Supplementary Information: Evaluating *eUniRep* and other protein feature representations for *in silico* directed evolution

Andrew Favor

Ivan Jayapurna

May 27, 2020

Supplementary Discussion

S1: Evotuning pre-training dataset curation

Pre-training data set curation follows the following steps:

1. Search up the wild-type protein sequence on InterPro and PFAM Protein Sequence Fetch to get what family / clan it is in. Download all these sequences as well as other related families / clans as suggested by PFAM. We recommend downloading via code snippets that fetch protein sequences using the InterPro API (these can be generated by InterPro).
2. If you have less sequences than desired (i.e. <50,000 for example, as note that some of these sequences may be duplicates or be invalid sequences) then search for keywords on InterPro and download those as outlined in the 2MS2 script. A JackHMMer search at this step may also optionally be attempted with the wild type sequence as the seed.
3. Clean the inputs by removing sequences with non-standard residues, anything with length greater than k, remove duplicate sequences.
4. Calculate Levenstein distances from highly desired mutant (if available, if not then from wild type sequence) on all clean sequences.
5. Generate training, in_domain validation and out_domain validation sets. out_domain validation set is generated first using a distribution proportional to distances^4 - taking 10% of the total sequences. Then 10% of the total (11.1% of the remainder) is taken for the in_domain validation set, with the remaining 80% of original = the training set.

S2: Ranking-error function: development and application

To optimize the parameters for the task of fitness score ranking, we were inspired to develop a new type of error function, E_r, defined as follows:

For a validation batch size of n, ranking-matrices $R \in \mathbb{R}^{n \times n}$ were constructed. For any two mutant variants in a given batch, $i, j \in B(n)$, the ranking-matrix elements were assigned binary values to represent whether the variant sequence of row i's fitness score was greater than, less than, or equal to the fitness score of the variant in column j.

$$a_{i,j} = \begin{cases} i > j : 1 \\ i \leq j : 0 \end{cases}$$

The purpose of allowing less-than and equal values to both be represented by $a_{i,j} = 0$ is two-fold: first, it was necessary, as some of the experimentally reported fitness scores had identical values for different mutants; second, in the Metropolis-Hastings selection steps, there is no need to accept a new mutation if it doesn't yield improvement over the current sequence. Thus, the ranking-matrix elements assigned values of $a_{i,j} = 1$ provide a simple way to represent the fitness rankings of variants, by showing which variants of columns j each variant of row i has fitness improvements over.

Two of these ranking-matrices were constructed: one for the experimentally-determined fitness scores of a given batch, Φ_E, and another for the model's predicted fitness scores for the input-representation of variants in that batch, Φ_P. A confusion matrix, Ψ, was produced from an inverse-truth comparison of the two ranking-matrices, such that it had values of 1 where the predicted fitness rankings were false with respect to experimental data, and 0 where the experimentally determined rankings were matched by predictions:

$$\Psi = \Phi_P \oplus \Phi_E$$

The sum of the elements in confusion matrix was computed, and normalized by the total number of elements, yielding our ranking-error function:

$$E_r = \frac{1}{n^2} \sum_{i,j} \psi_{i,j}$$

S3: Derivation of additive-type prediction criteria for linear kernel sequence representations:

Given the fitness score for a single amino acid substitution, $y_{i,u}$, where i denotes the backbone position of the mutation, and u denotes the new amino acid present, we can define the mutant fitness in terms of its difference relative to the wild type fitness, y_{wt}:

$$y_{(i,u)} = y_{wt} + \Delta y_{wt \to i,u}$$

where $\Delta y_{wt \to i,u} = y_{(i,u)} - y_{wt}$. Likewise, the predictive additive fitness of any of double-mutant, characterized by the substitution of amino acids u and v at positions i and j respectively, can be defined by:

$$\begin{aligned} \widehat{y}_{(i,u),(j,v)} &= y_{wt} + \Delta y_{wt \to i,u} + \Delta y_{wt \to j,v} \\ &= y_{(i,u)} + y_{(j,v)} - y_{wt} \end{aligned}$$

A one-hot encoding of a protein sequence composed on n amino acids, can represent it as the concatenation of n consecutive sparse binary row-vectors ϕ_u, with u assigning the index of the vector element that is equal to 1, based on a defined ordering of the canonical amino acids:

$$\boldsymbol{s}_{wt} = \begin{bmatrix} \boldsymbol{\phi}_1 & \cdots & \boldsymbol{\phi}_i & \cdots & \boldsymbol{\phi}_n \end{bmatrix}$$

Consider a wild type amino acid, w, at position i,

$$\boldsymbol{\phi}_i(w) = \begin{pmatrix} \cdots & 0 & 1 & \cdots & 0 & 0 & \cdots \end{pmatrix}$$

and a mutant amino acid, u, to substitute at position i:

$$\boldsymbol{\phi}_i(u) = \begin{pmatrix} \cdots & 0 & 0 & \cdots & 1 & 0 & \cdots \end{pmatrix}$$

we can define an additive operator to represent the change in one-hot representation when going from the wild-type sequence to the mutant sequence with amino acid u substituted at position i:

$$\begin{aligned} \Delta \boldsymbol{s}_{wt \to (i,u)} &= \boldsymbol{s}_{(i,u)} - \boldsymbol{s}_{wt} \\ &= \begin{bmatrix} (\boldsymbol{\phi}_1(w) - \boldsymbol{\phi}_1(w)) & \cdots & (\boldsymbol{\phi}_i(u) - \boldsymbol{\phi}_i(w)) & \cdots & (\boldsymbol{\phi}_n(w) - \boldsymbol{\phi}_n(w)) \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{0} & \cdots & (\cdots & 0 & -1 & \cdots & 1 & 0 & \cdots) & \cdots & \mathbf{0} \end{bmatrix} \end{aligned}$$

We can extend the use of this operator in the generation of one-hot sequence encodings for multiple amino acid substitutions. Again, considering a double-mutant with amino acids u and v substituted at positions i and j, a one-hot sequence representation can be produces as follows:

$$\begin{aligned}\boldsymbol{s}_{(i,u),(j,v)} &= \boldsymbol{s}_{wt} + \Delta\boldsymbol{s}_{wt\to(i,u)} + \Delta\boldsymbol{s}_{wt\to(j,v)} \\ &= \boldsymbol{s}_{(i,u)} + \boldsymbol{s}_{(j,v)} - \boldsymbol{s}_{wt}\end{aligned}$$

Given any linear model $F(s_{(i,u)}; \boldsymbol{c}, b)$, which maps a one-hot encoding $s_{(i,u)}$ to the corresponding fitness value $y_{(i,u)}$, a fitting operation will be performed to optimize the values of weights $\boldsymbol{c}$ and bias b, such that:

$$\begin{aligned}F(\boldsymbol{s}_{(i,u)}) &= \begin{bmatrix}\boldsymbol{s}_{(i,u)} & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{c}\\ b\end{bmatrix} \\ &= y_{(i,u)} + \sigma_{(i,u)}\end{aligned}$$

where $\sigma_{(i,u)}$ is the prediction error for a given mutation sample, which is to minimized by fitting the model's parameters. In the fitting operation, training elements include a feature-array, $\boldsymbol{S}$, and a fitness-vector, $\boldsymbol{y}$, corresponding to an experimentally derived data set of single amino acid substitutions.

The predictive model here is a linear operator, and thus abides by the property of being closed under additivity:

$$\begin{aligned}F\left(\sum_{(I,U)}\boldsymbol{s}_{(i,u)}\right) &= \sum_{(I,U)}F(\boldsymbol{s}_{(i,u)}) \\ &= \sum_{(I,U)}\begin{bmatrix}\boldsymbol{s}_{(i,u)} & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{c}\\ b\end{bmatrix} \\ &= \sum_{(I,U)} y_{(i,u)} + \sigma_{(i,u)}\end{aligned}$$

Providing the linear predictive model with the one-hot encoding of a double-mutation for any two amino acids, u and v, substituted at positions i and j, we produce additive fitness scores, offset linearly by the error values if their corresponding single-mutations:

$$\begin{aligned}F(\boldsymbol{s}_{(i,u),(j,v)}) &= F\left(\boldsymbol{s}_{(i,u)} + \boldsymbol{s}_{(j,v)} - \boldsymbol{s}_{wt}\right) \\ &= F(\boldsymbol{s}_{(i,u)}) + F(\boldsymbol{s}_{(j,v)}) - F(\boldsymbol{s}_{wt}) \\ &= \left(y_{(i,u)} + \sigma_{(i,u)}\right) + \left(y_{(j,v)} + \sigma_{(j,v)}\right) - (y_{wt} + \sigma_{wt}) \\ &= (y_{wt} + \Delta y_{wt\to i,u} + \Delta y_{wt\to j,v}) + \left(\sigma_{(i,u)} + \sigma_{(j,v)} - \sigma_{wt}\right) \\ &= \widehat{y}_{(i,u),(j,v)} + \left(\sigma_{(i,u)} + \sigma_{(j,v)} - \sigma_{wt}\right)\end{aligned}$$

If $\boldsymbol{s}_{(i,u)}$,$\boldsymbol{s}_{(j,v)}$, and $\boldsymbol{s}_{wt}$ are all in the training set used for fitting the parameters of the predictive model, and if the fitting occurs within an over-determined system where the number of samples is less than the number of elements in each one-hot representation ($n \times 20$), then the regression will fit closely to all training data, such that the values of $\sigma_{(i,u)}$, $\sigma_{(j,v)}$, and σ_{wt} are negligible.

In such a case, we find the predicted value of the one-hot encoding of a double-mutant to be:

$$F(\boldsymbol{s}_{(i,u),(j,v)}) \simeq \widehat{y}_{(i,u),(j,v)}$$

Therefore, the fitness value for a double-mutant, as predicted by a linear model that is unconstrained in fitting to a training set that includes the constituent single-mutations, is approximately equal to the additive fitness scores for that double-mutant.

Supplementary Figures:

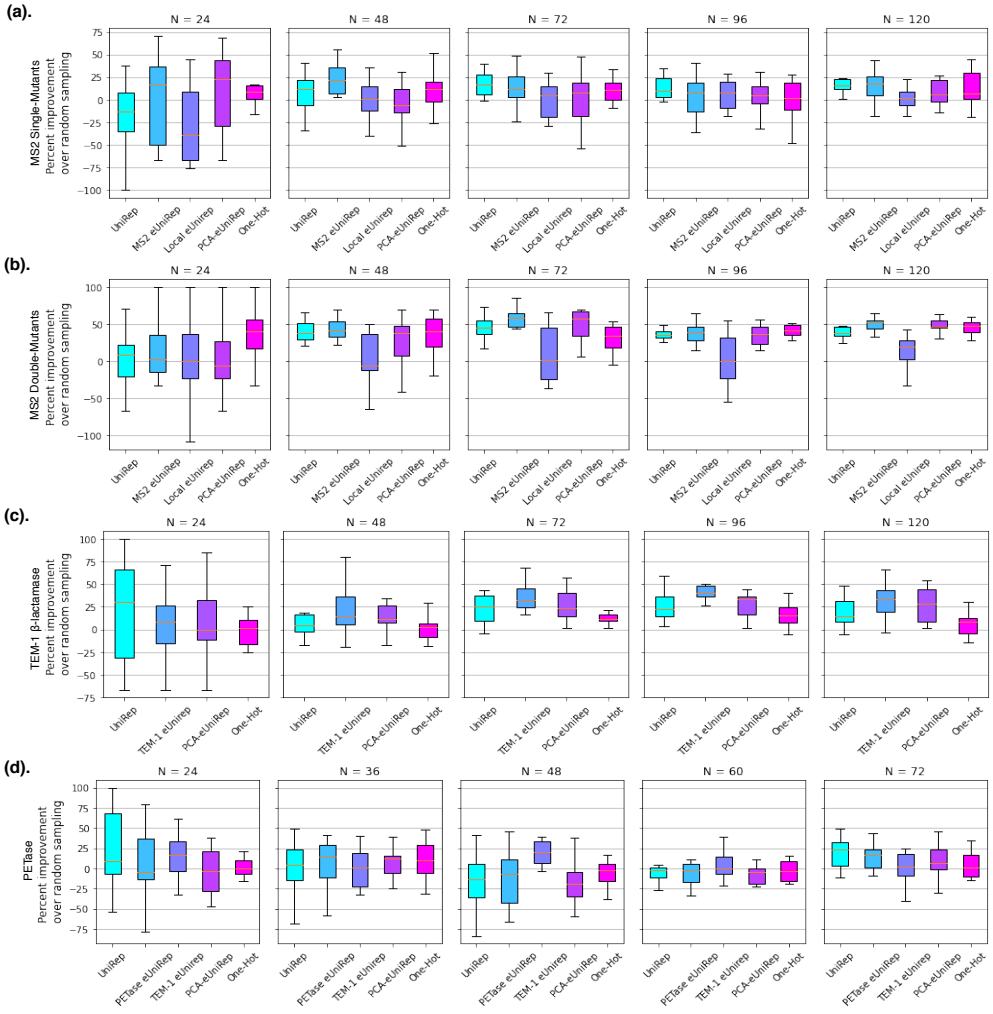

Figure 1: Comparison in percent ranking-error reduction when predicting fitness using different sequence representations, for several training batch sizes. Performance is shown for four data sets: MS2 single mutants, MS2 double mutants, TEM1 Beta-lactamase single mutants, and PETase single mutants.

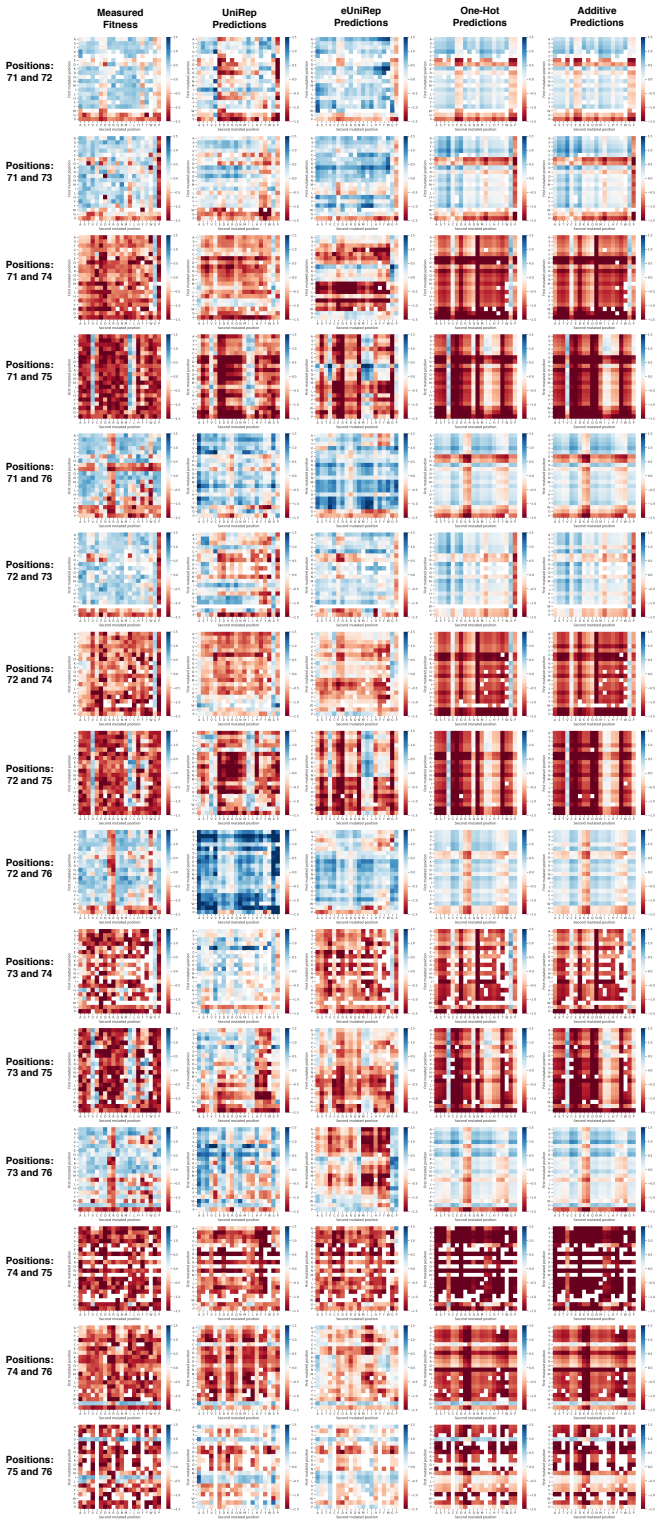

Figure 2: The predictions of MS2 double-mutant fitness, performed by a Ridge regression that was trained on MS2 single-mutant data. Rows correspond to all possible co-mutation combinations between amino acids within the range of residues 71 to 76. Column 1 displays the experimentally determined fitness values. Columns 2-5 correspond to different predictive methods tested.

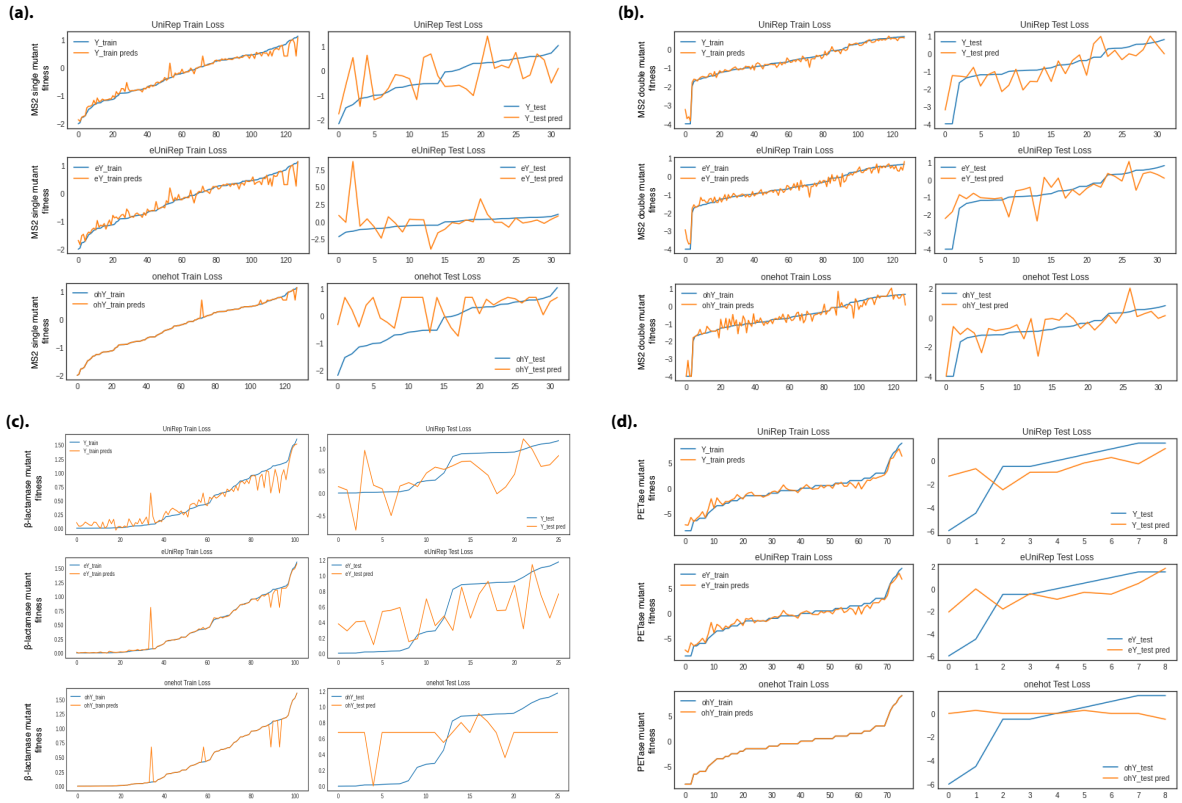

Figure 3: Predictions of top model (Ridge regression with 10-fold cross-validation) of training data and validation data. Data sets shown are MS2 single mutants **(a)**, MS2 double mutants **(b)**, TEM1 Beta-lactamase single mutants **(c)**, and PETase single mutants **(d)**. Training and validation sets were split with validation receiving 20% of the total batch size, which was randomly selected. Data sets sets are shown with predictive performance comparisons between three sequence representations: global UniRep, evotuned UniRep, and one-hot encodings.