Reconstructing Fitness Landscapes for Optogenetic Design
- ML4Science project -

Chloé Daul, Elia Mounier-Poulat, Kleon Karapas

In Collaboration with the Laboratory of the Physics of Biological Systems at EPFL, 2023

***Abstract*—In this project, we employ established machine learning models to reconstruct the fitness landscape of the EL222 protein. The objective is to create an in silico model that captures the evolutionary dynamics of this optogenetic protein, thereby enabling the design of variants with predefined functions.**

Background: Proteins are cellular components with various functions and roles. They are composed of a sequence of amino acid residues, all linked together. These residues possess various physicochemical properties, and their interactions shape the 3D structure of the protein, influencing how it interacts with other components and defining its function. A modification in the position of an amino acid residue is called a mutation and may lead to a different protein folding (3D configuration) and function.

I. Introduction

Optogenetic techniques have emerged as revolutionary tools in the field of biology. These methods employ the use of optoproteins, such as EL222, whose function can be switched on and off through light modulation, providing control over both temporal and spatial cellular regulation.

The objective of our investigation is to use machine learning techniques to predict which mutation in the EL222 sequence will result in an optimized optogenetic protein. In other words, we aim to find the sequence that will maximize the protein's fitness, i.e., its level of activity, when exposed to light while still maintaining low activity in darkness.

For this, our approach is to use an 'augmented model' from the paper *"Combining evolutionary and assay-labelled data for protein fitness prediction"* [1]. Their approach combines both supervised and unsupervised techniques to predict protein fitness based on an 'assay-labelled' and an 'evolutionary' dataset. One of their top performing models is called the 'Augmented Potts model,' which is the one we will use and train with the data provided by the Laboratory of the Physics of Biological Systems at EPFL *(LPBS)*. After model training, we will generate an artificial dataset for single mutants and double mutants of EL222 and use the trained model to predict their corresponding fitness. Finally, we will search for the sequence optimizing our optoprotein function. We will refer to this step as 'in silico evolution.'

II. Datasets and Data Preparation

The *LPBS* provided us with a labeled dataset that includes mutants names of the EL222 protein, along with their corresponding light sensitivities. However, the dataset's size is limited. To overcome this constraint and augment our data pool, we have access to a broad collection of similar proteins found in nature among diverse species, all sharing the same property as EL222. This constitutes the 'evolutionary dataset'[3].

Additionally, the code made available by the reference paper proposes a demo dataset, which we will include in our analysis as baseline comparisons for our research.

A. Exploration of the Assay-Labeled Dataset

The assay-labeled dataset contains 40 distinct EL222 mutants, with dark fitness values ranging from -4.45 to 3645.61 and light fitness values ranging from -5.71 to 7708.50. An important aspect is that all of the optoprotein's variants are 'positive mutants,' meaning the protein remains functional upon mutation.

In comparison, the labeled demo dataset comprises variants of a protein that is not an optogenetic protein, with a singular type of fitness value assigned, ranging from -3.74 to 0.35. It is mostly composed of negative mutants (where the protein loses its function due to mutation) and is considerably larger, consisting of 4807 labeled sequences.

B. Processing of the Assay-Labeled Dataset

The code we are using requires a 'seq' column for the sequence as a feature. The sequences should consist of a string of letters, each representing an amino acid: $s = (s_1, \ldots, s_L)$, identical for each variant as the EL222 wild-type sequence but with the corresponding amino acid mutated. For ease of code implementation, we selected only the single mutants, reducing the dataset to 35 variants.

Moreover, there should be only one label column, so we split the labeled dataset into two: the 'Light' and 'Darkness' datasets. In the code, the label 'log_fitness' corresponds to the log enrichment ratio. We applied the following transformation:

$$\log\_\text{fitness} = \log\left(\frac{\text{fitness\_mut} - \min(\text{fitness\_all\_mut}) + 1}{\text{fitness\_wt} - \min(\text{fitness\_all\_mut}) + 1}\right)$$

This label transformation allowed us to have a similar range as the demo dataset and reduce sparsity, which could affect a model with so few data points. We subtract the minimum fitness among all mutants and add 1 so that the log transformation is well-defined in our case.

C. Exploration of Evolutionary Data

The evolutionary dataset, which will be needed for the unsupervised part of our model, is composed of 6782 diverse protein sequences, similar to EL222. They'll constitute 'weak positive' labels, as evolutionarily related sequences share similar functions with the target.

D. Processing of Evolutionary Data

After aligning the sequences from the evolutionary dataset, we obtained what we called a Multiple Sequence Alignement (MSA), that will help the models compare them. We also formatted the dataset to a .a2m with the help of the hh-suite package.

Finally, to follow the lead of the reference paper, we removed sequences that had more than 30% gaps, resulting in a finite size of the evolutionary dataset of 4363, smaller but still comparable to the size of the demo evolutionary dataset, which has 9255 sequences.

III. Models and Methods

Our reference paper claims that combining both supervised and unsupervised approaches leads to great performance in predicting protein fitness. Indeed, unsupervised methods already show promising results in this field [2], and incorporating a supervised layer, such as a linear model that has low computational cost, improve their efficacy significantly [1]. In this section, we will describe the methods they used for the specific model they refer to as 'Augmented Potts,' which we will use to model EL222 fitness landscape.

A. Unsupervised EVMutation Potts

Evolutionary machine learning models capture protein's structural constraints encoded in multiple alignment sequences. For example, the frequency of observing a certain amino acid at one site provides information about its significance in the sequence. Similarly, the frequency of observing coevolving amino-acid pairs provide information about spatially close residues in protein 3D structures. EVMutation Potts uses these two statistics as a constraint for maximizing entropy in the probability distribution of a specific amino-acid sequence s: $P(s)$ [4] [5]. The sequence density $P(s)$ will then serve as a proxy for predicting fitness ranking [1]. A more detailed overview of this method is proposed in the appendix.

B. Adding a Supervised Ridge Regression layer

When considering only the assay-labeled dataset, a simple supervised model, such as a linear regression model, can be used. Since the sequence is represented as a string of letters for each amino acid: $s = (s_1, \dots, s_L)$ of length L, with $s_i \in \mathcal{A}$, $\mathcal{A}$ being all possible amino acids, we need to use one-hot encoded site-specific amino acid features. This technique preserves all the information stored in a protein sequence but has the drawback of overparameterizing the model, resulting in $|\mathcal{A}| \times L + 1$ parameters, where $+1$ comes from the bias term. The solution proposed in the reference paper is to impose constraints on the model and to favor ridge regression over the usual simple linear model [1].

However, due to the insufficient number of assay-labeled data points in protein fitness prediction problems, supervised methods often perform poorly. This is why unsupervised evolutionary models present a promising approach to 'augment' the features of ridge regression. The linear model will use both the evolutionary density score $\log(P(s))$, obtained from EV Potts, and the regularized one-hot encoding of a sequence s as features. The linear combination of these features is what we call the 'augmented' model [1].

Additionally, the reference paper [1] proposes a 5-fold cross-validation on the one-hot-encoded features of the training dataset to select the regularization hyper-parameter. However, we decided to change it to a 3-fold cross-validation since our training dataset is composed of only a few training points. Reducing the number of folds to an even smaller value would compromise the adequacy of the metric we are using.

C. Scoring Metric

The metric we use for validation and model assessment is the Spearman rank correlation coefficient, which is a common statistical metric in the field of protein landscaping [6] [7] [8] [9] [10]:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i is the difference between the rank of each observation ($\text{fitness}_{true}; \text{fitness}_{pred}$), and n is the total number of observations. It captures the monotonic relationship between the predicted and the true fitness values, and focuses on the relative order of values rather than their numerical differences. A score of -1 indicates a negative exact correlation, 0 implies no correlation, and 1 represents an exact positive correlation.

IV. In Silico Evolution

After model training and selection, the next step is to predict the mutations that will improve the sensitivity of EL222, aiming for high fitness under light conditions and low fitness in darkness. This problem can be formulated as follows:

$$\arg\max_{\text{seq}}(\text{fitness}_{\text{light}} - \text{fitness}_{\text{dark}}) \quad \text{s.t.} \quad \text{fitness}_{\text{dark}} < \sigma$$

where *seq* refers to an artificially generated amino acids sequence; $\textit{fitness}_{\textit{light}}$ and $\textit{fitness}_{\textit{dark}}$ represent the predicted outputs for this sequence, computed independently by a model trained once with labels from the darkness assay and another time with the light assay labels. σ represents a threshold based on biological significance, it is equal to the fitness value at which the protein activity is turned off. The above equation conveys important information in optogenetics, as the labeled fitness values were transformed into log fitness during the data processing phase. What is written as a subtraction actually corresponds to a division in the log domain, reflecting the *dynamic range* of EL222, which measures the light-inducibility of the protein.[11]

In our research, we focus solely on generating and evaluating single and double mutants due to the combinatorial explosion of mutant generation and limitations in memory. Specifically, generating k-mutations in the reference *wt* sequence results in a total of $(\mathcal{A}-1)^k \cdot \binom{\text{length(wt)}}{k}$ sequences, with $\mathcal{A}-1$ representing the possible amino acid substitutions at each position.

V. Experimental Research Process

A major part of our research involved understanding the code of the augmented model, learning to interact with multiple repositories and packages, getting familiar with high-performance computing resources and addressing setup issues.

Once prepared, the first step was to assess the augmented Potts model's performance on the demo dataset. We initially trained the supervised and unsupervised models separately and then combined both for 240 training points from the assay-labeled data. Note that we always used all possible points from the evolutionary datasets in our experiments. The test set used consists of 20% of the original size of the demo labeled dataset, that is 961 test points (see Fig. 1.a). We also ran the same experiment with a reduced set of 28 training points, keeping the number of test points constant. This simulation aimed to assess the model's performance under the constraint of a very limited number of labeled samples available for training, as in our case (see Fig. 1.b). For completeness, we also provide in the appendix, the models' performance on a 7-point test set, with 28 training points, for the demo dataset (see Fig. 4).

Then, as soon as we received the *LPBS*' dataset and completed the data preparations, we conducted the same model assessment analysis with the lab's datasets. The results are presented only for the Light dataset, although similar results were obtained with the Darkness dataset (see Fig. 2).

Afterwards, we wanted to try other augmented models proposed by the reference paper that could potentially offer better performance. Unfortunately, we did not manage to reproduce any of them, even with the demo dataset. Therefore, we decided to run a few more experiments to better understand the results we obtained. We observed a considerable variation in our results among different test runs and decided to investigate the impact of the test set's size on the demo data (see Fig. 3). This experiment was conducted with the supervised model for computational efficiency reasons.

At the end, we generated artificial sequences for single and double mutants and fed them into both the supervised model and the augmented one, trained separately with the Darkness and Light datasets with all 35 labeled samples available this time. The mutations optimizing the protein function, according to the models' predictions, are displayed in Table I.

VI. Results

Every seed value ensures that the same data points will be selected for the training and test sets every time the code is run, thus making the code reproducible.

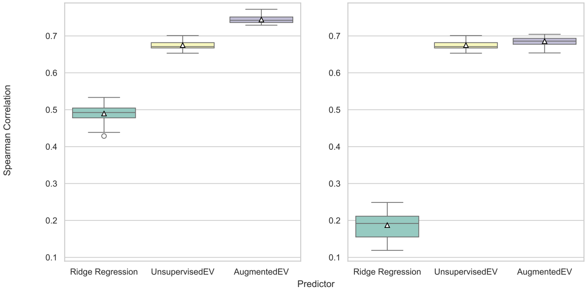

Fig. 1. Spearman correlation box plot for demo data. Ran on 20 seeds for 961 test points: 240 training points (left), 28 training points (right).

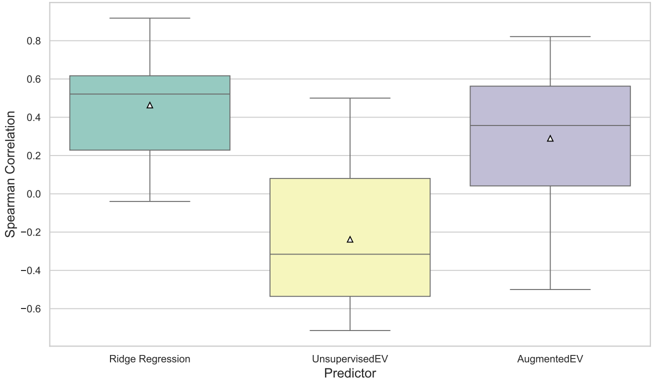

Fig. 2. Spearman correlation box plot for LPBS data. Ran on 20 seeds: 28 training points, 7 test points.

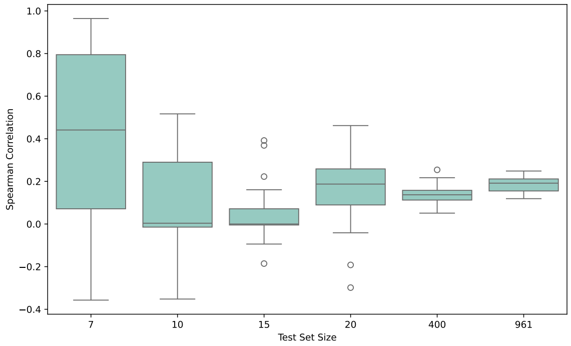

Fig. 3. Spearman correlation box plot for demo data. Ridge regression, ran on 20 seeds and varying test size (28 training points).

TABLE I
Optimal Variants obtained via in silico evolution

Predictor	Mutations	Dynamic Range
Ridge Regression	Asp92Tyr	2.447
Ridge Regression	Asp92Tyr ; Pro117Leu	4.986
Augmented model	Pro117Leu	1.515
Augmented model	Met1Thr ; Ala125Trp	10.672

Mutations are written in 3-letters code as Original-amino-acid / Position-of-mutation-on-the-wild-type-sequence / Mutated-amino-acid.

VII. Discussion

In Fig. 1.a, we can clearly observe the findings reported in the reference paper for the demo dataset. While using only a simple cross-validation on a ridge regression for 240 one-hot encoded sequences, the Spearman correlation between the predicted fitness and the true values is only ∼0.5, indicating a moderate degree of association between the predicted and true fitness ranks. The EV Potts unsupervised model outperforms the linear model, achieving a Spearman correlation of ∼0.67. The augmented model demonstrates an even stronger Spearman rank correlation of ∼0.74, providing strong evidence that this model is promising for our own research.

When using only 28 assay label data points from the demo dataset for training, we can still observe the same trend in the models' performance (Fig. 1.b). However, the supervised model shows very poor results, as expected with so few points and such a large degree of freedom in the one-hot encoded protein sequence. In contrast, the unsupervised model performs exactly as before (Fig. 1.a), which is also expected since the unsupervised model is independent of the number of labeled training points. Finally, the augmented model slightly outperforms the unsupervised model, but the difference in performance of ∼0.1 is not as significant as in the previous experiment. The augmented model's additional contribution to performance thus seem to depend significantly on the number of assay-labeled data points used for training.

Fig. 2 illustrates the performance of the models when trained with the *LPBS*' data. There is high variability in fitness rank predictions, and we cannot draw reliable conclusions from these results due to the absence of a clear tendency. Surprisingly, Ridge regression seems to achieve higher Spearman correlation values, while the unsupervised and augmented models performed more poorly, showing even negative Spearman correlation for some splits of a test set. These models' performance results are inconsistent with the findings from the demo dataset in Fig. 1 & 4, and contradict the claims of the reference paper. We have several hypotheses that can explain these results. One possibility is that we did not properly prepare the evolutionary dataset; the reference paper did not thoroughly document this aspect, relying solely on already existing .a2m datasets for its benchmarks. This could explain why the unsupervised methods don't perform well. Then, the variability observed probably comes from the small test size and the use of Spearman rank correlation as a metric for a dataset of this size. In fact, when using only 7 data points for model assessment, it is difficult to determine if an observed ranked correlation truly reflects a relationship or is simply due to chance. Moreover, this makes the assessment more sensitive to outliers.

Varying the test size of the demo data for a fixed number of training iterations in the supervised model confirmed our hypothesis that the variance in model performance results from the limited test size (Fig. 3). As the test set size increases, we observe a reduction in variation among different test runs, providing a more reliable estimate of the true scoring.

To analyze the in silico evolution, we provide an additional table in the appendix containing the observed optimal mutants and their corresponding dynamic range from the raw assay-labeled dataset (see Table II).

In our machine learning experiments, we found that the two single-mutants obtained through Ridge Regression or the augmented model were already the top two optimal variants present in the assay-labeled dataset (see Table I). The fact that mutants used for training were identified through in silico evolution strongly suggests overfitting, as supported by the poor results from the model assessment in Fig. 2. Additionally, we can notice that the predicted dynamic range differs from their true values. The difference in predicted dynamic range accuracy can be attributed to our metric: the Spearman rank correlation, which solely focuses on the protein's fitness rankings without considering the actual numerical values of fitness.

The double mutant predicted by Ridge regression is simply a combination of these two optimal training variants (Table I). This once again demonstrates overfitting and highlights the limitation of Ridge regression in capturing evolutionary information in a sequence. Regarding the augmented model, the double mutant appears very promising with a high dynamical range of 10.672 (Table I); however, one of the mutations is located at the first position of the sequence, affecting Methionine. From a biological perspective, it is unconventional to mutate the first amino acid of a sequence, Methionine, as protein synthesis universally initiates with this amino acid. This confirms that the performance of the augmented model is poor and not exploitable.

VIII. Conclusion

The project's aim was to use an augmented model proposed in the paper *"Combining evolutionary and assay-labelled data for protein fitness prediction"* [1] to predict which variant of EL222 would result in a revolutionary tool for optogenetics development. Although promising model performance was obtained on the demo dataset provided in their code, we did not manage to achieve similarly great results with the *LPBS* datasets. A better understanding of the evolutionary dataset preparation, another choice of metric, and an enrichment of the assay-labelled dataset would probably significantly help yield better results. Additionally, we decided to train our models for light fitness and dark fitness independently. Reflecting on it, it would have been better to directly transform the fitness values into a dynamic range so that both pieces of information could be captured simultaneously by the methods. Exploring alternative unsupervised methods, such as alpha folds2 combined with sequence clustering to predict multiple protein configurations, could also provide a more promising avenue for further investigation in this task [14].

IX. Ethical Risk Assesment

Risk Definition: The ethical risk in our project involves the potential use of fitness landscape machine learning methods for pathogenic protein editing purposes. For instance, one could amplify the fitness of a pathogen, to make it more harmful to humans.

Stakeholders Impacted: People directly impacted by this risk include the general public, healthcare systems, regulatory bodies, and the scientific community. The indirect potential harm extends to everyone, as even those not directly affected by the pathogen could bear the consequences in a scenario where the engineered pathogen evolves into a pandemic crisis.

Negative Impact: The negative impact involves an increased threat to public health, potential loss of life, strain on healthcare resources, a possible world crisis, and ethical challenges associated with the intentional enhancement of a protein pathogenicity.

Significance of the Risk: The risk is considered highly significant, with severe consequences that could affect the global population. The likelihood of occurrence is low but not negligible as this domain of research continues to develop.

How Did You Evaluate This Risk: To evaluate the risk, we conducted research and found an article by Watters et al. (2021) that discusses why introducing new harmful pathogens would be very dangerous [12].

How Have You Taken This Risk into Account in Your Project: We cannot directly address this risk, as it emphasizes the responsible and ethical use of protein editing technologies. It primarily depends on people's interest. Ethical guidelines and regulations governing gene editing exist, promoting transparent communication regarding the intended purposes of protein editing research. There also exists laws to prevent such things from happening. For example, you can read about the Biological Weapons Convention (BWC) on the Nuclear Threat Initiative website. The BWC mandates the elimination of existing biological weapons and prohibits the development, stockpiling, or use of biological and toxin weapons [13].

References

[1] Chloe Hsu, Hunter Nisonoff, Clara Fannjiang & Jennifer Listgarten. *Combining Evolutionary and Assay-Labeled Data.* [Online] Available: https://www.biorxiv.org/content/10.1101/2021.03.28.437402v1.abstract.

[2] Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; Fergus, R. (2021). *Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences.* [Online] Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8053943/.

[3] Glantz, S. T., Carpenter, E. J., Melkonian, M., Gardner, K. H., Boyden, E. S., Wong, G. K.-S., & Chow, B. Y. (2016). *Functional and topological diversity of LOV domain photoreceptors. Proceedings of the National Academy of Sciences of the United States of America, 113*(11), E1442-E1451. [Online] Available: https://doi.org/10.1073/pnas.1509428113.

[4] Thomas A Hopf, John B Ingraham, Frank J Poelwijk, Charlotta P I Schärfe, Michael Springer, Chris Sander & Debora S Marks. *Mutation effects predicted from sequence co-variation* . [Online] Available: https://www.nature.com/articles/nbt.3769.

[5] Marks, D. S.; Colwell, L. J.; Sheridan, R.; Hopf, T. A.; Pagnani, A.; Zecchina, R.; Sander, C. (2011). *Protein 3D Structure Computed from Evolutionary Sequence Variation. PLoS ONE*, Published: December 7, 2011. [Online] Available: https://doi.org/10.1371/journal.pone.0028766

[6] Barbero-Aparicio, José A. and Olivares-Gil, Alicia and Rodríguez, Juan J. and García-Osorio, César and Díez-Pastor & José F. (2023). *Addressing data scarcity in protein fitness landscape analysis: A study on semi-supervised and deep transfer learning techniques. Departamento de Ingeniería Informática, Universidad de Burgos, Burgos, 09001, Spain* [Online] Available: https://www.sciencedirect.com/science/article/pii/S1566253523003512

[7] Abakarova, Marina and Marquet, Céline and Rera, Michael and Rost, Burkhard and Laine, Elodie. (2023). *Alignment-based Protein Mutational Landscape Prediction: Doing More with Less. Genome Biology and Evolution*, Volume 15, Issue 11, November 2023, evad201. [Online] Available: https://academic.oup.com/gbe/article/15/11/evad201/7344676

[8] Laine, Elodie and Karami, Yasaman and Carbone, Alessandra. (2019). *GEMME: A Simple and Fast Global Epistatic Model Predicting Mutational Effects. Molecular Biology and Evolution*, Volume 36, Issue 11, November 2019, Pages 2604–2619. [Online] Available: https://academic.oup.com/mbe/article/36/11/2604/5548199

[9] Meier, Joshua; Rao, Roshan; Verkuil, Robert; Liu, Jason; Sercu, Tom; Rives, Alex. (2021). *Language models enable zero-shot prediction of the effects of mutations on protein function. Advances in Neural Information Processing Systems 34 (NeurIPS 2021).* [Online] Available: https://proceedings.neurips.cc/paper_files/paper/2021/hash/f51338d736f95dd42427296047067694-Abstract.html

[10] Notin, Pascal; Dias, Mafalda; Frazer, Jonathan; Marchena-Hurtado, Javier; Gomez, Aidan; Marks, Debora S.; Gal, Yarin. (2022). *Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval.* In: *International Conference on Machine Learning*. PMLR. p. 16990–17017. [Online] Available: https://arxiv.org/abs/2205.13760

[11] Niu, Jacqueline and Ben Johny, Manu and Dick, Ivy E. and Inoue, & Takanari (August 17, 2016). *Following Optogenetic Dimerizers and Quantitative Prospects.* [Online] Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5034304/.

[12] Watters, K. E., Kirkpatrick, J., Palmer, M. J., & Koblentz, G. D. (2021). The CRISPR revolution and its potential impact on global health security. *Pathogens and Global Health, 115*(2), 80–92. doi: 10.1080/20477724.2021.1880202. PMID: 33590814, PMCID: PMC8550201. [Online] Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8550201/

[13] Biological Weapons Convention (BWC). (n.d.). Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons (BTWC). [Online] Available: https://www.nti.org/education-center/treaties-and-regimes/convention-prohibition-development-production-and-stockpiling-\bacteriological-biological-and-toxin-weapons-btwc/

[14] Wayment-Steele, H. K., Ojoawo, A., Otten, R., Apitz, J. M., Pitsawong, W., Hömberger, M., Ovchinnikov, S., Colwell, L., Kern, D. (2023). *Predicting multiple conformations via sequence clustering and AlphaFold2. Nature*, Nov 13. doi: 10.1038/s41586-023-06832-9. [Online ahead of print] Available: https://pubmed.ncbi.nlm.nih.gov/37956700/.

[15] Bitbol Anne-Florence Raphaëlle *Randomness and information in biological data*. [Online] Available: https://edu.epfl.ch/coursebook/en/randomness-and-information-in-biological-data-BIO-369.

APPENDIX A: POTTS MODEL

In the following, we provide a deeper overview of the Potts Model, which is derived from the BIO-369 course and the Hopf paper [4] [5] [15].

In a protein sequence, the arrangement and interactions between amino acids determine the protein's 3D structure. The Potts model uses prior information from the multiple sequence alignment of the protein's family to derive the probability distribution for observing each possible sequence $s = (s_1, \dots, s_L)$. This probability density $P(s)$ captures the evolutionary dynamic of the target protein, revealing the likelihood of a given sequence in nature.

$P(s)$ Derivation:

Let's consider an amino acid sequence $s = (s_1, \dots, s_L)$ of length L, where $s_i \in \mathcal{A}$ and $\mathcal{A}$ represents all possible amino acids. Our goal is to find the probability of observing a particular sequence s, denoted as $P(s)$, consistent with the observed frequencies in the multiple sequence alignment.

$$f^i_\alpha = \sum_s P(s)\delta^\alpha_{s_i} \quad \text{with} \quad \delta^\alpha_{s_i} = \begin{cases} 1 & \text{if } s_i = \alpha \\ 0 & \text{otherwise} \end{cases}$$

the frequency of observing $s_i = \alpha$ at position i and

$$f^{ij}_{\alpha\beta} = \sum_s P(s)\delta^{\alpha\beta}_{s_i s_j} \quad \text{with} \quad \delta^{\alpha\beta}_{s_i s_j} = \begin{cases} 1 & \text{if } s_i, s_j = \alpha, \beta \\ 0 & \text{otherwise} \end{cases}$$

the frequency of observing $s_i = \alpha$ at position i and $s_j = \beta$ at position j.

Then, to account for unobserved interactions of amino acids in the sequence, the probability $P(s)$ can be determined by maximizing the entropy:

$$H = \sum_s P(s) \log P(s)$$

with the following prior information constraints:

$$\sum_s P(s) = 1, \quad f^i_\alpha = \sum_s P(s)\delta^\alpha_{s_i}, \quad f^{ij}_{\alpha\beta} = \sum_s P(s)\delta^{\alpha\beta}_{s_i s_j}$$

Whereby, we obtain with the Lagrange Multiplier method:

$$\begin{aligned}\hat{H} = \sum_s P(s)\log P(s) + \gamma(1 - \sum_s P(s)) \\ + \sum_{i\alpha} h^i_\alpha (f^i_\alpha - \sum_s P(s)\delta^\alpha_{s_i}) \\ + \sum_{ij\alpha\beta} J^{ij}_{\alpha\beta}(f^{ij}_{\alpha\beta} - \sum_s P(s)\delta^{\alpha\beta}_{s_i s_j})\end{aligned}$$

where γ, h^i_α, and $J^{ij}_{\alpha\beta}$ are the Lagrange multipliers.

Maximizing $\hat{H}$ with respect to $P(s)$:

$$\frac{\partial \hat{H}}{\partial P(s)} = 0$$

yiels the final expression:

$$P(s) = \frac{1}{e^{1-\gamma}} \exp\{-(\sum_{i=1}^{N} h^i_{s_i} + \sum_{1\leq i<j\leq N} J^{ij}_{s_i s_j})\} = \frac{1}{Z} e^{-E(s)}$$

where:

- h^i_α is a site-specific parameter that depends on the amino acid α at position i.
- $J^{ij}_{\alpha\beta}$ is the coupling parameter describing pairwise interactions for amino acids α and β at positions i and j.
- $E(s)$ is referred to as the 'Potts energy' function, and captures the statistical energy landscape to model the effect of mutation in a sequence s.
- Z is a normalization constant [4].

Determining the values of $J^{ij}_{\alpha\beta}$ and h^i_α is a challenging task. The plmc package for EVMutation employs undirected graph models to infer these parameters [1] [4] [5].

Appendix B: Additional Results

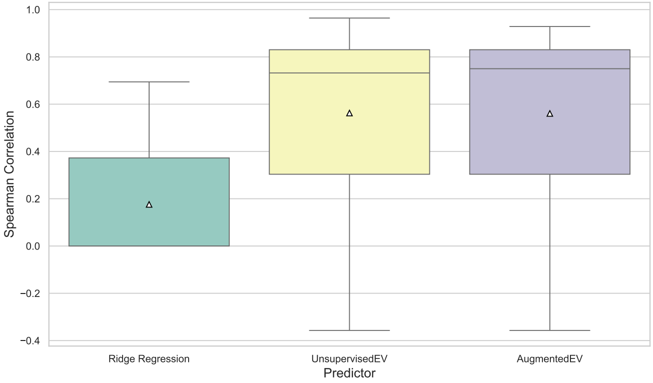

Fig. 4. Spearman correlation box plot for demo data. Ran on 20 seeds: 28 training points, 7 test points.

Appendix C: Optimal Mutants Identified in Assays

TABLE II
Dynamic Range obtained from the assay-labeled dataset

Mutant	Dynamic Range
Asp92Tyr	3.22
Pro117Leu	3.33

Mutations are written in 3-letters code as Original-amino-acid / Position-of-mutation-on-the-wild-type-sequence / Mutated-amino-acid.