nature methods

SUPPLEMENTARY INFORMATION

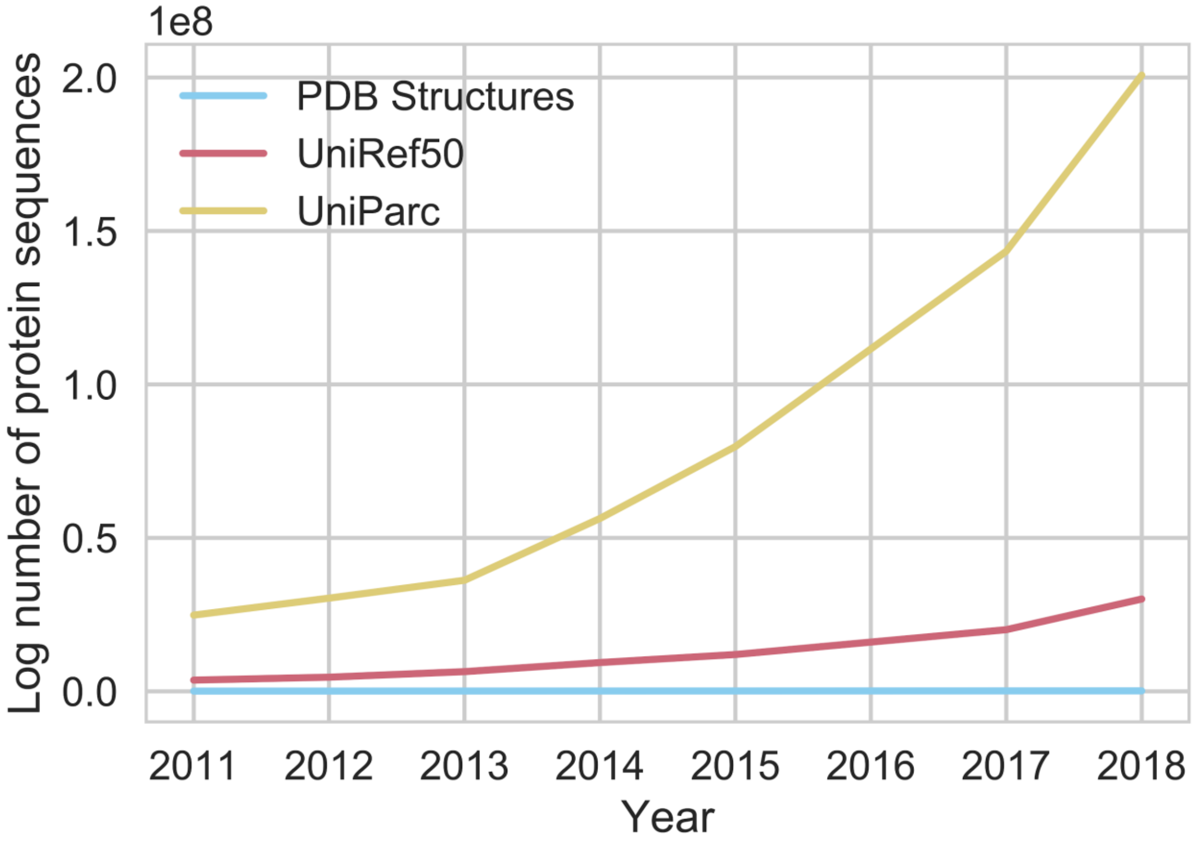
https://doi.org/10.1038/s41592-019-0598-1

In the format provided by the authors and unedited.

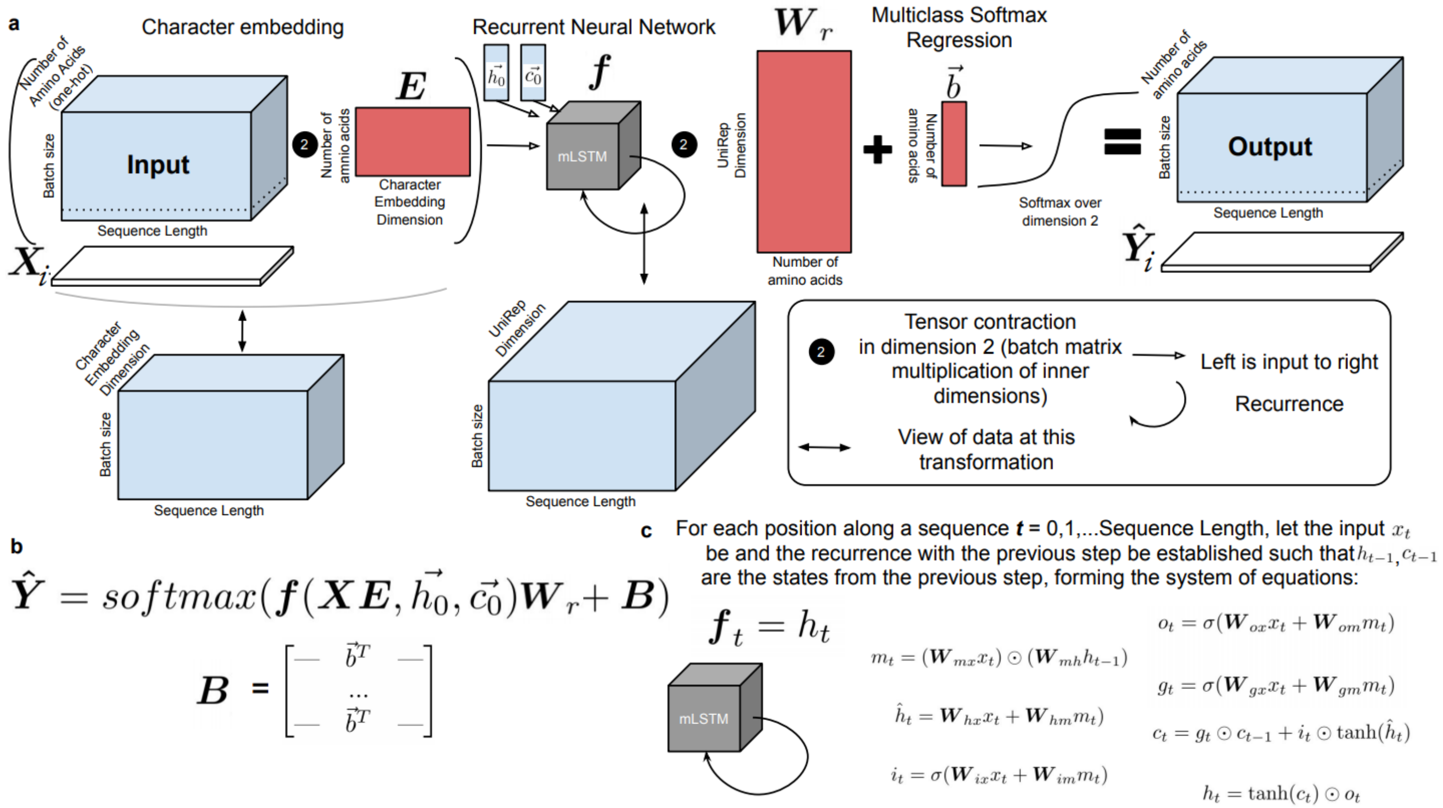
Unified rational protein engineering with sequence-based deep representation learning

Ethan C. Alley[1,2,6], Grigory Khimulya[6,7], Surojit Biswas[1,3,6], Mohammed AlQuraishi[4] and George M. Church[1,5]*

[1]Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA, USA. [2]MIT Media Laboratory, Cambridge, MA, USA. [3]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [4]Department of Systems Biology, Harvard Medical School, Boston, MA, USA. [5]Department of Genetics, Harvard Medical School, Boston, MA, USA. [6]These authors contributed equally: Ethan C. Alley, Grigory Khimulya, Surojit Biswas. [7]Unaffiliated: Grigory Khimulya. *e-mail: gchurch@genetics.med.harvard.edu



Supplemental Figure 1. Growth in sequence databases.

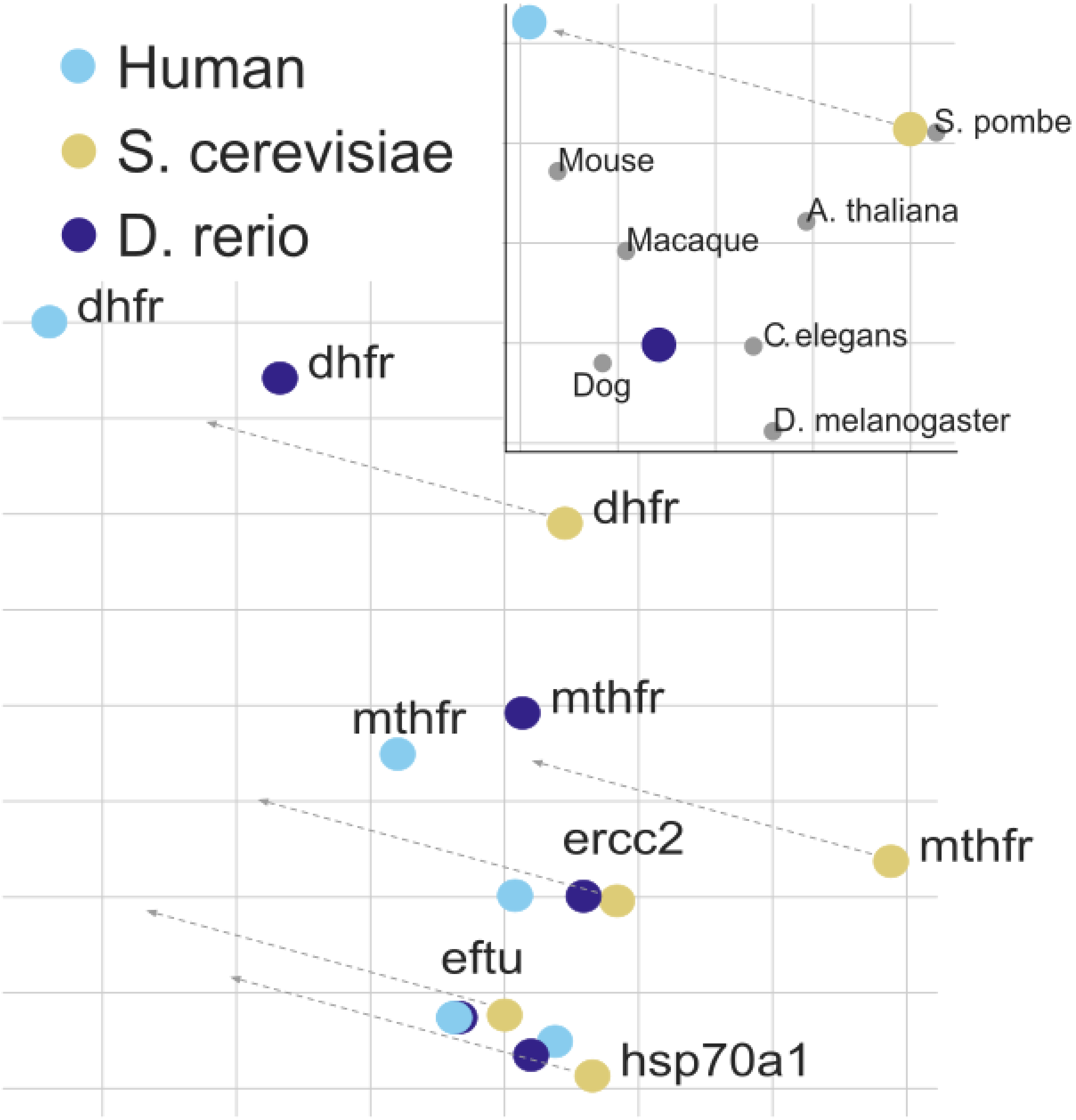

a
Character embedding
Number of Amino Acids (one-hot)
Batch size
Input
Sequence Length
X_i
2
Number of amino acids
E
Character Embedding Dimension
Recurrent Neural Network
$\vec{h_0}$
$\vec{c_0}$
f
mLSTM
2
W_r
UniRep Dimension
Number of amino acids
Multiclass Softmax Regression
+
$\vec{b}$
Number of amino acids
Softmax over dimension 2
=
Number of amino acids
Batch size
Output
Sequence Length
$\hat{Y}_i$
Character Embedding Dimension
Batch size
Sequence Length
UniRep Dimension
Batch size
Sequence Length
2 Tensor contraction in dimension 2 (batch matrix multiplication of inner dimensions)
Left is input to right
Recurrence
View of data at this transformation
b
$\hat{Y} = softmax(f(XE, \vec{h_0}, \vec{c_0})W_r + B)$
$B = \begin{bmatrix} — & \vec{b}^T & — \\ & \dots & \\ — & \vec{b}^T & — \end{bmatrix}$
c
For each position along a sequence t = 0,1,...Sequence Length, let the input x_t be and the recurrence with the previous step be established such that h_{t-1}, c_{t-1} are the states from the previous step, forming the system of equations:
$f_t = h_t$
mLSTM
$m_t = (W_{mx}x_t) \odot (W_{mh}h_{t-1})$
$\hat{h}_t = W_{hx}x_t + W_{hm}m_t)$
$i_t = \sigma(W_{ix}x_t + W_{im}m_t)$
$o_t = \sigma(W_{ox}x_t + W_{om}m_t)$
$g_t = \sigma(W_{gx}x_t + W_{gm}m_t)$
$c_t = g_t \odot c_{t-1} + i_t \odot \tanh(\hat{h}_t)$
$h_t = \tanh(c_t) \odot o_t$

Supplemental Figure 2. Model architecture. **a**. Diagram of information flow in the UniRep representation learning model. Transformations of data are shown in blue, learned variables/ parameters are in red. A batch of amino acid sequences is padded with zeros and one-hot encoded to form input (left). This multiplies a learned character embedding matrix on the left, which can also be seen as a lookup operation to replace each one-hot amino acid vector with the corresponding character embedding (see double-sided arrows to intermediate tensor below). This is input to the mLSTM, shown here as a black box, along with hidden states h and c, initially set to all zeros. The output of the mLSTM (shown below with double sided arrow) then multiplies learned weights $\mathbf{W}_r$ on the left and is summed with a learned bias b to produce logits, which are softmaxed. This output prediction tensor is a (batch size, sequence length, number of amino acids) tensor where the 2nd dimension, zero indexed, sums to one due to the softmax, and represents prediction confidences for the next amino acid. Note that for sequences longer than 128, we use the standard approach of truncated backpropagation through time (Methods). **b**. Mathematical description of $\hat{\mathbf{Y}}$ in a) shown for simplification on a single sequence (visualized as a slice of the input input tensors $\mathbf{X}_i$ and $\hat{\mathbf{Y}}_i$). The function $\boldsymbol{f}$ is defined as the fully unrolled application of the recurrent function $\boldsymbol{f}_t$ defined in b). This expression for a single sequence is repeated over a batch of sequences to produce the output tensor. **c**. Opening up the black box from a) to fully define the recurrent mLSTM function $\boldsymbol{f}_t$ which takes state vectors h_{t-1} and c_{t-1} with a character-embedded amino acid vector x_t and produces state vectors h_t and c_t output $\boldsymbol{f}_t = h_t$. For a complete treatment, theoretical motivations and performance characteristics of the mLSTM, see Krause et. al (2016)[61].

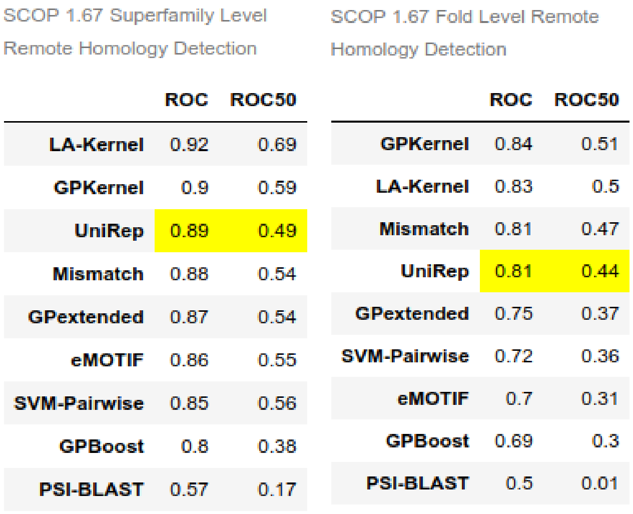
Organism Short Name	**Domain**	**Specific Name**
Halobacterium salinarum	Archaea	Archaea
Haloferax volcanii	Archaea	Archaea
Methanococcus maripaludis	Archaea	Archaea
Methanosarcina acetivorans	Archaea	Archaea
Sulfolobus solfataricus	Archaea	Archaea
Thermococcus kodakarensis	Archaea	Archaea
Escherichia coli (K12)	Bacteria	Bacteria
Aliivibrio fischeri	Bacteria	Bacteria
Azotobacter vinelandii	Bacteria	Bacteria
Bacillus subtilis (168)	Bacteria	Bacteria
Cyanothece (PCC7822)	Bacteria	Cyanobacteria
Mycoplasma genitalium	Bacteria	Bacteria
Mycobacterium tuberculosis	Bacteria	Bacteria
Prochlorococcus marinus	Bacteria	Cyanobacteria
Streptomyces coelicolor (A32)	Bacteria	Bacteria
Synechocystis (PCC 6803 Kazusa)	Bacteria	Cyanobacteria
Caenorhabditis elegans	Eukarya	Animalia
Drosophila melanogaster (fruit fly)	Eukarya	Animalia
Homo sapiens (human)	Eukarya	Mammalia
Mus musculus (mouse)	Eukarya	Mammalia
Saccharomyces cerevisiae (yeast)	Eukarya	Fungi
Anolis carolinensis	Eukarya	Animalia
Aspergillus nidulans	Eukarya	Fungi
Arabidopsis thaliana	Eukarya	Plantae
Cavia porcellus (guinea pig)	Eukarya	Mammalia
Gallus gallus domesticus (chicken)	Eukarya	Animalia
Coprinopsis cinerea	Eukarya	Fungi
Bos taurus (cow)	Eukarya	Mammalia
Chlamydomonas reinhardtii	Eukarya	Eukarya
Cryptococcus neoformans	Eukarya	Fungi

Canis familiaris (dog)	Eukarya	Mammalia
Emiliania huxleyi	Eukarya	Eukarya
Macaca mulatta (macaque)	Eukarya	Mammalia
Zea mays (corn)	Eukarya	Plantae
Heterocephalus glaber (naked mole rat)	Eukarya	Mammalia
Neurospora crassa	Eukarya	Fungi
Oryzias latipes	Eukarya	Animalia
Physcomitrella patens	Eukarya	Plantae
Columba liva (pigeon)	Eukarya	Animalia
Sus scrofa (pig)	Eukarya	Mammalia
Pristionchus pacificus	Eukarya	Animalia
Oryza sativa (Rice strain japonica)	Eukarya	Plantae
Schizosaccharomyces pombe	Eukarya	Fungi
Tetrahymena thermophila	Eukarya	Eukarya
Thalassiosira pseudonana	Eukarya	Eukarya
Ustilago maydis (corn smut)	Eukarya	Fungi
Xenopus tropicalis	Eukarya	Animalia
Danio rerio	Eukarya	Animalia
Phage lambda	Virus	Virus
SV40	Virus	Virus
T4 phage	Virus	Virus
T7 phage	Virus	Virus
Vaccinia virus (Copenhagen)	Virus	Virus

Supplemental Table 1. Reference proteomes used in the organism analysis in Fig. 2b. Species name and common/ subspecies/ strain in parentheses.



Supplemental Figure 3. Single-protein vector arithmetic in UniRep representation space. We suspected our organism vector clustering success may be explained by learning a measure of proteome content (e.g. abundance of various protein types). Surprisingly, after sourcing n=5 proteins conserved across n=3 model organisms (n=15 proteins total), we identify a common direction of variance, from Baker's Yeast to Human in the PCA projection space, which corresponds to the vector from Yeast to Human proteome representations, suggesting that organisms may have an arithmetic relationship in the representation space similar to that observed in Word Vectors [84]. Note that the direction of the vectors is invariant from the PCA in the upper right to the bottom left, but the length of the vector is meaningless in the bottom left. PC1 is the x-axis of both plots, PC2 is the y axis of both plots.



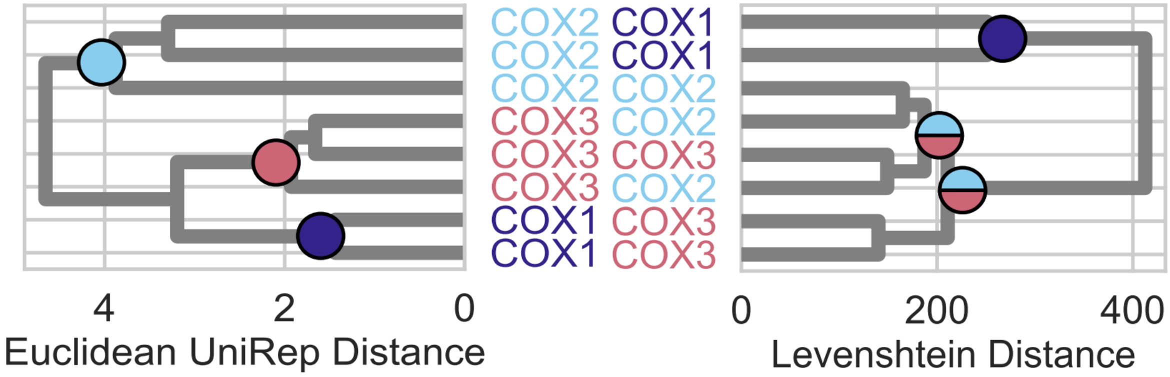
SCOP 1.67 Superfamily Level Remote Homology Detection

	ROC	ROC50
LA-Kernel	0.92	0.69
GPKernel	0.9	0.59
UniRep	0.89	0.49
Mismatch	0.88	0.54
GPextended	0.87	0.54
eMOTIF	0.86	0.55
SVM-Pairwise	0.85	0.56
GPBoost	0.8	0.38
PSI-BLAST	0.57	0.17

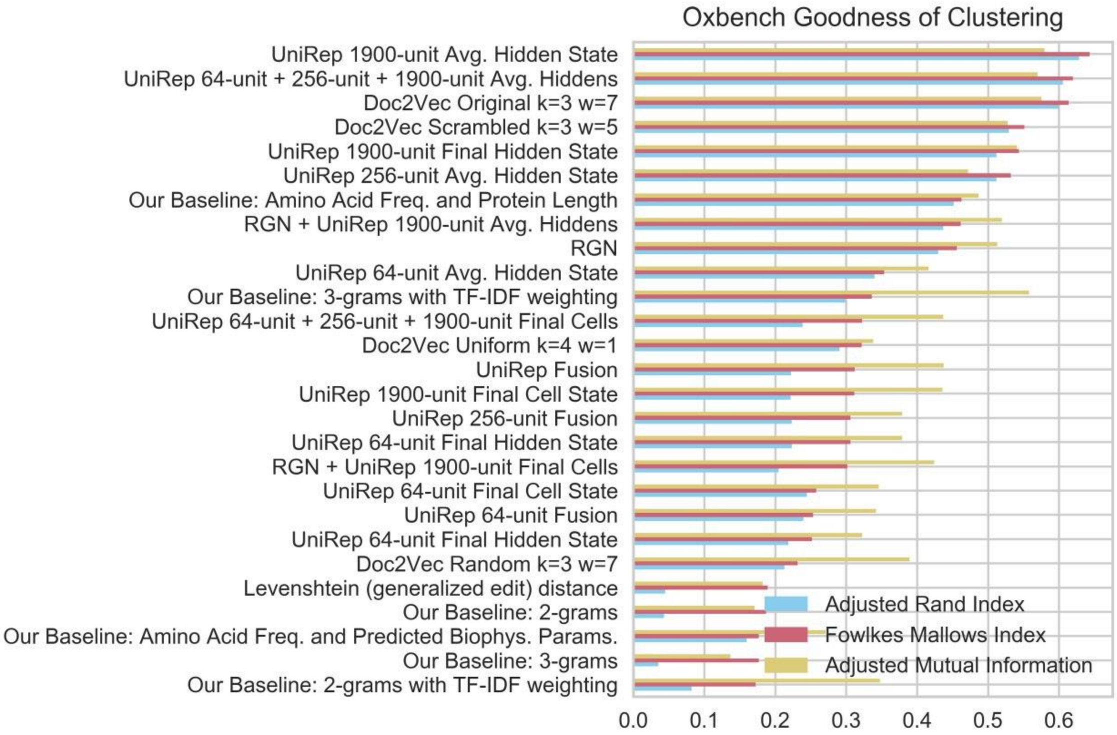
SCOP 1.67 Fold Level Remote Homology Detection

	ROC	ROC50
GPKernel	0.84	0.51
LA-Kernel	0.83	0.5
Mismatch	0.81	0.47
UniRep	0.81	0.44
GPextended	0.75	0.37
SVM-Pairwise	0.72	0.36
eMOTIF	0.7	0.31
GPBoost	0.69	0.3
PSI-BLAST	0.5	0.01

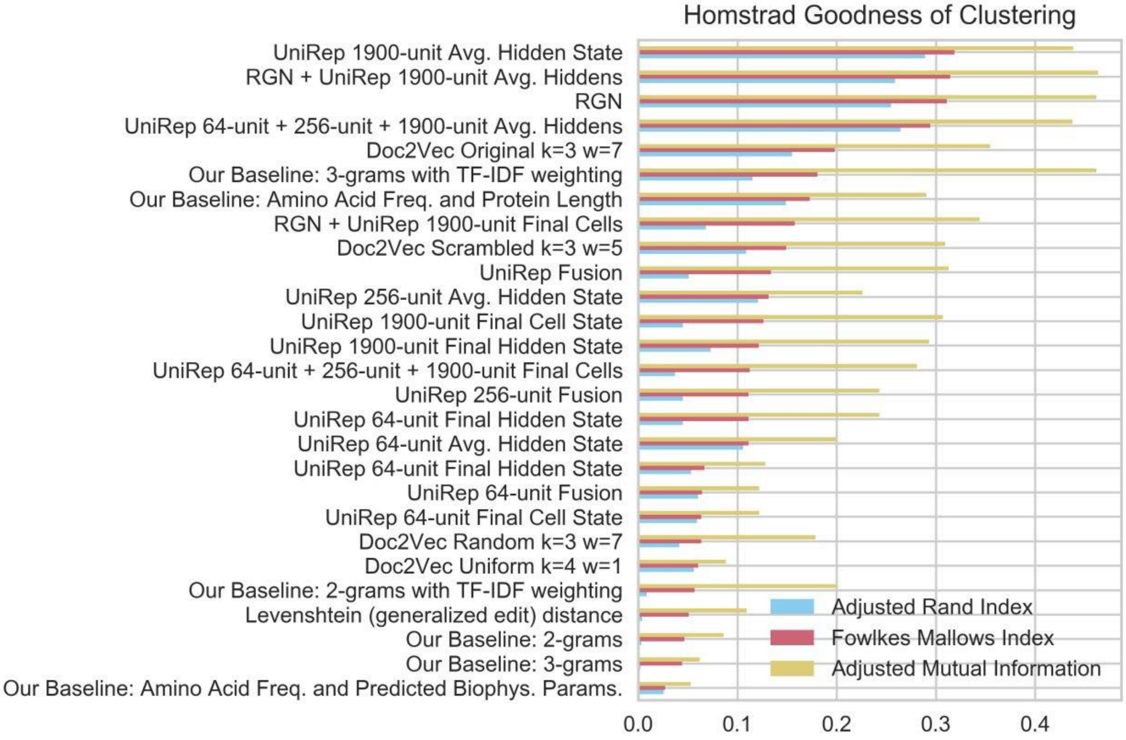
Supplemental Table 2. Unirep achieves competitive results on homology detection as measured by ROC-AUC and ROC50-AUC (sorted by ROC score). UniRep with RandomForest top model with Bayesian Hyperparameter optimization (Methods) achieves competitive performance with published sequence-only remote homology detection methods (Håndstad, 2007) on two most frequently used benchmark datasets (dataset sizes in Supp. Table 7).



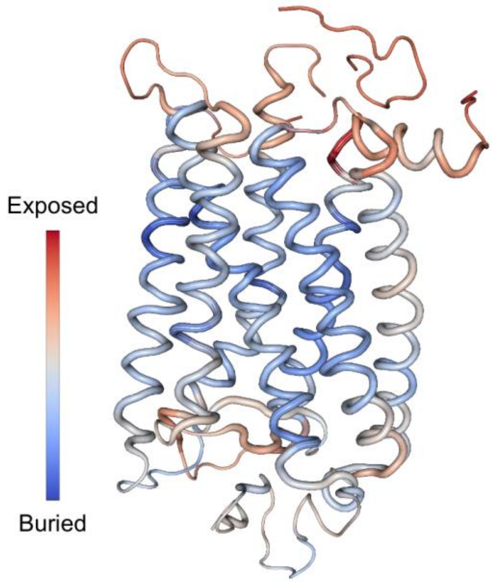
Supplemental Figure 4. Euclidean distance in UniRep space resolves protein clusters where generalized minimum edit distance fails. A representative clustering of Cytochrome-oxidase family enzymes (COX) 1, 2, and 3 from HOMSTRAD (Fig. 3e) further illustrates this result by reconstructing the correct monophyletic grouping of the true labels, where a sequence distance-based clustering fails.



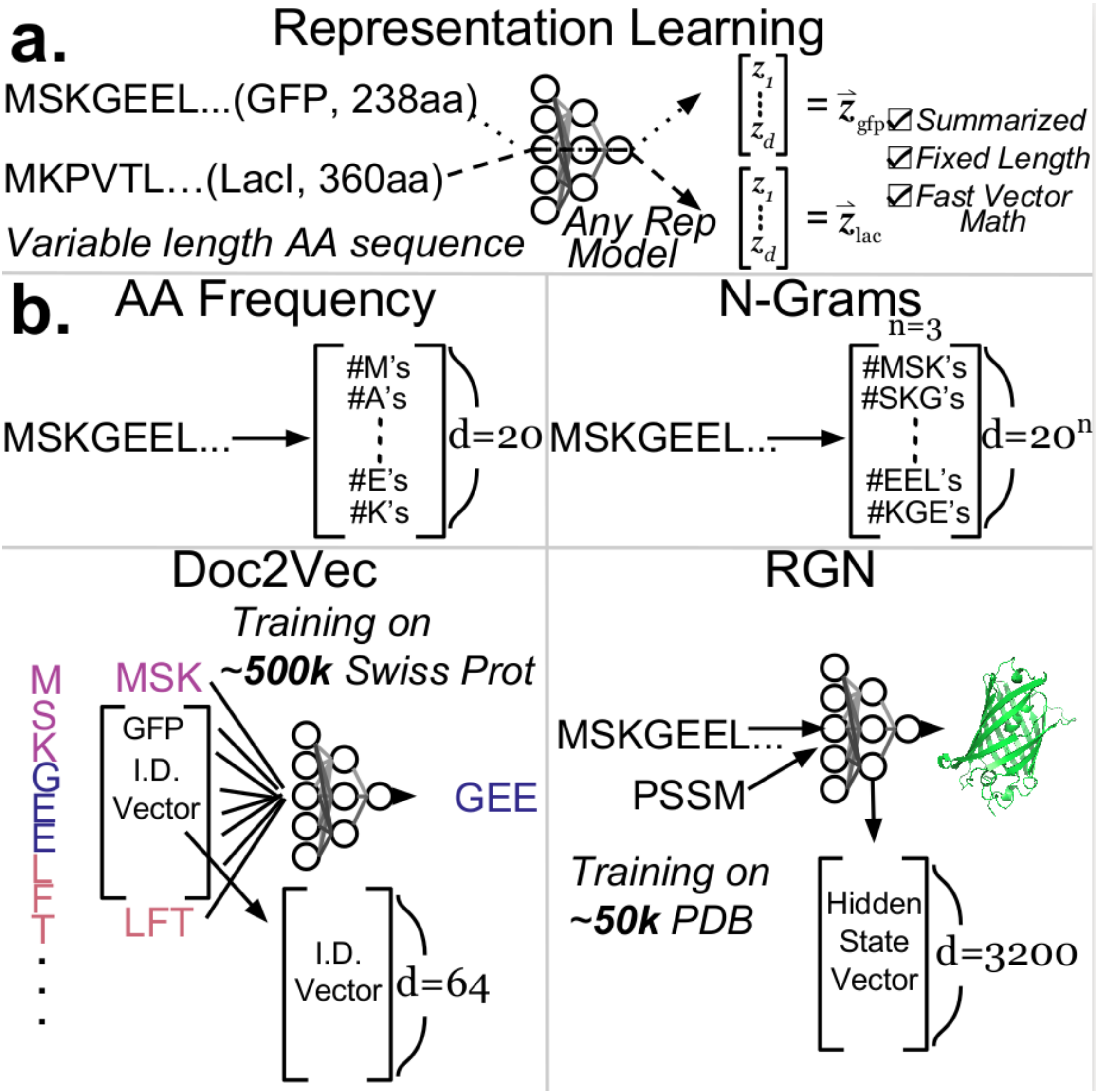
Supplemental Figure 5. OXBench unsupervised homology detection task, all results.



Supplemental Figure 6. HOMSTRAD unsupervised homology detection task, all representations results.



Supplemental Figure 7. Learned from a collection of PDB secondary structures, a linear combination of hidden state neurons identifies solvent accessible regions in the structure of Bovine Rhodopsin GPCR (PDB:1F88, left) oriented with the extracellular domain upwards (Methods; Pearson *r*=0.38). Pearson correlation was calculated between DSSP-computed solvent accessibility scores and the hidden state linear predictor across n=348 residues.



Supplemental Figure 8. Baseline representations. Representation learning inputs primary amino acid sequences and outputs fixed length vector representations. **a.** Schematic of representation learning in general. **b.** 4 baseline methods illustrated: simply counting amino acid occurrences (upper left) and occurrences of k-mers (upper right), Doc2Vec embeddings learned by Feed-Forward prediction of a central k-mer given the external context k-mers (lower left) [21], Recurrent Geometric Network hidden state, learned by recurrent processing of input sequences to predict crystal structure (lower right) [16].

	Representations being concatenated								
	UniRep State								RGN
	64-unit			256-unit		1900-unit			
Name	Avg. Hidden	Final Hidden	Final Cell	Avg. Hidden	Final Cell	Avg. Hidden	Final Hidden	Final Cell	
UniRep 64-unit Fusion	x	x	x						
UniRep 256-unit Fusion				x	x				
UniRep Fusion						x	x	x	
UniRep 64, 256, 1900-unit Avg. Hiddens	x			x		x			
64, 256, 1900-unit Final Cells			x		x			x	
RGN + UniRep 1900-unit Avg. Hiddens						x			x
RGN + UniRep 1900-unit Final Cells								x	x

Supplemental Table 3. Representation fusions (concatenations) analyzed.

	EEHEE	**HEEH**	**EHEE**	**HHH**
UniRep	**0.698**	**0.434**	**0.641**	**0.719**
Rosetta	0.597	0.057	0.472	0.596
Buried NPSA	0.544	0.044	0.509	0.632
Exposed NPSA	0.549	0.089	0.451	0.060

Supplemental Table 4. Stability prediction performance for *de novo* designed small proteins from Rocklin *et al.* (2017) for 4 different fold topologies (dataset sizes can be found in Supp. Table 7). Spearman correlation of UniRep-based stability predictions, Rosetta scores and structural/ biophysical quantities with experimental stability measurements. The *de novo* designed proteins are split by fold topology and a new UniRep top model was trained on the training set of each topology subset. Spearman correlation is computed for the held-out test set of each fold topology.

Performance of representations on 15 prediction tasks (Mean Squared Error) & stability ranking task (Spearman Correlation)- test

	Cytochrome P450 Thermostability	Rhodopsin Peak Absorption Wavelength	Epoxide Hydrolase Enantio selectivity	Channel rhodopsin Membrane Localization	TEM-1 Beta-lactamase	Ubiquitin (E1 Activity)	Protein G (IgG domain)
UniRep Fusion	**15.8**	**499**	189	1.25	**0.0545*****	**0.0421*****	**0.0233*****
Our Best Baseline	21.7	571	**93.2**	**0.912**	0.074	0.052	0.054
RGN	24.4	>2000	>1000	1.61	0.0904	0.054	0.0977
Best Doc2Vec	18.1	530	95.7	1	0.0881	0.0625	0.0724

	HSP90	Amino glycosidase (Kka2)	Pab1 (RRM domain)	PSD95 (Pdz3 domain)	Ubiquitin	Yap65 (WW domain)	Stability: 17 DMS datasets combined	Stability: *De Novo* Designed mini proteins	Spearman Rank Correlation Stability: *De Novo* Design Rounds
UniRep Fusion	**0.0258*****	**0.11****	**0.0265*****	**0.0208*****	**0.0323*****	**0.0415*****	**0.0304*****	**0.179*****	**ρ=0.59*****
Our Best Baseline	0.0344	0.115	0.0435	0.041	0.0515	0.0662	0.0398	0.201	-
RGN	0.0579	0.14	0.0596	0.0438	0.0601	0.0639	0.0338	0.189	-
Best Doc2Vec	0.0579	0.132	0.0495	0.046	0.064	0.0772	0.0473	0.258	-
								Rosetta Total Energy	ρ=0.42

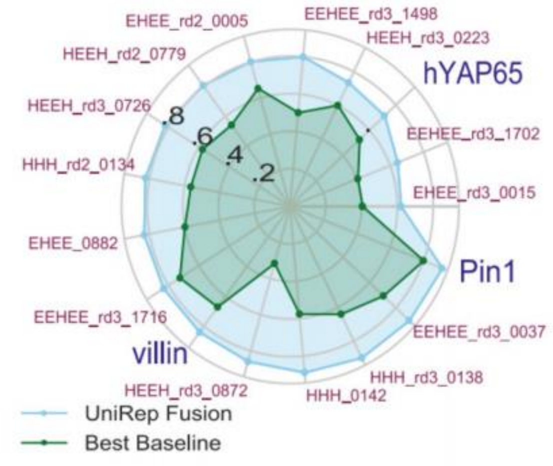
Supplemental Table 5. Regression results - test set metrics, with lowest MSE model or model class compared to the 2nd lowest MSE model or model class *p < 0.05, **p < 0.01, ***p<0.001 (Welch's two-tailed t-test for significance), standard deviations obtained through n=30 50% validation/test set resampling. All values are MSE except where specified. Validation set metrics can be found in Supp. Table 6. This table includes an extension of our analysis to 4 small datasets compiled previously for protein phenotype prediction using Doc2Vec representations (first 4 columns) [21]. We observed widely variable results and statistically insignificant results (caused by underpowered validation and test set), with UniRep or one of the baselines we developed outperforming previous state-of-the-art [21] (here and in Supp. Table 6), which underscored the importance of adequate data size for accurate estimation of performance.

Performance of representations on 15 prediction tasks (Mean Squared Error) & stability ranking task (Spearman Correlation)- validation

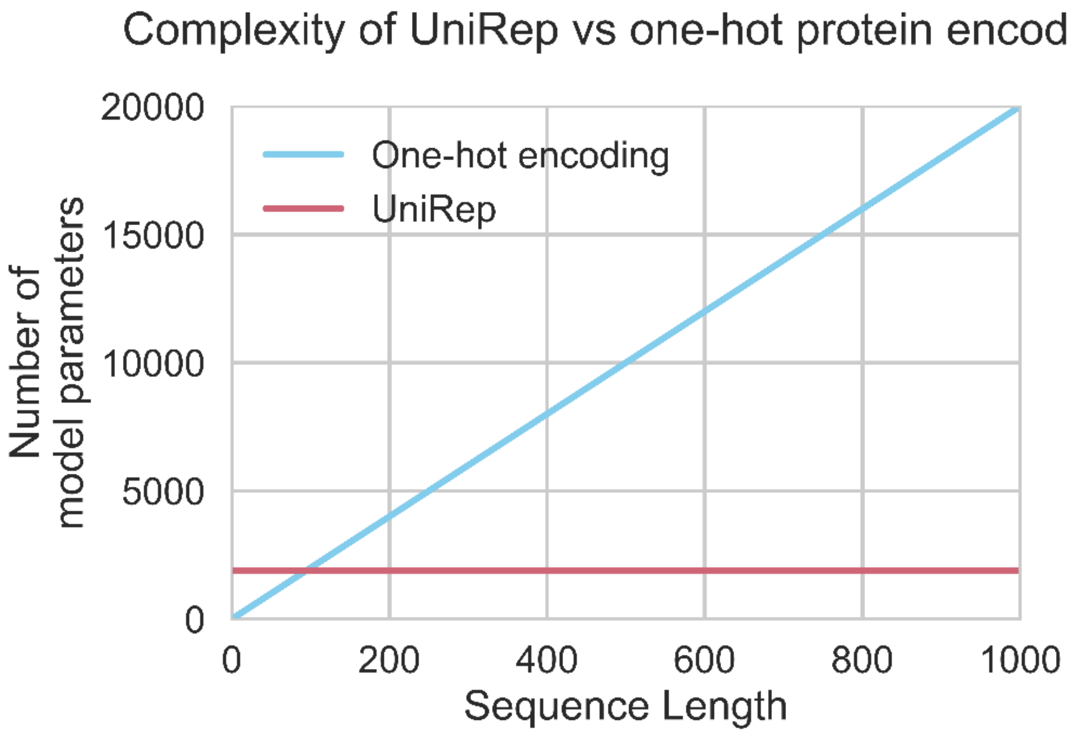
	Cytochrome P450 Thermostability	Rhodopsin Peak Absorption Wavelength	Epoxide Hydrolase Enantio selectivity	Channel rhodopsin Membrane Localization	TEM-1 Beta-lactamase	Ubiquitin (E1 Activity)	Protein G (IgG domain)
UniRep Fusion	8.3	130	444	1.47	**0.0471*** **	**0.0274*** **	**0.0307** **
Our Best Baseline	**8.27**	**79.4** **	**219*** **	1.4	0.0615	0.0365	0.0425
RGN	10.3	428	496	**1.35**	0.0724	0.0351	0.0952
Best Doc2Vec	8.68	97.9	320	1.4	0.08	0.0427	0.0629

	HSP90	Amino glycosidase (Kka2)	Pab1 (RRM domain)	PSD95 (Pdz3 domain)	Ubiquitin	Yap65 (WW domain)	Stability: 17 DMS Datasets Combined	Stability: *De Novo* Designed Mini Proteins	Ranking Stability: *De Novo* Design Rounds
UniRep Fusion	**0.0218*** **	**0.116** **	**0.0234*** **	**0.0183*** **	0.0521	0.0387	**0.031*** **	**0.185*** **	**ρ=0.62*** **
Our Best Baseline	0.0415	0.125	0.0497	0.0342	**0.0403*** **	**0.033*** **	0.0425	0.208	-
RGN	0.0541	0.129	0.0586	0.0488	0.0632	0.0504	0.0351	0.193	-
Best Doc2Vec	0.0626	0.145	0.047	0.0465	0.0612	0.0408	0.0511	0.266	-
								Rosetta Total Energy	ρ=0.42

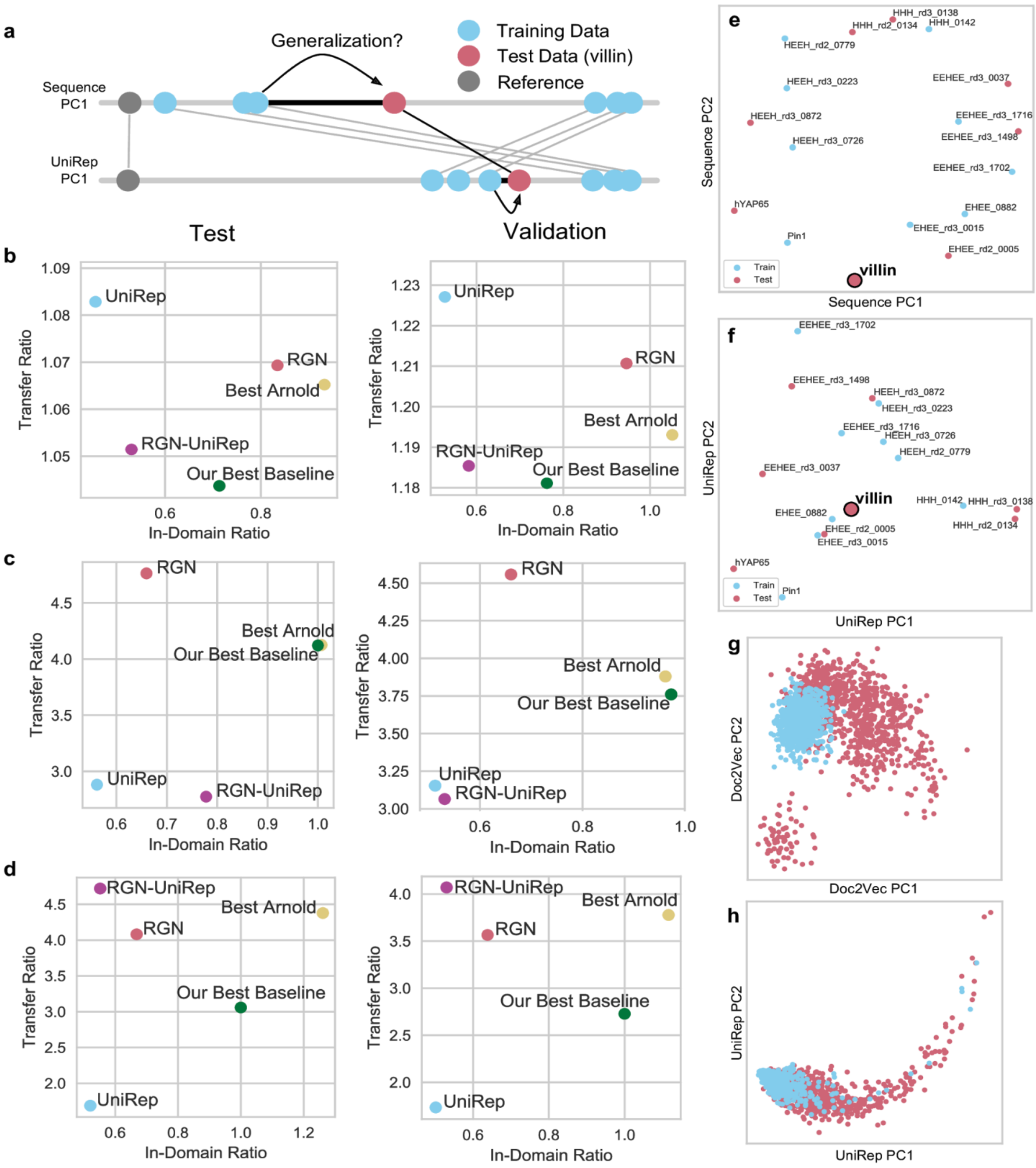
Supplemental Table 6. Regression results - validation set metrics, with lowest MSE model or model class compared to the 2nd lowest MSE model or model class *p < 0.05, **p < 0.01, ***p<0.001 (Welch's two-tailed t-test for significance), standard deviations obtained through n=30 50% validation/test set resampling. All values are MSE except where specified. Test set metrics and explanation of the first 4 column tasks can be found in Supp. Table 5.



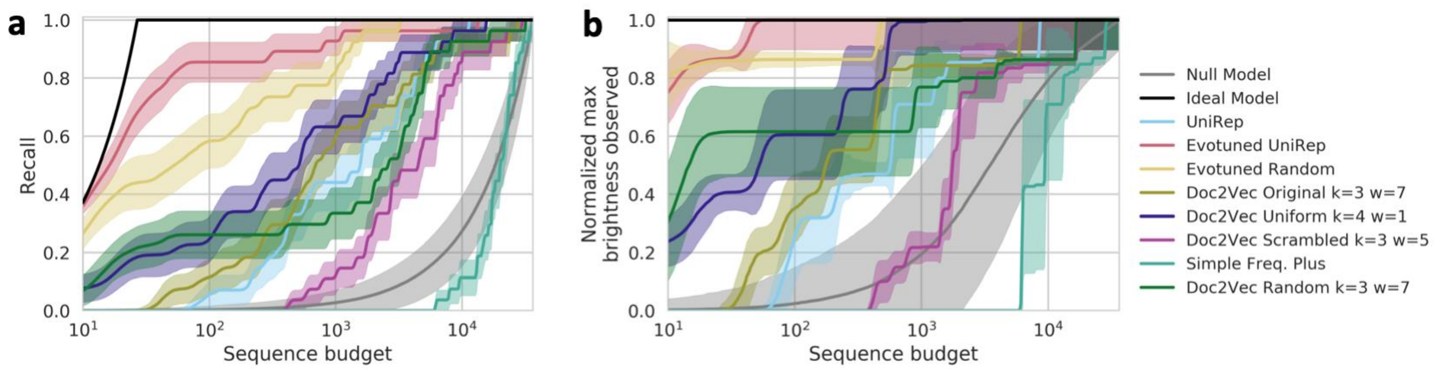
Supplemental Figure 9. Validation scores for main text Figure 3e, 17 DMS protein stability prediction datasets.



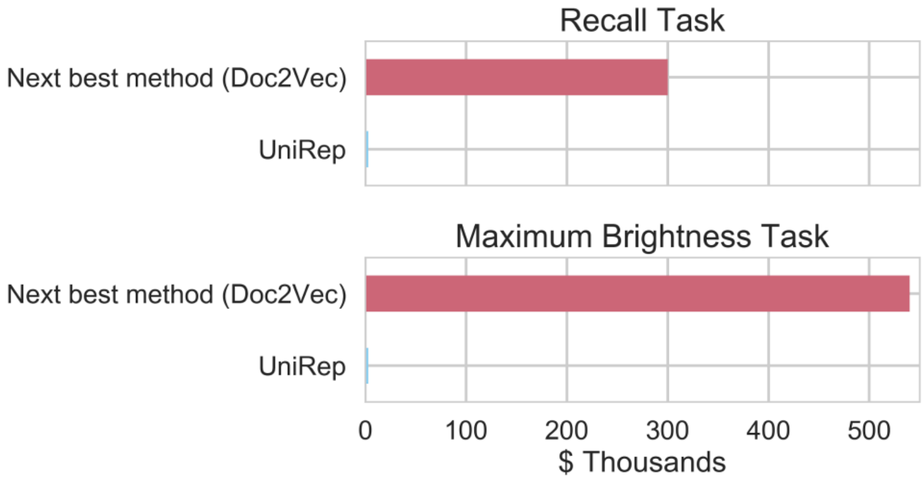
Supplemental Figure 10. Linear model on top of UniRep is simpler (has fewer parameters) than the same model using a standard One-Hot-Encoding if the sequence is longer than 95aa.



Supplemental Figure 11. Variant Effect and stability generalization tasks; hypothesized mechanism for transfer performance. We used a well-established generalization metrics from work in machine learning[69]. They quantify the error of local, "In-Domain" predictions as well as generalization error, or "transfer", relative to a baseline (Methods); *lower is better*. Unless stated otherwise, we use the Leave One Protein Out (LOPO) procedure, withholding one protein from the set at a time, training a model on all but that one protein, evaluating it on the single withheld protein, and taking an average of these generalization errors (Methods) [36]. **a.** Generalization is making accurate predictions on sequences that are distant from the training data. Here, the training data (blue), a distant reference sequence (grey) and the the test data (villin) are shown on the first Principle Coordinate of an MDS of the Levenshtein sequence distance. In sequence space, villin is far from the nearest training point, which makes generalization challenging. In UniRep space, shown here as the first Principal Component, the training and test data are rearranged so villin is much closer to the nearest training point, thereby enabling generalization. **b.** Generalization performance on the LOPO generalized variant effect prediction task as measured by In-Domain vs Transfer ratios for test set (left) and validation set (right). We used the function prediction dataset from Figure 3 with 8 proteins with 9 distinct functions measured in separate experiments[36,85]. It was previously shown that some of the regions most vulnerable to deleterious single mutations are functional and highly conserved or co-conserved in evolution[86]. Standard approaches therefore rely strongly on co-evolutionary data[15,36,85] or even structural data[36], which implicitly demarcate functional residues. Because it has neither of these as inputs and was trained on a corpus with at most 50% similarity between sequences, UniRep should find it challenging to identify such residues. Nevertheless, UniRep performs best in-Domain, successfully identifying functionally-important positions from some labeled mutation data for a given protein. However, as expected, UniRep does not generalize well to proteins for which it had no labeled training data, at least in this case. When we tested fusing UniRep to the RGN (RGN-Fusion), which was trained on a form of evolutionary data (PSSMs) and predicts protein structure, we see a boost in performance suggesting a good trade-off between In-Domain and Transfer Error. **c.** Generalization performance on the LOPO generalized DMS stability task, measured by In-Domain and Transfer ratios for test set (left) and validation set (right). Unlike a specific molecular function, stability is a property shared by all proteins. **d.** Generalization performance on the extrapolation DMS stability task, where the test set is selected from the most peripheral proteins in the set (e, red), measured by In-Domain and transfer ratios for test set (left) and validation set (right). We took the same DMS dataset as c) but instead of using the LOPO procedure, selected a single test set consisting of the most distant proteins in sequence space visualized with an MDS of the Levenshtein distance matrix (see e). This most closely represents the setup in a protein engineering task, where the engineer is exploring outwords in sequence space from local knowledge. Here we see the strongest performance of UniRep over baselines, suggesting that UniRep is well suited to this protein engineering formulation. **e.** MDS of the DMS stability dataset Levenshtein sequence matrix with the test set (red) and training set from the stability extrapolation task (d) indicated (n=17 proteins). Villin, shown in 1D earlier, is enlarged and bolded. **f.** PCA of the DMS stability datasets UniRep distance matrix on with the test set (red) and training set from the stability extrapolation task (d) indicated (n=17 proteins). Villin, shown in 1D earlier, is enlarged and bolded. **g.** PCA of the avGFP training set (blue) and FPbase test set (red) from Figure 4 (n=181,187 proteins, 49,129 avGFP training set proteins and 132,058 FPbase test set, downsampled randomly to 1000 points per set to aid visualization) with the best performing Doc2Vec model distance matrix . **h.** PCA of the avGFP training set (blue) and FPbase test set (red) from Figure 4 (n=181,187 proteins, 49,129 avGFP training set proteins and 132,058 FPbase test set, downsampled randomly to 1000 points per set to aid visualization) with the best performing UniRep model distance matrix.



Supplemental Figure 12. Efficiency curves for all baseline representations: recall (a) and maximum brightness (b). Note the variable performance of the best 4 variants of Doc2Vec presented in Yang *et al.* (2018) [21]. Error bands depict +/- 1 standard deviation calculated over n=100 bootstrap replicates.



Supplemental Figure 13. Estimated UniRep cost savings in GFP protein engineering tasks. Although real cost savings will vary depending on the number of non-functional sequences to sift through, Evotuned UniRep achieves 80% recall at ~60 sequences tested, or approximately $3,000 assuming the most competitive full gene synthesis price of $0.07/nt[87]. By contrast, to achieve the same level of recall, the best Doc2Vec baseline would require ~$300,000 (100x more). Random sampling still commonly used in this context would require $1,848,000 to achieve the same. Similarly, Evotuned UniRep captures the brightest sequence in the generalization set within the first $3,000 spent in testing, and the best Doc2Vec baseline would require ~$540,000 (180x more) to do the same. Assuming on-target assembly rates improve and full economies of scale, multiplex gene assembly methods such as DropSynth[56] could bring the cost of synthesizing a model proposed GFP down to ~$2. At these cost rates, Evotuned Unirep would enable high purity functional diversity capture and function optimization for just a few hundred dollars. Taken together, these results suggest that Evotuning UniRep enables generalization to distant parts of the fitness landscape and thereby facilitates protein engineering by drastically minimizing the cost required to capture functional diversity and optimize function.

Group	Protein(s) in the task	Size	Characteristic	Ref
Small-scale protein characteristics prediction	Cytochrome P450	261	Thermostability	[21]
	Bacterial Rhodopsin	81	Peak Absorption Wavelength	
	Epoxide Hydrolase	152	Enantioselectivity	
	Channelrhodopsin	248	Plasma Membrane Localization	
Large-scale function prediction	TEM1b-lactamase	5198	Function (diverse, see Fig. 3e)	[36]
	Ubiquitin - E1 activity	1085		
	Protein G (IgG domain)	1045		
	Pab1 (RRM domain)	1188		
	Ubiquitin	1195		
	PSD95 (Pdz3 domain)	1577		
	Yap65 (WW domain)	363		
	Aminoglycoside kinase	4234		
	Hsp90	4021		
Large-scale stability prediction	hYAP65	829	Stability (DMS data)	[5]
	HHH_0142	775		
	EEHEE_rd3_1716	775		
	HEEH_rd3_0726	775		

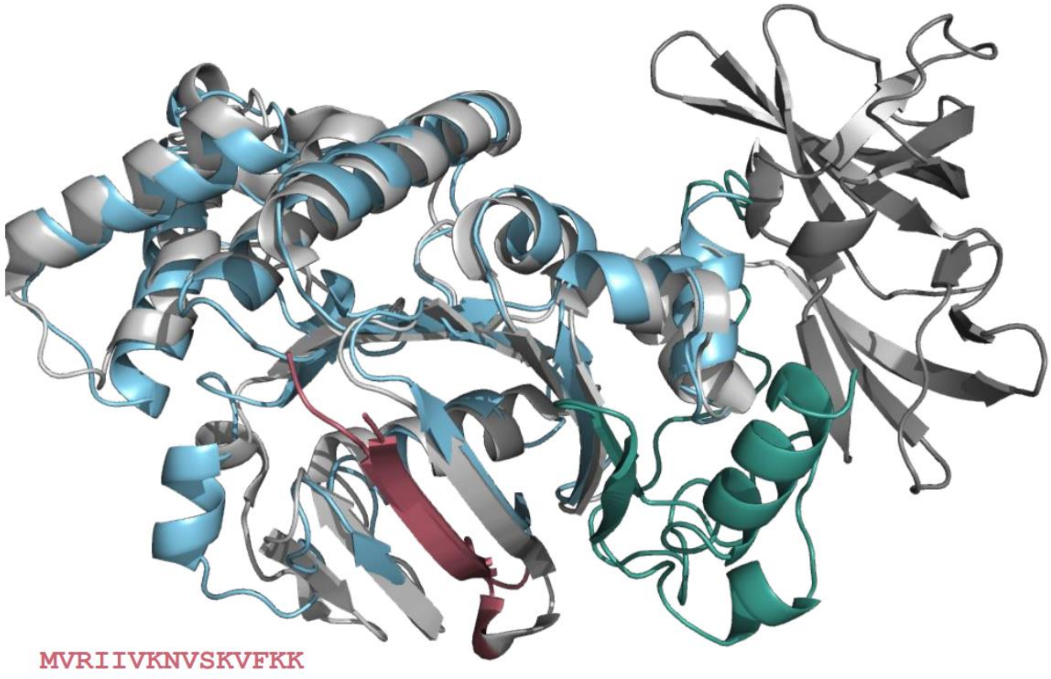
	EEHEE_rd3_1498	775		
	EEHEE_rd3_1702	775		
	HHH_rd2_0134	775		
	HHH_rd3_0138	775		
	HEEH_rd3_0872	775		
	HEEH_rd2_0779	775		
	HEEH_rd3_0223	775		
	EEHEE_rd3_0037	775		
	EHEE_rd3_0015	721		
	EHEE_rd2_0005	721		
	EHEE_0882	721		
	Pin1	703		
	villin	631		
	De Novo Proteins from Design Rounds	56083	Stability	
GFP engineering	avGFP	51715	Brightness	38
	A number of green fluorescent proteins	27		40

Dataset	Size	Ref	Comment
SCOP 1.67 Superfamily Remote Homology Detection	3802	70	Representing 102 superfamilies

SCOP 1.67 Fold-level Similarity Detection	3736		Representing 86 folds

Dataset	Size	Ref	Comment
OXBench Database Reference Alignment	811	26	Classified into 180 families
Protein Family Prediction from HOMSTRAD Database	3450	25	Classified into 1031 families

Supplemental Table 7. Analysis datasets and tasks.



Supplemental Figure 14. UniRep is a generative model of protein sequences. Homology model of a UniRep "babbled" sequence using a 15 amino acid seed (red) from a glucose ABC transporter (sequence from PDB:1oxx). Seed reference structure (PDB:1oxx) in grey. Modeled structure of babbled sequence shown in red (seed residues), blue (remaining N-terminus), and green (C-terminus which blasts to dipeptide ABC transporter with >40% identity). Alignment with the seed reference shows that UniRep has generated a sample which reconstructs structural regularities of the protein family. Full sequence blasts to ABC transporter family members with >50% similarity.

Data Source	Motivation	Applications	Mechanism	Refs
PDB crystal structures	Improve ab-initio structure prediction.	In-silico crystallography, structural comparisons	Joint training with the RGN, predicting next amino acid and structure to update joint weights by minimizing a hybrid averaged loss.	16,60
position sensitive scoring matrices (PSSMs)	Enrich representation of evolutionary relationships.	“Deep Homology”, semantic sequence comparisons, sequence search, phylogenomics.	Additional representation training input (predict next column of PSSM).	
experimentally characterized synthetic mutants	Tune for a specific engineering task by training on synthetic sequences which have been synthesized and tested.	High-throughput model guided protein engineering.	Evotuning, but replace extant sequences with synthetic sequences which have measured bright.	
ancestral sequence reconstruction	Tune for an engineering task by enriching training data with ancestral sequences near the target which have been shown to have desirable biotechnological properties.	Data-constrained model-guided protein engineering.	Evotuning, but replace extant sequences with high-confidence sequences from ancestral protein reconstruction methods applied to the target and nearby extant relatives.	88
Functional annotations (eg Pfam, GO, etc)	Incorporate existing functional data to improve representations for semantic sequence search.	Fast, vector math parallelized sequence annotation and homolog identification.	Auxiliary tasks requiring the network to predict annotations after finishing predicting next amino acid up to the end of a sequence.	18,60
Experimental epistasis data	Train UniRep to capture higher-	Study of theoretical and experimental	Evotuning to the context of a target	42,89

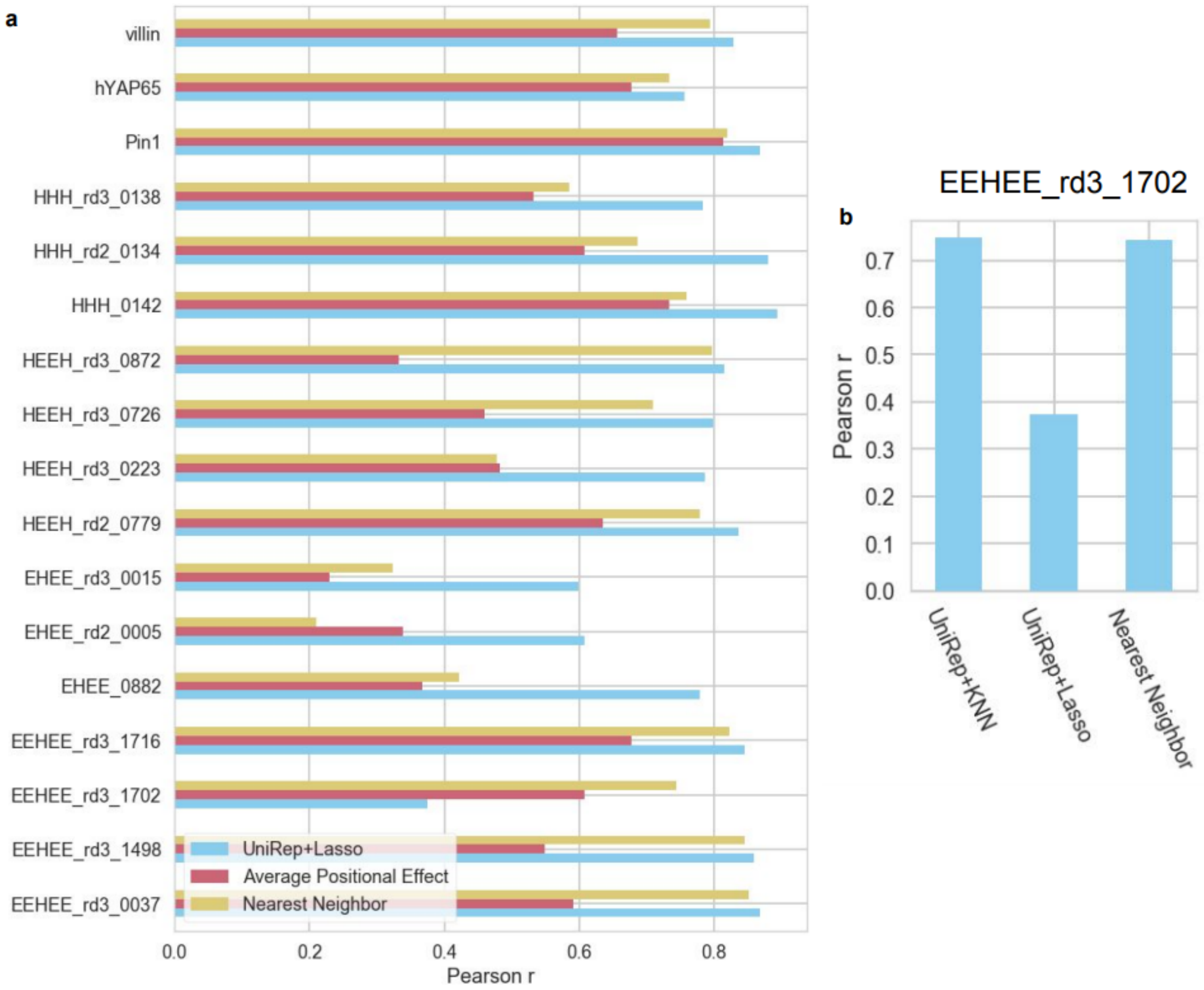
	order features of protein fitness landscapes to model the relationship between micro and macro scale and the accordance with theoretical results.	protein evolution.	epistasis model then subsequent top model training on evotuned representations to predict experimental fitness values (as in Fig. 4)	

Supplemental Table 8. Data augmentation of UniRep.

Name	**Type**	**# Dimensions**	**Ref**
RGN	RGN	3200	16
UniRep 64-unit Avg. Hidden State	UniRep	64	
UniRep 64-unit Final Hidden State	UniRep	64	
UniRep 64-unit Final Cell State	UniRep	64	
UniRep 256-unit Avg. Hidden State	UniRep	256	
UniRep 256-unit Final Hidden State	UniRep	256	
UniRep 1900-unit Avg. Hidden State	UniRep	1900	
UniRep 1900-unit Final Hidden State	UniRep	1900	
UniRep 1900-unit Final Cell State	UniRep	1900	
Doc2Vec Original k=3 w=7	Doc2Vec	64	21
Doc2Vec Scrambled k=3 w=5	Doc2Vec	64	
Doc2Vec Random k=3 w=7	Doc2Vec	64	21
Doc2Vec Uniform k=4 w=1	Doc2Vec	64	21
UniRep 64-unit Fusion	UniRep	192	
UniRep 256-unit Fusion	UniRep	512	
UniRep Fusion	UniRep	5700	
UniRep 64-unit + 256-unit + 1900-unit Avg. Hiddens	UniRep	2220	
UniRep 64-unit + 256-unit + 1900-unit Final Cells	UniRep	2220	
RGN + UniRep 1900-unit Avg. Hiddens	RGN+UniRep	5100	16
RGN + UniRep 1900-unit Final Cells	RGN+UniRep	5100	16

Our Baseline: Amino Acid Freq. and Predicted Biophys. Params.	Other baseline	26	
Our Baseline: Amino Acid Freq. and Protein Length	Other baseline	21	
Our Baseline: 2-grams	Other baseline	Dataset- dependent	66
Our Baseline: 3-grams	Other baseline	Dataset- dependent	66
Our Baseline: 2-grams with TF-IDF weighting	Other baseline	Dataset- dependent	66
Our Baseline: 3-grams with TF-IDF weighting	Other baseline	Dataset- dependent	
Our Baseline: Dataset Target Mean	Other baseline	1	

Supplemental Table 9. All the models we evaluated, including the baseline suite used for the majority of analyses in the manuscript. We additionally used Levenshtein distance (Needleman-Wunsch where all penalties are equal) for analysis in Fig. 2d and Rosetta total energy and NPSA measures for Fig. 3a (as described in Methods under “Stability Ranking Task”)



Supplemental Figure 15. a. DMS-specific non-representation non-parametric baseline comparison. **a.** Performance of UniRep with a linear Lasso top model on 17 stability deep mutational scanning datasets compared to two DMS-specific non-representation experimental data-based baselines - Average Positional Effect and Nearest Neighbor (Methods), dataset sizes can be found in Supp. Table 7. **b.** K Nearest Neighbors (KNN) top model (Methods) improves stability prediction performance of UniRep on EEHEE_rd3_1702 de novo designed protein.

Supplemental References

84. Mikolov, T., Yih, W.-T. & Zweig, G. Linguistic Regularities in Continuous Space Word Representations. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 746–751 (2013).

85. Hopf, T. A. *et al.* Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128–135 (2017).

86. Guo, H. H., Choe, J. & Loeb, L. A. Protein tolerance to random amino acid change. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 9205–9210 (2004).

87. Genes - Gene Synthesis | Twist Bioscience. Available at: https://www.twistbioscience.com/products/genes?gclid=Cj0KCQiA28nfBRCDARIsANc5BFAYK3MMQaN1ZZelOT-X3gKuAsIUXqeXbOwUZ17nYEPD5Rw6_nM_XegaAqAUEALw_wcB. (Accessed: 19th November 2018)

88. Thornton, J. W. Resurrecting ancient genes: experimental analysis of extinct molecules. *Nat. Rev. Genet.* **5**, 366 (2004).

89. Pokusaeva, V. *et al.* Experimental assay of a fitness landscape on a macroevolutionary scale. *bioRxiv* 222778 (2018). doi:10.1101/222778