FLOP: Tasks for Fitness Landscapes Of Protein wildtypes

Peter Mørch Groth[1,2], **Richard Michael**[1],
Jesper Salomon[2], **Pengfei Tian**[2], **Wouter Boomsma**[1]
[1]Department of Computer Science, University of Copenhagen
[2]Bioinformatics & Design, Enzyme Research, Novozymes
{petergroth,richard.michael,wb}@di.ku.dk
{pmg,jrsx,pfi}@novozymes.com

Abstract

Protein engineering has the potential to create optimized protein variants with improved properties and function. An initial step in the protein optimization process typically consists of a search among natural (wildtype) sequences to find the naturally occurring proteins with the most desirable properties. Promising candidates from this initial discovery phase then form the basis of the second step: a more local optimization procedure, exploring the space of variants separated from this candidate by a number of mutations. While considerable progress has been made on evaluating machine learning methods on single protein datasets, benchmarks of data-driven approaches for global fitness landscape exploration are still lacking. In this paper, we have carefully curated a representative benchmark dataset, which reflects industrially relevant scenarios for the initial wildtype discovery phase of protein engineering. We focus on exploration within a protein family, and investigate the downstream predictive power of various protein representation paradigms, i.e., protein language model-based representations, structure-based representations, and evolution-based representations. Our benchmark highlights the importance of coherent split strategies, and how we can be misled into overly optimistic estimates of the state of the field. The codebase and data can be accessed via `https://github.com/petergroth/FLOP`.

1 Introduction

The goal of protein engineering is to optimize proteins towards a particular trait of interest. This has applications both for industrial purposes and drug design. There is clear potential for machine learning to aid in this process. By predicting which protein sequences are most promising for experimental characterization, we can accelerate the exploration of the "fitness landscape" of the protein in question [1]. Regression of functional landscapes is challenging for multiple reasons. Typically a data-scarce problem, careful considerations of the experimental setup are required to avoid inadvertent data leakage. Concerning the functional landscapes of naturally occurring (also known as *wildtype*) proteins, the pairwise amino acid sequence identities can often vary significantly with some proteins differing by only a single amino acid while others might be less than ten percent similar. High-throughput experimental techniques are improving the data scarcity issue, while underlying structure typically exists in the datasets allowing for supervised learning despite the intrinsic challenges.

Submitted to the 37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks. Do not distribute.

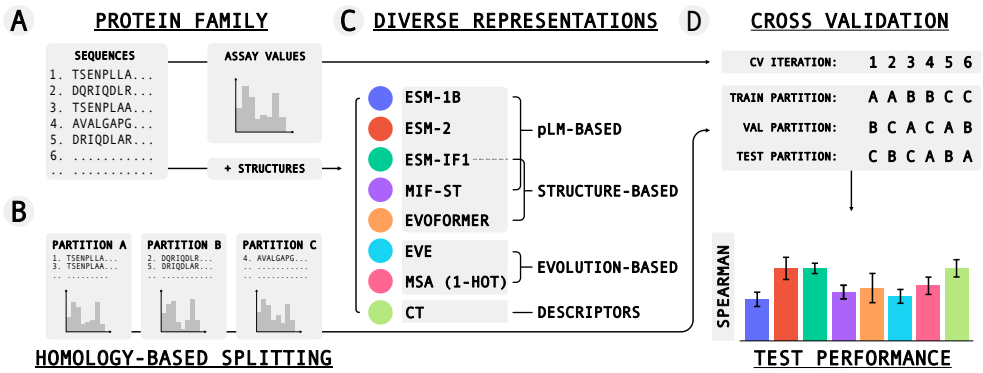

Figure 1: Schematic over dataset splitting, representations, and cross-validation process. A: A dataset with sequences from a single protein family and corresponding assay values is curated. B: A stratified sequence identity splitting procedure generates partitions A, B, and C, which are (1) homologically different from each other, (2) contain similar number of sequences, and (3) match the full dataset's target distribution. C: Eight types of protein representations are computed. D: Cross-validation using a random forest regressor is applied to obtain mean values and standard errors on the test partitions.

The optimization process of proteins and enzymes can typically be divided into multiple stages. An often employed initial step is to search for promising candidates among wildtype proteins, resulting in a set of proteins with desirable properties. We will refer to this as the *wildtype discovery* phase [2]. Since we are typically optimizing for a specific trait, we may often limit this initial exploration to a particular protein family, where the members share an evolutionary history which has resulted in a similar function. The selected set of wildtype proteins will then form the basis for a second phase in the engineering process: localized optimization, where novel variants of the wildtype proteins are examined through various assays [3]. Sometimes, the wildtype discovery phase is not only carried out once as several rounds might be required before an initial suitable candidate is found. Additionally, the resulting candidate might prove insufficient at a later stage of protein engineering, where conditions such as temperature are altered or where stress-factors are introduced.

In recent years, we have seen considerable efforts in defining benchmarks to help the machine learning community make progress in this field. However, these efforts have primarily focused on the second stage, i.e., variant effect prediction, where a dataset consist of thousands of variants from a single wildtype. In this paper, we argue for the importance of establishing well-defined benchmark tasks for the first stage as well. We present three challenging tasks and a careful analysis of the experimental design, demonstrating how poor choices can lead to dramatic overestimation of performance.

We conduct our experiments using a variety of fixed-size protein representations: sequence-based embeddings obtained through protein language models, structure-based representations from folding and inverse folding models, evolution-based representations obtained from multiple sequence alignments, as well as simple biologically-motivated sequence descriptors. In addition to the supervised approach, we include four zero-shot predictors to showcase a simpler approach to the task of identifying promising candidates. We show that the choice of representation can greatly affect the downstream predictive performance, and we therefore argue that more progress can be made by constructing meaningful representations and not solely in the construction of complex prediction models. Given the oftentimes limited dataset sizes, we therefore rely on a random forest regressor.

2 Related work

Benchmarks play an important role in driving progress in protein-related prediction tasks. The most well-known is perhaps the rolling CASP benchmark, which is arguably responsible for the recent breakthroughs in protein structure prediction [4–6]. For the prediction of protein stability and function, several studies have curated relevant experimental datasets for use as benchmarks. The TAPE benchmark was an early such example designed to test protein sequence representations on a

set of diverse downstream tasks [7]. Two of these tasks were related to protein engineering: stability prediction on variants of a set of 12 designed proteins [8] and characterization of the functional landscape of green fluorescent protein [9]. The PEER benchmark [10] expanded on the TAPE benchmark with many additional tasks. This included prediction of β-lactamase activity [11], and a binary solubility classification task on a diverse set of proteins. Focusing entirely on variant effects, the recent ProteinGym benchmark has assembled a large set of Deep Mutational Scanning (DMS) assays and made them available as substitution and insertion-deletion prediction tasks [12]. While the above all consider protein sequence inputs, the recent Atom3D benchmark [13] presents various prediction tasks using 3D structure as input, including predicting amino acid identity from structural environments (for general proteins), and mutation effects on protein binding, using data originating from the SKEMPI database [14, 15].

Most closely related to this current paper is the FLIP benchmark, which dedicates itself to the prediction of functional fitness landscapes of proteins for protein engineering [16]. FLIP introduces three tasks: one on the prediction of protein stability of wildtype proteins (distributed over many families) using data from the Meltome Atlas [17], and two tasks focused on mutations at specific sites of proteins GB1 [18] and AAV [19]. While the FLIP benchmark is of great value for protein engineering, there are key characteristics which make it unsuitable for wildtype discovery, e.g., the use of the Meltome Atlas, which consists of thousands of sequences from different organisms spanning many different protein families. The sequences in the GB1 dataset only have mutations at four fixed positions while the sequences in the AAV dataset only contain 39 mutation sites, both of which corresponds to mutations at less than 10% of the full-length proteins. Such datasets with very local fitness landscapes are not generalizable enough for wildtype discovery.

Most functional tasks in current benchmarks are thus concerned with protein sequences that are derived from a single wildtype sequence by one or more mutations. Characterizing the functional effects of such variants is critical for protein engineering. However, before engaging in the optimization process itself, it is important to select meaningful starting points. As a natural complement to the FLIP benchmark, we therefore present a novel benchmark titled FLOP. The tasks we present are the characterization of functional landscapes of wildtype proteins.

Our curated datasets all consist of functionally characterized wildtype sequences. For each dataset, we limit ourselves to a single family, and define our tasks as regression problems on the functional assay values. While mutational fitness landscape datasets are relatively abundant, few published datasets exist where the global fitness landscapes of wildtype proteins from single families are examined. This imposes limitations in the number and sizes of available datasets which are suitable for our considered problem. Given the low-data regime, the focus of our benchmark is thus to find representations of the protein input that makes few-shot or even zero-shot learning feasible. As a point of departure, we provide a set of state-of-the-art embeddings, reflecting different protein modalities.

3 Experimental setup

The domain we explore in this work is characterized by data scarcity, requiring special care in the design of the experimental setup. Figure 1 shows an overall schematic of the benchmarking process.

3.1 Dataset splitting

With the proliferation of large datasets and computationally demanding models, a common learning paradigm in machine learning is to rely on hold-out validation, whereby fixed training, validation, and testing sets are randomly generated. This method has several serious limitations when applied to biological datasets of limited sizes. Firstly, randomly splitting a dataset assumes that the data points are independent and identically distributed (i.i.d.). This is however not the case for members of a protein family which share common ancestors, leading to potential data leakage if protein sequences that are close in evolutionary space are placed in separate splits. Secondly, when splitting small datasets for a hold-out validation approach for supervised learning, the target values might not be well-balanced, resulting in dissimilar target distributions thus leading to bias and poor generalizability.

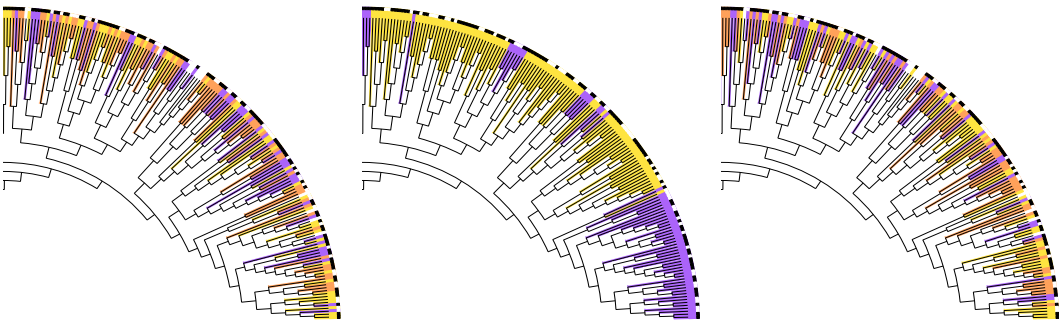

Figure 2: The same segment of a phylogenetic tree for the PPAT dataset. Branch color corresponds to its CV partition, while the outermost ring shows the target values (black indicates high and white indicates low values). The segments highlight the diversity found in wildtype protein families. Left: entries are colored according to the prescribed dataset splitting procedure which allows learning across subfamilies (indicated by the mix of colors). Middle: entries are colored by a clustering approach leading to wide regions, inhibiting learning across subfamilies. Right: entries are randomly assigned a color. While similar to the leftmost scheme, the random coloring allows near identical sequences to be placed in separate partitions leading to excessive data-leakage.

To handle these potential issues, we rely on a sequence identity-based, stratified cross-validation procedure ensuring that (1) partitions are generated such that any two proteins occurring in different partitions are guaranteed to be different at a pre-set homology cut-off, (2) cross-validation (CV) minimizes the potential bias which might occur during hold-out validation, (3) the target distribution is reflected by the generated partitions via stratification on discretized target values, and (4) the number of sequences in each partition is similar to reduce the variance.

To generate these high-quality data partitions, we use the four-phase procedure described in [20] and implemented in the GraphPart framework [21] to create three label-balanced partitions for each dataset. We begin the procedure from an initial sequence identity threshold, and increase the threshold until the generated partitions are of sufficient sizes (i.e., at least 25 % of sequences in all three partitions). The stratification is achieved by creating a binary label which indicates whether a protein has low or high target value, e.g., by fitting a two-component Gaussian mixture model. For the dataset-specific stratification boundaries, see Section A in the supplementary materials.

Figure 2 shows the same segment of a phylogenetic tree of the curated PPAT dataset, showing the evolutionary relationship between sequences. Large versions of the trees can be found in Section E. The colors indicate which CV partition each sequence belongs to while the black and white squares in the outer ring indicate the stratification labels. The segments show the diversity encountered in wildtype protein families. The left segment is colored by our splitting procedure and shows that it manages to create diverse partitions spanning the entire evolutionary tree to allow learning across protein subfamilies. The middle segment is colored by an MMseqs [22] clustering approach, leading to contiguous areas inhibiting learning across subfamilies. The entries in the rightmost segment are randomly assigned a color, corresponding to random splitting. While similar to the leftmost scheme, the random coloring allows near-identical sequences to be placed in separate partitions leading to excessive data-leakage.

3.2 Representations

To accurately reflect the current paradigms of state-of-the-art protein representations, we choose representatives from three main categories, the dimensionalities of which can be found in Section G in the supplementary materials.

Protein language models (pLMs) that are trained on hundreds of millions of protein sequences in an unsupervised fashion have been proven to be competitive for a multitude of tasks including supervised prediction of protein properties, residue contact prediction, variant effect prediction [16, 23–26], etc. We here choose the popular ESM-1B [24] and the more recent ESM-2 models [27]. To fix the

dimensionality for proteins of different lengths, we perform mean-pooling over the residue dimension. This operation is likely to filter out information encoded along the protein sequence and more optimal approaches will likely yield more informative representations and thus higher predictive performance (see Table 2 in [28]). Constructing fixed-size embeddings from sequences of variable lengths is however nontrivial and considered out of the scope of this study.

The second category we include is structure-based. We extract embeddings from the Evoformer-modules while folding proteins with AlphaFold2 [29] via ColabFold [30], which have been shown to perform well for structure-related prediction tasks [31]. Using the predicted structures, we then extract embeddings from the inverse-folding model ESM-IF1 (also known as the GVP-GNN) [32] which incorporates a pLM and graph neural network architecture. We similarly use embeddings from the MIF-ST model, which is an inverse folding model leveraging a pretrained convolutional pLM [33]. As with the pLMs, we apply mean pooling to achieve sequence-level embeddings.

The third category is evolution-based. As a baseline, we will use a one-hot encoded multiple sequence alignment (MSA) over the proteins of interest [34–36]. Since the MSA is independent of labels, we enrich the unaligned sequence pools with additional members from the respective protein families using UniProt [37] and InterPro [38]. Given MSAs, models can be designed which leverage the evolutionary history of the protein family (e.g., EVE and related models [12, 39–42]). For each curated dataset, we train EVE [39] on the corresponding protein family and extract the latent representations. Technical details on the training procedure can be found in Section H.

In addition to these groups of advanced representations, we include compositional and transitional (CT) physicochemical descriptors for each protein sequence as a simple baseline, which relate to overall polarizability, charge, hydrophobicity, polarity, secondary structure, solvent accessibility, and van der Waals volume of each sequence as predicted using the PyBioMed library [43].

With the exception of the physicochemical descriptors, all included representations rely on models which have been pretrained on thousands to hundreds of millions of proteins. While it is possible that a number of the sequences in the curated datasets also belong to the training sets of these models (which by design is the case for the evolution-based approaches), we do not consider this to be a fatal form of data leakage as it purely pertains to the un- or self-supervised pretraining phases and is independent of the sequence labels.

3.3 Regression

The purpose of this benchmark is to provide a structured procedure to evaluate the predictive performance on downstream regression tasks given protein representations. We believe that larger prediction improvements can be achieved by focusing on developing novel protein representations rather than more complex regression models. Due to the low-N setting in which we operate, the training of large, complex models is practically inhibited, which is why we have chosen to rely on a random forest regressor. For each combination of the generated CV partitions, we perform a hyperparameter optimization on the current validation partition and evaluate the best-performing predictor on the current test partition. The experiments were also carried out using alternate regressors. See Sections N.1 and K for these results and all hyperparameter grids, respectively.

3.4 Zero-shot predictors

To investigate the efficacy of unsupervised learning on the curated datasets, we evaluate four zero-shot predictors. Using EVE, we evaluate the evidence lower bound (ELBO) by sampling and obtain a proxy for sequence fitness, analogous to the evolutionary index in [39]. Second and third proxies are obtained by evaluating the log-likelihood of a sequence conditioned on its structure using the inverse folding models ESM-IF1 [32] and ProteinMPNN [44]. The fourth zero-shot estimator is obtained by using Tranception [45] to evaluate the log-likelihood of each sequence. Details for the use of ProteinMPNN and Tranception can be found in Sections I and J in the supplementary materials.

Table 1: Summary of datasets and splits.

	$\mathbf{N}_{tot}$	$\mathbf{N}_A$	$\mathbf{N}_B$	$\mathbf{N}_C$	**Split %ID**	**Target**	**Median %ID**	**Avg. length**
GH114	55	20	18	17	0.55	Activity	0.46	268.8
CM	855	341	259	255	0.40	Activity	0.40	91.1
PPAT	615	182	234	199	0.55	Fitness	0.51	161.6

4 Datasets

The three curated datasets and the corresponding fitness landscapes are here motivated and described. Despite the scarcity of available datasets described in Section 2, the curated datasets are representative examples of wildtype discovery campaigns in terms of size and diversity. For additional curation details on each dataset including specific thresholds for stratified splitting, see Section A.

4.1 GH114

Motivation. Accurately identifying enzymes with the highest activities towards a specific substrate is of central importance during enzyme engineering. To achieve this, it is essential to ensure that assay observations are directly comparable [46]. This includes maintaining identical experimental assay conditions, including evaluating enzymes at the same concentrations and purity levels. However, purifying enzymes requires significant work and resources, often resulting in assays composed of fewer sequences, which are in turn of higher experimental quality.

Landscape. This dataset includes purified and concentration normalized natural glycoside hydrolase 114 (GH114) alpha-1,4-polygalactosaminidase enzymes and corresponding catalytic activity values [47] which will act as the target of interest. GH114 enzymes degrade the exopolysaccharide PEL, which provides structure and protection in some biofilms [48]. Having measurements of purified enzymes avoids issues with background effects from other enzymes in the recombinant host background. We provide a curated version of the GH114 dataset which, to our knowledge, has not been used in previous work for function prediction purposes.

4.2 CM

Motivation. Identification of enzymes with high catalytic activities is essential for enzyme engineering campaigns. However, predicting the activity level of enzymes using physics-based methods remains a great challenge [49]. Recent progress in high throughput screening allows the measurement of enzyme activity of sequences with high diversity, but with low experimental cost.

Landscape. This dataset contains the catalytic activity of chorismate mutase (CM) homologous proteins, as well as artificial sequences which follow the same pattern of variations (e.g., conservation and co-evolution) [50]. The artificial sequences generated by Monte Carlo simulations at low and medium temperatures match the empirical first-, second-, and higher-order statistics of the natural homologs, while also exhibiting comparable catalytic levels when experimentally synthesized. These sequence have therefore been included given the similarity in both sequence and fitness landscape. See Section A.3 for further details. We perform an additional filtering of the dataset prior to the splitting procedure by removing sequences with target values less than 0.42, corresponding to inactive proteins [50]. This task thereby assumes that a preceding classification procedure has been carried out. For completeness, we include benchmark results for the CM dataset when only the natural homologs were used (see Section N.4) and classification results before the filtering step (see Section M), which supports this last assumption.

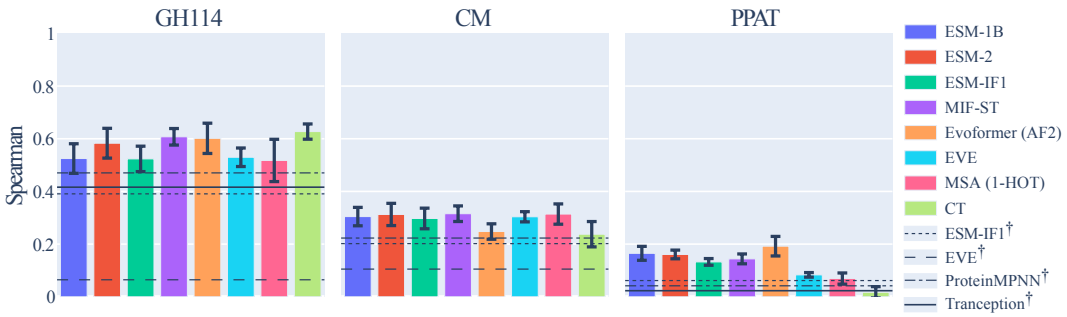

Figure 3: Average Spearman's rank correlation (and standard error) between predictions and targets over test partitions. Higher is better. †: Zero-shot correlations.

4.3 PPAT

Motivation. PPAT (phosphopantetheine adenylyltransferase) is an essential enzyme that catalyzes the second-to-last step in the CoA biosynthetic pathway. The target value for this prediction task is the fitness score, which reflects the ability of PPAT homologs to complement a knockout E. coli strain. The fitness of homologs can be affected by factors such as protein misfolding, mismatched metabolic flux, or environmental mismatches etc. [51].

Landscape. This dataset contains fitness scores of 615 different PPAT homologs obtained by a novel DNA synthesis/assembly technology, DropSynth, followed by a multiplexed functional assay to measure how well each PPAT homolog can rescue a knockout phenotype [51].

4.4 Summary

A summary of the curated datasets including their total sizes, partition sizes, between-partition sequence identity threshold (Split %ID), regression target, median pairwise sequence identity, and average sequence length can be seen in Table 1. Additional curation details are found in the supplementary materials (see Section A), while histograms of the regression targets for the three datasets as well as partition histograms can be seen in Sections C and D, respectively.

The low median sequence identity observed in Table 1 highlights the diversity – and difficulty – of wildtype datasets. Deep mutational scanning (DMS) datasets commonly used for variant effect prediction on average have median sequence identities greater than 0.99. For comparison, Section F shows both median, mean, and standard deviation for the curated datasets (see Table A1) contrasted to 48 tasks from the ProteinGym benchmark (see Table A2). A visual example highlighting the sequence diversity for a protein family compared to a DMS dataset can be seen in Figure 4 in [40]. The DMS data is localized to a small section of the protein family as it is composed of all single mutations of a single wildtype. Similarly, a DMS of a protein belonging to the PPAT family would all be positioned on a single branch of the phylogenetic tree in Figure 2.

5 Results

Spearman's rank correlation between the predictions and targets over the three datasets can be seen in Figure 3 and Table 2. The highest performing proteins are of interest making it the ranking and not the absolute predictions that indicate the performance, despite the regressors being optimized using the mean squared error. The RMSE can found in Table A4 in the supplementary materials.

GH114 The performance on the GH114 dataset highlights some of the peculiarities encountered with small wildtype protein datasets. The collection of physicochemical descriptors (CT), which is simpler than its competitors, achieves the highest score. Slightly below it are the structure-informed MIF-ST and Evoformer representations, indicating a structural signal which is however not picked

Table 2: Benchmark results. Mean Spearman correlation and standard error using cross-validation. †: Zero-shot correlation on full datasets. Highest value and values within 1 SE are bold.

	GH114	**CM**	**PPAT**
ESM-1B	0.52 ± 0.06	$\mathbf{0.30} \pm 0.03$	$\mathbf{0.16} \pm 0.03$
ESM-2	$\mathbf{0.58} \pm 0.06$	$\mathbf{0.31} \pm 0.04$	$\mathbf{0.16} \pm 0.02$
ESM-IF1	0.52 ± 0.05	$\mathbf{0.30} \pm 0.04$	0.13 ± 0.01
MIF-ST	$\mathbf{0.61} \pm 0.03$	$\mathbf{0.32} \pm 0.03$	0.14 ± 0.02
Evoformer (AF2)	$\mathbf{0.60} \pm 0.06$	0.25 ± 0.03	$\mathbf{0.19} \pm 0.04$
EVE	0.53 ± 0.04	$\mathbf{0.30} \pm 0.02$	0.08 ± 0.01
MSA (1-HOT)	0.52 ± 0.08	$\mathbf{0.31} \pm 0.04$	0.07 ± 0.02
CT	$\mathbf{0.63} \pm 0.03$	0.24 ± 0.05	0.02 ± 0.02
ESM-IF1†	0.39	0.20	0.06
EVE†	0.06	0.11	−0.01
ProteinMPNN†	0.47	0.22	0.04
Tranception†	0.42	−0.05	0.02

up by the ESM-IF1 embeddings. While the CT representation achieves the highest mean value, several others are within one standard error, giving no clear advantage to neither complex nor simple models. While all supervised approaches beat the zero-shot predictors, ProteinMPNN, Tranception and ESM-IF1 likelihoods correlate well with the targets.

CM The second prediction task can be considered more challenging given the results, despite the comparatively large size of the CM dataset. While similar to the first task, an abundance of data is not sufficient to increase the downstream capabilities if it comes at the cost of potentially noisier measurements, as compared to the concentration normalized GH114 dataset. Most representations fall within one standard error of the top performer such that, once again, no representation paradigm has a clear advantage.

PPAT The most challenging task of the datasets, the results on the PPAT task show different behaviour. The evolutionary signal, i.e., the amount of information which can be learned from evolutionary homologs, is weak as indicated by the low correlations from the one-hot encoded MSA and from EVE (both in the supervised and zero-shot settings as per Table 2). Furthermore, the physicochemical descriptors fail to correlate – as do the remanining zero-shot predictors. The pLM and structure-based representations achieve the highest scores with the Evoformer embeddings coming out slightly ahead.

6 Ablation study

The dataset splitting procedure and benchmark tasks have been carefully constructed to ensure reliable estimates of model performance. In this section we show three ablation studies – one for each dataset – whereby different choices of task-structuring might lead to great over-estimations of performance. The results can be seen in Table 3 and in Figure A8. The Δ columns in the table indicate differences to the benchmark results, where a positive/green value indicates *better* performance during ablation, i.e., over-estimation.

Hold-out validation For GH114, we perform hold-out validation by arbitrarily designating the three generated partitions as training, validation, and test sets and running the experiment only once. The correlations are significantly different to the benchmark results, with the ESM-2 correlation decreasing by 50 %. With no systematic pattern and decreased nuance given the lack of errorbars, it is easy to draw incorrect conclusions.

As the data-leakage between partitions has been controlled via the splitting procedure, the partitions are different from each other up to the sequence identity threshold. This implies that a model might

Table 3: Ablation results. Spearman correlation. *: Hold-out validation, **: Regression on both active and inactive proteins, ***: Repeated random splitting. Δ shows difference to benchmark results. Highest value and values within 1 SE are bold.

	GH114*	Δ	**CM****	Δ	**PPAT*****	Δ
ESM-1B	0.36	-0.16	0.64 ± 0.01	$+0.33$	**0.23** ± 0.05	$+0.07$
ESM-2	0.39	-0.20	**0.66** ± 0.01	$+0.35$	0.18 ± 0.02	$+0.02$
ESM-IF1	0.46	-0.06	0.58 ± 0.01	$+0.28$	0.19 ± 0.01	$+0.06$
MIF-ST	0.62	$+0.01$	0.60 ± 0.02	$+0.28$	**0.24** ± 0.04	$+0.09$
Evoformer (AF2)	0.64	$+0.04$	0.57 ± 0.01	$+0.33$	**0.24** ± 0.02	$+0.04$
EVE	0.38 ± 0.04	-0.15	0.62 ± 0.00	$+0.31$	0.16 ± 0.01	$+0.08$
MSA (1-HOT)	0.50	-0.02	0.61 ± 0.01	$+0.30$	0.06 ± 0.06	-0.01
CT	**0.65**	$+0.03$	0.52 ± 0.01	$+0.28$	0.13 ± 0.01	$+0.11$

perform well on, e.g., only a subset of the partitions. Choosing which partitions to use for training, validation, and testing is (in this case) arbitrary and can thereby lead to misleading results. To avoid this pitfall, cross-validation is needed such that the average predictive performance on all combinations of partitions can be estimated. An analogue ablation study for the CM and PPAT datasets can be found in Section L.1, where similar conclusions can be drawn.

Disregarding distinct target modalities For the CM dataset, we only included the active sequences in the benchmark. To demonstrate why, we have included the results of performing regression on both active *and* inactive sequences in the center of Table 3. These results are greatly overinflated compared to the benchmark results, with some representations more than doubling the correlation scores.

Regression performed on a dataset with a distinctly bimodal target distribution (such as the full CM dataset, see Figure C in the supplementary materials) can inflate the results significantly. The regressor is able to distinguish between the two target modalities, i.e., between the inactive cluster around 0 and the active cluster around 1, driving the ranking correlation to overly-optimistic values. The caveat to this preprocessing step is that it requires knowing the whether the proteins are active or not a priori, which assumes that a preceding classification-screening has been performed. The classification results of such a process can be seen in Section M.

Random partitioning for cross-validation To illustrate why random splitting of wildtype protein datasets is ill-advised, we applied repeated random splitting to the PPAT dataset. This was done by randomly assigning sequences to training, validation, and testing partitions without any consideration of sequence similarity. Given the randomized partitions, the predictive performance using the selected representations was evaluated using cross validation. This was repeated a total of three times with different seeds. While the results look similar to the benchmark results, we do see an increase in performance across the board.

With random sampling, we risk placing very similar sequences in separate partitions, thereby allowing extensive data-leakage, where we are essentially testing on training/validation data, thus overestimating the predictive performance [52]. The results for this ablation study carried out on the GH114 and CM datasets can be found in Section L.2, where we can once again draw similar conclusions.

7 Discussion

The choice of representation greatly affects the downstream predictive capabilities, with no consistent, clear edge given by any of the three representation paradigms. For CM, a one-hot encoded MSA acts as an impressive baseline proving difficult to convincingly beat. For GH114, physicochemical descriptors are sufficient to achieve top performance, while the PPAT dataset benefits from the complex, structure-informed Evoformer embeddings. While the specific top-scoring representation fluctuates, the ESM-2 embeddings are consistently within one standard error and can thus be considered a relatively

consistent baseline for future experiments, where others occasionally underperform. For the three tasks, we see supervised learning outperforming zero-shot predictions, while the inverse-folding estimators however offer decent zero-shot approaches for two of three tasks.

Despite similar overall patterns, some results stand out, e.g., the comparatively high performance on the GH114 dataset and the low performance on the PPAT dataset. Variations in experimental conditions and techniques can introduce different levels of noise. The CM and PPAT datasets are derived from tests on supernatants with complex backgrounds with potential side-activities from impurities, whereas the GH114 dataset uses purified samples with less expected noise. This can be a potential reason for the comparatively high performance of the latter. As for the low performance on the PPAT dataset, the reason might lie in the target values: the GH114 and CM datasets both measure enzymatic activities while the PPAT dataset measures fitness. The overall performance disparities suggest that enzyme activities, rather than a more complex and assay-specific fitness value, are easier to model given the available protein representation paradigms. The stark contrast in performance between the concentration normalized GH114 dataset and both the CM and PPAT datasets indicates that higher quality datasets are of central importance to learn accurate fitness landscapes – more so than the number of labelled sequences.

8 Conclusion

In this work we have presented a novel benchmark which investigates an unexplored domain of machine learning-driven protein engineering: the navigation of global fitness landscapes for single protein families. Wildtype exploration can be viewed as a predominantly explorative phase of protein optimization, which precedes the exploitation phase comprised of the subsequent protein engineering. Often, limited resources are allocated to wildtype exploration since it is inherently costly and considered wasteful as it tends to produce many poor candidates. This is unlikely to change unless we find ways to improve our wildtype search strategy, which will require better predictions. We therefore consider the limited dataset sizes as an inherent condition and limitation in this domain. This makes the collection and curation of relevant labelled datasets challenging and also necessitates the design of careful learning schemes and model evaluation to ensure reliable estimates of generalizability while avoiding inadvertently overestimating the results. We anticipate that the creation of this new set of comprehensive family-wide datasets will facilitate and improve future model development and applicability in this domain.

Given the limited dataset sizes, our focus has been on transfer learning and zero-shot prediction. Our results show that the supervised approaches outperform the zero-shot approaches, but that no one representation or representation paradigm consistently outperforms the others. This could suggest that the employed representations are not sufficiently informative. A key limitation for a number of the included representations is that we obtained protein-level representations as averages over the protein length to arrive at fixed-length embeddings, which is known to be suboptimal [28]. We encourage the community to experiment with novel aggregation strategies and new representation designs to improve performance on our benchmark. It is also conceivable that general-purpose protein representation models might not by themselves be sufficient to convincingly improve on the proposed tasks. One can imagine that further improvements can be obtained using pretrained models fine-tuned on a protein family of interest – or by developing weakly-supervised representation models incorporating relevant properties that correlate with the function of interest (e.g., thermostability).

Although the performance of current baselines on some of our test-cases is fairly low in absolute terms, even low correlations can provide useful guidance on selecting wildtype protein starting points and can have measurable real-world impacts. Any further improvements will enhance the importance of wildtype exploration relative to the subsequent local optimization step. In silico screenings of potential wildtype candidates can be scaled efficiently compared to expensive, time-consuming in vitro assays, significantly reducing the early costs of future protein engineering campaigns. We hope that FLOP will pave the way for these developments.

Acknowledgments and Disclosure of Funding

This work was funded in part by Innovation Fund Denmark (1044-00158A), the Novo Nordisk Foundation through the MLSS Center (Basic Machine Learning Research in Life Science, NNF20OC0062606), the Pioneer Centre for AI (DRNF grant number P1), and the Danish Data Science Academy, which is funded by the Novo Nordisk Foundation (NNF21SA0069429) and VILLUM FONDEN (40516).

References

[1] B. J. Wittmann, Y. Yue, and F. H. Arnold, "Informed training set design enables efficient machine learning-assisted directed protein evolution," *Cell systems*, vol. 12, no. 11, pp. 1026–1045, 2021.

[2] T. Davids, M. Schmidt, D. Böttcher, and U. T. Bornscheuer, "Strategies for the discovery and engineering of enzymes for biocatalysis," *Current Opinion in Chemical Biology*, vol. 17, no. 2, pp. 215–220, 2013.

[3] K. K. Yang, Z. Wu, and F. H. Arnold, "Machine-learning-guided directed evolution for protein engineering," *Nature methods*, vol. 16, no. 8, pp. 687–694, 2019.

[4] P. E. Bourne, "Casp and cafasp experiments and their findings," *Methods of Biochemical Analysis*, vol. 44, pp. 501–508, 2003.

[5] J. Moult, "A decade of casp: progress, bottlenecks and prognosis in protein structure prediction," *Current opinion in structural biology*, vol. 15, no. 3, pp. 285–289, 2005.

[6] A. Kryshtafovych, T. Schwede, M. Topf, K. Fidelis, and J. Moult, "Critical assessment of methods of protein structure prediction (casp)—round xiv," *Proteins: Structure, Function, and Bioinformatics*, vol. 89, no. 12, pp. 1607–1617, 2021.

[7] R. Rao, N. Bhattacharya, N. Thomas, Y. Duan, X. Chen, J. Canny, P. Abbeel, and Y. S. Song, "Evaluating Protein Transfer Learning with TAPE," *arXiv:1906.08230 [cs, q-bio, stat]*, Jun. 2019, arXiv: 1906.08230. [Online]. Available: http://arxiv.org/abs/1906.08230

[8] G. J. Rocklin, T. M. Chidyausiku, I. Goreshnik, A. Ford, S. Houliston, A. Lemak, L. Carter, R. Ravichandran, V. K. Mulligan, A. Chevalier, C. H. Arrowsmith, and D. Baker, "Global analysis of protein folding using massively parallel design, synthesis, and testing," *Science (New York, N.Y.)*, vol. 357, no. 6347, pp. 168–175, Jul. 2017.

[9] K. S. Sarkisyan, D. A. Bolotin, M. V. Meer, D. R. Usmanova, A. S. Mishin, G. V. Sharonov, D. N. Ivankov, N. G. Bozhanova, M. S. Baranov, O. Soylemez *et al.*, "Local fitness landscape of the green fluorescent protein," *Nature*, vol. 533, no. 7603, pp. 397–401, 2016.

[10] M. Xu, Z. Zhang, J. Lu, Z. Zhu, Y. Zhang, C. Ma, R. Liu, and J. Tang, "PEER: A Comprehensive and Multi-Task Benchmark for Protein Sequence Understanding," Jun. 2022, number: arXiv:2206.02096 arXiv:2206.02096 [cs]. [Online]. Available: http://arxiv.org/abs/2206.02096

[11] V. E. Gray, R. J. Hause, J. Luebeck, J. Shendure, and D. M. Fowler, "Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data," *Cell Systems*, vol. 6, no. 1, pp. 116–124.e3, Jan. 2018.

[12] P. Notin, M. Dias, J. Frazer, J. Marchena-Hurtado, A. Gomez, D. S. Marks, and Y. Gal, "Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval," May 2022, number: arXiv:2205.13760 arXiv:2205.13760 [cs]. [Online]. Available: http://arxiv.org/abs/2205.13760

[13] R. J. L. Townshend, M. Vögele, P. Suriana, A. Derry, A. Powers, Y. Laloudakis, S. Balachandar, B. Jing, B. Anderson, S. Eismann, R. Kondor, R. B. Altman, and R. O. Dror, "ATOM3D: Tasks On Molecules in Three Dimensions," *arXiv:2012.04035 [physics, q-bio]*, Jan. 2022, arXiv: 2012.04035. [Online]. Available: http://arxiv.org/abs/2012.04035

[14] I. H. Moal and J. Fernández-Recio, "SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models," *Bioinformatics*, vol. 28, no. 20, pp. 2600–2607, Oct. 2012. [Online]. Available: https://doi.org/10.1093/bioinformatics/bts489

[15] J. Jankauskaite, B. Jiménez-García, J. Dapkunas, J. Fernández-Recio, and I. H. Moal, "SKEMPI 2.0: an updated benchmark of changes in protein-protein binding energy, kinetics and thermodynamics upon mutation," *Bioinformatics (Oxford, England)*, vol. 35, no. 3, pp. 462–469, Feb. 2019.

[16] C. Dallago, J. Mou, K. E. Johnston, B. J. Wittmann, N. Bhattacharya, S. Goldman, A. Madani, and K. K. Yang, "FLIP: Benchmark tasks in fitness landscape inference for proteins," Jan. 2022, pages: 2021.11.09.467890 Section: New Results. [Online]. Available: https://www.biorxiv.org/content/10.1101/2021.11.09.467890v2

[17] A. Jarzab, N. Kurzawa, T. Hopf, M. Moerch, J. Zecha, N. Leijten, Y. Bian, E. Musiol, M. Maschberger, G. Stoehr, I. Becher, C. Daly, P. Samaras, J. Mergner, B. Spanier, A. Angelov, T. Werner, M. Bantscheff, M. Wilhelm, M. Klingenspor, S. Lemeer, W. Liebl, H. Hahne, M. M. Savitski, and B. Kuster, "Meltome atlas—thermal proteome stability across the tree of life," *Nature Methods*, vol. 17, no. 5, pp. 495–503, May 2020, number: 5 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41592-020-0801-4

[18] N. C. Wu, L. Dai, C. A. Olson, J. O. Lloyd-Smith, and R. Sun, "Adaptation in protein fitness landscapes is facilitated by indirect paths," *Elife*, vol. 5, p. e16965, 2016.

[19] D. H. Bryant, A. Bashir, S. Sinai, N. K. Jain, P. J. Ogden, P. F. Riley, G. M. Church, L. J. Colwell, and E. D. Kelsic, "Deep diversification of an aav capsid protein by machine learning," *Nature Biotechnology*, vol. 39, no. 6, pp. 691–696, 2021.

[20] M. H. Gíslason, H. Nielsen, J. J. Almagro Armenteros, and A. R. Johansen, "Prediction of gpi-anchored proteins with pointer neural networks," *Current Research in Biotechnology*, vol. 3, pp. 6–13, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590262821000010

[21] F. Teufel, M. H. Gíslason, J. J. A. Armenteros, A. R. Johansen, O. Winther, and H. Nielsen, "GraphPart: Homology partitioning for biological sequence analysis," Apr. 2023, pages: 2023.04.14.536886 Section: New Results. [Online]. Available: https://www.biorxiv.org/content/10.1101/2023.04.14.536886v1

[22] M. Steinegger and J. Söding, "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets," *Nature Biotechnology*, vol. 35, no. 11, pp. 1026–1028, Nov. 2017, number: 11 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/nbt.3988

[23] K. K. Yang, A. X. Lu, and N. Fusi, "Convolutions are competitive with transformers for protein sequence pretraining," May 2022, pages: 2022.05.19.492714 Section: New Results. [Online]. Available: https://www.biorxiv.org/content/10.1101/2022.05.19.492714v2

[24] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. Fergus, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," bioRxiv, Tech. Rep., Dec. 2020, section: New Results Type: article. [Online]. Available: https://www.biorxiv.org/content/10.1101/622803v4

[25] J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, and A. Rives, "Language models enable zero-shot prediction of the effects of mutations on protein function," bioRxiv, Tech. Rep., Jul. 2021, section: New Results Type: article. [Online]. Available: https://www.biorxiv.org/content/10.1101/2021.07.09.450648v1

[26] R. M. Rao, J. Liu, R. Verkuil, J. Meier, J. Canny, P. Abbeel, T. Sercu, and A. Rives, "MSA Transformer," in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, Jul. 2021, pp. 8844–8856, iSSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v139/rao21a.html

[27] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, A. d. Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives, "Language models of protein sequences at the scale of evolution enable accurate structure prediction," Synthetic Biology, preprint, Jul. 2022. [Online]. Available: http://biorxiv.org/lookup/doi/10.1101/2022.07.20.500902

[28] N. S. Detlefsen, S. Hauberg, and W. Boomsma, "Learning meaningful representations of protein sequences," *Nature Communications*, vol. 13, no. 1, p. 1914, Apr. 2022, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41467-022-29443-w

[29] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, Aug. 2021, number: 7873 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41586-021-03819-2

[30] M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, and M. Steinegger, "ColabFold: making protein folding accessible to all," *Nature Methods*, vol. 19, no. 6, pp. 679–682, Jun. 2022, number: 6 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41592-022-01488-1

[31] M. Hu, F. Yuan, K. K. Yang, F. Ju, J. Su, H. Wang, F. Yang, and Q. Ding, "Exploring evolution-based & -free protein language models as protein function predictors," Jun. 2022, number: arXiv:2206.06583 arXiv:2206.06583 [cs, q-bio]. [Online]. Available: http://arxiv.org/abs/2206.06583

[32] C. Hsu, R. Verkuil, J. Liu, Z. Lin, B. Hie, T. Sercu, A. Lerer, and A. Rives, "Learning inverse folding from millions of predicted structures," in *Proceedings of the 39th International Conference on Machine Learning*. PMLR, Jun. 2022, pp. 8946–8970, iSSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v162/hsu22a.html

[33] K. K. Yang, H. Yeh, and N. Zanichelli, "Masked inverse folding with sequence transfer for protein representation learning," Mar. 2023, pages: 2022.05.25.493516 Section: New Results. [Online]. Available: https://www.biorxiv.org/content/10.1101/2022.05.25.493516v3

[34] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, "Direct-coupling analysis of residue coevolution captures native contacts across many protein families," *Proceedings of the National Academy of Sciences*, vol. 108, no. 49, pp. E1293–E1301, 2011.

[35] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, "Protein 3d structure computed from evolutionary sequence variation," *PloS one*, vol. 6, no. 12, p. e28766, 2011.

[36] P. Tian and R. B. Best, "How many protein sequences fold to a given structure? a coevolutionary analysis," *Biophysical journal*, vol. 113, no. 8, pp. 1719–1730, 2017.

[37] The UniProt Consortium, "UniProt: the Universal Protein Knowledgebase in 2023," *Nucleic Acids Research*, vol. 51, no. D1, pp. D523–D531, Jan. 2023. [Online]. Available: https://doi.org/10.1093/nar/gkac1052

[38] M. Blum, H.-Y. Chang, S. Chuguransky, T. Grego, S. Kandasaamy, A. Mitchell, G. Nuka, T. Paysan-Lafosse, M. Qureshi, S. Raj, L. Richardson, G. A. Salazar, L. Williams, P. Bork, A. Bridge, J. Gough, D. H. Haft, I. Letunic, A. Marchler-Bauer, H. Mi, D. A. Natale, M. Necci, C. A. Orengo, A. P. Pandurangan, C. Rivoire, C. J. A. Sigrist, I. Sillitoe, N. Thanki, P. D. Thomas, S. C. E. Tosatto, C. H. Wu, A. Bateman, and R. D. Finn, "The interpro protein families and domains database: 20 years on," *Nucleic acids research*, vol. 49, no. D1, p. D344—D354, January 2021. [Online]. Available: https://europepmc.org/articles/PMC7778928

[39] J. Frazer, P. Notin, M. Dias, A. Gomez, J. K. Min, K. Brock, Y. Gal, and D. S. Marks, "Disease variant prediction with deep generative models of evolutionary data," *Nature*, vol. 599, no. 7883, pp. 91–95, Nov. 2021, number: 7883 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41586-021-04043-8

[40] A. J. Riesselman, J. B. Ingraham, and D. S. Marks, "Deep generative models of genetic variation capture the effects of mutations," *Nature Methods*, vol. 15, no. 10, pp. 816–822, Oct. 2018, number: 10 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41592-018-0138-4

[41] T. A. Hopf, A. G. Green, B. Schubert, S. Mersmann, C. P. I. Schärfe, J. B. Ingraham, A. Toth-Petroczy, K. Brock, A. J. Riesselman, P. Palmedo, C. Kang, R. Sheridan, E. J. Draizen, C. Dallago, C. Sander, and D. S. Marks, "The EVcouplings Python framework for coevolutionary sequence analysis," *Bioinformatics (Oxford, England)*, vol. 35, no. 9, pp. 1582–1584, May 2019.

[42] D. Hesslow, N. Zanichelli, P. Notin, I. Poli, and D. Marks, "RITA: a Study on Scaling Up Generative Protein Sequence Models," arXiv, Tech. Rep. arXiv:2205.05789, May 2022, arXiv:2205.05789 [cs, q-bio] type: article. [Online]. Available: http://arxiv.org/abs/2205.05789

[43] J. Dong, Z.-J. Yao, L. Zhang, F. Luo, Q. Lin, A.-P. Lu, A. F. Chen, and D.-S. Cao, "PyBioMed: a python library for various molecular representations of chemicals, proteins and DNAs and their interactions," *Journal of Cheminformatics*, vol. 10, no. 1, p. 16, Mar. 2018.

[44] J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. d. Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker, "Robust deep learning based protein sequence design using ProteinMPNN," Jun. 2022, pages: 2022.06.03.494563 Section: New Results. [Online]. Available: https://www.biorxiv.org/content/10.1101/2022.06.03.494563v1

[45] P. Notin, M. Dias, J. Frazer, J. M. Hurtado, A. N. Gomez, D. Marks, and Y. Gal, "Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval," in *International Conference on Machine Learning*. PMLR, 2022, pp. 16 990–17 017.

[46] R. Lonsdale, J. N. Harvey, and A. J. Mulholland, "A practical guide to modelling enzyme-catalysed reactions," *Chemical Society Reviews*, vol. 41, no. 8, pp. 3025–3038, Apr. 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3371381/

[47] M. Li, J. Salomon, D. R. Segura, M. A. Stringer, R. M. Vejborg, D. M. K. Klitgaard, D. Nissen, W. Peng, and T. Sun, "Polypeptides," Patent WO/2019/228 448, Dec., 2019. [Online]. Available: https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2019228448

[48] K. M. Colvin, V. D. Gordon, K. Murakami, B. R. Borlee, D. J. Wozniak, G. C. L. Wong, and M. R. Parsek, "The Pel Polysaccharide Can Serve a Structural and Protective Role in the Biofilm Matrix of Pseudomonas aeruginosa," *PLOS Pathogens*, vol. 7, no. 1, p. e1001264, Jan. 2011, publisher: Public Library of Science. [Online]. Available: https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1001264

[49] M. K. Tiwari, R. Singh, R. K. Singh, I.-W. Kim, and J.-K. Lee, "Computational approaches for rational design of proteins with novel functionalities," *Computational and structural biotechnology journal*, vol. 2, no. 3, p. e201204002, 2012.

[50] W. P. Russ, M. Figliuzzi, C. Stocker, P. Barrat-Charlaix, M. Socolich, P. Kast, D. Hilvert, R. Monasson, S. Cocco, M. Weigt, and R. Ranganathan, "An evolution-based model for designing chorismate mutase enzymes," *Science*, vol. 369, no. 6502, pp. 440–445, Jul. 2020, publisher: American Association for the Advancement of Science. [Online]. Available: https://www.science.org/doi/full/10.1126/science.aba3304

[51] C. Plesa, A. M. Sidore, N. B. Lubock, D. Zhang, and S. Kosuri, "Multiplexed gene synthesis in emulsions for exploring protein functional landscapes," *Science*, vol. 359, no. 6373, pp. 343–347, Jan. 2018, publisher: American Association for the Advancement of Science. [Online]. Available: https://www.science.org/doi/10.1126/science.aao5167

[52] P. Bork and E. V. Koonin, "Predicting functions from protein sequences–where are the bottlenecks?" *Nature Genetics*, vol. 18, no. 4, pp. 313–318, Apr. 1998.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No] We do not believe that there are any potential negative societal impacts of this work, since it concerns modelling naturally occurring wildtype proteins.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See supplementary materials.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See both Section 3.1 for data splits and supplementary materials Section K for hyperparameter grids.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See supplementary materials, Section B.1.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No] No specific license agreements were found. Consent was however given from authors of the datasets to include them in the benchmark, and we specify in the supplementary materials Section A that references to tasks in this benchmark should include references to the original data sources.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No] All datasets and methods are publicly available prior to this work. We do however provide thorough descriptions of how to access and use the original data as well as our processed/curated versions of it.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] Consent was given explicitly for all datasets.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]