FLOP: Tasks for Fitness Landscapes Of Protein families using sequence- and structure-based representations

Anonymous authors
Paper under double-blind review

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 protein with the most desirable properties. This chosen candidate is then the basis for the second step: a more local optimization procedure, exploring the space of variants separated from this candidate by a few mutations. While advances in protein representation learning promise to facilitate the exploration of wildtype space, results from real-life cases are often underwhelming, and progress in the area difficult to track. In this paper, we have carefully curated a representative benchmark dataset, which reflects industrially-relevant scenarios for the initial wildtype exploration of protein engineering. We focus on the exploration within a protein family or superfamily, and investigate the downstream predictive power of various dominating protein representation paradigms, i.e., transformer-based language 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. We hope our benchmark can drive further methodological developments in this important field.

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 (Wittmann et al., 2021). The optimization process of proteins and enzymes typically proceeds in multiple stages. An initial step is to navigate the protein landscape to find a suitable, naturally occurring protein (also denoted as a wildtype) with the desired properties. This process is often limited to a particular protein family, where the members share an evolutionary history which has resulted in a similar function. This chosen protein will then form the basis for a second phase in the engineering process: localized optimization, where novel variants of the wildtype are examined through various assays.

Regression of functional landscapes is challenging for multiple reasons. It is typically a data-scarce problem, although high-throughput experimental techniques are improving this, and underlying structure typically exists in the dataset, careful considerations of the experimental setup are required. In recent years, we have seen considerable efforts into defining benchmarks to help the Machine Learning community make progress in this field. These efforts have, however, primarily focused on the second stage, i.e., variant effect prediction. In this paper, we argue for the importance of creating clearly defined benchmark tasks for the first stage as well. We present three challenging tasks as well as a careful analysis of the experimental design, demonstrating how poor choices can lead to dramatic overestimation of performance. Finally, we make available our input data in a variety of representations: sequence-based embeddings obtained through recent state-of-the-art language models, structure-based representations from folding and inverse folding models, and evolutionary-

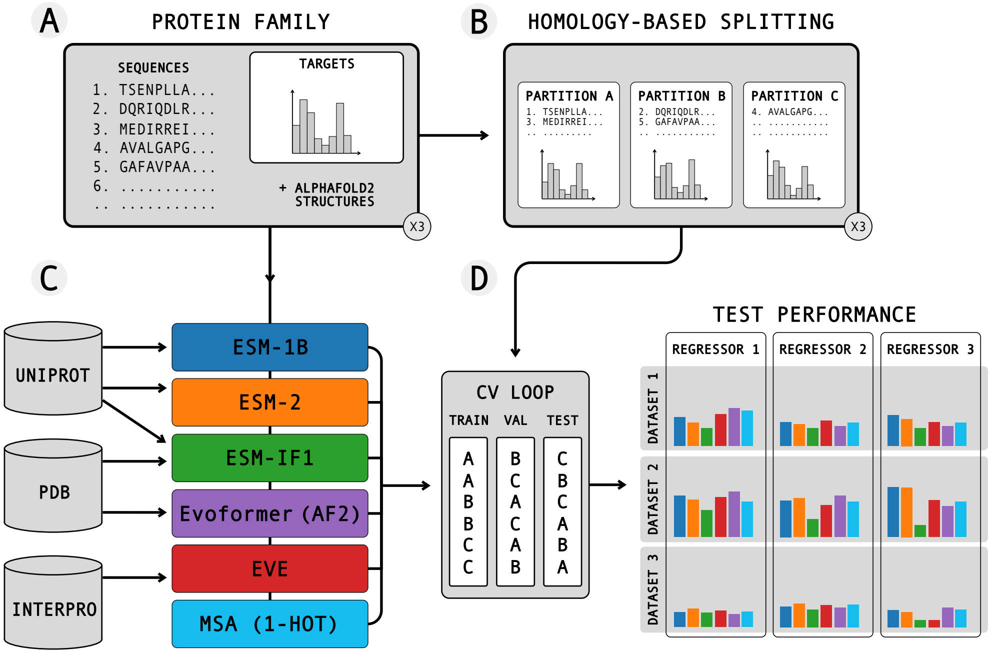

Figure 1: Schematic over dataset splitting, representations, and cross-validation process. A: Three datasets each containing sequences for a single protein family or superfamily with corresponding continuous assay values are collected. The protein structures are predicted with AlphaFold2. B: A stratified sequence identity splitting procedure generates partitions A, B, and C. These partitions are designed to (1) be homologically different from each other, (2) to contain similar number of sequences, and (3) to match the full datasets target distribution. C: Six protein representations are for all sequences generated across the three datasets. D: Exhaustive cross-validation is applied to obtain accurate test estimates across the datasets using three different regression models.

based representations obtained from multiple sequence alignments. Our benchmark thus also gives a perspective on the relative information content in these fundamentally different modalities.

2 Related work

Benchmarks have long played 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 (Bourne, 2003; Moult, 2005; Kryshtafovych et al., 2021). For the prediction of protein stability and function, several studies have in recent years curated relevant experimental datasets for use as benchmarks to the machine learning community. The TAPE benchmark was an early such example designed to test protein sequence representations on a set of diverse downstream tasks (Rao et al., 2019). Two of these tasks were related to protein engineering: stability prediction on variants of a set of 12 designed proteins (Rocklin et al., 2017) and characterization of the functional landscape of green fluorescent protein (Sarkisyan et al., 2016). The PEER benchmark (Xu et al., 2022) expanded on the TAPE benchmark with many additional tasks. In the function/stability category, this included prediction of β-lactamase activity (Gray et al., 2018), 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 (Notin et al., 2022)). While the above all consider protein sequence inputs, the recent Atom3D benchmark (Townshend et al., 2022) 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 (Moal & Fernández-Recio, 2012; Jankauskaite et al., 2019).

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. Their study includes three tasks: one on the prediction of protein stability of general proteins (i.e. distributed over many families) using data from the Meltome Atlas (Jarzab et al., 2020), and two tasks focused on mutations at specific, functionally relevant, sites of proteins GB1 (Wu et al., 2016) and AAV (Bryant et al., 2021)).

Most of the functional tasks in current benchmarks are concerned with protein sequences that are derived from a single naturally-occurring sequence by one or more mutations. Characterizing the functional effects of such variants is critical for protein engineering. However, in the protein engineering pipeline, the optimization of variants is typically preceded by an earlier stage of wildtype discovery, where the goal is to search among wildtype proteins to find the optimal starting point for the subsequent variant optimization process. Typically, this involves looking for a suitable protein with a particular function, and will thus restrict the wildtype search within a particular protein family or superfamily. The tasks we are interested in are thus the characterization of functional landscapes of wildtype proteins within a family or superfamily. This is not addressed in any of the current benchmarks. As a natural complement to the FLIP benchmark, we therefore here present a novel benchmark titled FLOP. Our curated datasets all consist of functionally characterized wildtype sequences from a single family or superfamily, where the tasks are defined as regression tasks. As inputs, we provide sequence embeddings from state-of-the-art protein language models such as ESM-1B and ESM-2 (Rives et al., 2020; Hsu et al., 2022b). We also provide structure-based representations obtained from 3D structures predicted by AlphaFold2 (Jumper et al., 2021). Finally, since we restrict ourselves to protein families, we also provide embeddings obtained from multiple sequence alignments of the input sequences, e.g. by using the EVE variational autoencoder model (Frazer et al., 2021).

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 final 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 holdout validation, whereby fixed training, validation, and testing sets are created. This is often achieved by randomly splitting the dataset into appropriately sized partitions. There are two limitations to this method when considering biological datasets of limited sizes. Firstly, randomly splitting a dataset into subsets assumes that the data points are sampled i.i.d. This is however not the case for members of a protein family which share common ancestors. Significant data leakage can thus inadvertently appear when protein sequences that are close in evolutionary space are placed in separate splits. Secondly, when the size of a dataset is limited to hundreds, a holdout validation approach will often lead to biased results, where the method might not generalize well. An additional issue with splitting small datasets for the purpose of supervised learning is that the target values might not be well-balanced. This can result in splits which follow different target distributions thus leading to poor generalizability.

A common method of avoiding the issue of non-i.i.d. protein sequences is to use identity-based clustering, whereby the sequence identity between clusters is limited to a certain threshold (Li et al., 2001; Gíslason et al., 2021a). These clusters can then be aggregated into separate training, validation, and test splits for holdout validation. The second point regarding the potential issues of holdout validation for small datasets is however often not explicitly dealt with when considering protein landscapes, since the sizes of popular datasets often diminish the negative effects hereof. A traditional way of estimating unbiased generalization errors – and hence directly deal with this issue

– is by the use of cross-validation (CV). If possible, stratification of the generated partitions on the target values might further reduce the variance of the estimated errors (Kohavi, 1995).

To handle these potential issues, we rely on a sequence identity-based, stratified cross-validation procedure which ensures the following:

- Partitions are generated that are guaranteed to be homologically different thereby minimising data leakage.
- Exhaustive cross-validation minimizes the potential bias which might occur when relying on holdout validation.
- The target distribution is accurately reflected by the generated partitions via stratification on discretized target values thus reducing the variance.
- The number of sequences in each partition is similar to reduce the variance between consecutive CV rounds.

To generate these high-quality data partitions, we use the four-phase procedure described in Gíslason et al. (2021a) to create label-balanced dataset splits. This procedure ensures that each generated partition does not share sequences that have global sequence identity above a dataset-specific threshold as determined using ggsearch36 (Pearson & Lipman, 1988). We concretely use the “GraphPart” implementation in Gíslason et al. (2021b). We begin the procedure from an initial sequence identity threshold using stratification labels, and increase the threshold until the generated partitions are of sufficient sizes, i.e., where each partition contains at least 25 % of the sequences (in the case of three partitions). The stratification is handled by creating a binary label which indicates whether a protein has low or high target value. In the case of uni- or multimodal datasets, this procedure can be analogously extended.

3.2 Representations

To accurately reflect the current paradigms of state-of-the-art protein representations, we choose representatives from three major categories.

As a baseline, we will rely on a multiple sequence alignment (MSA) over the proteins of interest. This allows for an alignment of proteins with different sequence lengths such that the dimensionality is fixed. The sequences in an MSA are assumed to share an evolutionary history, whereby the MSA-algorithm captures the coevolution of related sequences (Chatzou et al., 2016). Using this alignment for a protein family, the amino acids are encoded in a 1-of-20 fashion. The representation for each protein is thus an MSA-length by 20 dimensional matrix, which is flattened to create a fixed-sized vector input. The MSAs for each labelled protein family is generated using MAFFT (Katoh et al., 2002).

Protein language models (PLMs) that are trained on hundreds of millions of protein sequences in an unsupervised fashion have proved to be competitive representations for a multitude of tasks (Rives et al., 2020; Meier et al., 2021; Rao et al., 2021). We here choose the popular ESM-1B (Rives et al., 2020) and the more recent ESM-2 models (Lin et al., 2022). These models generate embeddings for each amino acid in a sequence. To fix the dimensionality for proteins of different lengths within a family, we perform mean-pooling over the residue dimension. Despite this coarsening operation, it has proved competitive with retaining the full residue-wise representation as seen in Dallago et al. (2021). We use the 650M and 3B parameter models for ESM-1B and ESM-2, respetively.

The second protein representation category we include is structure-based. With the rise of AlphaFold2 (Jumper et al., 2021) and the availability of protein structures (Tunyasuvunakool et al., 2021), representations can now be developed which directly use the protein structure. We include two such representations. We obtain the first of these by extracting the final embeddings from the Evoformer-modules of the AlphaFold2 model (Jumper et al., 2021) using ColabFold ((Mirdita et al., 2022)). We denote these representations as ”Evoformer (AF2)”. These embeddings have been shown to perform well for structure-related prediction tasks (Hu et al., 2022). For our second model, we include the ESM-IF1 (also known as the GVP-Transformer) from Hsu et al. (2022b). This model combines the geometric vector perceptron (Jing et al., 2021b;a) with a transformer model (Vaswani et al., 2017) to leverage both structural and sequential information. To fix the dimensionality of the variably-sized matrix representations that result from the Evoformer (AF2) and ESM-IF1, we use the same mean-pooling scheme as for the sequence-based PLM representations.

The third category of representations we investigate, we denote as evolutionary-based representations. These models can leverage the evolutionary history of proteins, often based on MSAs (e.g., EVE, (Frazer et al., 2021), DeepSequence, and EVcouplings (Hopf et al., 2019)). For each curated dataset, we train EVE on its protein family and extract the latent representations of our labelled sequences. We include EVE since it and its DeepSequence-precursor continue to show competitive performance (Hesslow et al., 2022; Hsu et al., 2022a). Since EVE is trained in a fully unsupervised manner, we enrich our limited collections of sequences with additional members of the respective protein families using the InterPro-database (Blum et al., 2021). This increases the available sequences during the structuring of the latent spaces from an order of hundreds to tens of thousands of sequences for each landscape. Additional technical details on EVE can be found in Appendix A.1.

3.3 Regressors

The purpose of this benchmark is to provide simple baselines for the regression tasks using different input representations. We have therefore chosen three simple yet representative regression models:

- A KNN-regressor as a non-parametric distance-based model.
- A ridge-regressor as a regularized linear model.
- A random forest-regressor as a tree-based ensemble method.

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 hyperparameter grids can be seen in Appendix A.2. More complex regressors are expected to have improved performance over these baseline models, and we thus hope for the creation of novel regressors for the task of protein fitness through this benchmark.

4 Datasets

4.1 CM

Motivation. Identification of an enzyme with high catalytic activity is of primary importance. However, predicting the activity level of enzymes using physics-based methods remains a great challenge (Tiwari et al., 2012). Recent progress in high throughput screening allows one to measure the activity of enzyme sequences of high diversity, but with low experimental cost. These kind of datasets are very valuable for learning the global fitness landscapes.

Landscape. This dataset contains the catalytic activity of natural chorismate mutase (CM) homologous proteins, as well as artificial sequences which follow the same pattern of variations, e.g., conservation and co-evolution of residues (Russ et al., 2020). The artificial sequences generated by Monte Carlo simulations at low temperatures exhibit catalytic levels comparable to the natural homologs and have therefore been included in the dataset. The target value of the prediction task is the catalytic activity of each CM sequence. The MSA used for EVE was generated by FAMSA (Deorowicz et al., 2016).

Split. We perform an additional filtering of the dataset, where we ignore sequences with target values less than 0.42, corresponding to inactive proteins (Russ et al., 2020). During the dataset splitting process, a number of sequences are removed to allow for partitioning at seqeunce identity threshold of 0.35. The final dataset size is thus 715 proteins. A summary can be found in Table 1.

4.2 PPAT

Motivation. PPAT (phosphopantetheine adenylyltransferase) is composed of around 160 amino acids and 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.

Landscape. This dataset contains fitness scores of 615 different PPAT homologs, which are 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 (Plesa et al., 2018).

Split. The dataset is split using the described procedure resulting in a threshold of 0.55. The MSA is generated and processed in the same way for EVE as the CM dataset. Details can be seen in Table 1.

4.3 TIM

Motivation. One of the most important properties for enzymes in industrial applications is thermostability, since running applications at higher temperatures allow for faster reaction rates. Many efforts have been made to identify naturally thermostable enzymes, e.g. from thermophiles. However, predicting stability of individual proteins accurately has proven challenging; the fitness landscape is non-smooth and a single mutation can lead to a significant shift in thermostability (Pezeshgi Modarres et al., 2018).

Landscape. The Meltome Atlas contains thermostability measurements of proteins across the tree of life, captured in thermal profiling experiments using mass spectrometry (Jarzab et al., 2020). Prediction tasks of the full dataset and of individual proteomes has been attempted (Dallago et al., 2021). As reported in (Jarzab et al., 2020; Gado et al., 2020), the dominant factor has been shown to be the growth temperature of the donor organism. However, structural features are known to be important determinants for thermostability, since thermostability *is* the preservation of structure under thermal stress (Ahmed et al., 2022; Kumwenda et al., 2013). Thus, proteins sharing the same structural fold may provide an easier starting point and help identify common structural determinants of thermostability. Members of the Aldolase-type TIM barrel superfamily (InterPro entry IPR013785) form a β/α barrel, and include a wide range of enzymatic functions such as aldolases, synthases, etc. (Blum et al., 2021). The TIM barrel is typically the dominant structural domain in the member proteins, and expected to be the main signal in the thermal profiling experiments (Jarzab et al., 2020).

Split. Due to the broad diversity of the included proteins, sequence identity-based partitioning can be done at a threshold of 0.10. This however had the adverse affect of creating very dissimilar partitions. The threshold was therefore artificially set to 0.5, which ensured proper mixing of proteins within the generated partitions. To generate MSAs, normal sequence alignment tools are unable to handle the distant homologs and produce meaningful alignments, so we used Foldseek to construct a query based structural alignment (Kempen et al., 2022) for EVE. For the one-hot encoded MSA, we used Caretta (Akdel et al., 2020). We attempted to use Caretta to align the extended superfamily. This however proved infeasible given the number of proteins.

4.4 Summary

A summary of the datasets including their total sizes, partition sizes (A, B, and C refer to the generated partitions), between-partition sequence identity threshold (%ID), target value, and median pair-wise sequence identity can be seen in Table 1, the latter of which highlights the diversity – and difficulty – of the datasets.

5 Results

Spearman's rank correlation between the predictions and targets over the three datasets using our CV procedure can be seen in Figure 2 and Table 2. During the first phase of protein engineering,

Table 1: Summary of datasets and splits.

	# Total	**# in A**	**# in B**	**# in C**	**%ID**	**Target value**	**Median %ID**
CM	715	265	204	246	0.35	Catalytic activity	0.43
PPAT	615	182	234	199	0.55	Fitness score	0.45
TIM	218	72	72	72	0.50	Thermostability	0.12

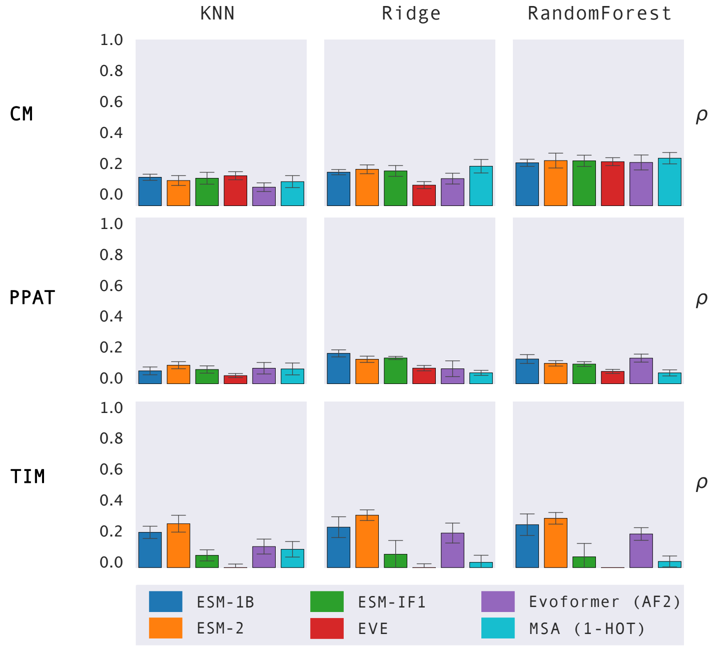

Figure 2: Spearman's correlation coefficient between predictions and targets over test partitions, grouped by dataset and regressor. Standard error is shown as vertical bars.

Table 2: Benchmark results (Spearman's rank correlation). We report the mean over the test partitions and the standard error. The best result has been highlighted for each dataset.

		KNN	Ridge	RandomForest
CM	ESM-1B	0.17 ± 0.02	0.20 ± 0.02	0.26 ± 0.02
	ESM-2	0.15 ± 0.03	0.22 ± 0.03	0.27 ± 0.04
	ESM-IF1	0.17 ± 0.04	0.21 ± 0.03	0.27 ± 0.03
	EVE	0.18 ± 0.02	0.12 ± 0.02	0.27 ± 0.02
	Evoformer (AF2)	0.11 ± 0.03	0.16 ± 0.03	0.26 ± 0.04
	MSA (1-HOT)	0.14 ± 0.04	0.24 ± 0.04	**0.29** ± 0.03
PPAT	ESM-1B	0.08 ± 0.02	**0.18** ± 0.02	0.15 ± 0.03
	ESM-2	0.11 ± 0.02	0.15 ± 0.02	0.12 ± 0.02
	ESM-IF1	0.09 ± 0.02	0.16 ± 0.01	0.12 ± 0.01
	EVE	0.05 ± 0.01	0.09 ± 0.02	0.07 ± 0.01
	Evoformer (AF2)	0.09 ± 0.03	0.09 ± 0.05	0.15 ± 0.02
	MSA (1-HOT)	0.09 ± 0.04	0.07 ± 0.02	0.07 ± 0.02
TIM	ESM-1B	0.21 ± 0.04	0.24 ± 0.06	0.26 ± 0.06
	ESM-2	0.26 ± 0.05	**0.32** ± 0.03	0.30 ± 0.03
	ESM-IF1	0.07 ± 0.03	0.08 ± 0.08	0.06 ± 0.08
	EVE	−0.02 ± 0.04	−0.01 ± 0.04	−0.08 ± 0.04
	Evoformer (AF2)	0.13 ± 0.04	0.21 ± 0.06	0.20 ± 0.04
	MSA (1-HOT)	0.11 ± 0.05	0.03 ± 0.04	0.04 ± 0.03

the highest performing proteins are of interest. It is thus the ranking, and not the absolute values of the errors that indicate the performance, despite the regressors being optimized using the mean squared error. We observe that the regression tasks on our datasets are challenging with maximum averaged correlations of only 0.32. We generally see similar performance from the three regressors in all configurations with only minor differences, such as the KNN slightly underperforming for all three landscapes. While the performance using the different protein representations is similar, we do observe minor patterns. We see stable and competitive results from the sequence-based PLMs (ESM-1B and ESM-2) as well as from the AlphaFold2-extracted Evoformer embeddings.

5.1 CM

The performance on the CM dataset is not strongly dependent on the input representations as these provide similar results for the three regressors, respectively. It is however clear that the random forest regressor has an advantage when compared to the neighbour-based and linear models. This indicates that further exploration of complex non-linear regression models are viable to improve the results. The only representation that has a slight edge is the one-hot encoded MSA representation.

5.2 PPAT

The PLM representations have a clear advantage on the PPAT dataset, where we see both ESM-1B, ESM-2, and ESM-IF1 performing well, the first outperforming all other representations. We furthermore see competitive performance from the AlphaFold2 embeddings.

5.3 TIM

The performance on the TIM dataset is strikingly different when compared to the previous datasets. Here, the sequence-based PLM embeddings significantly outperform both the structure- and alignment-based models with ESM-2 coming out well ahead. The poor performances of the one-hot encoded MSA and especially the EVE representations indicate that either the alignment of the superfamily is inaccurate or the method is not optimal for such distinct proteins, as also reflected by the low pair-wise sequence identity in Table 1.

6 Design choices

For the CM dataset, we chose to only include the active sequences in our modeling. This is potentially questionable given the inherent difficulties of being in the small-data regime. Two alternatives would be (1) to cast the prediction task as classification and then use all sequences (as seen in Nijkamp et al. (2022)) or (2) to simply include all sequences in the regression task. The issue with (1) is that in the wildtype exploration phase, knowing whether a protein is inactive or active is not sufficient information to proceed to the local optimization phase. A higher resolution is favorable, which is why regression is preferred. To demonstrate why we do not report the results of (2) as our primary task, we have included them in the top row of Figure 3. This shows that if we perform regression on a distinctly bimodal dataset (such as CM), the results will be inflated. The models are able to distinguish between the modalities of the target distribution which drives the Spearman correlation to overly-optimistic values. The caveat here, is that this requires knowing the target labels a priori. This process thus assumes that a preceding classification-screening has been performed to obtain only active sequences.

In the second row of Figure 3, we see the predictive performance on the PPAT dataset when using random splitting instead of cross-validation. The random splitting was achieved by generating three random, equally-size permutations of the proteins (with no repetition). Training, validation, and testing was then performed on these three partitions. This was repeated a total of five times to decrease bias and obtain error bars. While the results overall look similar to the CV-results from Figure 2, we do see slightly better results. That the results are not significantly better is not surprising. The stratified, sequence identity-based CV procedure that we employ, can be considered a type of “informed random splitting”. What we avoid with this procedure when comparing to random splitting is placing two very similar sequences in separate partitions, which can be considered data leakage. Similarly, we avoid imbalance in the partitions, which becomes more important when the dataset sizes decrease, avoiding partitions with only inactive sequences. This however also shows that repeated random splitting of wildtype protein datasets is not as detrimental as is often the case for variant datasets (Bork & Koonin, 1998).

In the third row of Figure 3 we see the results of performing holdout validation on the TIM dataset. We used the generated CV partitions once, instead of performing the CV loop. This shows that creating stratified, homologically different partitions is not sufficient to achieve nuanced results. A

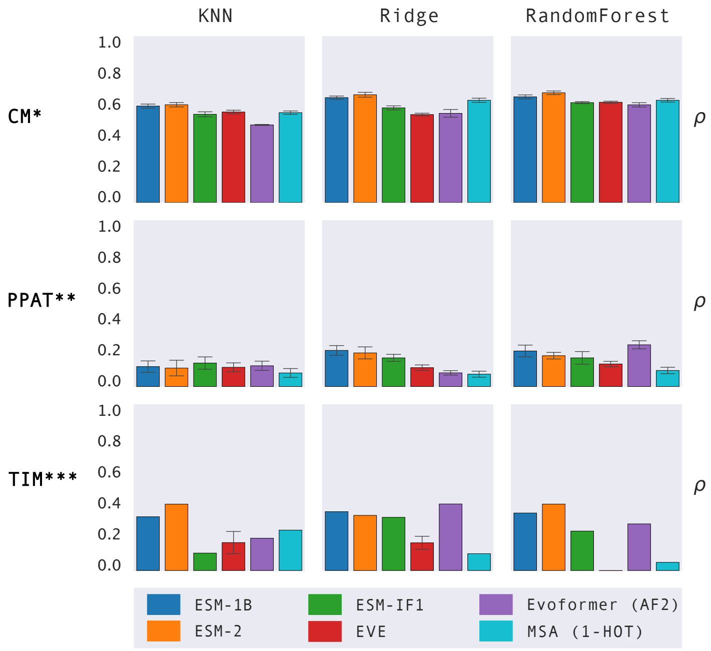

Figure 3: Spearman's correlation coefficient between predictions and targets over test partitions, grouped by dataset and regressor. Standard error is shown as vertical bars. *: Regression on both active and inactive proteins. **: Repeated random splitting. ***: Holdout validation.

clear difference between the third row of Figures 2 and 3 is the lack of error bars in the latter (with the exception of EVE, see Appendix A.1), which gives an overly optimistic impression of the results. We furthermore observe that the results are generally better than when performing CV. Interestingly, we see that the difference between regression models is emphasized in the given partition which can again give a biased impression of the actual predictive performance.

For completeness, the tabulated values of Figure 3 can be seen in Table 3 in Appendix A.3.

7 Conclusion

In this work we have presented a novel benchmark dataset which investigates an unexplored domain of machine learning-driven protein engineering: the navigation of global fitness landscapes for single protein families. The curation itself of relevant labelled datasets proved challenging and the creation of new comprehensive family-wide datasets is bound to ease and improve future model development and applicability. We have here highlighted the importance of careful learning schemes in the evaluation of our models to ensure reliable estimates of generalizability. While the sequence-based PLM representations consistently yielded good results, the difference in performance is not significant among all the representations. We also saw cases where a simple one-hot encoded MSA was competitive, highlighting the intricacies of creating meaningful representations. Our attempt at creating useful superfamily alignments proved challenging and highlights an area where improvements can be made.

We hope that FLOP will pave the way for modeling global fitness landscapes of protein families and accelerate the further development of complex regression models and novel protein representations, the combination of which is will drive protein engineering for drug discovery and sustainable biotechnology forward.

8 REPRODUCIBILITY STATEMENT

All code and data used in the this work will be made publicly available. This includes (but is not limited to) code that has been used to

- preprocess raw data files
- perform splitting of the datasets according to the described procedure
- generate all representations used
- train, validate, and test all regression models.

We will additionally provide all representations and data files for further use, e.g., the representations themselves, AlphaFold predicted structures, csv-files mapping proteins to CV-partitions and target values, etc.

All code will be made available via a Github-repository, while all data will be made available for download through a hosted service.

REFERENCES

Zahoor Ahmed, Hasan Zulfiqar, Lixia Tang, and Hao Lin. A statistical analysis of the sequence and structure of thermophilic and non-thermophilic proteins. *International Journal of Molecular Sciences*, 23(17), 2022. ISSN 1422-0067. doi: 10.3390/ijms231710116. URL `https://www.mdpi.com/1422-0067/23/17/10116`.

Mehmet Akdel, Janani Durairaj, Dick de Ridder, and Aalt D. J. van Dijk. Caretta – A multiple protein structure alignment and feature extraction suite. *Computational and Structural Biotechnology Journal*, 18:981–992, January 2020. ISSN 2001-0370. doi: 10.1016/j.csbj.2020.03.011. URL `https://www.sciencedirect.com/science/article/pii/S2001037019304477`.

Matthias Blum, Hsin-Yu Chang, Sara Chuguransky, Tiago Grego, Swaathi Kandasaamy, Alex Mitchell, Gift Nuka, Typhaine Paysan-Lafosse, Matloob Qureshi, Shriya Raj, Lorna Richardson, Gustavo A Salazar, Lowri Williams, Peer Bork, Alan Bridge, Julian Gough, Daniel H Haft, Ivica Letunic, Aron Marchler-Bauer, Huaiyu Mi, Darren A Natale, Marco Necci, Christine A Orengo, Arun P Pandurangan, Catherine Rivoire, Christian J A Sigrist, Ian Sillitoe, Narmada Thanki, Paul D Thomas, Silvio C E Tosatto, Cathy H Wu, Alex Bateman, and Robert D Finn. The interpro protein families and domains database: 20 years on. *Nucleic acids research*, 49(D1):D344—D354, January 2021. ISSN 0305-1048. doi: 10.1093/nar/gkaa977. URL `https://europepmc.org/articles/PMC7778928`.

P. Bork and E. V. Koonin. Predicting functions from protein sequences–where are the bottlenecks? *Nature Genetics*, 18(4):313–318, April 1998. ISSN 1061-4036. doi: 10.1038/ng0498-313.

Philip E Bourne. Casp and cafasp experiments and their findings. *Methods of Biochemical Analysis*, 44:501–508, 2003.

Drew H Bryant, Ali Bashir, Sam Sinai, Nina K Jain, Pierce J Ogden, Patrick F Riley, George M Church, Lucy J Colwell, and Eric D Kelsic. Deep diversification of an aav capsid protein by machine learning. *Nature Biotechnology*, 39(6):691–696, 2021.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122, 2013.

Maria Chatzou, Cedrik Magis, Jia-Ming Chang, Carsten Kemena, Giovanni Bussotti, Ionas Erb, and Cedric Notredame. Multiple sequence alignment modeling: methods and applications. *Briefings in Bioinformatics*, 17(6):1009–1023, November 2016. ISSN 1467-5463. doi: 10.1093/bib/bbv099. URL `https://doi.org/10.1093/bib/bbv099`.

Christian Dallago, Jody Mou, Kadina E. Johnston, Bruce Wittmann, Nick Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K. Yang. FLIP: Benchmark tasks in fitness landscape inference for proteins. August 2021. URL https://openreview.net/forum?id=p2dMLEwL8tF.

Sebastian Deorowicz, Agnieszka Debudaj-Grabysz, and Adam Gudyś. FAMSA: Fast and accurate multiple sequence alignment of huge protein families. *Scientific Reports*, 6(1):33964, September 2016. ISSN 2045-2322. doi: 10.1038/srep33964. URL https://www.nature.com/articles/srep33964. Number: 1 Publisher: Nature Publishing Group.

Nicki Skafte Detlefsen, Søren Hauberg, and Wouter Boomsma. Learning meaningful representations of protein sequences. *Nature Communications*, 13(1):1914, April 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-29443-w. URL https://www.nature.com/articles/s41467-022-29443-w. Number: 1 Publisher: Nature Publishing Group.

Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K. Min, Kelly Brock, Yarin Gal, and Debora S. Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, November 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-04043-8. URL https://www.nature.com/articles/s41586-021-04043-8. Number: 7883 Publisher: Nature Publishing Group.

Japheth E. Gado, Gregg T. Beckham, and Christina M. Payne. Improving enzyme optimum temperature prediction with resampling strategies and ensemble learning. *Journal of Chemical Information and Modeling*, 60(8):4098–4107, 2020. doi: 10.1021/acs.jcim.0c00489. URL https://doi.org/10.1021/acs.jcim.0c00489. PMID: 32639729.

Vanessa E. Gray, Ronald J. Hause, Jens Luebeck, Jay Shendure, and Douglas M. Fowler. Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data. *Cell Systems*, 6(1): 116–124.e3, January 2018. ISSN 2405-4712. doi: 10.1016/j.cels.2017.11.003.

Magnús Halldór Gíslason, Henrik Nielsen, José Juan Almagro Armenteros, and Alexander Rosenberg Johansen. Prediction of gpi-anchored proteins with pointer neural networks. *Current Research in Biotechnology*, 3:6–13, 2021a. ISSN 2590-2628. doi: https://doi.org/10.1016/j.crbiot.2021.01.001. URL https://www.sciencedirect.com/science/article/pii/S2590262821000010.

Magnús Halldór Gíslason, Felix Teufel, José Juan Almagro Armenteros, Ole Winther, and Henrik Nielsen. Protein dataset partitioning pipeline. 2021b. URL https://github.com/graph-part/graph-part.

Daniel Hesslow, Niccoló Zanichelli, Pascal Notin, Iacopo Poli, and Debora Marks. RITA: a Study on Scaling Up Generative Protein Sequence Models. Technical Report arXiv:2205.05789, arXiv, May 2022. URL http://arxiv.org/abs/2205.05789. arXiv:2205.05789 [cs, q-bio] type: article.

Thomas A. Hopf, Anna G. Green, Benjamin Schubert, Sophia Mersmann, Charlotta P. I. Schärfe, John B. Ingraham, Agnes Toth-Petroczy, Kelly Brock, Adam J. Riesselman, Perry Palmedo, Chan Kang, Robert Sheridan, Eli J. Draizen, Christian Dallago, Chris Sander, and Debora S. Marks. The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics (Oxford, England)*, 35(9):1582–1584, May 2019. ISSN 1367-4811. doi: 10.1093/bioinformatics/bty862.

Chloe Hsu, Hunter Nisonoff, Clara Fannjiang, and Jennifer Listgarten. Learning protein fitness models from evolutionary and assay-labeled data. *Nature Biotechnology*, pp. 1–9, January 2022a. ISSN 1546-1696. doi: 10.1038/s41587-021-01146-5. URL https://www.nature.com/articles/s41587-021-01146-5. Publisher: Nature Publishing Group.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 8946–8970. PMLR, June 2022b. URL https://proceedings.mlr.press/v162/hsu22a.html. ISSN: 2640-3498.

Mingyang Hu, Fajie Yuan, Kevin K. Yang, Fusong Ju, Jin Su, Hui Wang, Fei Yang, and Qiuyang Ding. Exploring evolution-based & -free protein language models as protein function predictors, June 2022. URL http://arxiv.org/abs/2206.06583. Number: arXiv:2206.06583 arXiv:2206.06583 [cs, q-bio].

Justina Jankauskaite, Brian Jiménez-García, Justas Dapkunas, Juan Fernández-Recio, and Iain H. Moal. SKEMPI 2.0: an updated benchmark of changes in protein-protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics (Oxford, England)*, 35(3):462–469, February 2019. ISSN 1367-4811. doi: 10.1093/bioinformatics/bty635.

Anna Jarzab, Nils Kurzawa, Thomas Hopf, Matthias Moerch, Jana Zecha, Niels Leijten, Yangyang Bian, Eva Musiol, Melanie Maschberger, Gabriele Stoehr, Isabelle Becher, Charlotte Daly, Patroklos Samaras, Julia Mergner, Britta Spanier, Angel Angelov, Thilo Werner, Marcus Bantscheff, Mathias Wilhelm, Martin Klingenspor, Simone Lemeer, Wolfgang Liebl, Hannes Hahne, Mikhail M. Savitski, and Bernhard Kuster. Meltome atlas—thermal proteome stability across the tree of life. *Nature Methods*, 17(5):495–503, May 2020. ISSN 1548-7105. doi: 10.1038/s41592-020-0801-4. URL https://www.nature.com/articles/s41592-020-0801-4. Number: 5 Publisher: Nature Publishing Group.

Bowen Jing, Stephan Eismann, Pratham N. Soni, and Ron O. Dror. Equivariant Graph Neural Networks for 3D Macromolecular Structure. *arXiv:2106.03843 [cs, q-bio]*, July 2021a. URL http://arxiv.org/abs/2106.03843. arXiv: 2106.03843.

Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael J. L. Townshend, and Ron Dror. Learning from Protein Structure with Geometric Vector Perceptrons. *arXiv:2009.01411 [cs, q-bio, stat]*, May 2021b. URL http://arxiv.org/abs/2009.01411. arXiv: 2009.01411.

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

Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14):3059–3066, July 2002. ISSN 0305-1048. doi: 10.1093/nar/gkf436. URL https://doi.org/10.1093/nar/gkf436.

Michel van Kempen, Stephanie Kim, Charlotte Tumescheit, Milot Mirdita, Cameron L. M. Gilchrist, Johannes Soeding, and Martin Steinegger. Foldseek: fast and accurate protein structure search, September 2022. URL https://www.biorxiv.org/content/10.1101/2022.02.07.479398v4. Pages: 2022.02.07.479398 Section: New Results.

Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, IJCAI'95, pp. 1137–1143, San Francisco, CA, USA, August 1995. Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-363-9.

Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. Critical assessment of methods of protein structure prediction (casp)—round xiv. *Proteins: Structure, Function, and Bioinformatics*, 89(12):1607–1617, 2021.

Benjamin Kumwenda, Derek Litthauer, Özlem Tastan Bishop, and Oleg Reva. Analysis of protein thermostability enhancing factors in industrially important thermus bacteria species. *Evolutionary Bioinformatics*, 9:EBO.S12539, 2013. doi: 10.4137/EBO.S12539. URL https://doi.org/10.4137/EBO.S12539. PMID: 24023508.

Weizhong Li, Lukasz Jaroszewski, and Adam Godzik. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, 17(3):282–283, 2001.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, and Alexander Rives. Language models of protein sequences at the scale of evolution enable accurate structure prediction. preprint, Synthetic Biology, July 2022. URL `http://biorxiv.org/lookup/doi/10.1101/2022.07.20.500902`.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. Technical report, bioRxiv, July 2021. URL `https://www.biorxiv.org/content/10.1101/2021.07.09.450648v1`. Section: New Results Type: article.

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

Iain H. Moal and Juan Fernández-Recio. SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. *Bioinformatics*, 28(20):2600–2607, October 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts489. URL `https://doi.org/10.1093/bioinformatics/bts489`.

John Moult. A decade of casp: progress, bottlenecks and prognosis in protein structure prediction. *Current opinion in structural biology*, 15(3):285–289, 2005.

Erik Nijkamp, Jeffrey Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. ProGen2: Exploring the Boundaries of Protein Language Models, June 2022. URL `http://arxiv.org/abs/2206.13517`. Number: arXiv:2206.13517 arXiv:2206.13517 [cs, q-bio].

Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena-Hurtado, Aidan Gomez, Debora S. Marks, and Yarin Gal. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval, May 2022. URL `http://arxiv.org/abs/2205.13760`. Number: arXiv:2205.13760 arXiv:2205.13760 [cs].

William R Pearson and David J Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, 85(8):2444–2448, 1988.

Hassan Pezeshgi Modarres, Mohammad R. Mofrad, and Amir Sanati-Nezhad. Protdatatherm: A database for thermostability analysis and engineering of proteins. *PLOS ONE*, 13(1):1–9, 01 2018. doi: 10.1371/journal.pone.0191222. URL `https://doi.org/10.1371/journal.pone.0191222`.

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. doi: 10.1126/science.aao5167. URL `https://www.science.org/doi/10.1126/science.aao5167`. Publisher: American Association for the Advancement of Science.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S. Song. Evaluating Protein Transfer Learning with TAPE. *arXiv:1906.08230 [cs, q-bio, stat]*, June 2019. URL `http://arxiv.org/abs/1906.08230`. arXiv: 1906.08230.

Roshan M. Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. MSA Transformer. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8844–8856. PMLR, July 2021. URL `https://proceedings.mlr.press/v139/rao21a.html`. ISSN: 2640-3498.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Technical report, bioRxiv, December 2020. URL `https://www.biorxiv.org/content/10.1101/622803v4`. Section: New Results Type: article.

Gabriel J. Rocklin, Tamuka M. Chidyausiku, Inna Goreshnik, Alex Ford, Scott Houliston, Alexander Lemak, Lauren Carter, Rashmi Ravichandran, Vikram K. Mulligan, Aaron Chevalier, Cheryl H. Arrowsmith, and David Baker. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science (New York, N.Y.)*, 357(6347):168–175, July 2017. ISSN 1095-9203. doi: 10.1126/science.aan0693.

William P. Russ, Matteo Figliuzzi, Christian Stocker, Pierre Barrat-Charlaix, Michael Socolich, Peter Kast, Donald Hilvert, Remi Monasson, Simona Cocco, Martin Weigt, and Rama Ranganathan. An evolution-based model for designing chorismate mutase enzymes. *Science*, 369(6502):440–445, July 2020. doi: 10.1126/science.aba3304. URL `https://www.science.org/doi/full/10.1126/science.aba3304`. Publisher: American Association for the Advancement of Science.

Karen S Sarkisyan, Dmitry A Bolotin, Margarita V Meer, Dinara R Usmanova, Alexander S Mishin, George V Sharonov, Dmitry N Ivankov, Nina G Bozhanova, Mikhail S Baranov, Onuralp Soylemez, et al. Local fitness landscape of the green fluorescent protein. *Nature*, 533(7603):397–401, 2016.

Manish Kumar Tiwari, Ranjitha Singh, Raushan Kumar Singh, In-Won Kim, and Jung-Kul Lee. Computational approaches for rational design of proteins with novel functionalities. *Computational and structural biotechnology journal*, 2(3):e201204002, 2012.

Raphael J. L. Townshend, Martin Vögele, Patricia Suriana, Alexander Derry, Alexander Powers, Yianni Laloudakis, Sidhika Balachandar, Bowen Jing, Brandon Anderson, Stephan Eismann, Risi Kondor, Russ B. Altman, and Ron O. Dror. ATOM3D: Tasks On Molecules in Three Dimensions. *arXiv:2012.04035 [physics, q-bio]*, January 2022. URL `http://arxiv.org/abs/2012.04035`. arXiv: 2012.04035.

Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, Sameer Velankar, Gerard J. Kleywegt, Alex Bateman, Richard Evans, Alexander Pritzel, Michael Figurnov, Olaf Ronneberger, Russ Bates, Simon A. A. Kohl, Anna Potapenko, Andrew J. Ballard, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Ellen Clancy, David Reiman, Stig Petersen, Andrew W. Senior, Koray Kavukcuoglu, Ewan Birney, Pushmeet Kohli, John Jumper, and Demis Hassabis. Highly accurate protein structure prediction for the human proteome. *Nature*, 596 (7873):590–596, August 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03828-1. URL `https://www.nature.com/articles/s41586-021-03828-1`. Number: 7873 Publisher: Nature Publishing Group.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *arXiv:1706.03762 [cs]*, December 2017. URL `http://arxiv.org/abs/1706.03762`. arXiv: 1706.03762.

Bruce J Wittmann, Yisong Yue, and Frances H Arnold. Informed training set design enables efficient machine learning-assisted directed protein evolution. *Cell systems*, 12(11):1026–1045, 2021.

Nicholas C Wu, Lei Dai, C Anders Olson, James O Lloyd-Smith, and Ren Sun. Adaptation in protein fitness landscapes is facilitated by indirect paths. *Elife*, 5:e16965, 2016.

Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Chang Ma, Runcheng Liu, and Jian Tang. PEER: A Comprehensive and Multi-Task Benchmark for Protein Sequence Understanding, June 2022. URL `http://arxiv.org/abs/2206.02096`. Number: arXiv:2206.02096 arXiv:2206.02096 [cs].

A Appendix

A.1 EVE for protein families

Due to the stochastic training process, we train EVE on each fitness landscape using three different random seeds. The reported performance will thus be the average over the predictions using the three different representations for each sequence.

While EVE was originally used to predict variant effects of single wildtype proteins, it can be used on any MSA. This however means that the built-in preprocessing requires a reference wildtype sequence. This query sequence is then used to trim and otherwise clean the remaining sequences in the MSA. Since no single wildtype is representative for our protein families, we instead generate an artificial query sequence. Given the full-length MSA, we iterate through all of our labelled sequences (a minor part of the full MSA), and create a query sequence which has any amino acid (we chose 'A') at any position in the MSA, where any of the labelled sequences also have an amino acid. The remaining positions are filled with gaps. For example, say that sequences `--A-T-H` and `-AT--J-` are two labelled sequences from the MSA. The corresponding query sequence would thus be `-AA-AAA`. The query sequence is only used in the preprocessing, e.g., conserve the columns, where the labelled sequences have occupancy, and to remove columns where none do. The query sequence is thus not included in the actual model training. This results in a query-centric bias (Detlefsen et al., 2022) of our latent spaces which ensures the preservation of our limited labelled sequences. Alternative preprocessing is equally viable which can avoid the creation of the artificial query sequence.

A.2 Regressor hyperparameters

In each CV iteration, the three regressors were optimized via a grid search. The regressors were trained with all configurations on the training set, and the models providing the lowest mean squared error on the validation set was used to predict on the test set. The following hyperparameters were used:

- `KNeigborsRegressor()`: The number of neighbours was chosen among: 1, 2, 5, 10, 25.
- `Ridge(random_state=0)`: Regularization strength was chosen among: 0.0001, 0.001, 0.01, 0.1, 1, 10, 25
- `RandomForestRegressor(random_state=0, max_features="sqrt")`: Minimum samples to split was chosen among 2, 5.

We use the scikit-learn implementations of the regressors (Buitinck et al., 2013). The parameters not explicitly defined above were thus the default parameters. Several other grids for the three models were examined but provided no significant performance increases.

A.3 Supplementary experiments

Table 3: Misleading test results (Spearman correlation, *: Regression on both active and inactive proteins. **: Repeated random splitting. ***: Holdout validation.)

		KNN	**Ridge**	**RandomForest**
CM*	ESM-1B	0.58 ± 0.01	0.63 ± 0.01	0.64 ± 0.01
	ESM-2	0.59 ± 0.01	0.65 ± 0.01	**0.66** ± 0.01
	ESM-IF1	0.53 ± 0.01	0.57 ± 0.01	0.60 ± 0.01
	EVE	0.55 ± 0.01	0.53 ± 0.01	0.60 ± 0.01
	Evoformer (AF2)	0.47 ± 0.0	0.54 ± 0.02	0.59 ± 0.01
	MSA (1-HOT)	0.54 ± 0.01	0.62 ± 0.01	0.62 ± 0.01
PPAT**	ESM-1B	0.12 ± 0.03	0.22 ± 0.03	0.21 ± 0.04
	ESM-2	0.11 ± 0.05	0.20 ± 0.04	0.19 ± 0.02
	ESM-IF1	0.14 ± 0.04	0.17 ± 0.02	0.17 ± 0.04
	EVE	0.12 ± 0.03	0.11 ± 0.02	0.14 ± 0.02
	Evoformer (AF2)	0.13 ± 0.03	0.08 ± 0.01	**0.25** ± 0.02
	MSA (1-HOT)	0.08 ± 0.03	0.07 ± 0.02	0.10 ± 0.02
TIM***	ESM-1B	0.32	0.35	0.35
	ESM-2	**0.40**	0.33	**0.40**
	ESM-IF1	0.10	0.32	0.24
	EVE	0.17 ± 0.07	0.17 ± 0.04	-0.17 ± 0.05
	Evoformer (AF2)	0.19	**0.40**	0.28
	MSA (1-HOT)	0.24	0.10	0.05