generated through HT approaches can support the training of powerful machine learning (ML) models to learn the corresponding fitness landscape and further optimize the target properties [2, 3]. However, while HT assays have become more accessible for certain properties such as thermostability [4, 5] or fluorescence [6], it can still be prohibitively expensive, time-consuming, or altogether infeasible to produce large-scale functional measurements for many other properties.

Doing away with experimental annotations entirely, deep generative models such as Protein Language Models (pLMs) trained on large sets of natural sequences from protein data banks (eg., UniRef [7], MGnify [8]) have offered a promising avenue to address these shortcomings [9–11]. However, although the zero-shot predictions they provide can be remarkably effective for protein design in certain settings [12], they are still insufficiently accurate for many practical applications, providing rough starting points that need to be further optimized [13]. As a result, *few-shot* learning has emerged as a critical challenge in protein engineering: how can we accurately predict or optimize protein fitness from only a handful of experimental observations? Recent studies have attempted to tackle that problem by combining zero-shot scores with minimal labeled data [14, 15, 3]. While this approach offers improved performance, such supervised methods often still demand a separate validation set to

Published in the AI for Science Workshop (NeurIPS 2025).

prevent overfitting, which can easily exceed the available budget in a strict few-shot setting (e.g., a single 96-well plate) and may fail to accommodate more complex variant types such as insertions and deletions.

To address these challenges, we introduce **PRIMO (PRotein In-context Mutation Oracle)**, a transformer-based framework that integrates *in-context learning* with *test-time training* to deliver highly accurate protein fitness predictions with only a handful of labeled samples per assay. PRIMO treats each sequence and any available measurements (including zero-shot predictions) as a unified token set in a masked language modeling paradigm, employing a preference-based loss to rank variants correctly. Pre-training over many assays allows PRIMO to adapt rapidly to new proteins and properties, circumventing the need for extensive labeled datasets or large validation sets. Moreover, PRIMO handles both substitution and indel mutations, broadening its applicability across a diverse range of protein engineering tasks.

Our main contributions are summarized as follows:

- **A new few-shot prediction framework**. We present PRIMO, a transformer architecture that combines in-context learning with test-time training, enabling accurate ranking of protein variants under extreme data scarcity.
- **State-of-the-art few-shot performance**. We demonstrate that PRIMO significantly outperforms both zero-shot baselines and fully supervised models in low-data regimes, achieving superior fitness predictions even with a limited number of labeled examples, across diverse assays and protein families from the ProteinGym benchmark.
- **Broad applicability**. Unlike many existing methods, PRIMO accommodates both single-substitution and indel variants and does not require a separate validation set, making it more practical for real-world protein design scenarios.
- **A novel natural evolution benchmark**. We curate a benchmark comprised of several high-throughput assays that each characterize broad fitness landscapes spanned by natural sequences from a given protein family. This benchmark allows to assess models in challenging settings where train and test sequences are farther apart in sequence space.

Our results highlight the promise of large-scale pre-training on diverse deep mutational scans, followed by efficient test-time adaptation, in tackling challenging protein design tasks where experimental resources and labeled data are severely constrained.

2 Related Work

As experimental budgets in biomolecule research can be highly constrained, few-shot learning for property prediction has been a long-standing challenge with high practical relevance.

2.1 Zero-shot fitness prediction

Given that learning from few-shot observations is challenging, a popular alternative approach is *zero-shot* fitness prediction, where the likelihood of a model trained on evolutionarily observed sequences is used to score variant effects. Family specific models learn the distribution of evolutionary sequence within a protein family by leveraging Multiple Sequence Alignments (MSAs) [16, 17]. Protein language models trained on large protein sequence databases capture the fitness distribution of many protein families in a single model without the need for MSAs [9, 10]. Augmenting sequence models with protein structure has been shown to further improve zero-shot prediction performance [18, 19].

Hybrid methods seek to unify the family-specific and protein language models. MSA Transformer [20] learns from millions of MSAs across protein families. TranceptEVE [21] complements the pre-trained protein language model with an alignment-based model at inference. To overcome the limitations of MSAs, PoET [22] proposed a retrieval-augmented model that scores sequences given a context set of related sequences, by concatenating and performing self-attention over the context sequences.

2.2 Supervised learning of protein fitness

Shallow ML approaches such as ridge regression models on one-hot encoded amino acids or pLM embeddings have been successfully applied for few-shot protein engineering [14], and see further performance boosts when incorporating zero-shot scores as additional features [23]. Using DMS-scale datasets of protein fitness, ProteinNPT [2] proposed to train transformer models for substitution variant effect prediction, also incorporating zero-shot scores as input features.

To further bridge the gap between unsupervised zero-shot prediction and supervised learning from experimental labels, few-shot *likelihood-based fine-tuning* of generative pLMs using a preference objective has been proposed by Hawkins-Hooker et al. [15]. However, a key limitation of both fine-tuning and training ProteinNPT-scale models is the requirement of a validation set to prevent overfitting: Likelihood-based fine-tuning in [15] requires 128 observations for validation, which precludes true few-shot usage, and already exceeds the budget afforded by e.g. a single 96-well plate.

Alternatively, rather than treating zero-shot scores as features, *Kermut* [3] proposed to use a zero-shot predictor as the prior mean function in a Gaussian process (GP) with a dedicated kernel for substitution variant effect prediction. As GPs are customarily trained using the likelihood of the training data, the architecture choice also circumvents the need for a validation dataset.

Machine learning guided Directed evolution [24] can also effectively explore protein fitness landscape with small number of wet-lab experiments. EVOLVEpro [25] augments experimental directed evolution by combining Protein Language Models with few-shot learning. The efficient exploration of fitness landscape with machine learning could further be powered by automated robotic system for wet-lab experiments [26].

Recently, Beck et al. [27] proposed *Metalic*, an in-context learning (ICL) approach using a model trained on many different proteins and fitness assays. As pure in-context learning proved too limiting to model new proteins, a supervised fine-tuning approach was adopted that again required 128 validation data points. Moreover, as Metalic reused the ProteinNPT transformer architecture, it cannot model indel variants. Lastly, Metalic was trained using a data split that is inadequate for transfer learning and yields high overlap between training and evaluation assays, as we will demonstrate in Section 5.

3 The PRIMO Model

3.1 Architecture

3.1.1 Inputs

PRIMO is a transformer-based masked language model that processes labeled sequence variant sets of size N. Each sample i consists of an amino acid (AA) sequence x_i^{AA} of length L, a target quantitative fitness label x_i^{Fitness}, a categorical property type ID x_i^{ID} denoting the assay type of the measured fitness, and one or multiple auxiliary zero-shot labels $x_i^{\text{Auxiliary}}$. In PRIMO, we use autoregressive zero-shot scores from the ProGen pLM [11], as they can be computed for both indel and substitution variants. All tokens are embedded and concatenated to a sequence h_i, using ESM-2 [28] for the AA sequence and learned embeddings for all other inputs:

$$
\begin{aligned}
h_i^{\text{AA}} &= \text{ESM}(x_i^{\text{AA}}) \\
h_i^{\text{ID}} &= \text{Embed}(x_i^{\text{ID}}) \\
h_i^{\text{Fitness}} &= \text{Linear}(x_i^{\text{Fitness}}) \\
h_i^{\text{Auxiliary}} &= \text{Linear}(x_i^{\text{Auxiliary}}) \\
h_i &= \text{Concat}(h_i^{\text{AA}}, h_i^{\text{Fitness}}, h_i^{\text{ID}}, h_i^{\text{Auxiliary}})
\end{aligned} \tag{1}
$$

3.1.2 Transformer stack

To overcome the computational complexity of full sequence-of-sequences self-attention, which suffers from quadratic scaling by the product of the sequence length and set size, we employ a lightweight

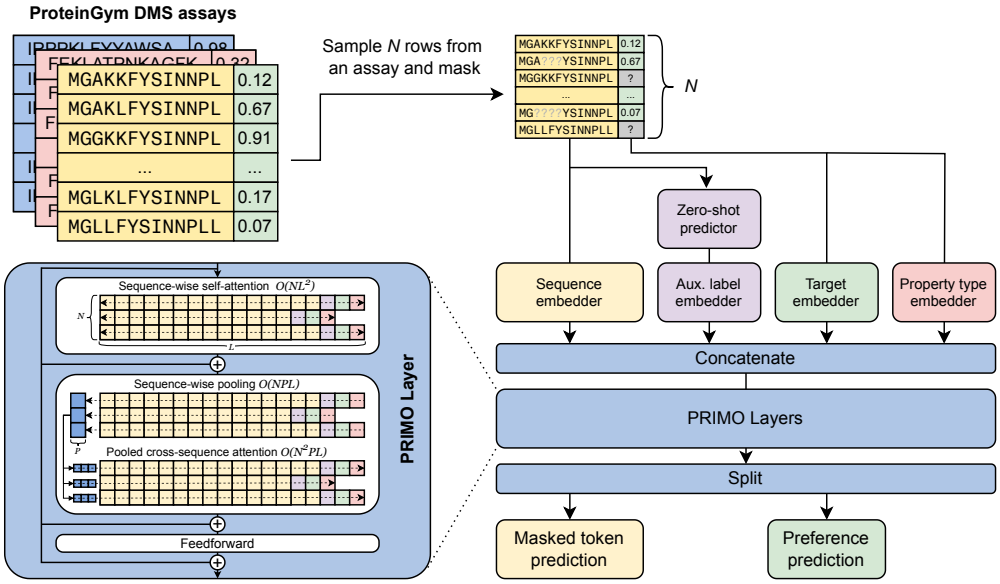

Figure 1: **The PRIMO architecture and training approach**. PRIMO processes labeled sets of proteins drawn from ProteinGym DMS assays. After processing the set with a transformer stack that allows for exchange of information between samples, it performs preference prediction on samples with masked fitness, and masked token prediction on amino acids.

pooling attention mechanism to allow sequences to exchange information. Note that, as PRIMO processes sequences with indel mutations, column attention as used in ProteinNPT is less suitable, since insertions and deletions result in variable sequence lengths and misaligned positions. We first perform standard self-attention on each sequence h_i individually,

$$h_i = \text{MHA}(h_i, h_i, h_i) \quad \forall i \in 1, \ldots, N. \tag{2}$$

where $\text{MHA}(q, k, v)$ denotes standard multi-head self-attention with query q, key k and value v. Next, we perform PRIMO's inter-sequence attention operation. We first pool each sequence h_i into a representation p_i of fixed size P using attention pooling [29] (Equation 3). We concatetate the pooled representations of all sequences in the set (Equation 4), and let each sequence h_i cross-attend to the pooled representations of all the sequences (Equation 5):

$$p_i = \text{MHA}(\text{Mean}(h_i), h_i, h_i) \quad \forall i \in 1, \ldots, N \tag{3}$$

$$p = \text{Concat}(p_1, p_2, \ldots, p_N) \tag{4}$$

$$h_i = \text{MHA}(h_i, p, p) \quad \forall i \in 1, \ldots, N. \tag{5}$$

As we exchange information between individual sequences only using pool representations, we can overcome the limiting scaling complexity of sequence-of-sequences self-attention, $\mathcal{O}(N^2L^2)$. The self-attention on each sequence has complexity of $\mathcal{O}(NL^2)$, while the sequence pooling and cross-attention has complexity of $\mathcal{O}(NPL)$ and $\mathcal{O}(N^2PL)$. As the size of the pooled representation used for cross-attention will typically be smaller than the sequence length, $NP << L$, the limiting attention scaling remains $\mathcal{O}(NL^2)$. Together with a final feedfoward layer, the sequence-wise and cross-sequence attention operations constitute one PRIMO layer. As in PoET, we employ skip connections, pre-LayerNorm and rotary positional embeddings [30]. PRIMO uses 6 layers with a hidden size of 400.

3.1.3 Prediction heads

After the sequences have been processed by the PRIMO layers, we make predictions based on the updated hidden states h_i^{AA} and h_i^{Fitness}. We reuse ESM's pre-trained prediction head, and train a linear layer for fitness prediction.

$$h_i^{\text{AA}}, h_i^{\text{Fitness}} = \text{Split}(h_i) \tag{6}$$

$$\hat{x}_i^{\text{AA}} = \text{ESMHead}(h_i^{\text{AA}}) \tag{7}$$

$$\hat{y}_i = \text{Linear}(h_i^{\text{Fitness}}) \tag{8}$$

3.2 Pre-training

PRIMO is pre-trained using a hybrid masked token reconstruction objective [2]: during pre-training, either the label x_i^{Fitness} or a span in x_i^{AA} may be masked. We mask labels with 33% probability, and mask spans in the remaining samples with 20% probability. As a set of sequences from a given DMS assay will be highly identical beyond a few mutations, making masked reconstruction mostly trivial, spans are placed such that the mutated regions are covered. For AA token reconstruction, we use a simple cross-entropy masked token prediction loss. For label reconstruction, following [15], we use a preference-based loss that tasks the model with correctly ranking the fitness of the Q masked samples in the set (non-masked samples are ignored from the loss calculation and serve as support context only):

$$L = \sum_{i=1}^{Q} \sum_{j=1}^{Q} -\mathbb{I}(x_i^{\text{Fitness}} > x_j^{\text{Fitness}}) \log \sigma(\hat{y}_i - \hat{y}_j). \tag{9}$$

$\mathbb{I}$ denotes the indicator function, and σ is the sigmoid activation function. The loss is therefore equivalent to $Q \times Q$ binary classifications. For pre-training, we sample sets of size $N = 32$ from one assay at a time, using a batch size of 12. Sequences are cropped to 512 AAs for computational efficiency. When cropping, we ensure that all relevant mutated positions are still present. The ESM-2 pLM for sequence embedding remains frozen during training.

3.3 Test-time training

After pre-training, we wish to predict the fitness of novel proteins given few-shot observations. As this usage may result in a distribution shift that makes in-context learning infeasible, we adopt a test-time training (TTT) [31, 32] approach for PRIMO. In TTT, instead of directly making predictions after conditioning on the context, the model's weights are first adapted to the task using fine-tuning on the context data. After predicting the test problem, the updated weights are discarded. Rather than using a fully self-supervised objective for TTT, we take advantage of available few-shot observations and use PRIMO's hybrid sequence and preference label reconstruction loss, sampling masked sets from the few-shot data, as typically done for TTT in in-context learning scenarios [33]. We perform gradient descent for a fixed number of 25 steps, using same loss function and learning rate as in pre-training. The pLM weights also remain frozen during TTT.

3.4 Prediction

For inference, we condition on Q unmasked samples, and predict each test sample by masking it. The unnormalized preference score of each sample serves as its fitness prediction. As PRIMO is a model that was designed to predict *relative* fitness, we also need to provide context in zero-shot prediction that serves as the reference. We therefore provide an arbitrary sequence with an arbitrary fitness of 0.5 (center of a min-max scale) as unmasked context when scoring. Note that this is conceptually similar to zero-shot prediction with masked LMs, where a second sequence (usually the wild type) is required as reference.

We annotate all results obtained from predictions using direct conditioning on Q samples as PRIMO (ICL), and any that first uses the same samples to perform TTT prior to conditioning as PRIMO (TTT).

3.5 Training data

We pre-train PRIMO on DMS assays from ProteinGym [34] that had a fitness readout that falls into the following categories: *Stability, Enzymatic activity, Abundance, Fluorescence*, and *Binding*.

Table 1: **Zero-shot prediction performance (Spearman correlation) of PRIMO when training on either the PRIMO or Metalic split**. The presence of training samples with high sequence identity to the test set in the Metalic split leads to inflated zero-shot performances.

Test dataset	Closest protein in Metalic training set	Identity	Zero-shot performance	
			PRIMO split	Metalic split
BLAT_ECOLX_Jacquier_2013	BLAT_ECOLX	100 %	0.23	0.65
DYR_ECOLI_Thompson_2019	DYR_ECOLI	100 %	0.45	0.50
DLG4_RAT_McLaughlin_2012	DLG4_HUMAN	99 %	0.30	0.37
RL40A_YEAST_Roscoe_2013	RL40A_YEAST	100 %	0.30	0.74
GFP_AEQVI_Sarkisyan_2016	Q8WTC7_9CNID	62 %	0.20	0.44

All assays that cover more specific aspects of function that do not fall into this categorization were excluded. We train on both substitution and indel variants. To overcome different experimental scales and units, raw DMS fitness values are min-max normalized for each assay.

4 Results

4.1 Baselines

Following Notin et al. [2], we use zero-shot fitness augmented ridge regression models as the baseline. As demonstrated before [23, 14], such models can be used for few-shot learning without requiring a separate validation set to prevent overfitting during training. We also evaluate the EvolvePro [25] random forest (RF) regression model that does not leverage zero-shot scores. Additionally, we consider a GP with a zero-shot prior mean function and embedding kernel based on Kermut [3]. We omit Kermut's structure kernel to enable modeling of indels. To match the setup of PRIMO, all baseline models also use ESM-2 embeddings and ProGen zero-shot scores.

4.2 Fitness prediction performance

We draw an increasing number of N labeled few-shot observations for learning, and report performance by predicting all samples in an assay and computing the Spearman rank correlation to the experimental fitness. We also perform zero-shot prediction at $N = 0$, where we do not condition on any experimental measurement. We use a hold out subset of ProteinGym for evaluation that was designed to control for overlap to the training data on the protein level (Table A1). We first trained two PRIMO models on this holdout ("PRIMO split") and the Metalic split that did not control for overlap. In direct comparison, we find that the high sequence similarity overlap between Metalics' training and testing set can cause inflated zero-shot performances. In the most drastic case, on RL40A_YEAST, where the Metalic pre-training set contains 2,633 observations of the same protein in two other assays, the 100% similarity train-test overlap results in an apparent increase of 0.4 over the zero-shot performance obtained with the PRIMO split.

Using the PRIMO split, we first evaluate PRIMO's in-context learning capability. While zero-shot prediction shows an average improvement (0.51) over ProGen (0.41), performance mostly stays flat with increasing N (Table 2). To rule out the possibility that PRIMO fails to process context, effectively working as a single-sample model (a highly accurate single-sequence regression model would also perform well on a ranking metric), we ablate the inputs and only provide one unlabeled sequence at a time. This context-free prediction results in catastrophic failure, with performance becoming worse than random (Table A4). This confirms that PRIMO did in fact learn to perform context-based prediction, and fundamentally works by *comparing* sequences. We therefore consider it more likely that ICL is ineffective due to the pre-training data available in ProteinGym being too limiting to mitigate (expected) distribution shifts when testing on new assays.

When using TTT to adapt PRIMO's parameters to the unseen test assays, we observe a gradual performance improvement with increasing data, going from a zero-shot average Spearman correlation of 0.51 to 0.67 with 128 shots. The GP and ridge regression baseline methods exhibit the same trend, but are outperformed by PRIMO with TTT on all levels of N. Especially at extreme low-N of up to

Table 2: **Few-shot prediction of held-out DMS assays**. Average Spearman correlation coefficient over all held out assays is shown, aggregated over five replicates. Per-assay performances are reported in Table A9 and Table A10. The zero-shot performances of the GP and Ridge regression models marked with * are zero-shot predictions from ProGen that either serve as the prior mean (GP) or an additional input feature (Ridge). The best value per level of N is highlighted in bold, unless zero-shot prediction is superior. Error bars are computed over different draws for each N.

Method	Shots 0	4	8	16	32	64	128
Overall							
GP	0.42*	0.24 ± 0.02	0.32 ± 0.03	0.39 ± 0.02	0.45 ± 0.01	0.51 ± 0.01	0.56 ± 0.00
Ridge	0.42*	0.26 ± 0.02	0.33 ± 0.02	0.41 ± 0.01	0.48 ± 0.01	0.56 ± 0.01	0.63 ± 0.00
RF	-	0.23 ± 0.01	0.32 ± 0.01	0.39 ± 0.01	0.45 ± 0.02	0.52 ± 0.00	0.59 ± 0.00
PRIMO (ICL)	**0.51 ± 0.01**	**0.53 ± 0.00**	**0.53 ± 0.00**	0.53 ± 0.00	0.53 ± 0.00	0.53 ± 0.00	0.53 ± 0.00
PRIMO (TTT)	**0.51 ± 0.01**	0.49 ± 0.01	0.51 ± 0.01	**0.54 ± 0.01**	**0.58 ± 0.02**	**0.63 ± 0.00**	**0.67 ± 0.01**
Stability							
GP	0.41	0.36 ± 0.05	0.47 ± 0.04	0.55 ± 0.02	0.61 ± 0.01	0.66 ± 0.01	0.71 ± 0.00
Ridge	0.41	0.39 ± 0.03	0.46 ± 0.04	0.55 ± 0.02	0.61 ± 0.01	0.69 ± 0.01	0.75 ± 0.00
RF	-	0.36 ± 0.03	0.46 ± 0.02	0.53 ± 0.01	0.58 ± 0.02	0.65 ± 0.01	0.71 ± 0.00
PRIMO (ICL)	**0.61 ± 0.01**	**0.62 ± 0.00**	**0.62 ± 0.00**	0.62 ± 0.00	0.62 ± 0.00	0.62 ± 0.00	0.62 ± 0.00
PRIMO (TTT)	0.59 ± 0.01	0.59 ± 0.02	0.62 ± 0.02	**0.65 ± 0.01**	**0.69 ± 0.01**	**0.73 ± 0.01**	**0.77 ± 0.01**
Enzymatic activity							
GP	0.52*	0.11 ± 0.06	0.19 ± 0.04	0.25 ± 0.03	0.31 ± 0.01	0.37 ± 0.02	0.42 ± 0.01
Ridge	0.52*	0.11 ± 0.05	0.21 ± 0.03	0.26 ± 0.02	0.35 ± 0.03	0.44 ± 0.02	0.51 ± 0.01
RF	-	0.07 ± 0.05	0.17 ± 0.02	0.23 ± 0.04	0.3 ± 0.02	0.38 ± 0.02	0.47 ± 0.01
PRIMO (ICL)	0.49 ± 0.05	**0.57 ± 0.00**	**0.57 ± 0.00**	**0.57 ± 0.00**	**0.57 ± 0.00**	**0.57 ± 0.00**	0.57 ± 0.00
PRIMO (TTT)	**0.53 ± 0.02**	0.44 ± 0.05	0.49 ± 0.01	0.51 ± 0.01	0.54 ± 0.03	0.56 ± 0.01	**0.61 ± 0.01**
Fluorescence							
GP	0.09*	0.01 ± 0.08	0.02 ± 0.09	0.09 ± 0.01	0.14 ± 0.01	0.17 ± 0.02	0.24 ± 0.01
Ridge	0.09*	0.08 ± 0.05	0.08 ± 0.07	0.12 ± 0.05	0.16 ± 0.02	0.21 ± 0.02	0.28 ± 0.02
RF	-	0.10 ± 0.08	**0.12 ± 0.04**	**0.15 ± 0.03**	**0.21 ± 0.03**	0.24 ± 0.01	**0.32 ± 0.01**
PRIMO (ICL)	0.07 ± 0.00	0.06 ± 0.00	0.06 ± 0.00	0.06 ± 0.00	0.06 ± 0.00	0.06 ± 0.00	0.06 ± 0.00
PRIMO (TTT)	**0.11 ± 0.02**	**0.14 ± 0.06**	0.11 ± 0.02	**0.15 ± 0.05**	0.20 ± 0.04	**0.25 ± 0.02**	0.30 ± 0.02
Binding							
GP	0.47*	0.18 ± 0.04	0.22 ± 0.05	0.28 ± 0.03	0.34 ± 0.05	0.41 ± 0.02	0.46 ± 0.01
Ridge	0.47*	0.17 ± 0.05	0.24 ± 0.04	0.33 ± 0.04	0.42 ± 0.03	0.51 ± 0.01	0.58 ± 0.01
RF	-	0.15 ± 0.05	0.22 ± 0.02	0.3 ± 0.03	0.38 ± 0.05	0.46 ± 0.01	0.55 ± 0.01
PRIMO (ICL)	**0.49 ± 0.00**	**0.50 ± 0.00**	**0.50 ± 0.00**	**0.50 ± 0.00**	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00
PRIMO (TTT)	0.47 ± 0.01	0.42 ± 0.05	0.42 ± 0.04	0.48 ± 0.02	**0.56 ± 0.04**	**0.63 ± 0.02**	**0.69 ± 0.01**

32 shots, baseline methods prove ineffective, with the exception of the EvolvePro RF model on GFP fluorescence prediction. A detailed breakdown by substitution and indel mutations is provided in the appendix.

4.3 Performance on the natural evolution benchmark

The previous analysis focused on test assays from ProteinGym in which mutated sequences vary from the reference wild-type sequence by at most a handful of mutations (typically singles or doubles). In order to assess the ability of models to extrapolate farther away in sequence space, we curate a new benchmark, comprised of three high-throughput assays that each characterize broad fitness landscapes spanned by natural sequences for Chorismate mutase [12], Rubisco [35] and PPAT [36] respectively. We find that PRIMO outperforms all other baselines (Table 3). However, in that context, it is critical to make use of test-time training to allow the model to adapt more flexibly to the broader landscapes characterized by the test set.

Table 3: **Performance (Spearman correlation) on the natural evolution benchmark**. Values are reported as mean and standard deviation over five replicates for three assays. The zero-shot performances of the GP and Ridge regression models marked with * are zero-shot predictions from ProGen that either serve as the prior mean (GP) or an additional input feature (Ridge).

Shots	0	4	8	16	32
PRIMO ICL	**0.10** $\pm$ **0.01**	**0.09** $\pm$ **0.02**	0.09 $\pm$ 0.02	0.07 $\pm$ 0.03	0.06 $\pm$ 0.01
PRIMO TTT	0.09 $\pm$ 0.01	0.07 $\pm$ 0.03	**0.10** $\pm$ 0.03	**0.19** $\pm$ **0.06**	**0.30** $\pm$ **0.02**
GP	0.04*	0.02 $\pm$ 0.04	0.04 $\pm$ 0.04	0.07 $\pm$ 0.02	0.18 $\pm$ 0.11
Ridge	0.04*	0.00 $\pm$ 0.03	0.08 $\pm$ 0.04	0.14 $\pm$ 0.02	0.24 $\pm$ 0.12
RF	–	0.03 $\pm$ 0.03	0.08 $\pm$ 0.06	0.13 $\pm$ 0.06	0.24 $\pm$ 0.07
MLP	–	0.00 $\pm$ 0.03	0.07 $\pm$ 0.04	0.13 $\pm$ 0.01	0.24 $\pm$ 0.11

5 Discussion

Inappropriate splitting inflates prediction performance. Data partitioning strategies are critical to ensure reliable performance reporting in ML on biological sequences, as widely recognized in the field [37–39]. When exploring a new paradigm such as pre-training on DMS data, followed by ICL evaluation, previous ProteinGym test subsets are inadequate, as they were only designed for single-assay learning and evaluation. As one may expect from finding cases of 100% sequence identity overlap between partitions, we observed that apparent "zero-shot" performances can be highly inflated, as they in fact are driven by thousands of training observations of the same property of the same protein, just measured in a different experiment. While Metalic did not report per-assay performances, we believe that its claimed average performance of 0.484 suffers from the issue demonstrated in Table 1, and find that Metalic underperforms PRIMO when training on a clean split (Table A11).

TTT can boost prediction performance. In our experiments, we find that ICL often fails to efficiently use the provided context, with performance staying flat and there not being a marginal benefit from providing more labeled data. While it remains to some degree unclear why ICL is not effective, it needs to be recognized that using ProteinGym as the pre-training dataset is limiting, only exposing PRIMO to context sets sampled from 116 distinct experiments. However, we find that TTT can be an effective remedy to the unavoidable distribution shift at test time, allowing the model to adapt to new tasks in weight space as more data becomes available.

6 Conclusion and Outlook

In this work, we have introduced PRIMO, a transformer model for in-context learning with test-time training that enables few-shot protein fitness prediction. By training on a large number of protein fitness assays, PRIMO learns to extract information from labeled samples and perform relative protein fitness prediction. While the direct application of ICL can be limiting and fail to efficiently use the available labeled information, TTT enables PRIMO to adapt to unseen assays, making it an efficient few-shot learner that achieves state of the art performance. PRIMO demonstrates that while the problem remains challenging, progress can be made by leveraging existing experimental data.

From a modeling perspective, assuming such a high-diversity, low-N data resource can be established, a more flexible encoding of property types could be applied to widen the scope. Given sufficient metadata of the actual experimental protocols, natural language encoding could be considered. Moreover, future work may find it useful to consider leveraging other pLMs than ESM-2 in PRIMO, or selectively fine-tune the pLM, as it has been demonstrated that doing so can aid prediction performance in some cases [40].

In summary, our work demonstrates the potential of the ICL paradigm for few-shot fitness prediction, which we believe to become increasingly relevant in the future, given the success of ICL and large-scale pre-training in general across domains [41, 42]. We highlight the need for dedicated fit-for-purpose data splitting regimes and detailed performance reporting, which we consider to be crucial for further advancement of the field.

7 Availability

Code and data is available at `https://github.com/fteufel/PRIMO`.

Acknowledgements

We thank Peter Mørch Groth for helpful discussions regarding Kermut.

References

[1] Douglas M. Fowler and Stanley Fields. Deep mutational scanning: a new style of protein science. *Nature Methods*, 11:801–807, 2014.

[2] Pascal Notin, Ruben Weitzman, Debora Marks, and Yarin Gal. Proteinnpt: Improving protein property prediction and design with non-parametric transformers. *Advances in Neural Information Processing Systems*, 36:33529–33563, 2023.

[3] Peter Mørch Groth, Mads Herbert Kerrn, Lars Olsen, Jesper Salomon, and Wouter Boomsma. Kermut: Composite kernel regression for protein variant effects. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[4] Kotaro Tsuboyama, Justas Dauparas, Jonathan Chen, Élodie 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*, 620:434 – 444, 2023.

[5] Antoni Beltran, Xianger Jiang, Yue Shen, and Ben Lehner. Site-saturation mutagenesis of 500 human protein domains. *Nature*, 637:885 – 894, 2024.

[6] 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, 2021.

[7] Baris E. Suzek, Yuqi Wang, Hongzhan Huang, Peter B. McGarvey, and Cathy H. Wu. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31:926 – 932, 2014.

[8] Lorna J. Richardson, Ben Allen, Germana Baldi, Martin Beracochea, Maxwell L. Bileschi, Tony Burdett, Josephine Burgin, Juan Caballero, Guy Cochrane, Lucy J. Colwell, Tom Curtis, Alejandra Escobar-Zepeda, Tatiana A. Gurbich, Varsha Kale, Anton I. Korobeynikov, Shriya Raj, Alexander B. Rogers, Ekaterina A. Sakharova, Santiago Sanchez, Darren J. Wilkinson, and Robert D. Finn. Mgnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Research*, 51:D753 – D759, 2022.

[9] 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. *Advances in neural information processing systems*, 34:29287–29303, 2021.

[10] 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 *International Conference on Machine Learning*, pages 16990–17017. PMLR, 2022.

[11] Erik Nijkamp, Jeffrey A Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. Progen2: exploring the boundaries of protein language models. *Cell systems*, 14(11):968–978, 2023.

[12] William P. Russ, Matteo Figliuzzi, Christian Stocker, Pierre Barrat-Charlaix, Michael Socolich, Peter Kast, Donald Hilvert, Rémi Monasson, Simona Cocco, Martin Weigt, and Rama Ranganathan. An evolution-based model for designing chorismate mutase enzymes. *Science*, 369:440 – 445, 2020.

[13] Pascal Notin, Nathan J. Rollins, Yarin Gal, Chris Sander, and Debora Marks. Machine learning for functional protein design. *Nature Biotechnology*, 42:216–228, 2024.

[14] Surojit Biswas, Grigory Khimulya, Ethan C Alley, Kevin M Esvelt, and George M Church. Low-n protein engineering with data-efficient deep learning. *Nature methods*, 18(4):389–396, 2021.

[15] Alex Hawkins-Hooker, Jakub Kmec, Oliver Bent, and Paul Duckworth. Likelihood-based fine-tuning of protein language models for few-shot fitness prediction and design. In *ICML 2024 Workshop on Efficient and Accessible Foundation Models for Biological Discovery*, 2024.

[16] Thomas A Hopf, John B Ingraham, Frank J Poelwijk, Charlotta PI Schärfe, Michael Springer, Chris Sander, and Debora S Marks. Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.*, 35(2):128, 2017.

[17] Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K Min, Kelly P. Brock, Yarin Gal, and Debora S. Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 2021.

[18] Zuobai Zhang, Pascal Notin, Yining Huang, Aurelie Lozano, Vijil Vijil, Debora Marks, Payel Das, and Jian Tang. Multi-scale representation learning for protein fitness prediction. *Annual Conference on Neural Information Processing Systems*, 2024.

[19] Mingchen Li, Pan Tan, Xinzhu Ma, Bozitao Zhong, Huiqun Yu, Ziyi Zhou, Wanli Ouyang, Bingxin Zhou, Liang Hong, and Yang Tan. ProSST: Protein language modeling with quantized structure and disentangled attention. *bioRxiv*, page 2024.04.15.589672, April 2024.

[20] Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. MSA transformer. In *International Conference on Machine Learning*, pages 8844–8856. PMLR, July 2021.

[21] Pascal Notin, Lood Van Niekerk, Aaron W Kollasch, Daniel Ritter, Yarin Gal, and Debora S Marks. TranceptEVE: Combining family-specific and family-agnostic models of protein sequences for improved fitness prediction. *36th Conference on Neural Information Processing Systems (NeurIPS 2022), LMRL workshop*, page 2022.12.07.519495, December 2022.

[22] Timothy F Truong, Jr and Tristan Bepler. PoET: A generative model of protein families as sequences-of-sequences. *37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, June 2023.

[23] 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, 2022.

[24] Jason Yang, Ravi G Lal, James C Bowden, Raul Astudillo, Mikhail A Hameedi, Sukhvinder Kaur, Matthew Hill, Yisong Yue, and Frances H Arnold. Active learning-assisted directed evolution. *Nat. Commun.*, 16(1):714, January 2025.

[25] Kaiyi Jiang, Zhaoqing Yan, Matteo Di Bernardo, Samantha R Sgrizzi, Lukas Villiger, Alisan Kayabolen, B J Kim, Josephine K Carscadden, Masahiro Hiraizumi, Hiroshi Nishimasu, Jonathan S Gootenberg, and Omar O Abudayyeh. Rapid in silico directed evolution by a protein language model with EVOLVEpro. *Science*, 387(6732):eadr6006, January 2025.

[26] Jacob T Rapp, Bennett J Bremer, and Philip A Romero. Self-driving laboratories to autonomously navigate the protein fitness landscape. *Nat Chem Eng*, 1(1):97–107, January 2024.

[27] Jacob Beck, Shikha Surana, Manus McAuliffe, Oliver Bent, Thomas D Barrett, Juan Jose Garau Luis, and Paul Duckworth. Metalic: Meta-learning in-context with protein language models. *arXiv preprint arXiv:2410.08355*, 2024.

[28] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[29] Zihang Dai, Guokun Lai, Yiming Yang, and Quoc Le. Funnel-transformer: Filtering out sequential redundancy for efficient language processing. *Advances in neural information processing systems*, 33:4271–4282, 2020.

[30] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[31] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9229–9248. PMLR, 13–18 Jul 2020.

[32] Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei A Efros. Test-time training with masked autoencoders. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[33] Ekin Akyürek, Mehul Damani, Linlu Qiu, Han Guo, Yoon Kim, and Jacob Andreas. The surprising effectiveness of test-time training for abstract reasoning, 2024.

[34] Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood Van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, et al. Proteingym: Large-scale benchmarks for protein fitness prediction and design. *Advances in Neural Information Processing Systems*, 36, 2024.

[35] Dan Davidi, Melina Shamshoum, Zhijun Guo, Y. Bar-On, Noam Prywes, Aia Oz, Jagoda Jabłońska, Avi I. Flamholz, David G. Wernick, Niv Antonovsky, Benoit de 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, 2020.

[36] Calin Plesa, Angus M. Sidore, Nathan B. Lubock, Di Zhang, and Sriram Kosuri. Multiplexed gene synthesis in emulsions for exploring protein functional landscapes. *Science*, 359(6373):343–347, January 2018.

[37] Felix Teufel, Magnús Halldór Gíslason, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, Ole Winther, and Henrik Nielsen. Graphpart: homology partitioning for biological sequence analysis. *NAR genomics and bioinformatics*, 5(4):lqad088, 2023.

[38] Judith Bernett, David B Blumenthal, Dominik G Grimm, Florian Haselbeck, Roman Joeres, Olga V Kalinina, and Markus List. Guiding questions to avoid data leakage in biological machine learning applications. *Nature Methods*, 21(8):1444–1453, 2024.

[39] Yasha Ektefaie, Andrew Shen, Daria Bykova, Maximillian G Marin, Marinka Zitnik, and Maha Farhat. Evaluating generalizability of artificial intelligence models for molecular datasets. *Nature Machine Intelligence*, 6(12):1512–1524, 2024.

[40] Cade W Gordon, Amy X. Lu, and Pieter Abbeel. Protein language model fitness is a matter of preference. In *The Thirteenth International Conference on Learning Representations*, 2025.

[41] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning, 2024.

[42] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

A Few-shot benchmark DMS assay selection

We select a representative collection of assays from ProteinGym for benchmarking few-shot prediction performance. As opposed to previous selections, such as the one used by ProteinNPT and Metalic, we also ensure that the assays are reasonably independent with respect to the training data, so that no close homolog or identical protein was trained on.

We avoid benchmarking on assays that reported abundance, as we consider it an ill-defined target with less relevance for real-world protein optimization efforts. Abundance, as measured by DMS, could be understood as a convolution of biological processes such as expression, stability and degradation. In few-shot protein engineering campaigns, these properties would typically be evaluated one-by-one in real units. However, we still consider abundance assays to be a useful data resource for pre-training on diverse experiments.

We excluded both HIS7_YEAST_Pokusaeva_2019 and CAPSD_AAV2S_Sinai_2021 as their scale proved prohibitive for experimentation given our available GPU resources.

Table A1: The hold out set of ProteinGym DMS assays. Similarity to train is computed as pairwise Needleman-Wunsch global sequence identity of the wild type proteins.

Assay	Type	Similarity to train	Closest protein
AMFR_HUMAN_Tsuboyama_2023_4G3O	Stability	23%	CUE1_YEAST
RCD1_ARATH_Tsuboyama_2023_5OAO	Stability	20%	NUSG_MYCTU
SR43C_ARATH_Tsuboyama_2023_2N88	Stability	39%	CBX4_HUMAN
FECA_ECOLI_Tsuboyama_2023_2D1U	Stability	20%	RPC1_BP434
PKN1_HUMAN_Tsuboyama_2023_1URF	Stability	21%	DN7A_SACS2
CSN4_MOUSE_Tsuboyama_2023_1UFM	Stability	20%	UBE4B_HUMAN
SPA_STAAU_Tsuboyama_2023_1LP1	Stability	22%	HECD1_HUMAN
NKX31_HUMAN_Tsuboyama_2023_2L9R	Stability	32%	PITX2_HUMAN
EPHB2_HUMAN_Tsuboyama_2023_1F0M	Stability	25%	PR40A_HUMAN
SQSTM_MOUSE_Tsuboyama_2023_2RRU	Stability	29%	OTU7A_HUMAN
MAFG_MOUSE_Tsuboyama_2023_1K1V	Stability	24%	RPB1_HUMAN
SCIN_STAAR_Tsuboyama_2023_2QFF	Stability	23%	HVP_LAMBD
DNJA1_HUMAN_Tsuboyama_2023_2LO1	Stability	23%	HECD1_HUMAN
VRPI_BPT7_Tsuboyama_2023_2WNM	Stability	19%	MYO3_YEAST
ESTA_BACSU_Nutschel_2020	Stability	15%	CALM1_HUMAN
CASP3_HUMAN_Roychowdhury_2020	Enz. Activity	48%	CASP7_HUMAN
BLAT_ECOLX_Deng_2012	Enz. Activity	18%	CD19_HUMAN
BLAT_ECOLX_Jacquier_2013	Enz. Activity	18%	CD19_HUMAN
BLAT_ECOLX_Stiffler_2015	Enz. Activity	18%	CD19_HUMAN
BLAT_ECOLX_Firnberg_2014	Enz. Activity	18%	CD19_HUMAN
VKOR1_HUMAN_Chiasson_2020_activity	Enz. Activity	13%	RPC1_BP434
VKOR1_HUMAN_Chiasson_2020_abundance	Abundance	13%	RPC1_BP434
Q8WTC7_9CNID_Somermeyer_2022	Fluoresence	18%	Q6WV13_9MAXI
D7PM05_CLYGR_Somermeyer_2022	Fluoresence	19%	MTH3_HAEAE
GFP_AEQVI_Sarkisyan_2016	Fluoresence	18%	Q6WV13_9MAXI
DLG4_RAT_McLaughlin_2012	Binding	19%	PSAE_SYNP2
RL40A_YEAST_Roscoe_2013	Binding	20%	SPG2_STRSG
GRB2_HUMAN_Faure_2021	Binding	27%	SRBS1_HUMAN
DYR_ECOLI_Thompson_2019	Enz. Activity	15%	NUD15_HUMAN
DLG4_HUMAN_Faure_2021	Binding	25%	EPHB2_HUMAN
RL40A_YEAST_Mavor_2016	Binding	20%	SPG2_STRSG
DYR_ECOLI_Nguyen_2023	Enz. Activity	15%	NUD15_HUMAN
RL40A_YEAST_Roscoe_2014	Binding	20%	SPG2_STRSG

Table A2: The training set of ProteinGym DMS assays.

Assay	Type
A4GRB6_PSEAI_Chen_2020	Enz. Activity
AACC1_PSEAI_Dandage_2018	Enz. Activity
ACE2_HUMAN_Chan_2020	Binding
AICDA_HUMAN_Gajula_2014_3cycles	Enz. Activity
AMIE_PSEAE_Wrenbeck_2017	Enz. Activity
ANCSZ_Hobbs_2022	Enz. Activity
ARGR_ECOLI_Tsuboyama_2023_1AOY	Stability
B2L11_HUMAN_Dutta_2010_binding-Mcl-1	Binding
BBC1_YEAST_Tsuboyama_2023_1TG0	Stability
BCHB_CHLTE_Tsuboyama_2023_2KRU	Stability
CALM1_HUMAN_Weile_2017	Binding
CAS9_STRP1_Spencer_2017_positive	Enz. Activity
CASP7_HUMAN_Roychowdhury_2020	Enz. Activity
CATR_CHLRE_Tsuboyama_2023_2AMI	Stability
CBPA2_HUMAN_Tsuboyama_2023_1O6X	Stability
CBS_HUMAN_Sun_2020	Enz. Activity
CBX4_HUMAN_Tsuboyama_2023_2K28	Stability
CD19_HUMAN_Klesmith_2019_FMC_singles	Binding
CP2C9_HUMAN_Amorosi_2021_abundance	Abundance
CP2C9_HUMAN_Amorosi_2021_activity	Binding
CUE1_YEAST_Tsuboyama_2023_2MYX	Stability
DN7A_SACS2_Tsuboyama_2023_1JIC	Stability
DOCK1_MOUSE_Tsuboyama_2023_2M0Y	Stability
ENVZ_ECOLI_Ghose_2023	Enz. Activity
ERBB2_HUMAN_Elazar_2016	Abundance
F7YBW8_MESOW_Aakre_2015	Binding
F7YBW8_MESOW_Ding_2023	Binding
FKBP3_HUMAN_Tsuboyama_2023_2KFV	Stability
GDIA_HUMAN_Silverstein_2021	Binding
GLPA_HUMAN_Elazar_2016	Abundance
HCP_LAMBD_Tsuboyama_2023_2L6Q	Stability
HECD1_HUMAN_Tsuboyama_2023_3DKM	Stability
HMDH_HUMAN_Jiang_2019	Enz. Activity
HXK4_HUMAN_Gersing_2022_activity	Enz. Activity
HXK4_HUMAN_Gersing_2023_abundance	Abundance
ILF3_HUMAN_Tsuboyama_2023_2L33	Stability
ISDH_STAAW_Tsuboyama_2023_2LHR	Stability
KKA2_KLEPN_Melnikov_2014	Enz. Activity
LGK_LIPST_Klesmith_2015	Enz. Activity
LYAM1_HUMAN_Elazar_2016	Abundance
MBD11_ARATH_Tsuboyama_2023_6ACV	Stability
MET_HUMAN_Estevam_2023	Enz. Activity
MK01_HUMAN_Brenan_2016	Enz. Activity
MSH2_HUMAN_Jia_2020	Enz. Activity
MTH3_HAEAE_RockahShmuel_2015	Enz. Activity
MTHR_HUMAN_Weile_2021	Enz. Activity
MYO3_YEAST_Tsuboyama_2023_2BTT	Stability
NUD15_HUMAN_Suiter_2020	Enz. Activity
NUSA_ECOLI_Tsuboyama_2023_1WCL	Stability
NUSG_MYCTU_Tsuboyama_2023_2MI6	Stability
OBSCN_HUMAN_Tsuboyama_2023_1V1C	Stability
ODP2_GEOSE_Tsuboyama_2023_1W4G	Stability
OPSD_HUMAN_Wan_2019	Abundance
OTC_HUMAN_Lo_2023	Enz. Activity
OTU7A_HUMAN_Tsuboyama_2023_2L2D	Stability

Assay	Type
OXDA_RHOTO_Vanella_2023_activity	Enz. Activity
P84126_THETH_Chan_2017	Enz. Activity
PAI1_HUMAN_Huttinger_2021	Binding
PIN1_HUMAN_Tsuboyama_2023_1I6C	Stability
PITX2_HUMAN_Tsuboyama_2023_2L7M	Stability
POLG_PESV_Tsuboyama_2023_2MXD	Stability
PPARG_HUMAN_Majithia_2016	Binding
PR40A_HUMAN_Tsuboyama_2023_1UZC	Stability
PRKN_HUMAN_Clausen_2023	Abundance
PSAE_PICP2_Tsuboyama_2023_1PSE	Abundance
PTEN_HUMAN_Mighell_2018	Enz. Activity
Q53Z42_HUMAN_McShan_2019_binding-TAPBPR	Binding
Q59976_STRSQ_Romero_2015	Enz. Activity
Q6WV12_9MAXI_Somermeyer_2022	Fluorescence
RAD_ANTMA_Tsuboyama_2023_2CJJ	Stability
RAF1_HUMAN_Zinkus-Boltz_2019	Binding
RASH_HUMAN_Bandaru_2017	Enz. Activity
RASK_HUMAN_Weng_2022_abundance	Abundance
RASK_HUMAN_Weng_2022_binding-DARPin_K55	Binding
RBP1_HUMAN_Tsuboyama_2023_2KWH	Stability
RCRO_LAMBD_Tsuboyama_2023_1ORC	Stability
RD23A_HUMAN_Tsuboyama_2023_1IFY	Stability
RFAH_ECOLI_Tsuboyama_2023_2LCL	Stability
RL20_AQUAE_Tsuboyama_2023_1GYZ	Stability
RNC_ECOLI_Weeks_2023	Enz. Activity
RPC1_BP434_Tsuboyama_2023_1R69	Stability
RS15_GEOSE_Tsuboyama_2023_1A32	Stability
SAV1_MOUSE_Tsuboyama_2023_2YSB	Stability
SBI_STAAM_Tsuboyama_2023_2JVG	Stability
SDA_BACSU_Tsuboyama_2023_1PV0	Stability
SERC_HUMAN_Xie_2023	Enz. Activity
SHOC2_HUMAN_Kwon_2022	Binding
SOX30_HUMAN_Tsuboyama_2023_7JJK	Stability
SPG1_STRSG_Olson_2014	Binding
SPG1_STRSG_Wu_2016	Binding
SPG2_STRSG_Tsuboyama_2023_5UBS	Stability
SPIKE_SARS2_Starr_2020_binding	Binding
SPIKE_SARS2_Starr_2020_expression	Abundance
SPTN1_CHICK_Tsuboyama_2023_1TUD	Stability
SRBS1_HUMAN_Tsuboyama_2023_2O2W	Stability
SRC_HUMAN_Ahler_2019	Enz. Activity
SRC_HUMAN_Chakraborty_2023_binding-DAS_25uM	Enz. Activity
TCRG1_MOUSE_Tsuboyama_2023_1E0L	Stability
THO1_YEAST_Tsuboyama_2023_2WQG	Stability
TNKS2_HUMAN_Tsuboyama_2023_5JRT	Stability
TPK1_HUMAN_Weile_2017	Enz. Activity
TPMT_HUMAN_Matreyek_2018	Abundance
UBC9_HUMAN_Weile_2017	Enz. Activity
UBE4B_HUMAN_Tsuboyama_2023_3L1X	Stability
UBE4B_MOUSE_Starita_2013	Enz. Activity
UBR5_HUMAN_Tsuboyama_2023_1I2T	Stability
VG08_BPP22_Tsuboyama_2023_2GP8	Stability
VILI_CHICK_Tsuboyama_2023_1YU5	Stability
YAIA_ECOLI_Tsuboyama_2023_2KVT	Stability
YAP1_HUMAN_Araya_2012	Binding
YNZC_BACSU_Tsuboyama_2023_2JVD	Stability

B Experimental details

We follow ProteinNPT [2] for the Ridge regression baseline. Specifically, we process the mean-pooled ESM embedding using a linear layer, and the zero-shot score from ProGen using another linear layer without bias which is initialized with weight 1.0. We apply an L2 penalty of 5×10^{-3} for the embedding linear layer and 1×10^{-8} for the zero-shot linear layer. The models are trained for 1500 steps at a learning rate of 0.01. The Gaussian process (GP) model was adapted from Kermut [3]. We reuse model components and the training loop from the official Kermut code release, omitting the structure kernel from the GP's composite kernel. GP models are trained for 150 steps at a learning rate of 3×10^{-4}. The random forest (RF) baseline including all its hyperparameters was taken from the official EvolvePro codebase [25].

Table A3: Hyperparameters of PRIMO. Unless stated otherwise, the architectural configuration of the transformer is based on the ESM-2 attention block, with the PRIMO layer adaptations as discussed in the main text. For the TTT phase, we only list parameters that differ from the pre-training setup. PRIMO was trained on a single RTX 6000, using gradient accumulation to enable the specified batch size.

Parameter	Value
Layers	6
Hidden size	400
Attention heads	8
Feedforward factor	4
Pooling vectors per sequence	3
AA Embedder	ESM-2 650M [28]
Zero-shot predictor	ProGen-2 medium [11]
Dropout	0.1
Weight decay	0.01
Gradient norm clipping	1.0
Learning rate	0.0001
Learning rate schedule	Triangular
Warmup steps	1000
Total sets	150,000
Set size	32
Batch size	12
AA sequence length	512
Start MLM loss factor	0.5
End MLM loss factor	0.05
MLM loss factor schedule	Cosine
Label masking probability	0.33
Span masking probability[1]	0.2
Minimum span fraction	0.05
Maximum span fraction	0.15
Span length distribution	Uniform
Mask all variant positions[2]	True
TTT learning rate	0.0001
TTT learning rate schedule	Flat
TTT training set size	min(N, 32)
TTT sequence length	L
TTT total steps	25
TTT/ICL inference set size	$N + 1$

[1]The span masking probability is applied after the label masking probability, so that the effective probability becomes $(1 - 0.33) * 0.2$.

[2]This means that the sampled span length will be split over k mutation positions as needed.

C Additional results

C.1 Ablations

We ablate the following components of PRIMO:

- Conditioning approach for zero-shot prediction (Table A4)
- Loss function (Table A5)
- Auxiluary inputs (Property type and pLM zero-shot score) (Table A6)
- TTT protocol (Table A7)
- Attention mechanism (Table A8)

For efficiency, ablation experiments are performed at a constant set size of $N = 16$ (with the exception of the zero-shot ablation). Overall, the ablations demonstrate that the modeling choices of PRIMO are sensible, and all individually contribute to improved performance.

Table A4: Effect of providing an arbitrary reference datapoint with fitness set to 0.5 (midpoint of the min-max scale) for calibration when performing 0-shot prediction with PRIMO. When no context sequence is provided, the prediction performance of the logits returned by PRIMO collapses.

Approach	Average Spearman
Unlabeled sequence only	- 0.257
Reference sequence with arbitrary fitness	0.522

Table A5: Performance of PRIMO (TTT) when training using different loss functions at N=16.

Loss	Average Spearman
Preference	0.538
MSE	0.528

Table A6: Performance of PRIMO (TTT) when ablating additional inputs at N=16.

Inputs	Average Spearman
All	0.538
w/o Property ID	0.532
w/o zero-shot score	0.529

Table A7: Performance of PRIMO (TTT) with different numbers of TTT steps at N=16.

Steps	Avg. Spearman
0 (ICL)	0.513
10	0.526
25	0.539
50	0.526
75	0.518
100	0.515

Table A8: Performance of PRIMO models using different attention mechanisms at N=16. *Tiered* refers to PoET's tiered attention, *Pooled* is the attention mechanism used by the final PRIMO model described in the main text. The tiered attention model was trained at a pre-training set size of 16 due to increased computational complexity. Results were obtained on a preliminary data split before training the final PRIMO model.

	Tiered	Pooled
ICL	0.45	0.48
TTT	0.43	0.47

C.2 Assay level performance

Table A9: Per-assay results on the hold out set (1/2).

Shots	Method	AMFR_HUMAN_Tsuboyama_2023_4G3O	BLAT_ECOLX_Deng_2012	BLAT_ECOLX_Firnberg_2014	BLAT_ECOLX_Gonzalez_2019	BLAT_ECOLX_Jacquier_2013	BLAT_ECOLX_Stiffler_2015	CASP3_HUMAN_Roychowdhury_2020	CSN4_MOUSE_Tsuboyama_2023_1UFM	D7PM05_CLYGR_Sarkermeyer_2022	DLG4_HUMAN_Faure_2021	DLG4_RAT_McLaughlin_2012	DNJA1_HUMAN_Tsuboyama_2023_2LO1	DYR_ECOLI_Nguyen_2023	DYR_ECOLI_Thompson_2019	EPHB2_HUMAN_Tsuboyama_2023_1F0M	ESTA_BACSU_Nutschel_2020	FECA_ECOLI_Tsuboyama_2023_2D1U
0	ProGen	0.01	0.43	0.64	0.62	0.64	0.19	0.51	0.57	0.13	0.61	0.41	0.75	0.44	0.39	0.65	0.26	0.50
	PRIMO (ICL)	0.73 ± 0.03	0.44 ± 0.02	0.54 ± 0.17	0.56 ± 0.03	0.48 ± 0.25	0.67 ± 0.01	0.43 ± 0.21	0.66 ± 0.01	0.06 ± 0.0	0.58 ± 0.0	0.4 ± 0.0	0.84 ± 0.01	0.45 ± 0.06	0.41 ± 0.07	0.87 ± 0.01	0.24 ± 0.01	0.66 ± 0.01
	PRIMO (TTT)	0.71 ± 0.02	0.42 ± 0.01	0.66 ± 0.01	0.57 ± 0.02	0.6 ± 0.01	0.65 ± 0.02	0.47 ± 0.14	0.59 ± 0.06	0.18 ± 0.01	0.53 ± 0.01	0.38 ± 0.0	0.8 ± 0.02	0.44 ± 0.06	0.44 ± 0.01	0.85 ± 0.02	0.27 ± 0.01	0.62 ± 0.03
4	GP	0.36 ± 0.24	0.12 ± 0.17	0.21 ± 0.13	0.13 ± 0.51	0.16 ± 0.11	0.17 ± 0.15	0.09 ± 0.18	0.21 ± 0.28	-0.1 ± 0.05	0.27 ± 0.07	0.14 ± 0.25	0.61 ± 0.24	-0.01 ± 0.13	0.04 ± 0.09	0.7 ± 0.04	0.02 ± 0.06	0.38 ± 0.07
	PRIMO (ICL)	0.77 ± 0.01	0.45 ± 0.0	0.68 ± 0.0	0.61 ± 0.0	0.62 ± 0.0	0.68 ± 0.0	0.59 ± 0.0	0.66 ± 0.02	0.06 ± 0.0	0.57 ± 0.0	0.41 ± 0.0	0.85 ± 0.0	0.5 ± 0.0	0.45 ± 0.0	0.87 ± 0.0	0.24 ± 0.0	0.67 ± 0.0
	PRIMO (TTT)	0.54 ± 0.18	0.36 ± 0.06	0.57 ± 0.05	0.62 ± 0.06	0.48 ± 0.12	0.46 ± 0.1	0.48 ± 0.05	0.52 ± 0.18	0.14 ± 0.04	0.5 ± 0.11	0.39 ± 0.04	0.79 ± 0.09	0.42 ± 0.07	0.33 ± 0.16	0.86 ± 0.03	0.22 ± 0.04	0.62 ± 0.05
	RF	0.19 ± 0.2	0.05 ± 0.07	0.12 ± 0.12	0.34 ± 0.32	0.07 ± 0.13	0.07 ± 0.1	0.1 ± 0.19	0.27 ± 0.32	-0.01 ± 0.15	0.27 ± 0.08	0.05 ± 0.17	0.66 ± 0.18	0.07 ± 0.09	-0.01 ± 0.07	0.67 ± 0.04	0.01 ± 0.08	0.17 ± 0.11
	Ridge	0.19 ± 0.21	0.08 ± 0.14	0.13 ± 0.13	0.34 ± 0.26	0.1 ± 0.18	0.1 ± 0.17	0.21 ± 0.25	0.24 ± 0.34	0.03 ± 0.08	0.25 ± 0.17	0.11 ± 0.24	0.68 ± 0.2	0.09 ± 0.1	0.03 ± 0.09	0.73 ± 0.05	0.03 ± 0.07	0.32 ± 0.07
8	GP	0.5 ± 0.05	0.11 ± 0.13	0.28 ± 0.06	0.34 ± 0.39	0.28 ± 0.07	0.19 ± 0.17	0.19 ± 0.19	0.32 ± 0.24	-0.05 ± 0.07	0.31 ± 0.06	0.17 ± 0.26	0.66 ± 0.14	0.11 ± 0.09	0.19 ± 0.1	0.74 ± 0.05	0.09 ± 0.09	0.44 ± 0.07
	PRIMO (ICL)	0.8 ± 0.0	0.45 ± 0.0	0.68 ± 0.0	0.61 ± 0.0	0.62 ± 0.0	0.68 ± 0.0	0.59 ± 0.0	0.66 ± 0.01	0.06 ± 0.0	0.57 ± 0.0	0.41 ± 0.0	0.85 ± 0.0	0.5 ± 0.0	0.45 ± 0.0	0.87 ± 0.0	0.24 ± 0.0	0.67 ± 0.0
	PRIMO (TTT)	0.66 ± 0.18	0.35 ± 0.05	0.65 ± 0.04	0.64 ± 0.03	0.54 ± 0.05	0.53 ± 0.07	0.48 ± 0.09	0.55 ± 0.36	0.07 ± 0.07	0.51 ± 0.06	0.32 ± 0.12	0.82 ± 0.02	0.46 ± 0.04	0.4 ± 0.04	0.86 ± 0.02	0.25 ± 0.04	0.64 ± 0.08
	RF	0.43 ± 0.09	0.09 ± 0.13	0.27 ± 0.1	0.34 ± 0.12	0.21 ± 0.09	0.19 ± 0.08	0.24 ± 0.18	0.29 ± 0.25	0.03 ± 0.08	0.28 ± 0.06	0.15 ± 0.18	0.66 ± 0.16	0.1 ± 0.11	0.1 ± 0.07	0.75 ± 0.05	0.09 ± 0.05	0.42 ± 0.06
	Ridge	0.4 ± 0.14	0.1 ± 0.11	0.27 ± 0.09	0.38 ± 0.16	0.25 ± 0.11	0.2 ± 0.16	0.27 ± 0.25	0.33 ± 0.39	0.03 ± 0.08	0.29 ± 0.14	0.16 ± 0.21	0.66 ± 0.2	0.17 ± 0.12	0.21 ± 0.08	0.76 ± 0.04	0.06 ± 0.04	0.43 ± 0.08
16	GP	0.58 ± 0.06	0.15 ± 0.11	0.32 ± 0.08	0.57 ± 0.04	0.35 ± 0.02	0.21 ± 0.14	0.31 ± 0.15	0.45 ± 0.18	-0.02 ± 0.03	0.33 ± 0.01	0.3 ± 0.01	0.77 ± 0.02	0.18 ± 0.07	0.2 ± 0.04	0.77 ± 0.04	0.16 ± 0.06	0.46 ± 0.08
	PRIMO (ICL)	0.8 ± 0.0	0.45 ± 0.0	0.68 ± 0.0	0.6 ± 0.0	0.62 ± 0.0	0.68 ± 0.0	0.59 ± 0.0	0.66 ± 0.02	0.06 ± 0.0	0.57 ± 0.0	0.41 ± 0.0	0.85 ± 0.0	0.5 ± 0.0	0.45 ± 0.0	0.87 ± 0.0	0.24 ± 0.0	0.67 ± 0.0
	PRIMO (TTT)	0.74 ± 0.05	0.4 ± 0.04	0.66 ± 0.02	0.61 ± 0.06	0.58 ± 0.04	0.63 ± 0.03	0.54 ± 0.02	0.73 ± 0.05	0.15 ± 0.08	0.55 ± 0.06	0.41 ± 0.06	0.84 ± 0.02	0.41 ± 0.07	0.38 ± 0.07	0.87 ± 0.01	0.26 ± 0.04	0.68 ± 0.04
	RF	0.53 ± 0.07	0.12 ± 0.11	0.34 ± 0.07	0.32 ± 0.07	0.3 ± 0.08	0.22 ± 0.12	0.36 ± 0.06	0.51 ± 0.14	0.05 ± 0.07	0.36 ± 0.05	0.23 ± 0.07	0.78 ± 0.03	0.11 ± 0.08	0.16 ± 0.06	0.78 ± 0.03	0.14 ± 0.06	0.44 ± 0.08
	Ridge	0.58 ± 0.07	0.12 ± 0.1	0.35 ± 0.07	0.44 ± 0.11	0.33 ± 0.05	0.23 ± 0.1	0.39 ± 0.06	0.57 ± 0.13	0.09 ± 0.1	0.38 ± 0.06	0.28 ± 0.02	0.8 ± 0.03	0.2 ± 0.09	0.23 ± 0.05	0.77 ± 0.05	0.15 ± 0.06	0.47 ± 0.1
32	GP	0.65 ± 0.05	0.22 ± 0.04	0.39 ± 0.06	0.62 ± 0.02	0.38 ± 0.04	0.33 ± 0.18	0.31 ± 0.11	0.57 ± 0.1	0.05 ± 0.03	0.36 ± 0.04	0.32 ± 0.01	0.8 ± 0.02	0.23 ± 0.11	0.31 ± 0.03	0.79 ± 0.02	0.17 ± 0.05	0.54 ± 0.12
	PRIMO (ICL)	0.8 ± 0.0	0.45 ± 0.0	0.68 ± 0.0	0.61 ± 0.0	0.62 ± 0.0	0.68 ± 0.0	0.59 ± 0.0	0.66 ± 0.02	0.06 ± 0.0	0.57 ± 0.0	0.41 ± 0.0	0.85 ± 0.0	0.5 ± 0.0	0.45 ± 0.0	0.87 ± 0.0	0.24 ± 0.0	0.67 ± 0.0
	PRIMO (TTT)	0.78 ± 0.03	0.4 ± 0.05	0.69 ± 0.02	0.72 ± 0.04	0.58 ± 0.04	0.67 ± 0.02	0.54 ± 0.03	0.8 ± 0.02	0.19 ± 0.06	0.59 ± 0.04	0.52 ± 0.05	0.86 ± 0.04	0.45 ± 0.06	0.42 ± 0.07	0.88 ± 0.01	0.31 ± 0.04	0.72 ± 0.04
	RF	0.64 ± 0.09	0.23 ± 0.05	0.41 ± 0.06	0.4 ± 0.16	0.36 ± 0.05	0.33 ± 0.09	0.4 ± 0.04	0.6 ± 0.05	0.11 ± 0.06	0.45 ± 0.11	0.32 ± 0.02	0.79 ± 0.03	0.13 ± 0.11	0.21 ± 0.08	0.76 ± 0.07	0.15 ± 0.04	0.5 ± 0.07
	Ridge	0.63 ± 0.09	0.22 ± 0.06	0.45 ± 0.06	0.48 ± 0.09	0.43 ± 0.04	0.37 ± 0.08	0.42 ± 0.05	0.69 ± 0.06	0.1 ± 0.06	0.48 ± 0.1	0.34 ± 0.01	0.8 ± 0.04	0.25 ± 0.13	0.3 ± 0.05	0.81 ± 0.01	0.19 ± 0.04	0.62 ± 0.03
64	GP	0.72 ± 0.03	0.28 ± 0.04	0.42 ± 0.03	0.64 ± 0.01	0.45 ± 0.02	0.38 ± 0.12	0.35 ± 0.08	0.67 ± 0.03	0.11 ± 0.03	0.36 ± 0.02	0.33 ± 0.02	0.82 ± 0.01	0.3 ± 0.08	0.38 ± 0.04	0.82 ± 0.01	0.21 ± 0.04	0.65 ± 0.05
	PRIMO (ICL)	0.8 ± 0.0	0.45 ± 0.0	0.68 ± 0.0	0.6 ± 0.0	0.62 ± 0.0	0.68 ± 0.0	0.59 ± 0.0	0.66 ± 0.01	0.06 ± 0.0	0.57 ± 0.0	0.41 ± 0.0	0.85 ± 0.0	0.5 ± 0.0	0.45 ± 0.0	0.87 ± 0.0	0.24 ± 0.0	0.67 ± 0.0
	PRIMO (TTT)	0.82 ± 0.02	0.44 ± 0.05	0.72 ± 0.01	0.78 ± 0.03	0.61 ± 0.02	0.7 ± 0.01	0.54 ± 0.03	0.86 ± 0.01	0.24 ± 0.05	0.67 ± 0.04	0.57 ± 0.03	0.88 ± 0.01	0.47 ± 0.04	0.47 ± 0.03	0.89 ± 0.01	0.33 ± 0.02	0.78 ± 0.01
	RF	0.73 ± 0.02	0.33 ± 0.05	0.5 ± 0.03	0.56 ± 0.07	0.45 ± 0.04	0.45 ± 0.08	0.45 ± 0.03	0.65 ± 0.04	0.2 ± 0.04	0.5 ± 0.04	0.36 ± 0.03	0.81 ± 0.02	0.18 ± 0.12	0.32 ± 0.05	0.82 ± 0.02	0.21 ± 0.04	0.66 ± 0.02
	Ridge	0.75 ± 0.04	0.33 ± 0.05	0.53 ± 0.02	0.57 ± 0.1	0.52 ± 0.04	0.49 ± 0.07	0.46 ± 0.03	0.77 ± 0.03	0.2 ± 0.02	0.54 ± 0.03	0.37 ± 0.02	0.85 ± 0.02	0.36 ± 0.06	0.39 ± 0.03	0.83 ± 0.01	0.25 ± 0.04	0.72 ± 0.03
128	GP	0.78 ± 0.02	0.32 ± 0.03	0.45 ± 0.04	0.66 ± 0.01	0.49 ± 0.01	0.49 ± 0.06	0.4 ± 0.06	0.73 ± 0.02	0.18 ± 0.02	0.37 ± 0.02	0.34 ± 0.02	0.84 ± 0.01	0.39 ± 0.04	0.4 ± 0.03	0.84 ± 0.01	0.27 ± 0.03	0.75 ± 0.01
	PRIMO (ICL)	0.8 ± 0.0	0.45 ± 0.0	0.68 ± 0.0	0.61 ± 0.0	0.62 ± 0.0	0.68 ± 0.0	0.59 ± 0.0	0.66 ± 0.01	0.06 ± 0.0	0.57 ± 0.0	0.41 ± 0.0	0.85 ± 0.0	0.5 ± 0.0	0.45 ± 0.0	0.87 ± 0.0	0.24 ± 0.0	0.67 ± 0.0
	PRIMO (TTT)	0.83 ± 0.01	0.48 ± 0.05	0.74 ± 0.01	0.82 ± 0.02	0.66 ± 0.02	0.73 ± 0.02	0.6 ± 0.01	0.87 ± 0.01	0.29 ± 0.06	0.71 ± 0.04	0.63 ± 0.01	0.91 ± 0.0	0.54 ± 0.04	0.52 ± 0.02	0.91 ± 0.01	0.42 ± 0.03	0.81 ± 0.0
	RF	0.78 ± 0.02	0.39 ± 0.04	0.59 ± 0.03	0.67 ± 0.04	0.55 ± 0.02	0.59 ± 0.03	0.52 ± 0.02	0.71 ± 0.04	0.28 ± 0.04	0.58 ± 0.01	0.44 ± 0.02	0.85 ± 0.0	0.27 ± 0.05	0.38 ± 0.03	0.85 ± 0.01	0.33 ± 0.03	0.74 ± 0.04
	Ridge	0.82 ± 0.02	0.37 ± 0.04	0.61 ± 0.03	0.63 ± 0.04	0.57 ± 0.03	0.61 ± 0.02	0.5 ± 0.02	0.83 ± 0.03	0.27 ± 0.04	0.58 ± 0.02	0.41 ± 0.02	0.89 ± 0.01	0.44 ± 0.03	0.46 ± 0.03	0.86 ± 0.01	0.32 ± 0.04	0.78 ± 0.03

Table A10: Per-assay results on the hold out set (2/2).

	Shots	GFP_AEQVI_Sarkisyan_2016	GRB2_HUMAN_Faure_2021	MAFG_MOUSE_Tsuboyama_2023_1K1V	NKX31_HUMAN_Tsuboyama_2023_2L9R	PKN1_HUMAN_Tsuboyama_2023_1URF	Q8WTC7_9CNID_Sommermeyer_2022	RCD1_ARATH_Tsuboyama_2023_5OAO	RL40A_YEAST_Mavor_2016	RL40A_YEAST_Roscoe_2013	RL40A_YEAST_Roscoe_2014	SCIN_STAAR_Tsuboyama_2023_2QFF	SPA_STAAU_Tsuboyama_2023_1LP1	SQSTM_MOUSE_Tsuboyama_2023_2RRU	SR43C_ARATH_Tsuboyama_2023_2N88	VKOR1_HUMAN_Chiasson_2020_abundance	VKOR1_HUMAN_Chiasson_2020_activity	VRPI_BPT7_Tsuboyama_2023_2WNM
0	ProGen	0.15	0.53	0.51	0.58	0.43	0.01	0.38	0.42	0.48	.38	0.13	-0.09	0.57	0.56	0.50	0.41	0.11
	PRIMO (ICL)	0.08 ± 0.0	0.6 ± 0.0	0.2 ± 0.04	0.75 ± 0.0	0.71 ± 0.0	0.05 ± 0.0	0.72 ± 0.0	0.4 ± 0.0	0.45 ± 0.0	0.54 ± 0.0	0.51 ± 0.01	0.2 ± 0.0	0.76 ± 0.01	0.78 ± 0.04	0.45 ± 0.0	0.35 ± 0.01	0.78 ± 0.07
	PRIMO (TTT)	0.08 ± 0.05	0.59 ± 0.01	0.34 ± 0.07	0.65 ± 0.11	0.55 ± 0.02	0.08 ± 0.01	0.66 ± 0.04	0.38 ± 0.01	0.45 ± 0.02	0.47 ± 0.03	0.57 ± 0.02	0.26 ± 0.02	0.73 ± 0.02	0.81 ± 0.02	0.42 ± 0.01	0.34 ± 0.01	0.76 ± 0.01
4	GP	0.08 ± 0.22	0.23 ± 0.12	0.24 ± 0.3	0.39 ± 0.36	0.4 ± 0.05	0.06 ± 0.02	0.42 ± 0.11	0.11 ± 0.13	0.19 ± 0.19	0.13 ± 0.14	0.29 ± 0.15	0.34 ± 0.09	0.59 ± 0.04	0.62 ± 0.15	0.23 ± 0.16	0.06 ± 0.1	0.44 ± 0.32
	PRIMO (ICL)	0.07 ± 0.0	0.6 ± 0.0	0.23 ± 0.01	0.75 ± 0.01	0.71 ± 0.0	0.05 ± 0.0	0.74 ± 0.0	0.4 ± 0.0	0.45 ± 0.0	0.54 ± 0.0	0.52 ± 0.0	0.18 ± 0.01	0.75 ± 0.08	0.81 ± 0.0	0.45 ± 0.0	0.36 ± 0.0	0.82 ± 0.0
	PRIMO (TTT)	0.24 ± 0.14	0.56 ± 0.06	0.46 ± 0.07	0.7 ± 0.05	0.65 ± 0.03	0.06 ± 0.03	0.69 ± 0.02	0.33 ± 0.11	0.39 ± 0.12	0.34 ± 0.13	0.52 ± 0.04	0.29 ± 0.15	0.76 ± 0.02	0.8 ± 0.08	0.43 ± 0.06	0.25 ± 0.05	0.8 ± 0.02
	RF	0.25 ± 0.14	0.17 ± 0.12	0.18 ± 0.26	0.3 ± 0.43	0.35 ± 0.06	0.05 ± 0.04	0.51 ± 0.07	0.15 ± 0.1	0.17 ± 0.22	0.09 ± 0.12	0.35 ± 0.09	0.31 ± 0.1	0.51 ± 0.06	0.65 ± 0.08	0.18 ± 0.09	0.04 ± 0.09	0.67 ± 0.1
	Ridge	0.23 ± 0.12	0.24 ± 0.08	0.22 ± 0.17	0.42 ± 0.26	0.26 ± 0.16	-0.0 ± 0.03	0.5 ± 0.05	0.12 ± 0.17	0.15 ± 0.17	0.16 ± 0.12	0.37 ± 0.2	0.37 ± 0.1	0.53 ± 0.05	0.65 ± 0.09	0.23 ± 0.12	0.06 ± 0.07	0.68 ± 0.11
8	GP	0.08 ± 0.24	0.36 ± 0.04	0.51 ± 0.19	0.57 ± 0.09	0.45 ± 0.05	0.03 ± 0.03	0.53 ± 0.07	0.16 ± 0.12	0.19 ± 0.13	0.12 ± 0.15	0.42 ± 0.18	0.33 ± 0.13	0.62 ± 0.04	0.75 ± 0.07	0.25 ± 0.2	0.03 ± 0.06	0.64 ± 0.11
	PRIMO (ICL)	0.06 ± 0.0	0.6 ± 0.0	0.22 ± 0.01	0.75 ± 0.01	0.71 ± 0.0	0.06 ± 0.0	0.74 ± 0.0	0.4 ± 0.0	0.45 ± 0.0	0.54 ± 0.0	0.53 ± 0.0	0.17 ± 0.0	0.75 ± 0.06	0.81 ± 0.0	0.45 ± 0.0	0.36 ± 0.0	0.82 ± 0.0
	PRIMO (TTT)	0.22 ± 0.08	0.55 ± 0.05	0.6 ± 0.19	0.7 ± 0.06	0.67 ± 0.04	0.04 ± 0.05	0.68 ± 0.04	0.34 ± 0.1	0.43 ± 0.09	0.35 ± 0.12	0.54 ± 0.05	0.33 ± 0.06	0.74 ± 0.02	0.82 ± 0.06	0.38 ± 0.11	0.24 ± 0.08	0.8 ± 0.02
	RF	0.34 ± 0.08	0.33 ± 0.08	0.47 ± 0.1	0.51 ± 0.14	0.44 ± 0.02	-0.02 ± 0.04	0.59 ± 0.04	0.21 ± 0.09	0.21 ± 0.12	0.15 ± 0.11	0.43 ± 0.11	0.3 ± 0.16	0.58 ± 0.04	0.72 ± 0.07	0.17 ± 0.14	0.1 ± 0.07	0.7 ± 0.07
	Ridge	0.19 ± 0.18	0.35 ± 0.07	0.52 ± 0.16	0.56 ± 0.09	0.38 ± 0.24	0.0 ± 0.04	0.54 ± 0.05	0.22 ± 0.14	0.25 ± 0.19	0.17 ± 0.14	0.47 ± 0.13	0.28 ± 0.25	0.48 ± 0.07	0.76 ± 0.08	0.23 ± 0.17	0.04 ± 0.07	0.73 ± 0.04
16	GP	0.22 ± 0.04	0.42 ± 0.03	0.56 ± 0.1	0.66 ± 0.04	0.5 ± 0.05	0.05 ± 0.02	0.6 ± 0.05	0.2 ± 0.09	0.25 ± 0.04	0.22 ± 0.13	0.54 ± 0.08	0.42 ± 0.1	0.65 ± 0.03	0.81 ± 0.01	0.32 ± 0.12	0.08 ± 0.12	0.74 ± 0.04
	PRIMO (ICL)	0.06 ± 0.0	0.6 ± 0.0	0.23 ± 0.01	0.75 ± 0.0	0.71 ± 0.0	0.06 ± 0.0	0.74 ± 0.0	0.4 ± 0.0	0.45 ± 0.0	0.54 ± 0.0	0.53 ± 0.0	0.17 ± 0.0	0.76 ± 0.03	0.81 ± 0.0	0.45 ± 0.0	0.36 ± 0.0	0.82 ± 0.0
	PRIMO (TTT)	0.23 ± 0.07	0.6 ± 0.04	0.61 ± 0.12	0.76 ± 0.02	0.67 ± 0.04	0.05 ± 0.03	0.72 ± 0.02	0.42 ± 0.08	0.46 ± 0.08	0.45 ± 0.11	0.59 ± 0.06	0.34 ± 0.1	0.67 ± 0.09	0.85 ± 0.01	0.43 ± 0.14	0.24 ± 0.06	0.8 ± 0.02
	RF	0.33 ± 0.05	0.39 ± 0.07	0.51 ± 0.14	0.65 ± 0.04	0.49 ± 0.04	0.08 ± 0.04	0.62 ± 0.03	0.26 ± 0.08	0.32 ± 0.06	0.24 ± 0.12	0.52 ± 0.09	0.41 ± 0.12	0.6 ± 0.02	0.8 ± 0.02	0.27 ± 0.07	0.16 ± 0.08	0.75 ± 0.02
	Ridge	0.23 ± 0.09	0.45 ± 0.06	0.58 ± 0.1	0.66 ± 0.03	0.53 ± 0.08	0.04 ± 0.02	0.6 ± 0.05	0.29 ± 0.1	0.29 ± 0.14	0.29 ± 0.13	0.53 ± 0.07	0.36 ± 0.15	0.53 ± 0.06	0.81 ± 0.01	0.35 ± 0.12	0.12 ± 0.12	0.74 ± 0.02
32	GP	0.32 ± 0.04	0.43 ± 0.06	0.64 ± 0.04	0.71 ± 0.03	0.61 ± 0.04	0.06 ± 0.01	0.66 ± 0.01	0.29 ± 0.14	0.38 ± 0.1	0.24 ± 0.13	0.57 ± 0.07	0.46 ± 0.06	0.7 ± 0.04	0.83 ± 0.01	0.4 ± 0.1	0.24 ± 0.04	0.77 ± 0.01
	PRIMO (ICL)	0.06 ± 0.0	0.6 ± 0.0	0.23 ± 0.01	0.75 ± 0.0	0.71 ± 0.0	0.06 ± 0.0	0.74 ± 0.0	0.4 ± 0.0	0.45 ± 0.0	0.54 ± 0.0	0.53 ± 0.0	0.17 ± 0.0	0.77 ± 0.02	0.81 ± 0.0	0.45 ± 0.0	0.36 ± 0.0	0.82 ± 0.0
	PRIMO (TTT)	0.31 ± 0.05	0.62 ± 0.04	0.69 ± 0.09	0.77 ± 0.02	0.7 ± 0.02	0.1 ± 0.04	0.73 ± 0.04	0.49 ± 0.08	0.61 ± 0.05	0.51 ± 0.1	0.59 ± 0.07	0.34 ± 0.09	0.77 ± 0.03	0.87 ± 0.01	0.46 ± 0.08	0.31 ± 0.05	0.82 ± 0.01
	RF	0.41 ± 0.03	0.45 ± 0.08	0.59 ± 0.13	0.68 ± 0.02	0.58 ± 0.05	0.1 ± 0.05	0.63 ± 0.06	0.39 ± 0.07	0.41 ± 0.05	0.28 ± 0.13	0.59 ± 0.01	0.46 ± 0.07	0.67 ± 0.03	0.8 ± 0.02	0.37 ± 0.1	0.24 ± 0.06	0.76 ± 0.02
	Ridge	0.32 ± 0.06	0.5 ± 0.06	0.7 ± 0.07	0.7 ± 0.03	0.65 ± 0.05	0.05 ± 0.05	0.63 ± 0.06	0.38 ± 0.11	0.44 ± 0.04	0.36 ± 0.12	0.56 ± 0.09	0.44 ± 0.11	0.69 ± 0.06	0.83 ± 0.01	0.39 ± 0.11	0.25 ± 0.02	0.75 ± 0.04
64	GP	0.35 ± 0.03	0.5 ± 0.03	0.73 ± 0.03	0.73 ± 0.03	0.65 ± 0.03	0.05 ± 0.04	0.69 ± 0.01	0.41 ± 0.06	0.49 ± 0.08	0.37 ± 0.06	0.63 ± 0.03	0.53 ± 0.03	0.76 ± 0.03	0.85 ± 0.01	0.43 ± 0.04	0.31 ± 0.06	0.8 ± 0.01
	PRIMO (ICL)	0.06 ± 0.0	0.6 ± 0.0	0.23 ± 0.01	0.76 ± 0.0	0.71 ± 0.0	0.06 ± 0.0	0.74 ± 0.0	0.4 ± 0.0	0.45 ± 0.0	0.54 ± 0.0	0.53 ± 0.0	0.17 ± 0.0	0.78 ± 0.01	0.81 ± 0.0	0.45 ± 0.0	0.36 ± 0.0	0.82 ± 0.0
	PRIMO (TTT)	0.35 ± 0.06	0.66 ± 0.03	0.76 ± 0.05	0.79 ± 0.03	0.73 ± 0.02	0.14 ± 0.05	0.75 ± 0.04	0.56 ± 0.05	0.69 ± 0.04	0.61 ± 0.05	0.68 ± 0.03	0.44 ± 0.08	0.81 ± 0.03	0.88 ± 0.01	0.51 ± 0.01	0.35 ± 0.07	0.85 ± 0.02
	RF	0.43 ± 0.03	0.5 ± 0.04	0.73 ± 0.03	0.74 ± 0.03	0.63 ± 0.04	0.1 ± 0.01	0.67 ± 0.02	0.49 ± 0.02	0.53 ± 0.02	0.41 ± 0.06	0.65 ± 0.02	0.51 ± 0.05	0.72 ± 0.04	0.82 ± 0.02	0.43 ± 0.03	0.37 ± 0.04	0.81 ± 0.02
	Ridge	0.37 ± 0.06	0.6 ± 0.02	0.8 ± 0.02	0.75 ± 0.01	0.68 ± 0.04	0.05 ± 0.07	0.69 ± 0.04	0.51 ± 0.04	0.56 ± 0.04	0.49 ± 0.04	0.67 ± 0.03	0.56 ± 0.05	0.77 ± 0.03	0.86 ± 0.01	0.46 ± 0.03	0.37 ± 0.05	0.81 ± 0.01
128	GP	0.42 ± 0.03	0.54 ± 0.03	0.8 ± 0.01	0.76 ± 0.02	0.71 ± 0.02	0.11 ± 0.04	0.74 ± 0.01	0.49 ± 0.03	0.57 ± 0.03	0.43 ± 0.04	0.69 ± 0.01	0.62 ± 0.01	0.81 ± 0.02	0.87 ± 0.0	0.51 ± 0.04	0.37 ± 0.02	0.82 ± 0.02
	PRIMO (ICL)	0.06 ± 0.0	0.6 ± 0.0	0.24 ± 0.0	0.75 ± 0.0	0.71 ± 0.0	0.06 ± 0.0	0.74 ± 0.0	0.4 ± 0.0	0.45 ± 0.0	0.54 ± 0.0	0.53 ± 0.0	0.17 ± 0.0	0.78 ± 0.0	0.81 ± 0.0	0.45 ± 0.0	0.36 ± 0.0	0.82 ± 0.0
	PRIMO (TTT)	0.42 ± 0.03	0.71 ± 0.01	0.77 ± 0.07	0.83 ± 0.03	0.75 ± 0.03	0.2 ± 0.04	0.81 ± 0.02	0.65 ± 0.02	0.75 ± 0.02	0.68 ± 0.02	0.71 ± 0.02	0.56 ± 0.03	0.85 ± 0.01	0.9 ± 0.01	0.56 ± 0.03	0.47 ± 0.04	0.86 ± 0.01
	RF	0.49 ± 0.01	0.59 ± 0.02	0.79 ± 0.02	0.78 ± 0.01	0.68 ± 0.02	0.18 ± 0.03	0.74 ± 0.0	0.58 ± 0.02	0.62 ± 0.03	0.48 ± 0.03	0.71 ± 0.02	0.61 ± 0.02	0.78 ± 0.01	0.85 ± 0.01	0.54 ± 0.03	0.45 ± 0.04	0.83 ± 0.03
	Ridge	0.46 ± 0.04	0.66 ± 0.02	0.85 ± 0.02	0.81 ± 0.02	0.74 ± 0.03	0.13 ± 0.05	0.8 ± 0.01	0.59 ± 0.02	0.67 ± 0.02	0.58 ± 0.04	0.77 ± 0.02	0.71 ± 0.03	0.84 ± 0.02	0.89 ± 0.01	0.57 ± 0.02	0.43 ± 0.04	0.85 ± 0.02

C.3 PRIMO vs Metalic

Table A11: Performance of a PRIMO and Metalic model on substitution few-shot prediction of held-out DMS assays, using the PRIMO split.

Method	Shots						
	0	4	8	16	32	64	128
Overall							
PRIMO (ICL)	0.43 ± 0.0	0.44 ± 0.0	0.44 ± 0.0	0.44 ± 0.0	0.44 ± 0.0	0.44 ± 0.0	0.45 ± 0.01
PRIMO (TTT)	0.46 ± 0.0	0.46 ± 0.01	0.47 ± 0.02	0.52 ± 0.02	0.56 ± 0.01	0.61 ± 0.01	0.66 ± 0.01
Metalic	0.34 ± 0.02	0.34 ± 0.01	0.38 ± 0.01	0.39 ± 0.01	0.43 ± 0.00	0.47 ± 0.00	0.54 ± 0.00
Stability							
PRIMO (ICL)	0.58 ± 0.0	0.58 ± 0.0	0.59 ± 0.0	0.59 ± 0.0	0.59 ± 0.0	0.59 ± 0.0	0.59 ± 0.0
PRIMO (TTT)	0.55 ± 0.01	0.59 ± 0.02	0.62 ± 0.03	0.67 ± 0.02	0.71 ± 0.02	0.74 ± 0.02	0.79 ± 0.01
Metalic	0.50 ± 0.01	0.54 ± 0.01	0.60 ± 0.01	0.64 ± 0.00	0.70 ± 0.00	0.73 ± 0.00	0.78 ± 0.00
Enzymatic activity							
PRIMO (ICL)	0.36 ± 0.0	0.36 ± 0.0	0.36 ± 0.0	0.36 ± 0.0	0.36 ± 0.0	0.36 ± 0.0	0.36 ± 0.0
PRIMO (TTT)	0.47 ± 0.0	0.39 ± 0.02	0.4 ± 0.02	0.44 ± 0.02	0.48 ± 0.01	0.52 ± 0.02	0.57 ± 0.02
Metalic	0.27 ± 0.02	0.25 ± 0.02	0.25 ± 0.02	0.26 ± 0.02	0.28 ± 0.02	0.29 ± 0.02	0.35 ± 0.01
Fluorescence							
PRIMO (ICL)	0.13 ± 0.0	0.13 ± 0.0	0.13 ± 0.0	0.13 ± 0.0	0.13 ± 0.0	0.13 ± 0.0	0.13 ± 0.0
PRIMO (TTT)	0.14 ± 0.03	0.15 ± 0.05	0.13 ± 0.03	0.15 ± 0.06	0.21 ± 0.02	0.26 ± 0.03	0.32 ± 0.01
Metalic	0.32 ± 0.02	0.30 ± 0.02	0.34 ± 0.01	0.31 ± 0.01	0.36 ± 0.01	0.40 ± 0.01	0.45 ± 0.01
Binding							
PRIMO (ICL)	0.32 ± 0.0	0.33 ± 0.0	0.33 ± 0.0	0.33 ± 0.0	0.33 ± 0.0	0.33 ± 0.0	0.38 ± 0.03
PRIMO (TTT)	0.36 ± 0.0	0.39 ± 0.04	0.39 ± 0.05	0.45 ± 0.03	0.53 ± 0.03	0.59 ± 0.03	0.66 ± 0.03
Metalic	0.28 ± 0.03	0.31 ± 0.03	0.37 ± 0.02	0.40 ± 0.02	0.46 ± 0.01	0.52 ± 0.01	0.60 ± 0.01

C.4 Performance on Substitution and Indel

Table A12: Substitution few-shot prediction of held-out DMS assays.

Method	Shots						
	0	4	8	16	32	64	128
Overall							
GP	-	0.25 ± 0.01	0.32 ± 0.02	0.38 ± 0.02	0.44 ± 0.01	0.5 ± 0.01	0.55 ± 0.0
Ridge	-	0.25 ± 0.02	0.32 ± 0.02	0.4 ± 0.01	0.48 ± 0.01	0.56 ± 0.01	0.63 ± 0.0
RF	-	0.22 ± 0.01	0.31 ± 0.01	0.38 ± 0.01	0.45 ± 0.01	0.52 ± 0.01	0.59 ± 0.0
PRIMO (ICL)	0.5 ± 0.01	0.53 ± 0.0	0.53 ± 0.0	0.53 ± 0.0	0.53 ± 0.0	0.53 ± 0.0	0.53 ± 0.0
PRIMO (TTT)	0.5 ± 0.01	0.48 ± 0.01	0.5 ± 0.01	0.53 ± 0.01	0.58 ± 0.02	0.62 ± 0.0	0.67 ± 0.01
Stability							
GP	-	0.37 ± 0.03	0.47 ± 0.03	0.54 ± 0.02	0.6 ± 0.01	0.66 ± 0.01	0.71 ± 0.0
Ridge	-	0.38 ± 0.03	0.45 ± 0.04	0.55 ± 0.02	0.62 ± 0.01	0.7 ± 0.01	0.76 ± 0.0
RF	-	0.35 ± 0.02	0.45 ± 0.02	0.53 ± 0.01	0.58 ± 0.01	0.65 ± 0.0	0.71 ± 0.0
PRIMO (ICL)	0.6 ± 0.01	0.61 ± 0.0	0.61 ± 0.0	0.61 ± 0.0	0.61 ± 0.0	0.62 ± 0.0	0.62 ± 0.0
PRIMO (TTT)	0.58 ± 0.01	0.58 ± 0.02	0.61 ± 0.02	0.64 ± 0.01	0.68 ± 0.01	0.72 ± 0.01	0.76 ± 0.01
Enzymatic activity							
GP	-	0.11 ± 0.06	0.19 ± 0.04	0.25 ± 0.03	0.31 ± 0.01	0.37 ± 0.02	0.42 ± 0.01
Ridge	-	0.11 ± 0.05	0.21 ± 0.03	0.26 ± 0.02	0.35 ± 0.03	0.44 ± 0.02	0.51 ± 0.01
RF	-	0.07 ± 0.05	0.17 ± 0.02	0.23 ± 0.04	0.3 ± 0.02	0.38 ± 0.02	0.47 ± 0.01
PRIMO (ICL)	0.49 ± 0.05	0.57 ± 0.0	0.57 ± 0.0	0.57 ± 0.0	0.57 ± 0.0	0.57 ± 0.0	0.57 ± 0.0
PRIMO (TTT)	0.53 ± 0.02	0.44 ± 0.05	0.49 ± 0.01	0.51 ± 0.01	0.54 ± 0.03	0.56 ± 0.01	0.61 ± 0.01
Fluorescence							
GP	-	0.01 ± 0.08	0.02 ± 0.09	0.09 ± 0.01	0.14 ± 0.01	0.17 ± 0.02	0.24 ± 0.01
Ridge	-	0.08 ± 0.05	0.08 ± 0.07	0.12 ± 0.05	0.16 ± 0.02	0.21 ± 0.02	0.28 ± 0.02
RF	-	0.1 ± 0.08	0.12 ± 0.04	0.15 ± 0.03	0.21 ± 0.03	0.24 ± 0.01	0.32 ± 0.01
PRIMO (ICL)	0.07 ± 0.0	0.06 ± 0.0	0.06 ± 0.0	0.06 ± 0.0	0.06 ± 0.0	0.06 ± 0.0	0.06 ± 0.0
PRIMO (TTT)	0.11 ± 0.02	0.14 ± 0.06	0.11 ± 0.02	0.15 ± 0.05	0.2 ± 0.04	0.25 ± 0.02	0.3 ± 0.02
Binding							
GP	-	0.18 ± 0.04	0.22 ± 0.05	0.28 ± 0.03	0.34 ± 0.05	0.41 ± 0.02	0.46 ± 0.01
Ridge	-	0.17 ± 0.05	0.24 ± 0.04	0.33 ± 0.04	0.42 ± 0.03	0.51 ± 0.01	0.58 ± 0.01
RF	-	0.15 ± 0.05	0.22 ± 0.02	0.3 ± 0.03	0.38 ± 0.05	0.46 ± 0.01	0.55 ± 0.01
PRIMO (ICL)	0.49 ± 0.0	0.5 ± 0.0	0.5 ± 0.0	0.5 ± 0.0	0.5 ± 0.0	0.5 ± 0.0	0.5 ± 0.0
PRIMO (TTT)	0.47 ± 0.01	0.42 ± 0.05	0.42 ± 0.04	0.48 ± 0.02	0.56 ± 0.04	0.63 ± 0.02	0.69 ± 0.01

Table A13: Indel few-shot prediction of held-out DMS assays.

Method	Shots						
	0	4	8	16	32	64	128
Overall/Stability							
GP	-	0.29 ± 0.09	0.43 ± 0.09	0.5 ± 0.02	0.54 ± 0.03	0.59 ± 0.02	0.64 ± 0.02
Ridge	-	0.42 ± 0.06	0.49 ± 0.07	0.56 ± 0.04	0.59 ± 0.02	0.63 ± 0.02	0.71 ± 0.03
RF	-	0.39 ± 0.05	0.48 ± 0.06	0.55 ± 0.05	0.56 ± 0.03	0.64 ± 0.02	0.69 ± 0.02
PRIMO (ICL)	0.64 ± 0.0	0.65 ± 0.01	0.65 ± 0.01	0.64 ± 0.0	0.64 ± 0.0	0.64 ± 0.0	0.64 ± 0.0
PRIMO (TTT)	0.65 ± 0.01	0.65 ± 0.02	0.66 ± 0.03	0.67 ± 0.01	0.69 ± 0.01	0.71 ± 0.01	0.75 ± 0.02

Table A14: Substititution per-assay results on the hold out set (1/2).

Shots	Method	AMFR_HUMAN_Tsuboyama_2023_4G3O	BLAT_ECOLX_Deng_2012	BLAT_ECOLX_Firnberg_2014	BLAT_ECOLX_Gonzalez_2019	BLAT_ECOLX_Jacquier_2013	BLAT_ECOLX_Stiffler_2015	CASP3_HUMAN_Roychowdhury_2020	CSN4_MOUSE_Tsuboyama_2023_1UFM	D7PM05_CLYGR_Somermeyer_2022	DLG4_HUMAN_Faure_2021	DLG4_RAT_McLaughlin_2012	DNJA1_HUMAN_Tsuboyama_2023_2LO1	DYR_ECOLI_Nguyen_2023	DYR_ECOLI_Thompson_2019	EPHB2_HUMAN_Tsuboyama_2023_1F0M	ESTA_BACSU_Nutschel_2020	FECA_ECOLI_Tsuboyama_2023_2D1U
1	PRIMO (ICL)	0.73 ± 0.03	0.44 ± 0.02	0.54 ± 0.17	-	0.48 ± 0.25	0.67 ± 0.01	0.43 ± 0.21	0.64 ± 0.01	0.06 ± 0.0	0.58 ± 0.0	0.4 ± 0.0	0.84 ± 0.0	0.45 ± 0.06	0.41 ± 0.07	0.87 ± 0.0	0.24 ± 0.01	0.63 ± 0.01
	PRIMO (TTT)	0.7 ± 0.02	0.42 ± 0.01	0.66 ± 0.01	-	0.6 ± 0.01	0.65 ± 0.02	0.47 ± 0.14	0.56 ± 0.06	0.18 ± 0.01	0.53 ± 0.01	0.38 ± 0.0	0.81 ± 0.02	0.44 ± 0.06	0.44 ± 0.01	0.85 ± 0.02	0.27 ± 0.01	0.59 ± 0.04
4	GP	0.42 ± 0.25	0.12 ± 0.17	0.21 ± 0.13	-	0.16 ± 0.11	0.17 ± 0.15	0.09 ± 0.18	0.14 ± 0.31	-0.1 ± 0.05	0.27 ± 0.07	0.14 ± 0.25	0.61 ± 0.25	-0.01 ± 0.13	0.04 ± 0.09	0.7 ± 0.04	0.02 ± 0.06	0.31 ± 0.08
	PRIMO (ICL)	0.77 ± 0.02	0.45 ± 0.0	0.68 ± 0.0	-	0.62 ± 0.0	0.68 ± 0.0	0.59 ± 0.0	0.63 ± 0.02	0.06 ± 0.0	0.57 ± 0.0	0.41 ± 0.0	0.85 ± 0.0	0.5 ± 0.0	0.45 ± 0.0	0.88 ± 0.0	0.24 ± 0.0	0.65 ± 0.0
	PRIMO (TTT)	0.53 ± 0.18	0.36 ± 0.06	0.57 ± 0.05	-	0.48 ± 0.12	0.46 ± 0.1	0.48 ± 0.05	0.48 ± 0.2	0.14 ± 0.04	0.5 ± 0.11	0.39 ± 0.04	0.79 ± 0.1	0.42 ± 0.07	0.33 ± 0.16	0.86 ± 0.03	0.22 ± 0.04	0.6 ± 0.06
	RF	0.18 ± 0.19	0.05 ± 0.07	0.12 ± 0.12	-	0.07 ± 0.13	0.07 ± 0.1	0.1 ± 0.19	0.25 ± 0.3	-0.01 ± 0.15	0.27 ± 0.08	0.05 ± 0.17	0.66 ± 0.17	0.07 ± 0.09	-0.01 ± 0.07	0.67 ± 0.05	0.01 ± 0.08	0.16 ± 0.08
	Ridge	0.18 ± 0.19	0.08 ± 0.14	0.13 ± 0.13	-	0.1 ± 0.18	0.1 ± 0.17	0.21 ± 0.25	0.2 ± 0.35	0.03 ± 0.08	0.25 ± 0.17	0.11 ± 0.24	0.68 ± 0.2	0.09 ± 0.1	0.03 ± 0.09	0.73 ± 0.05	0.03 ± 0.07	0.28 ± 0.07
8	GP	0.56 ± 0.06	0.11 ± 0.13	0.28 ± 0.06	-	0.28 ± 0.07	0.19 ± 0.17	0.19 ± 0.19	0.26 ± 0.27	-0.05 ± 0.07	0.31 ± 0.06	0.17 ± 0.26	0.66 ± 0.14	0.11 ± 0.09	0.19 ± 0.1	0.74 ± 0.05	0.09 ± 0.09	0.38 ± 0.08
	PRIMO (ICL)	0.79 ± 0.0	0.45 ± 0.0	0.68 ± 0.0	-	0.62 ± 0.0	0.68 ± 0.0	0.59 ± 0.0	0.63 ± 0.01	0.06 ± 0.0	0.57 ± 0.0	0.41 ± 0.0	0.85 ± 0.0	0.5 ± 0.0	0.45 ± 0.0	0.88 ± 0.0	0.24 ± 0.0	0.65 ± 0.0
	PRIMO (TTT)	0.66 ± 0.19	0.35 ± 0.05	0.65 ± 0.04	-	0.54 ± 0.05	0.53 ± 0.07	0.48 ± 0.09	0.53 ± 0.38	0.07 ± 0.07	0.51 ± 0.06	0.32 ± 0.12	0.83 ± 0.02	0.46 ± 0.04	0.4 ± 0.04	0.87 ± 0.02	0.25 ± 0.04	0.62 ± 0.1
	RF	0.44 ± 0.11	0.09 ± 0.13	0.27 ± 0.1	-	0.21 ± 0.09	0.19 ± 0.08	0.24 ± 0.18	0.28 ± 0.25	0.03 ± 0.08	0.28 ± 0.06	0.15 ± 0.18	0.66 ± 0.16	0.1 ± 0.11	0.1 ± 0.07	0.75 ± 0.05	0.09 ± 0.05	0.41 ± 0.07
	Ridge	0.38 ± 0.15	0.1 ± 0.11	0.27 ± 0.09	-	0.25 ± 0.11	0.2 ± 0.16	0.27 ± 0.25	0.31 ± 0.41	0.03 ± 0.08	0.29 ± 0.14	0.16 ± 0.21	0.66 ± 0.18	0.17 ± 0.12	0.21 ± 0.08	0.77 ± 0.04	0.06 ± 0.04	0.4 ± 0.09
16	GP	0.65 ± 0.06	0.15 ± 0.11	0.32 ± 0.08	-	0.35 ± 0.02	0.21 ± 0.14	0.31 ± 0.15	0.41 ± 0.19	-0.02 ± 0.03	0.33 ± 0.01	0.3 ± 0.01	0.77 ± 0.02	0.18 ± 0.07	0.2 ± 0.04	0.78 ± 0.04	0.16 ± 0.06	0.4 ± 0.1
	PRIMO (ICL)	0.79 ± 0.0	0.45 ± 0.0	0.68 ± 0.0	-	0.62 ± 0.0	0.68 ± 0.0	0.59 ± 0.0	0.64 ± 0.02	0.06 ± 0.0	0.57 ± 0.0	0.41 ± 0.0	0.85 ± 0.0	0.5 ± 0.0	0.45 ± 0.0	0.88 ± 0.0	0.24 ± 0.0	0.65 ± 0.0
	PRIMO (TTT)	0.74 ± 0.06	0.4 ± 0.04	0.66 ± 0.02	-	0.58 ± 0.04	0.63 ± 0.03	0.54 ± 0.02	0.72 ± 0.05	0.15 ± 0.08	0.55 ± 0.06	0.41 ± 0.06	0.85 ± 0.02	0.41 ± 0.07	0.38 ± 0.07	0.88 ± 0.01	0.26 ± 0.04	0.67 ± 0.05
	RF	0.53 ± 0.09	0.12 ± 0.11	0.34 ± 0.07	-	0.3 ± 0.08	0.22 ± 0.12	0.36 ± 0.06	0.52 ± 0.11	0.05 ± 0.07	0.36 ± 0.05	0.23 ± 0.07	0.77 ± 0.03	0.11 ± 0.08	0.16 ± 0.06	0.78 ± 0.04	0.14 ± 0.06	0.42 ± 0.07
	Ridge	0.58 ± 0.09	0.12 ± 0.1	0.35 ± 0.07	-	0.33 ± 0.05	0.23 ± 0.1	0.39 ± 0.06	0.58 ± 0.11	0.09 ± 0.1	0.38 ± 0.06	0.28 ± 0.02	0.8 ± 0.02	0.2 ± 0.09	0.23 ± 0.05	0.78 ± 0.03	0.15 ± 0.06	0.47 ± 0.09
32	GP	0.72 ± 0.06	0.22 ± 0.04	0.39 ± 0.06	-	0.38 ± 0.04	0.33 ± 0.18	0.31 ± 0.11	0.54 ± 0.11	0.05 ± 0.03	0.36 ± 0.04	0.32 ± 0.01	0.8 ± 0.02	0.23 ± 0.11	0.31 ± 0.03	0.8 ± 0.01	0.17 ± 0.05	0.5 ± 0.14
	PRIMO (ICL)	0.8 ± 0.0	0.45 ± 0.0	0.68 ± 0.0	-	0.62 ± 0.0	0.68 ± 0.0	0.59 ± 0.0	0.64 ± 0.02	0.06 ± 0.0	0.57 ± 0.0	0.41 ± 0.0	0.85 ± 0.0	0.5 ± 0.0	0.45 ± 0.0	0.88 ± 0.0	0.24 ± 0.0	0.65 ± 0.0
	PRIMO (TTT)	0.78 ± 0.03	0.4 ± 0.05	0.69 ± 0.02	-	0.58 ± 0.04	0.67 ± 0.02	0.54 ± 0.03	0.8 ± 0.03	0.19 ± 0.06	0.59 ± 0.04	0.52 ± 0.05	0.86 ± 0.05	0.45 ± 0.06	0.42 ± 0.07	0.89 ± 0.01	0.31 ± 0.04	0.72 ± 0.05
	RF	0.66 ± 0.09	0.23 ± 0.05	0.41 ± 0.06	-	0.36 ± 0.05	0.33 ± 0.09	0.4 ± 0.04	0.62 ± 0.04	0.11 ± 0.06	0.45 ± 0.11	0.32 ± 0.02	0.79 ± 0.03	0.13 ± 0.11	0.21 ± 0.08	0.76 ± 0.06	0.15 ± 0.04	0.49 ± 0.07
	Ridge	0.64 ± 0.1	0.22 ± 0.06	0.45 ± 0.06	-	0.43 ± 0.04	0.37 ± 0.08	0.42 ± 0.05	0.69 ± 0.06	0.1 ± 0.06	0.48 ± 0.1	0.34 ± 0.01	0.8 ± 0.04	0.25 ± 0.13	0.3 ± 0.05	0.82 ± 0.01	0.19 ± 0.04	0.62 ± 0.04
64	GP	0.78 ± 0.03	0.28 ± 0.04	0.42 ± 0.03	-	0.45 ± 0.02	0.38 ± 0.12	0.35 ± 0.08	0.66 ± 0.03	0.11 ± 0.03	0.36 ± 0.02	0.33 ± 0.02	0.81 ± 0.02	0.3 ± 0.08	0.38 ± 0.04	0.83 ± 0.01	0.21 ± 0.04	0.63 ± 0.07
	PRIMO (ICL)	0.8 ± 0.0	0.45 ± 0.0	0.68 ± 0.0	-	0.62 ± 0.0	0.68 ± 0.0	0.59 ± 0.0	0.64 ± 0.01	0.06 ± 0.0	0.57 ± 0.0	0.41 ± 0.0	0.85 ± 0.0	0.5 ± 0.0	0.45 ± 0.0	0.88 ± 0.0	0.24 ± 0.0	0.65 ± 0.0
	PRIMO (TTT)	0.82 ± 0.02	0.44 ± 0.05	0.72 ± 0.01	-	0.61 ± 0.02	0.7 ± 0.01	0.54 ± 0.03	0.85 ± 0.02	0.24 ± 0.05	0.67 ± 0.04	0.57 ± 0.03	0.89 ± 0.01	0.47 ± 0.04	0.47 ± 0.03	0.9 ± 0.01	0.33 ± 0.02	0.79 ± 0.02
	RF	0.74 ± 0.01	0.33 ± 0.05	0.5 ± 0.03	-	0.45 ± 0.04	0.45 ± 0.08	0.45 ± 0.03	0.67 ± 0.02	0.2 ± 0.04	0.5 ± 0.04	0.36 ± 0.03	0.81 ± 0.02	0.18 ± 0.12	0.32 ± 0.05	0.82 ± 0.02	0.21 ± 0.04	0.65 ± 0.03
	Ridge	0.76 ± 0.04	0.33 ± 0.05	0.53 ± 0.02	-	0.52 ± 0.04	0.49 ± 0.07	0.46 ± 0.03	0.78 ± 0.03	0.2 ± 0.02	0.54 ± 0.03	0.37 ± 0.02	0.85 ± 0.02	0.36 ± 0.06	0.39 ± 0.03	0.84 ± 0.01	0.25 ± 0.04	0.72 ± 0.02
128	GP	0.82 ± 0.02	0.32 ± 0.03	0.45 ± 0.04	-	0.49 ± 0.01	0.49 ± 0.06	0.4 ± 0.06	0.72 ± 0.02	0.18 ± 0.02	0.37 ± 0.02	0.34 ± 0.02	0.84 ± 0.01	0.39 ± 0.04	0.4 ± 0.03	0.84 ± 0.01	0.27 ± 0.03	0.75 ± 0.01
	PRIMO (ICL)	0.8 ± 0.0	0.45 ± 0.0	0.68 ± 0.0	-	0.62 ± 0.0	0.68 ± 0.0	0.59 ± 0.0	0.64 ± 0.01	0.06 ± 0.0	0.57 ± 0.0	0.41 ± 0.0	0.85 ± 0.0	0.5 ± 0.0	0.45 ± 0.0	0.88 ± 0.0	0.24 ± 0.0	0.65 ± 0.0
	PRIMO (TTT)	0.84 ± 0.01	0.48 ± 0.05	0.74 ± 0.01	-	0.66 ± 0.02	0.73 ± 0.02	0.6 ± 0.01	0.88 ± 0.01	0.29 ± 0.06	0.71 ± 0.04	0.63 ± 0.01	0.92 ± 0.0	0.54 ± 0.04	0.52 ± 0.02	0.91 ± 0.01	0.42 ± 0.03	0.81 ± 0.0
	RF	0.79 ± 0.02	0.39 ± 0.04	0.59 ± 0.03	-	0.55 ± 0.02	0.59 ± 0.03	0.52 ± 0.02	0.72 ± 0.03	0.28 ± 0.04	0.58 ± 0.01	0.44 ± 0.02	0.85 ± 0.0	0.27 ± 0.05	0.38 ± 0.03	0.85 ± 0.01	0.33 ± 0.03	0.74 ± 0.03
	Ridge	0.83 ± 0.02	0.37 ± 0.04	0.61 ± 0.03	-	0.57 ± 0.03	0.61 ± 0.02	0.5 ± 0.02	0.83 ± 0.02	0.27 ± 0.04	0.58 ± 0.02	0.41 ± 0.02	0.89 ± 0.01	0.44 ± 0.03	0.46 ± 0.03	0.87 ± 0.01	0.32 ± 0.04	0.78 ± 0.03

Table A15: Substitution per-assay results on the hold out set (2/2).

	Shots	GFP_AEQVI_Sarkisyan_2016	GRB2_HUMAN_Faure_2021	MAFG_MOUSE_Tsuboyama_2023_1K1V	NKX31_HUMAN_Tsuboyama_2023_2L9R	PKN1_HUMAN_Tsuboyama_2023_1URF	Q8WTC7_9CNID_Sonermeyer_2022	RCD1_ARATH_Tsuboyama_2023_5OAO	RL40A_YEAST_Mavor_2016	RL40A_YEAST_Roscoe_2013	RL40A_YEAST_Roscoe_2014	SCIN_STAAR_Tsuboyama_2023_2QFF	SPA_STAAU_Tsuboyama_2023_1LP1	SQSTM_MOUSE_Tsuboyama_2023_2RRU	SR43C_ARATH_Tsuboyama_2023_2N88	VKOR1_HUMAN_Chiasson_2020_abundance	VKOR1_HUMAN_Chiasson_2020_activity	VRPI_BPT7_Tsuboyama_2023_2WNM
1	PRIMO (ICL)	0.08 ± 0.0	0.6 ± 0.0	0.17 ± 0.04	0.74 ± 0.0	0.65 ± 0.01	0.05 ± 0.0	0.72 ± 0.0	0.4 ± 0.0	0.45 ± 0.0	0.54 ± 0.0	0.51 ± 0.01	0.2 ± 0.0	0.76 ± 0.01	0.78 ± 0.05	0.45 ± 0.0	0.35 ± 0.01	0.75 ± 0.09
	PRIMO (TTT)	0.08 ± 0.05	0.59 ± 0.01	0.34 ± 0.08	0.64 ± 0.12	0.43 ± 0.02	0.08 ± 0.01	0.64 ± 0.04	0.38 ± 0.01	0.45 ± 0.02	0.47 ± 0.03	0.57 ± 0.02	0.26 ± 0.02	0.72 ± 0.03	0.8 ± 0.02	0.42 ± 0.01	0.34 ± 0.01	0.73 ± 0.02
4	GP	0.08 ± 0.22	0.23 ± 0.12	0.25 ± 0.29	0.39 ± 0.34	0.25 ± 0.06	0.06 ± 0.02	0.5 ± 0.07	0.11 ± 0.13	0.19 ± 0.19	0.13 ± 0.14	0.29 ± 0.15	0.34 ± 0.09	0.57 ± 0.05	0.64 ± 0.15	0.23 ± 0.16	0.06 ± 0.1	0.5 ± 0.26
	PRIMO (ICL)	0.07 ± 0.0	0.6 ± 0.0	0.21 ± 0.01	0.75 ± 0.01	0.65 ± 0.0	0.05 ± 0.0	0.73 ± 0.0	0.4 ± 0.0	0.45 ± 0.0	0.54 ± 0.0	0.52 ± 0.0	0.18 ± 0.01	0.76 ± 0.05	0.81 ± 0.0	0.45 ± 0.0	0.36 ± 0.0	0.79 ± 0.0
	PRIMO (TTT)	0.24 ± 0.14	0.56 ± 0.06	0.47 ± 0.07	0.69 ± 0.04	0.58 ± 0.04	0.06 ± 0.03	0.68 ± 0.02	0.33 ± 0.11	0.39 ± 0.12	0.34 ± 0.13	0.52 ± 0.04	0.29 ± 0.15	0.76 ± 0.02	0.79 ± 0.09	0.43 ± 0.06	0.25 ± 0.05	0.77 ± 0.02
	RF	0.25 ± 0.14	0.17 ± 0.12	0.19 ± 0.23	0.31 ± 0.42	0.24 ± 0.07	0.05 ± 0.04	0.48 ± 0.07	0.15 ± 0.1	0.17 ± 0.22	0.09 ± 0.12	0.35 ± 0.09	0.31 ± 0.1	0.48 ± 0.07	0.65 ± 0.09	0.18 ± 0.09	0.04 ± 0.09	0.63 ± 0.1
	Ridge	0.23 ± 0.12	0.24 ± 0.08	0.23 ± 0.17	0.42 ± 0.24	0.21 ± 0.14	-0.0 ± 0.03	0.5 ± 0.02	0.12 ± 0.17	0.15 ± 0.17	0.16 ± 0.12	0.37 ± 0.2	0.37 ± 0.1	0.5 ± 0.05	0.66 ± 0.08	0.23 ± 0.12	0.06 ± 0.07	0.65 ± 0.13
8	GP	0.08 ± 0.24	0.36 ± 0.04	0.53 ± 0.2	0.56 ± 0.07	0.32 ± 0.06	0.03 ± 0.03	0.56 ± 0.06	0.16 ± 0.12	0.19 ± 0.13	0.12 ± 0.15	0.42 ± 0.18	0.33 ± 0.13	0.6 ± 0.04	0.76 ± 0.07	0.25 ± 0.2	0.03 ± 0.06	0.67 ± 0.08
	PRIMO (ICL)	0.06 ± 0.0	0.6 ± 0.0	0.2 ± 0.01	0.75 ± 0.01	0.65 ± 0.0	0.06 ± 0.0	0.74 ± 0.0	0.4 ± 0.0	0.45 ± 0.0	0.54 ± 0.0	0.53 ± 0.0	0.17 ± 0.0	0.76 ± 0.04	0.82 ± 0.0	0.45 ± 0.0	0.36 ± 0.0	0.79 ± 0.0
	PRIMO (TTT)	0.22 ± 0.08	0.55 ± 0.05	0.61 ± 0.2	0.69 ± 0.06	0.61 ± 0.05	0.04 ± 0.05	0.67 ± 0.04	0.34 ± 0.1	0.43 ± 0.09	0.35 ± 0.12	0.54 ± 0.05	0.33 ± 0.06	0.74 ± 0.02	0.82 ± 0.06	0.38 ± 0.11	0.24 ± 0.08	0.78 ± 0.01
	RF	0.34 ± 0.08	0.33 ± 0.08	0.5 ± 0.1	0.5 ± 0.14	0.31 ± 0.03	-0.02 ± 0.04	0.56 ± 0.04	0.21 ± 0.09	0.21 ± 0.12	0.15 ± 0.11	0.43 ± 0.11	0.3 ± 0.16	0.55 ± 0.04	0.73 ± 0.07	0.17 ± 0.14	0.1 ± 0.07	0.66 ± 0.07
	Ridge	0.19 ± 0.18	0.35 ± 0.07	0.55 ± 0.16	0.56 ± 0.09	0.33 ± 0.18	0.0 ± 0.04	0.54 ± 0.03	0.22 ± 0.14	0.25 ± 0.19	0.17 ± 0.14	0.47 ± 0.13	0.28 ± 0.25	0.46 ± 0.08	0.76 ± 0.08	0.23 ± 0.17	0.04 ± 0.07	0.7 ± 0.05
16	GP	0.22 ± 0.04	0.42 ± 0.03	0.57 ± 0.11	0.66 ± 0.03	0.38 ± 0.06	0.05 ± 0.02	0.62 ± 0.05	0.2 ± 0.09	0.25 ± 0.04	0.22 ± 0.13	0.54 ± 0.08	0.42 ± 0.1	0.64 ± 0.02	0.81 ± 0.02	0.32 ± 0.12	0.08 ± 0.12	0.73 ± 0.04
	PRIMO (ICL)	0.06 ± 0.0	0.6 ± 0.0	0.2 ± 0.01	0.75 ± 0.0	0.65 ± 0.0	0.06 ± 0.0	0.74 ± 0.0	0.4 ± 0.0	0.45 ± 0.0	0.54 ± 0.0	0.53 ± 0.0	0.17 ± 0.0	0.77 ± 0.02	0.82 ± 0.0	0.45 ± 0.0	0.36 ± 0.0	0.79 ± 0.0
	PRIMO (TTT)	0.23 ± 0.07	0.6 ± 0.04	0.62 ± 0.14	0.75 ± 0.03	0.62 ± 0.03	0.05 ± 0.03	0.71 ± 0.03	0.42 ± 0.08	0.46 ± 0.08	0.45 ± 0.11	0.59 ± 0.06	0.34 ± 0.1	0.67 ± 0.11	0.85 ± 0.01	0.43 ± 0.14	0.24 ± 0.06	0.78 ± 0.02
	RF	0.33 ± 0.05	0.39 ± 0.07	0.53 ± 0.15	0.65 ± 0.04	0.38 ± 0.05	0.08 ± 0.04	0.59 ± 0.04	0.26 ± 0.08	0.32 ± 0.06	0.24 ± 0.12	0.52 ± 0.09	0.41 ± 0.12	0.58 ± 0.03	0.8 ± 0.02	0.27 ± 0.07	0.16 ± 0.08	0.72 ± 0.03
	Ridge	0.23 ± 0.09	0.45 ± 0.06	0.6 ± 0.11	0.66 ± 0.03	0.45 ± 0.09	0.04 ± 0.02	0.6 ± 0.05	0.29 ± 0.1	0.29 ± 0.14	0.29 ± 0.13	0.53 ± 0.07	0.36 ± 0.15	0.53 ± 0.05	0.82 ± 0.01	0.35 ± 0.12	0.12 ± 0.12	0.72 ± 0.02
32	GP	0.32 ± 0.04	0.43 ± 0.06	0.65 ± 0.05	0.7 ± 0.04	0.53 ± 0.05	0.06 ± 0.01	0.67 ± 0.03	0.29 ± 0.14	0.38 ± 0.1	0.24 ± 0.13	0.57 ± 0.07	0.46 ± 0.06	0.7 ± 0.05	0.83 ± 0.01	0.4 ± 0.1	0.24 ± 0.04	0.77 ± 0.02
	PRIMO (ICL)	0.06 ± 0.0	0.6 ± 0.0	0.2 ± 0.01	0.75 ± 0.0	0.65 ± 0.0	0.06 ± 0.0	0.74 ± 0.0	0.4 ± 0.0	0.45 ± 0.0	0.54 ± 0.0	0.53 ± 0.0	0.17 ± 0.0	0.77 ± 0.01	0.82 ± 0.0	0.45 ± 0.0	0.36 ± 0.0	0.79 ± 0.0
	PRIMO (TTT)	0.31 ± 0.05	0.62 ± 0.04	0.7 ± 0.1	0.77 ± 0.02	0.65 ± 0.01	0.1 ± 0.04	0.72 ± 0.04	0.49 ± 0.08	0.61 ± 0.05	0.51 ± 0.1	0.59 ± 0.07	0.34 ± 0.09	0.76 ± 0.04	0.87 ± 0.02	0.46 ± 0.08	0.31 ± 0.05	0.79 ± 0.01
	RF	0.41 ± 0.03	0.45 ± 0.08	0.61 ± 0.12	0.68 ± 0.02	0.49 ± 0.07	0.1 ± 0.05	0.61 ± 0.06	0.39 ± 0.07	0.41 ± 0.05	0.28 ± 0.13	0.59 ± 0.01	0.46 ± 0.07	0.66 ± 0.04	0.81 ± 0.02	0.37 ± 0.1	0.24 ± 0.06	0.73 ± 0.03
	Ridge	0.32 ± 0.06	0.5 ± 0.06	0.72 ± 0.07	0.7 ± 0.02	0.6 ± 0.05	0.05 ± 0.05	0.63 ± 0.04	0.38 ± 0.11	0.44 ± 0.04	0.36 ± 0.12	0.56 ± 0.09	0.44 ± 0.11	0.7 ± 0.07	0.84 ± 0.02	0.39 ± 0.11	0.25 ± 0.02	0.72 ± 0.05
64	GP	0.35 ± 0.03	0.5 ± 0.03	0.75 ± 0.03	0.72 ± 0.03	0.58 ± 0.04	0.05 ± 0.04	0.69 ± 0.02	0.41 ± 0.06	0.49 ± 0.08	0.37 ± 0.06	0.63 ± 0.03	0.53 ± 0.03	0.75 ± 0.03	0.85 ± 0.01	0.43 ± 0.04	0.31 ± 0.06	0.8 ± 0.02
	PRIMO (ICL)	0.06 ± 0.0	0.6 ± 0.0	0.21 ± 0.01	0.75 ± 0.0	0.65 ± 0.0	0.06 ± 0.0	0.74 ± 0.0	0.4 ± 0.0	0.45 ± 0.0	0.54 ± 0.0	0.53 ± 0.0	0.17 ± 0.0	0.78 ± 0.0	0.82 ± 0.0	0.45 ± 0.0	0.36 ± 0.0	0.79 ± 0.0
	PRIMO (TTT)	0.35 ± 0.06	0.66 ± 0.03	0.77 ± 0.05	0.79 ± 0.02	0.67 ± 0.03	0.14 ± 0.05	0.74 ± 0.03	0.56 ± 0.05	0.69 ± 0.04	0.61 ± 0.05	0.68 ± 0.03	0.44 ± 0.08	0.81 ± 0.03	0.89 ± 0.01	0.51 ± 0.01	0.35 ± 0.07	0.83 ± 0.01
	RF	0.43 ± 0.03	0.5 ± 0.04	0.76 ± 0.03	0.73 ± 0.03	0.55 ± 0.05	0.1 ± 0.01	0.66 ± 0.02	0.49 ± 0.02	0.53 ± 0.02	0.41 ± 0.06	0.65 ± 0.02	0.51 ± 0.05	0.7 ± 0.05	0.83 ± 0.02	0.43 ± 0.03	0.37 ± 0.04	0.78 ± 0.02
	Ridge	0.37 ± 0.06	0.6 ± 0.02	0.83 ± 0.02	0.75 ± 0.01	0.62 ± 0.03	0.05 ± 0.07	0.71 ± 0.03	0.51 ± 0.04	0.56 ± 0.04	0.49 ± 0.04	0.67 ± 0.03	0.56 ± 0.05	0.78 ± 0.03	0.87 ± 0.01	0.46 ± 0.03	0.37 ± 0.05	0.79 ± 0.01
128	GP	0.42 ± 0.03	0.54 ± 0.03	0.82 ± 0.01	0.75 ± 0.02	0.64 ± 0.03	0.11 ± 0.04	0.75 ± 0.01	0.49 ± 0.03	0.57 ± 0.03	0.43 ± 0.04	0.69 ± 0.01	0.62 ± 0.01	0.81 ± 0.02	0.87 ± 0.0	0.51 ± 0.04	0.37 ± 0.02	0.81 ± 0.02
	PRIMO (ICL)	0.06 ± 0.0	0.6 ± 0.0	0.21 ± 0.0	0.75 ± 0.0	0.65 ± 0.0	0.06 ± 0.0	0.74 ± 0.0	0.4 ± 0.0	0.45 ± 0.0	0.54 ± 0.0	0.53 ± 0.0	0.17 ± 0.0	0.78 ± 0.0	0.82 ± 0.0	0.45 ± 0.0	0.36 ± 0.0	0.79 ± 0.0
	PRIMO (TTT)	0.42 ± 0.03	0.71 ± 0.01	0.78 ± 0.07	0.82 ± 0.02	0.71 ± 0.03	0.2 ± 0.04	0.8 ± 0.02	0.65 ± 0.02	0.75 ± 0.02	0.68 ± 0.02	0.71 ± 0.02	0.56 ± 0.03	0.85 ± 0.01	0.9 ± 0.01	0.56 ± 0.03	0.47 ± 0.04	0.85 ± 0.01
	RF	0.49 ± 0.01	0.59 ± 0.02	0.81 ± 0.02	0.78 ± 0.01	0.61 ± 0.03	0.18 ± 0.03	0.73 ± 0.01	0.58 ± 0.02	0.62 ± 0.03	0.48 ± 0.03	0.71 ± 0.02	0.61 ± 0.02	0.77 ± 0.01	0.86 ± 0.01	0.54 ± 0.03	0.45 ± 0.04	0.81 ± 0.03
	Ridge	0.46 ± 0.04	0.66 ± 0.02	0.87 ± 0.0	0.8 ± 0.02	0.69 ± 0.04	0.13 ± 0.05	0.81 ± 0.02	0.59 ± 0.02	0.67 ± 0.02	0.58 ± 0.04	0.77 ± 0.02	0.71 ± 0.03	0.86 ± 0.02	0.89 ± 0.0	0.57 ± 0.02	0.43 ± 0.04	0.83 ± 0.02

Table A16: Indel per-assay results on the hold out set (1/2).

Shots	Method	AMFR_HUMAN_Tsuboyama_2023_4G3O	BLAT_ECOLX_Deng_2012	BLAT_ECOLX_Firnberg_2014	BLAT_ECOLX_Gonzalez_2019	BLAT_ECOLX_Jacquier_2013	BLAT_ECOLX_Stiffler_2015	CASP3_HUMAN_Roychowdhury_2020	CSN4_MOUSE_Tsuboyama_2023_1UFM	D7PM05_CLYGR_Somermeyer_2022	DLG4_HUMAN_Faure_2021	DLG4_RAT_McLaughlin_2012	DNJA1_HUMAN_Tsuboyama_2023_2LO1	DYR_ECOLI_Nguyen_2023	DYR_ECOLI_Thompson_2019	EPHB2_HUMAN_Tsuboyama_2023_1F0M	ESTA_BACSU_Nutschel_2020	FECA_ECOLI_Tsuboyama_2023_2D1U
1	PRIMO (ICL)	0.69 ± 0.02	-	-	0.56 ± 0.03	-	-	-	0.81 ± 0.0	-	-	-	0.76 ± 0.03	-	-	0.7 ± 0.02	-	0.47 ± 0.01
	PRIMO (TTT)	0.7 ± 0.02	-	-	0.57 ± 0.02	-	-	-	0.82 ± 0.03	-	-	-	0.75 ± 0.03	-	-	0.68 ± 0.06	-	0.46 ± 0.02
4	GP	-0.45 ± 0.21	-	-	0.13 ± 0.51	-	-	-	0.65 ± 0.05	-	-	-	0.65 ± 0.2	-	-	0.49 ± 0.1	-	0.41 ± 0.04
	PRIMO (ICL)	0.71 ± 0.01	-	-	0.61 ± 0.0	-	-	-	0.8 ± 0.0	-	-	-	0.78 ± 0.0	-	-	0.69 ± 0.0	-	0.47 ± 0.01
	PRIMO (TTT)	0.45 ± 0.4	-	-	0.62 ± 0.06	-	-	-	0.81 ± 0.04	-	-	-	0.78 ± 0.01	-	-	0.76 ± 0.03	-	0.4 ± 0.11
	RF	0.24 ± 0.57	-	-	0.34 ± 0.32	-	-	-	0.4 ± 0.53	-	-	-	0.53 ± 0.37	-	-	0.48 ± 0.1	-	-0.02 ± 0.22
	Ridge	0.21 ± 0.51	-	-	0.34 ± 0.26	-	-	-	0.49 ± 0.28	-	-	-	0.57 ± 0.34	-	-	0.61 ± 0.1	-	0.19 ± 0.2
8	GP	-0.31 ± 0.24	-	-	0.34 ± 0.39	-	-	-	0.71 ± 0.03	-	-	-	0.7 ± 0.14	-	-	0.5 ± 0.08	-	0.43 ± 0.04
	PRIMO (ICL)	0.71 ± 0.01	-	-	0.61 ± 0.0	-	-	-	0.79 ± 0.0	-	-	-	0.78 ± 0.0	-	-	0.69 ± 0.0	-	0.46 ± 0.0
	PRIMO (TTT)	0.66 ± 0.09	-	-	0.64 ± 0.03	-	-	-	0.67 ± 0.28	-	-	-	0.77 ± 0.05	-	-	0.73 ± 0.08	-	0.43 ± 0.04
	RF	0.31 ± 0.37	-	-	0.34 ± 0.12	-	-	-	0.37 ± 0.36	-	-	-	0.58 ± 0.26	-	-	0.61 ± 0.06	-	0.21 ± 0.12
	Ridge	0.53 ± 0.08	-	-	0.38 ± 0.16	-	-	-	0.42 ± 0.28	-	-	-	0.53 ± 0.39	-	-	0.62 ± 0.12	-	0.27 ± 0.1
16	GP	-0.27 ± 0.2	-	-	0.57 ± 0.04	-	-	-	0.72 ± 0.04	-	-	-	0.8 ± 0.03	-	-	0.52 ± 0.07	-	0.44 ± 0.04
	PRIMO (ICL)	0.69 ± 0.02	-	-	0.6 ± 0.0	-	-	-	0.79 ± 0.01	-	-	-	0.78 ± 0.0	-	-	0.69 ± 0.0	-	0.46 ± 0.0
	PRIMO (TTT)	0.66 ± 0.05	-	-	0.61 ± 0.06	-	-	-	0.75 ± 0.06	-	-	-	0.83 ± 0.03	-	-	0.7 ± 0.08	-	0.45 ± 0.09
	RF	0.42 ± 0.3	-	-	0.32 ± 0.07	-	-	-	0.51 ± 0.49	-	-	-	0.78 ± 0.06	-	-	0.62 ± 0.03	-	0.27 ± 0.1
	Ridge	0.49 ± 0.17	-	-	0.44 ± 0.11	-	-	-	0.54 ± 0.34	-	-	-	0.76 ± 0.11	-	-	0.57 ± 0.18	-	0.31 ± 0.09
32	GP	-0.22 ± 0.28	-	-	0.62 ± 0.02	-	-	-	0.75 ± 0.03	-	-	-	0.84 ± 0.02	-	-	0.54 ± 0.03	-	0.47 ± 0.05
	PRIMO (ICL)	0.69 ± 0.01	-	-	0.61 ± 0.0	-	-	-	0.79 ± 0.0	-	-	-	0.78 ± 0.0	-	-	0.69 ± 0.0	-	0.46 ± 0.0
	PRIMO (TTT)	0.56 ± 0.13	-	-	0.72 ± 0.04	-	-	-	0.72 ± 0.09	-	-	-	0.84 ± 0.01	-	-	0.69 ± 0.06	-	0.4 ± 0.06
	RF	0.31 ± 0.3	-	-	0.4 ± 0.16	-	-	-	0.48 ± 0.22	-	-	-	0.78 ± 0.04	-	-	0.57 ± 0.07	-	0.32 ± 0.16
	Ridge	0.38 ± 0.14	-	-	0.48 ± 0.09	-	-	-	0.67 ± 0.18	-	-	-	0.77 ± 0.1	-	-	0.65 ± 0.08	-	0.4 ± 0.12
64	GP	-0.09 ± 0.28	-	-	0.64 ± 0.01	-	-	-	0.77 ± 0.02	-	-	-	0.85 ± 0.03	-	-	0.65 ± 0.11	-	0.53 ± 0.07
	PRIMO (ICL)	0.7 ± 0.01	-	-	0.6 ± 0.0	-	-	-	0.78 ± 0.0	-	-	-	0.78 ± 0.0	-	-	0.69 ± 0.0	-	0.46 ± 0.0
	PRIMO (TTT)	0.55 ± 0.16	-	-	0.78 ± 0.03	-	-	-	0.82 ± 0.04	-	-	-	0.84 ± 0.03	-	-	0.71 ± 0.1	-	0.39 ± 0.05
	RF	0.47 ± 0.09	-	-	0.56 ± 0.07	-	-	-	0.57 ± 0.17	-	-	-	0.8 ± 0.04	-	-	0.65 ± 0.09	-	0.45 ± 0.12
	Ridge	0.46 ± 0.13	-	-	0.57 ± 0.1	-	-	-	0.7 ± 0.12	-	-	-	0.79 ± 0.06	-	-	0.66 ± 0.12	-	0.45 ± 0.09
128	GP	0.03 ± 0.22	-	-	0.66 ± 0.01	-	-	-	0.8 ± 0.03	-	-	-	0.87 ± 0.02	-	-	0.72 ± 0.08	-	0.63 ± 0.07
	PRIMO (ICL)	0.69 ± 0.0	-	-	0.61 ± 0.0	-	-	-	0.78 ± 0.0	-	-	-	0.78 ± 0.0	-	-	0.69 ± 0.0	-	0.46 ± 0.0
	PRIMO (TTT)	0.55 ± 0.14	-	-	0.82 ± 0.02	-	-	-	0.84 ± 0.04	-	-	-	0.87 ± 0.02	-	-	0.75 ± 0.06	-	0.51 ± 0.08
	RF	0.5 ± 0.09	-	-	0.67 ± 0.04	-	-	-	0.68 ± 0.11	-	-	-	0.82 ± 0.05	-	-	0.69 ± 0.08	-	0.58 ± 0.11
	Ridge	0.51 ± 0.13	-	-	0.63 ± 0.04	-	-	-	0.79 ± 0.04	-	-	-	0.84 ± 0.04	-	-	0.7 ± 0.1	-	0.58 ± 0.11

Table A17: Indel per-assay results on the hold out set (2/2).

	Shots	GFP_AEQVI_Sarkisyan_2016	GRB2_HUMAN_Faure_2021	MAFG_MOUSE_Tsuboyama_2023_1K1V	NKX31_HUMAN_Tsuboyama_2023_2L9R	PKN1_HUMAN_Tsuboyama_2023_1URF	Q8WTC7_9CNID_Somermeyer_2022	RCD1_ARATH_Tsuboyama_2023_5OAO	RL40A_YEAST_Mavor_2016	RL40A_YEAST_Roscoe_2013	RL40A_YEAST_Roscoe_2014	SCIN_STAAR_Tsuboyama_2023_2QFF	SPA_STAAU_Tsuboyama_2023_1LP1	SQSTM_MOUSE_Tsuboyama_2023_2RRU	SR43C_ARATH_Tsuboyama_2023_2N88	VKOR1_HUMAN_Chiasson_2020_abundance	VKOR1_HUMAN_Chiasson_2020_activity	VRPI_BPT7_Tsuboyama_2023_2WNM
1	PRIMO (ICL)	-	-	0.11 ± 0.03	0.75 ± 0.01	0.84 ± 0.01	-	0.52 ± 0.01	-	-	-	-	-	0.61 ± 0.02	0.66 ± 0.02	-	-	0.85 ± 0.02
	PRIMO (TTT)	-	-	0.01 ± 0.05	0.69 ± 0.01	0.85 ± 0.01	-	0.71 ± 0.03	-	-	-	-	-	0.61 ± 0.03	0.72 ± 0.02	-	-	0.84 ± 0.02
4	GP	-	-	0.18 ± 0.35	0.39 ± 0.57	0.79 ± 0.05	-	-0.29 ± 0.35	-	-	-	-	-	0.32 ± 0.1	0.48 ± 0.27	-	-	0.05 ± 0.53
	PRIMO (ICL)	-	-	0.12 ± 0.01	0.76 ± 0.01	0.83 ± 0.01	-	0.53 ± 0.03	-	-	-	-	-	0.61 ± 0.08	0.66 ± 0.02	-	-	0.87 ± 0.0
	PRIMO (TTT)	-	-	0.23 ± 0.12	0.76 ± 0.07	0.78 ± 0.09	-	0.69 ± 0.04	-	-	-	-	-	0.59 ± 0.05	0.74 ± 0.07	-	-	0.83 ± 0.05
	RF	-	-	0.09 ± 0.36	0.26 ± 0.56	0.6 ± 0.19	-	0.63 ± 0.04	-	-	-	-	-	0.26 ± 0.05	0.53 ± 0.11	-	-	0.75 ± 0.07
	Ridge	-	-	0.17 ± 0.24	0.39 ± 0.51	0.42 ± 0.31	-	0.48 ± 0.32	-	-	-	-	-	0.28 ± 0.06	0.59 ± 0.11	-	-	0.75 ± 0.06
8	GP	-	-	0.42 ± 0.11	0.59 ± 0.37	0.79 ± 0.04	-	0.06 ± 0.28	-	-	-	-	-	0.32 ± 0.09	0.68 ± 0.14	-	-	0.4 ± 0.23
	PRIMO (ICL)	-	-	0.11 ± 0.0	0.77 ± 0.01	0.83 ± 0.0	-	0.52 ± 0.03	-	-	-	-	-	0.61 ± 0.07	0.66 ± 0.01	-	-	0.87 ± 0.0
	PRIMO (TTT)	-	-	0.34 ± 0.1	0.76 ± 0.07	0.82 ± 0.04	-	0.58 ± 0.2	-	-	-	-	-	0.57 ± 0.04	0.75 ± 0.03	-	-	0.81 ± 0.08
	RF	-	-	0.32 ± 0.08	0.52 ± 0.26	0.7 ± 0.08	-	0.62 ± 0.1	-	-	-	-	-	0.28 ± 0.09	0.62 ± 0.12	-	-	0.78 ± 0.07
	Ridge	-	-	0.32 ± 0.11	0.63 ± 0.14	0.49 ± 0.32	-	0.47 ± 0.4	-	-	-	-	-	0.24 ± 0.05	0.69 ± 0.07	-	-	0.8 ± 0.01
16	GP	-	-	0.45 ± 0.09	0.73 ± 0.12	0.74 ± 0.12	-	0.12 ± 0.23	-	-	-	-	-	0.28 ± 0.1	0.73 ± 0.06	-	-	0.62 ± 0.1
	PRIMO (ICL)	-	-	0.12 ± 0.0	0.76 ± 0.0	0.84 ± 0.0	-	0.53 ± 0.03	-	-	-	-	-	0.61 ± 0.03	0.64 ± 0.01	-	-	0.87 ± 0.0
	PRIMO (TTT)	-	-	0.34 ± 0.1	0.8 ± 0.03	0.75 ± 0.13	-	0.68 ± 0.04	-	-	-	-	-	0.53 ± 0.08	0.77 ± 0.03	-	-	0.8 ± 0.07
	RF	-	-	0.36 ± 0.08	0.7 ± 0.11	0.69 ± 0.07	-	0.65 ± 0.08	-	-	-	-	-	0.33 ± 0.04	0.68 ± 0.04	-	-	0.78 ± 0.03
	Ridge	-	-	0.4 ± 0.09	0.71 ± 0.12	0.68 ± 0.14	-	0.54 ± 0.16	-	-	-	-	-	0.27 ± 0.11	0.73 ± 0.07	-	-	0.78 ± 0.03
32	GP	-	-	0.48 ± 0.03	0.78 ± 0.04	0.77 ± 0.07	-	0.26 ± 0.2	-	-	-	-	-	0.3 ± 0.14	0.78 ± 0.03	-	-	0.69 ± 0.02
	PRIMO (ICL)	-	-	0.12 ± 0.01	0.76 ± 0.0	0.84 ± 0.0	-	0.53 ± 0.03	-	-	-	-	-	0.62 ± 0.02	0.63 ± 0.01	-	-	0.87 ± 0.0
	PRIMO (TTT)	-	-	0.41 ± 0.12	0.8 ± 0.04	0.83 ± 0.05	-	0.71 ± 0.1	-	-	-	-	-	0.62 ± 0.06	0.79 ± 0.03	-	-	0.83 ± 0.05
	RF	-	-	0.42 ± 0.09	0.72 ± 0.14	0.76 ± 0.04	-	0.67 ± 0.05	-	-	-	-	-	0.41 ± 0.09	0.7 ± 0.06	-	-	0.8 ± 0.02
	Ridge	-	-	0.44 ± 0.09	0.69 ± 0.17	0.7 ± 0.12	-	0.48 ± 0.32	-	-	-	-	-	0.42 ± 0.1	0.75 ± 0.03	-	-	0.8 ± 0.04
64	GP	-	-	0.55 ± 0.04	0.8 ± 0.02	0.81 ± 0.03	-	0.28 ± 0.17	-	-	-	-	-	0.37 ± 0.13	0.78 ± 0.04	-	-	0.72 ± 0.01
	PRIMO (ICL)	-	-	0.12 ± 0.0	0.76 ± 0.0	0.84 ± 0.0	-	0.53 ± 0.01	-	-	-	-	-	0.63 ± 0.01	0.63 ± 0.01	-	-	0.87 ± 0.0
	PRIMO (TTT)	-	-	0.5 ± 0.13	0.8 ± 0.06	0.85 ± 0.03	-	0.61 ± 0.14	-	-	-	-	-	0.7 ± 0.06	0.78 ± 0.03	-	-	0.85 ± 0.02
	RF	-	-	0.5 ± 0.02	0.79 ± 0.02	0.81 ± 0.03	-	0.66 ± 0.09	-	-	-	-	-	0.53 ± 0.1	0.68 ± 0.09	-	-	0.82 ± 0.03
	Ridge	-	-	0.52 ± 0.09	0.75 ± 0.05	0.78 ± 0.1	-	0.38 ± 0.26	-	-	-	-	-	0.54 ± 0.09	0.77 ± 0.02	-	-	0.8 ± 0.04
128	GP	-	-	0.57 ± 0.05	0.81 ± 0.02	0.84 ± 0.03	-	0.45 ± 0.1	-	-	-	-	-	0.47 ± 0.11	0.79 ± 0.02	-	-	0.74 ± 0.02
	PRIMO (ICL)	-	-	0.12 ± 0.0	0.76 ± 0.0	0.84 ± 0.0	-	0.52 ± 0.01	-	-	-	-	-	0.63 ± 0.01	0.63 ± 0.01	-	-	0.87 ± 0.0
	PRIMO (TTT)	-	-	0.55 ± 0.14	0.82 ± 0.02	0.87 ± 0.02	-	0.79 ± 0.04	-	-	-	-	-	0.73 ± 0.01	0.79 ± 0.05	-	-	0.85 ± 0.02
	RF	-	-	0.54 ± 0.03	0.81 ± 0.02	0.81 ± 0.03	-	0.71 ± 0.04	-	-	-	-	-	0.61 ± 0.09	0.73 ± 0.04	-	-	0.82 ± 0.02
	Ridge	-	-	0.54 ± 0.12	0.8 ± 0.04	0.86 ± 0.02	-	0.64 ± 0.06	-	-	-	-	-	0.62 ± 0.08	0.83 ± 0.02	-	-	0.84 ± 0.02

C.5 Performance by MSA Depth

Table A18: Performance grouped by different levels of MSA depth as defined in ProteinGym.

Depth	Model	0	4	8	16	32	64	128
High	GP	0.51	0.3 ± 0.03	0.38 ± 0.04	0.44 ± 0.03	0.49 ± 0.01	0.54 ± 0.01	0.59 ± 0.01
	Ridge	0.51	0.3 ± 0.02	0.38 ± 0.04	0.45 ± 0.02	0.52 ± 0.02	0.59 ± 0.01	0.65 ± 0.01
	RF	-	0.26 ± 0.03	0.36 ± 0.03	0.43 ± 0.02	0.48 ± 0.01	0.55 ± 0.01	0.62 ± 0.01
	PRIMO (ICL)	0.6 ± 0.03	0.64 ± 0.0	0.64 ± 0.0	0.64 ± 0.0	0.64 ± 0.0	0.64 ± 0.0	0.64 ± 0.0
	PRIMO (TTT)	0.6 ± 0.01	0.57 ± 0.02	0.59 ± 0.02	0.62 ± 0.01	0.65 ± 0.01	0.68 ± 0.01	0.71 ± 0.01
Medium	GP	0.37	0.26 ± 0.03	0.34 ± 0.03	0.41 ± 0.03	0.48 ± 0.02	0.56 ± 0.01	0.62 ± 0.01
	Ridge	0.37	0.27 ± 0.04	0.34 ± 0.03	0.43 ± 0.02	0.51 ± 0.02	0.61 ± 0.01	0.69 ± 0.01
	RF	-	0.24 ± 0.02	0.33 ± 0.03	0.41 ± 0.01	0.48 ± 0.02	0.56 ± 0.01	0.63 ± 0.0
	PRIMO (ICL)	0.52 ± 0.01	0.53 ± 0.0	0.53 ± 0.0	0.53 ± 0.0	0.53 ± 0.0	0.53 ± 0.0	0.53 ± 0.0
	PRIMO (TTT)	0.51 ± 0.01	0.48 ± 0.02	0.5 ± 0.03	0.54 ± 0.02	0.6 ± 0.03	0.65 ± 0.01	0.71 ± 0.0
Low	GP	0.26	0.09 ± 0.05	0.11 ± 0.05	0.18 ± 0.0	0.22 ± 0.01	0.24 ± 0.01	0.29 ± 0.01
	Ridge	0.26	0.12 ± 0.01	0.14 ± 0.04	0.2 ± 0.03	0.26 ± 0.03	0.31 ± 0.01	0.37 ± 0.01
	RF	-	0.12 ± 0.04	0.16 ± 0.03	0.21 ± 0.02	0.28 ± 0.03	0.32 ± 0.01	0.39 ± 0.0
	PRIMO (ICL)	0.24 ± 0.0	0.23 ± 0.0	0.23 ± 0.0	0.23 ± 0.0	0.23 ± 0.0	0.23 ± 0.0	0.23 ± 0.0
	PRIMO (TTT)	0.25 ± 0.01	0.26 ± 0.03	0.23 ± 0.03	0.28 ± 0.03	0.33 ± 0.03	0.39 ± 0.01	0.45 ± 0.01

Software and Data

Code is available at `https://anonymous.4open.science/r/PRIMO-D3F9/README.md`