NeurIPS 2023

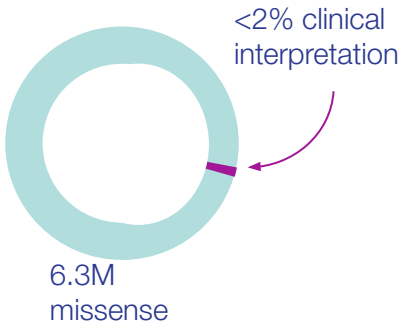
ProteinGym: Large-Scale Benchmarks for Protein Design and Fitness Prediction

Motivations

Accurately modeling the fitness landscape of protein sequences is critical to:

Mutation effects prediction

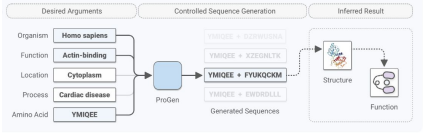
- The large majority of human variants[1] have no known interpretation



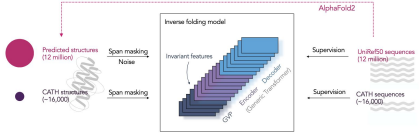
- Example: **EVE**[2], protein-specific alignment-based generative models for mutation effects prediction

Protein design

- Generating **novel** yet **fit** sequences, conditioning on:
  - **Labels**[3]



  - **Structure** (Inverse folding)[4,5]



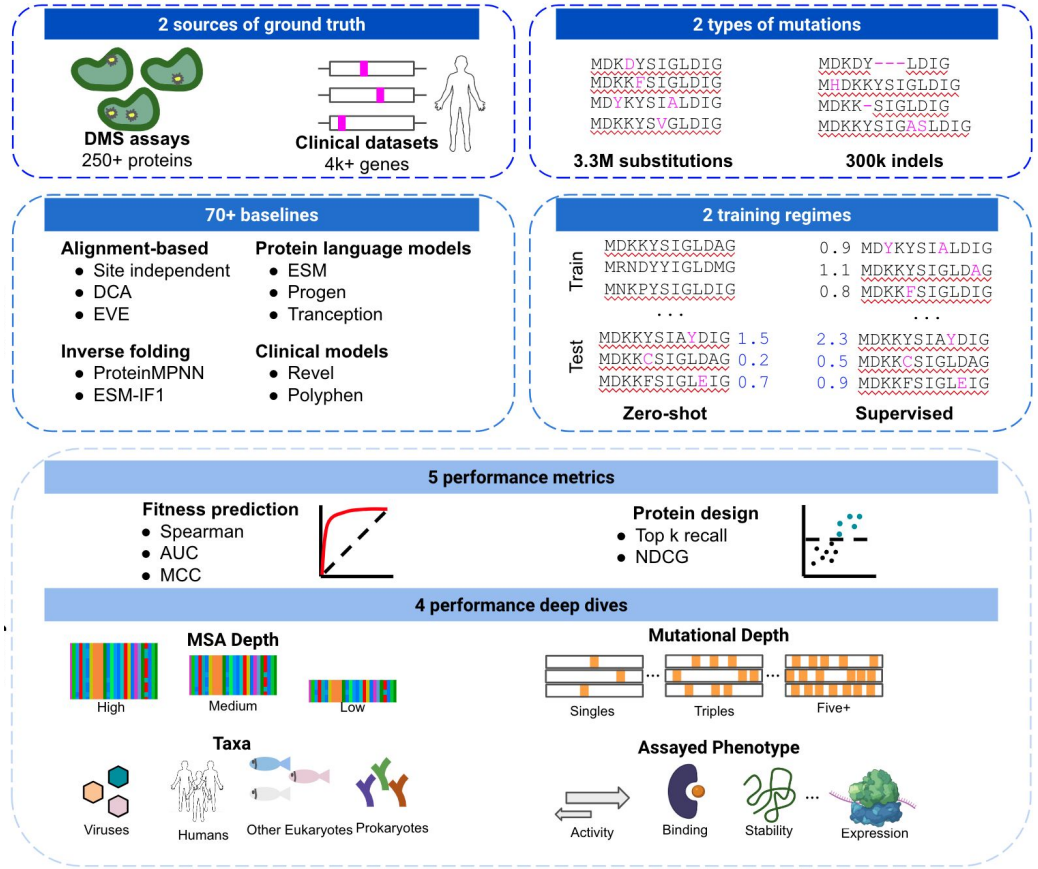
Challenges

- **A wide range of protein models** for fitness prediction and design have emerged in recent years (eg., alignment-based models, protein language models, inverse folding)
- **Prior protein benchmarks**[6,7] have been **critical** to support **initial** assessments, but are **limited to a handful of proteins**, and **there is** significant performance variation observed across assays[8]
- **Robust analysis** to drive the development of the **next generation of models** requires **scale**

1. Landrum & Kattman. ClinVar at five years: Delivering on the promise.
2. Frazer et al. Disease variant prediction with deep generative models of evolutionary data
3. Madani et al. ProGen: Language Modeling for Protein Generation.
4. Ingraham et al. Generative Models for Graph-Based Protein Design.
5. Hsu et al. Learning inverse folding from millions of predicted structures. 2022
6. Rao et al., Evaluating Protein Transfer Learning with TAPE
7. Dallago et al. FLIP: Benchmark tasks in fitness landscape inference for proteins
8. Riesselman et al., Deep generative models of genetic variation capture the effects of mutations

Overview of the ProteinGym benchmarks

1. **Datasets**
2. **Models**
3. **Performance metrics**
4. **Deep dives**



1 Two types of datasets to serve as ground truth in ProteinGym

Deep mutational scanning (DMS) assays

- **Large number of labels** (2.8M) for a limited number of proteins (200+)
- Labels are **experimentally determined**

Clinical datasets

- **Sparse collection of labels** (60k+) for a large number of proteins (3k+)
- Labels are **based on manual annotation from clinical experts**

Dataset	Description	Mutation type	# Proteins	# Mutants
DMS	High-throughput assays evaluating the functional impact of a wide range of protein mutations	Substitutions Indels	217 66	2.5M 0.3M
Clinical	Expert-curated clinical annotations across a wide range of human genes	Substitutions Indels	2,525 1,555	63k 3k
Total			3,422	2.8M

2 We implemented / compiled scores for 70+ baselines across two different model training regimes

70+ Baselines

- **Alignment-based** (e.g., DCA, EVE)
- **Protein language models** (e.g., ESM, RITA, Progen)
- **Hybrid models** (e.g., Tranception/TranceptEVE)
- **Inverse folding** (e.g., ProteinMPNN, ESM-IF1)
- **Clinical effect predictors** (e.g., PolyPhen-2, REVEL)

2 core training regimes

- **Zero-shot:** labels are only used for evaluation
- **Supervised:** labels used for training & evaluation → We created various cross validation schemes to assess ability to extrapolate across positions

3 We report 5 performance metrics to assess the ability of the various baselines to support fitness prediction of design initiatives

Fitness-focused metrics

- Spearman, AUC & MCC
- Assess overall performance of the model to classify / rank order all possible mutants

Design-focused metrics

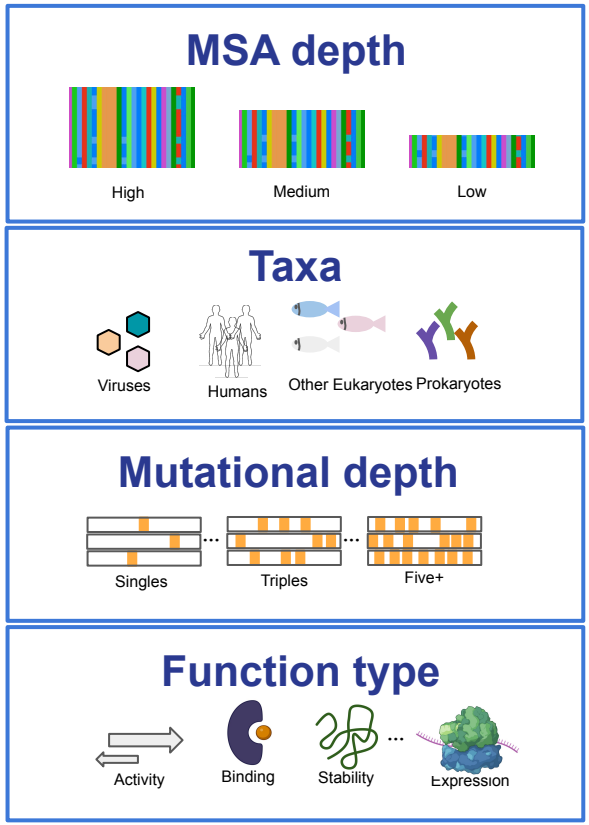
- NDCG & Recall
- Quantify the ability of the model to properly identify the top mutants for the phenotype of interest

Example: DMS zero-shot substitution benchmark

Model type	Model name	Spearman	AUC	MCC	NDCG	Recall
Alignment-based models	Site independent	0.35	0.692	0.278	0.731	0.195
	WaveNet	0.212	0.624	0.172	0.676	0.155
	EVmutation	0.39	0.715	0.301	0.762	0.217
	DeepSequence (ens.)	0.403	0.723	0.316	0.758	0.219
	EVE (ens.)	0.431	0.738	0.334	0.768	0.226
	GEMME	0.445	0.745	0.341	0.764	0.208
Protein language models	UniRep	0.166	0.595	0.131	0.63	0.135
	ESM-1b	0.381	0.714	0.298	0.731	0.196
	ESM2 (15B)	0.400	0.723	0.312	0.746	0.206
	RITA (ens.)	0.365	0.705	0.286	0.735	0.198
	ESM-1v (ens.)	0.366	0.720	0.309	0.734	0.207
	ProGen2 (ens.)	0.385	0.716	0.302	0.747	0.202
	VESPA	0.437	0.746	0.345	0.764	0.202
	CARP (640M)	0.353	0.696	0.273	0.727	0.194
Inverse Folding	ProteinMPNN	0.244	0.634	0.184	0.698	0.182
	ESM-IF1	0.405	0.722	0.315	0.728	0.216
	MIF-ST	0.389	0.712	0.298	0.750	0.219
Hybrid models	UniRep (evotuned)	0.324	0.700	0.257	0.720	0.176
	MSA Transformer (ens.)	0.427	0.745	0.333	0.766	0.223
	Tranception L	0.421	0.753	0.329	0.764	0.216
	TranceptEVE	**0.445**	**0.767**	**0.346**	**0.772**	**0.227**

Table 2: **ProteinGym - Zero-shot substitution DMS benchmark** Average Spearman's rank correlation, AUC, MCC, NDCG@10%, and top 10% recall between model scores and experimental measurements on the ProteinGym substitution benchmark. We use 'ens.' as a shorthand for ensemble.

4 Several deep dives allow us to assess the relative benefits of various architectures in different settings



Example: DMS zero-shot substitution performance by MSA depth

Model type	Model name	Spearman by MSA depth (↑) Low	Medium	High	All
Alignment-based models	Site-Independent	0.405	0.376	0.353	0.350
	WaveNet	0.276	0.372	0.489	0.212
	EVmutation	0.386	0.403	0.487	0.390
	DeepSequence (ensemble)	0.364	0.407	0.535	0.403
	EVE (ensemble)	0.408	0.44	0.532	0.431
	GEMME	0.418	0.45	0.508	0.445
Protein language models	UniRep	0.167	0.153	0.178	0.166
	ESM-1b	0.352	0.326	0.493	0.381
	ESM2 (15B)	0.370	0.376	0.440	0.400
	RITA (ensemble)	0.330	0.412	0.410	0.365
	ESM-1v (ensemble)	0.370	0.381	0.533	0.394
	ProGen2 (ensemble)	0.363	0.419	0.463	0.385
	VESPA	0.425	0.431	0.548	0.437
Hybrid models	UniRep evotuned	0.300	0.360	0.387	0.324
	MSA Transformer (ensemble)	0.377	0.432	0.514	0.427
	Tranception L	0.416	0.433	0.504	0.421
	TranceptEVE	0.432	0.461	0.543	0.445

Table A5: **ProteinGym - Zero-shot substitution DMS benchmark by MSA depth** Average Spearman's rank correlation between model scores and experimental measurements by MSA depth on the ProteinGym substitution benchmark. Alignment depth is measured by the ratio of the effective number of sequences N_{eff} in the MSA, following Hopf et al. [2017], by the length covered L (Low: $N_{\text{eff}}/L < 1$; Medium: $1 < N_{\text{eff}}/L < 100$; High: $N_{\text{eff}}/L > 100$)

A few insights that emerged from our analyses

For mutation effect prediction, SOTA performance still necessitates the use of alignments	• All **protein language models** of single-sequence input are currently **relatively far from SOTA** • The best performance is achieved by **hybrid** models (Tranception, TranceptEVE) or **alignment-based** models (GEMME, EVE, VESPA)
While they do not perform very well in aggregate, inverse folding models achieve the best performance on stability assays	• Certain modeling biases are best adapted to predicting specific properties • For a deeper analysis on this, you may want to check our workshop paper **"Combining Structure and Sequence for Superior Fitness Prediction"** to be presented at the **MLSB** and **GenBio** workshops
The best zero-shot fitness models rival their supervised counterparts on the clinical benchmarks	• The best **zero-shot baselines** (eg., TranceptEVE, EVE) perform **on par** with the best supervised baselines on the clinical benchmarks, **without being subject to the same label biases**

Resources to get started with ProteinGym

GitHub repo

github.com/OATML-Markslab/ProteinGym

- **Models:** all code for running zero-shot and supervised baselines
- **Metrics:** all code to compute performance metrics and the various deep dives
- **Data:** DMS assays (raw & processed files), model scores for all 2.8M mutants, Multiple Sequence Alignments, predicted 3D structures, processed ClinVar & gnomAD datasets

Website

www.proteingym.org/home

- **Performance summaries:** DMS Vs clinical benchmarks; for zero-shot vs supervised; for substitutions vs indels
- **Performance deep dives:** DMS level, by segmentation variable (eg., MSA depth, taxa, function grouping)
- **Quick links to resources** (paper & GitHub)

See you at NeurIPS!

Poster - Great Hall & Hall B1+B2 #326

Thanks to the broader ProteinGym team...


Pascal


Aaron


Daniel


Lood


Steffanie


Han


Nathan


Ada

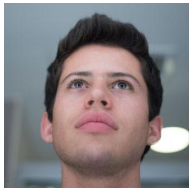

Ruben


Jonny


Mafalda


Dinko


Rose


Yarin


Debbie

& our sponsors!

GSK

CHAN ZUCKERBERG INITIATIVE

UKRI Engineering and Physical Sciences Research Council

NIH National Institutes of Health

The Alan Turing Institute