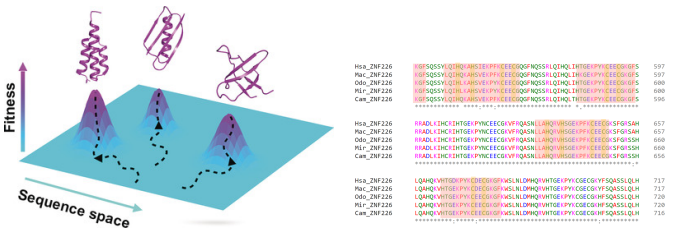
ML4Science Project: Exploration for mutants with enhanced protein fitness

Sara Zatezalo
Section of
Electrical Engineering

Marija Zelic
Section of
Life Sciences Engineering

Elena Mrdja
Section of
Life Sciences Engineering

***Abstract*—Efficient exploration of the protein fitness landscape can be crucial for finding proteins with desired functions. Our project aimed to find machine learning methods that can help the LPBS laboratory in finding suitable mutants of the optogenetic protein EL222. Since laboratory testing of protein function poses financial and time limitations, it was important to find a way to unravel the EL222 fitness landscape and find proteins with the most suitable optic traits, that would serve as candidates for further testing. Toward this goal, we combined various approaches such as AlphaFold (AF), AF Cluster, XGBoost, and BERT into a reliable method.**



(a) Protein fitness landscape. (b) Multiple Sequence Alignment

Fig. 1

I. Introduction

A. Background

Proteins are the machinery of life, constructed as unique sequences arranged from a pool of 20 amino acids (residues). These molecular entities are crucial in biological systems, serving as vital contributors to both the construction of cellular structures and the facilitation of biochemical reactions. They also act as mediators in intercellular communication, thus defining their distinct functionalities. Apart from their unique role, the magnitude of their functional execution, the protein fitness, holds significant scientific value. Researchers aim to assess the impact of various amino acid mutations on protein fitness, focusing on sequence alterations. This assessment entails creating a protein fitness landscape 1, not only to determine fitness across natural sequences and their variations but also to explore novel mutations, serving as a tool for discovery. As it turns out, this is a very extensive task since the space of possible alterations is related to the size of a specific protein sequence and a range of 20 possible amino acids, so, for a protein with L amino acids, the total number of mutants scales exponentially, with 20^L. Extensive and costly laboratory procedures have rendered the direct evaluation of mutations impractical. Consequently, researchers have shifted their methodology toward leveraging machine learning tools to anticipate a protein's fitness function based on its inherent properties.

Thus far, all attempts to address this challenge have centered on examining the primary structure of proteins - the raw amino acid sequence, primarily sourced from evolutionary data encompassing sequences of proteins in related species exhibiting similar functions. However, a protein's function is not solely dictated by its primary structure; the tertiary structure, a specific spatial conformation of the amino acid sequence necessary for the protein to attain its active state and execute its function, also significantly influences its functional capacity.

To establish a link between the primary structure of a protein and its function, there was an ongoing necessity to develop a tool that would predict protein spatial conformation solely based on its amino acid sequence. Predicting the three-dimensional structure that a protein will adopt based on its sequence—the structure prediction component of the 'protein folding problem'—has been an important open research problem for more than 50 years [1]. The breakthrough and solution for this problem has been established by Deep Mind's discovery of Alpha Fold, a deep learning model that gives astounding results in the prediction of protein folding. Alpha Fold employs the Multiple Sequence Alignment (MSA)1 to predict protein structures, which is a method used to align and compare multiple biological sequences, revealing similarities and differences to understand their evolutionary and functional relationships. Thus, AF infers patterns of interactions between related sequences, supposing that amino acids exist and co-evolve in the context of their 3D structure in patterns reflective of their underlying structure [2]. Considering that Alpha Fold represents the state-of-the-art method in predicting the protein structure from its sequence, and consequently its function, we decided to explore its capabilities in our concrete problem.

B. Problem

The research field of the Laboratory of the Physics of Biological Systems, which hosted our ML4Science project, is optogenetics - proteins that can detect light and transmit that signal further into the cell. Furthermore, they have discovered through directed evolution experiments that certain mutations have an impact on the protein that makes it more sensitive to light. Since, as we have previously addressed, finding

mutants of proteins with specific fitness takes a lot of time and resources, they have turned to us to help them develop an AI model that would systematically search for mutations that result in more favorable protein fitness properties.

The protein of interest, EL222, is a bacterial Light-Oxygen-Voltage (LOV) protein, containing LOV domain by which it can sense light. [3]. What is specific about this protein is that it takes on two conformations, the OFF conformation in darkness and ON upon light absorption. Only in the ON state, the EL222 protein binds the DNA and performs its biological functions. The laboratory provided us with a list of a 40, mostly single-point mutants that are more sensitive to light, likely to take on the ON conformation, compared to the wild-type EL222, the typical form of a species as it occurs in nature. Our task was to find other such mutants, that had not previously been assessed in their fitness landscape, potentially including multiple-point mutations, that would make the protein more sensitive to light, i.e. have a higher fitness function value.

To explore the entire landscape1 and identify mutants maximizing the fitness function, researchers rely on datasets significantly larger than the provided data. Even so, the models that are used do not provide highly accurate results, concretely, at highest 55% accuracy on the dataset of 48 mutants, especially for multiple-point mutations [4]. Furthermore, the benchmark models do not take into account the mutations that change the protein conformation, which is a specific situation in our problem.

C. Our approach

So far, Alpha Fold has been successful in predicting single structures from the protein's sequence, but not in predicting all structures of proteins that can adopt several conformational states. Therefore, we turned to a newly published approach - Alpha Fold Cluster [2], which by clustering the MSA by its sequence similarity enables Alpha Fold to sample alternate states of such protein. We categorized our predictions as ON and OFF conformations, based on how closely they resembled the ground-truth OFF EL222 structure. By examining sequences within clusters contributing to the ON conformation and comparing them to those in the OFF group, we have the potential to design protein mutants that preferentially adopt the ON conformation, exhibiting higher fitness.

Finally, we entertained the notion of forecasting a protein's conformational state exclusively from its sequence, bypassing Alpha Fold and its structural predictions. In pursuit of this objective, we employed two distinct machine learning models—XGBoost and the transformer-based deep learning model BERT—to categorize proteins into their respective conformational states solely based on their sequences.

II. Methods and results

The first step that our project required was to collect the data that would be useful for the Alpha Fold model to predict the structures of the EL222 protein. The collection of sequences of proteins that have a similar sequence and potentially a function as our observed protein is acquired by feeding the specific LOV domain sequence to the ColabFold [5] that searches against sequences from the UniRef30 database [6]. The LOV domain is a protein domain that is involved in sensing and responding to light stimuli, and in EL222 it spans from 16th to 167th residue. The sequences of these proteins form a MSA. Visual depictions reveal mutation events, including point mutations and insertion/deletion mutations, providing insights into sequence conservation within protein domains and tertiary structures.

After obtaining MSA related to the query EL222 domain from Alpha Fold, these sequences were fed to the AF Cluster [2], which first clusters MSA sequences using DBSCAN [7], and then generates one protein structure prediction for each cluster using Alpha Fold.

Predictions of Alpha Fold outputs for clusters were evaluated against the ground truth EL222 structure, specifically its LOV domain which was deprived of some amino acid regions that are considered to be unresolved in the original chain [8]. We used the Root Mean Square Deviation metric (RMSD), which measures the average distance between atoms of two structurally aligned sequences, in our case, protein tertiary structures. The RMSD is calculated as follows:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2} \tag{1}$$

where δ_i is the distance between the observed atom i and its equivalent in the reference structure. Lower RMSD values indicate higher structural similarity, signifying closer alignment between the atomic coordinates of the compared protein structures. In this specific context, where the reference structure that was used is the experimentally determined OFF conformation of the EL222 protein, our hypothesis centered on utilizing the RMSD value as a measure to assess the proximity of the conformation predicted by Alpha Fold within the identified cluster to either the ON or OFF conformations of the protein.

Two different ways of computing the RMSD were employed to attain this goal. First, we observed the values of RMSD computed for each cluster on all residues used for alignment. The histogram we obtained 4, showed a clear separation of two accumulated lobes of lower and higher values of RMSD, which would ideally correspond to the OFF and ON conformations, respectively. Next, to validate that the peak with higher RMSD indeed corresponds to the ON state of the protein, we focused only on the residues that are known to shift upon light absorption, i.e. upon the change from the OFF to the ON conformation [9], and compute the RMSD for those residues solely. The major conformational switch is known to be at the residue Gln138 in the EL222 protein [9], such that upon light absorption the amide group points away from the cysteine. We also took into account residue 93 which is near the chromophore, which could additionally indicate changes between conformations within the value of RMSD. The chromophore is a molecule that is an essential part of the LOV domain since it absorbs light and undergoes a conformational change upon absorption. After experimenting with the

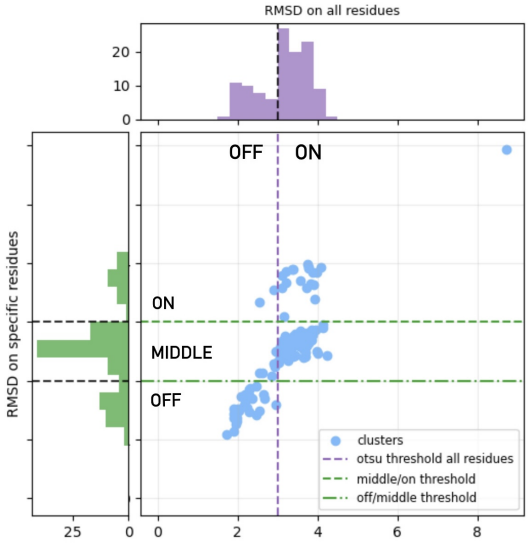


Fig. 2: Histograms of computed RMSD values for two different approaches; the histogram on top shows results for RMSD computation on all residues, while the left corresponds to specific residue computation.

different combinations of residues to take into account, we found that the best results in terms of cluster discrimination were obtained for residues 93, 108, and 138, which gave us the output represented in Figure 4.

Observing the histograms obtained by calculating the RMSD in these two different approaches, we can see that for the specific residue calculation, we end up having three categories, which we name ON, OFF, and MIDDLE. Upon analyzing the structures within each cluster category, we found that Gln 138, which is switching in the ON state, is indeed pointing toward the chromophore in OFF clusters and away from it in the ON clusters, which illuminates the known ON and OFF conformations assumed by the EL222 protein [9]. For further analysis and application, we designate clusters and their associated structures categorized as ON for both RMSD calculations as representatives of ON conformations, while conversely classifying those falling into the OFF category as representatives of OFF conformations.

From the 149 initially acquired clusters from the original EL222 MSA, we ended up with 19 ON clusters (105 sequences from the MSA in total) and 28 OFF clusters (3950 sequences). The high imbalance in sequences from the original MSA is expected since the OFF conformation is more likely to be found in nature.

To evaluate the effectiveness of classification when trained on sequences we extracted using Alpha Fold, we compared the performances of XGBoost, an ensemble learning method, and BERT, a transformer-based deep learning model.

XGBoost (eXtreme Gradient Boosting) is an implementation of gradient-boosted decision trees [10]. Similarly to other ensemble learning methods, XGBoost combines decisions from multiple machine learning models to reduce errors and improve prediction compared to a single model. It does so through gradient boosting by adding a new decision tree one at a time to a model that minimizes the loss using gradient descent. Boosting reduces overfitting by never changing the previous decision trees and only correcting the next predictor by learning from mistakes. Previous work suggests that XGBoost performs well on protein classification tasks based on positional relationships between amino acids [11], but to the extent of our knowledge, it has not yet been paired with MSA clustering to classify protein conformations. To evaluate the performance of XGBoost on this task, we tested it using 10-fold cross-validation. The final testing accuracy obtained was 69.7%, which we aimed to improve by using a transformer-based method.

BERT (Bidirectional Encoder Representations from Transformers) [12] is a leading pretraining model based on the Transformer architecture, utilizing bidirectional processing to capture intricate token dependencies. Trained with the masked language modeling (MLM) objective on a large amount of data, BERT leverages contextual information to predict original tokens, enhancing its comprehension of token contexts and interrelationships within sequences. Given its exceptional efficacy in addressing natural language processing (NLP) challenges, we opted to harness the capabilities of BERT for a non-NLP-oriented problem, specifically in the context of protein sequences, as we anticipate its ability to capture intricate contextual dependencies and enhance performance in this domain.

We seized the chance to fine-tune BERT for our specific downstream task, where the objective is to predict whether a given protein sequence aligns with an ON or OFF cluster. This process involved initially applying the MLM approach, with slight modifications, which will be addressed. The protein sequences employed for this supplementary task were derived from previously mentioned datasets associated with all clusters (ON, OFF, and MIDDLE), from which 80% was contributed for training and 20% for testing. These were tokenized following the strategy in which every two consecutive amino acids are taken as a token, with an overlap of one amino acid between adjacent tokens. Additionally, [CLS] (classification) and [SEP] (separation) tokens are added to the beginning and end of each sequence, respectively. The output corresponding to the [CLS] token is later used as a pooled representation that encapsulates the information from the entire sequence, which is of immense importance for the downstream classification task. In this context, the [SEP] token, typically used for sentence separation in NLP problems, is retained though not crucial for the current task. Besides this, each token is assigned with its sinusoidal positional encoding [13] to indicate its position in the sequence. The masking rate was configured at 5%, and to avoid the trivial task of randomly masking individual tokens, we opted to mask two consecutive tokens, aligning with the tokenization approach we employed. The model was trained for 250 epochs and in such configuration yielded 53% of testing accuracy.

The BERT model fine-tuned for the MLM problem provided

us with the embeddings for each sequence through the previously mentioned [CLS] token. For the final classification of the sequences towards the ON and OFF clusters, we developed a fully connected classifier and adopted two approaches for obtaining datasets, both of them trying to address the pronounced imbalance in sequence counts between ON and OFF clusters. The first one attempts to mitigate this discrepancy by subsampling the majority class, which results in 210 sequences in total (105 sequences for each class). While acknowledging a potential modest loss of information, this strategy will prove to be essential for enhancing the model's overall predictive efficacy. The other one constructs the so-called consensus sequence for each ON and OFF cluster by determining the most frequently occurring amino acid at each position across all sequences within the respective cluster, resulting in a total of 47 sequences (28 belonging to OFF class and 19 belonging to ON class), which is significantly smaller than in the first case.

The best classifiers for each dataset were determined by conducting hyperparameter tuning across a grid encompassing different fully connected architectures, learning rates, and regularization parameters within the optimizer (one used was AdamW), which served as a crucial factor in preventing overfitting and minimizing redundancy. Given the limited size of both datasets, we integrated a 5-fold cross-validation approach during hyperparameter tuning to maximize the utilization of available data. We report the following results:

Method	Test accuracy
BERT model classifier on balanced sequences	82%
BERT model classifier on consensus sequences	60%
XGBoost	69.7%

TABLE I: Performance of the models

The BERT model-based classifier achieves higher accuracy values than XGBoost on the balanced sequences (Table I). Its performance is comparable to the XGBoost one regardless of the negligible amount of available consensus sequences. This may be attributed to the fine-tuned BERT model's ability to extract key sequence features through the [CLS] token representation. It is important to note that we also performed an ablation study to evaluate how this classifier reacts to changes in its hyperparameters. The presence of the regularization parameter played a crucial role by preventing overfitting and facilitating smoother convergence in our observations.

Along with a tool to accurately predict the conformation of a given EL222 mutant based on its sequence, as mentioned before, the LPBS lab was looking for a method to find EL222 mutants that could be suitable candidates for further laboratory testing. For this purpose, we used the sequences extracted during AlphaFold Clustering to analyze the differences between the amino acids at positions of interest in ON and OFF sequences. Locating these differences would allow us to narrow down the choice of proteins that are likely to take on the ON conformation, and test if this is truly the case in laboratory conditions.

Given a set of sequences, the amino acid that is most frequent among these sequences at a fixed position is called the consensus amino acid. We visually represent the consensus amino acids of the entire LOV domain of the extracted ON sequences and the LOV domain in the extracted OFF sequences (Supplementary Fig. 5). Comparing the consensus ON sequence with the consensus OFF at positions 40-80 of the LOV domain, we can notice they differ at positions 43, 46, 47, 53, 68, and 74 (Fig. 3). Taking the mutants with sequences of amino acids that are consensus in ON confirmations but not present at the same position in the OFF consensus, could be an effective strategy for expanding the EL222 fitness landscape. Proteins that are likely to be in the ON conformation could be found by choosing a sequence with some of the mutations we have found in our analysis (Table II) and confirming its optogenetic properties through testing.

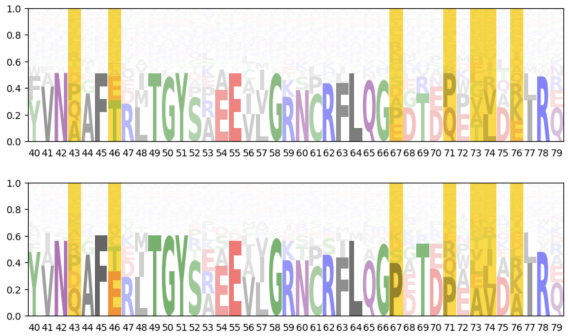


Fig. 3: Most frequent amino acids at positions 56-96 in EL222 (equivalent to position 40-80 in the LOV domain) of (upper) ON sequences, (bottom) OFF sequences. Differences in consensus amino acids for ON and OFF sequences are highlighted in yellow.

TABLE II: Positions where consensus in amino acids differs between ON and OFF sequences in the LOV domain

Pos	ON	OFF	Pos	ON	OFF	Pos	ON	OFF
3	I	V	39	I	V	121	I	V
6	R	K	43	A	Q	123	N	R
9	Q	E	46	T	E	125	N	R
11	Q	E	67	E	P	135	E	R
13	Q	R	71	Q	P	136	E	R
19	V	L	73	T	A	139	Y	E
21	S	A	74	L	V	140	A	Q
22	A	S	76	E	A	142	R	N
26	I	V	108	T	S	143	L	H
34	A	P	121	I	V	150	T	A

III. Conclusion

Our project successfully integrated and combined state-of-the-art machine learning methods, such as AlphaFold Cluster, XGBoost, and BERT, to explore the protein fitness landscape. We discovered that the combination of these approaches provides a promising methodology for discovering candidates for further experimental evaluation.

IV. Ethical risks

As computational methods that aid gene editing and protein tweaking are becoming more advanced, there's a crucial need to ponder the ethical implications of these advancements. Manipulating the genetic makeup of proteins, as we did with EL222, raises questions about unintended consequences, potential environmental impacts, and the responsible use of such technology [14]. The dialogue around these ethical considerations becomes even more significant when we think about the broader implications – how what we discover in a yeast protein might have an effect that reaches humans.

It is crucial to delve into the various stakeholders who may hold an interest in or be affected by our project, which centers around exploring the fitness landscape of a yeast protein. This acknowledgment is vital given the potential implications our research might have on the broader biological landscape, including human biology [15].

Firstly, our attention should be directed towards two distinct categories of project stakeholders, the research industry and the broader public. The scientific community and researchers are actively engaged in the exploration of the fitness landscape of various proteins, including those active in humans. Striking a balance between scientific curiosity and ethical responsibility becomes important in this context, ensuring that our research adheres to ethical standards and contributes positively to scientific knowledge[16].

Secondly, the broader public, including non-scientific individuals, and advocacy groups forms another essential category of stakeholders. Their interest in the ethical implications of our research necessitates transparency and engagement. It is imperative to acknowledge the potential impact on human biology and the larger biological picture, aligning our communication with societal values to address ethical concerns.

Beyond these primary stakeholders, consideration should be given to indirect stakeholders, such as media outlets, which play a crucial role in shaping public perception. Proactive engagement with the media becomes a valuable strategy to ensure accurate representation and mitigate potential misinformation. Furthermore, collaboration with environmental protection rights groups is essential to prevent any misuse of our research, given the potential ecological implications of our findings[17]. Communications with human rights organizations would also be important, in order to not breach personal privacy or endanger human health with our discoveries.

It's a great responsibility in research to have moral integrity and acknowledge the ethical risks of one's findings, and essential as we venture further into the exploration of protein analysis and its potential influence on life as we know it.

V. Appendix

A. Project pipeline

The flowchart of the methods used in the project for better understanding and visualization is given in the following figure:

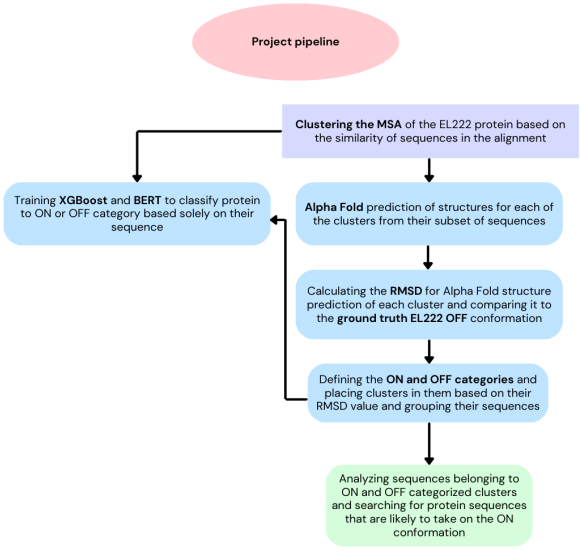


Fig. 4: Flowchart of the methods we used in the project.

B. Regarding the calculation of RMSD and analyzing ON, OFF, and MIDDLE categories

Even though the residue Cys 75 is the crucial residue in LOV domains, since it changes its chemical structure upon illumination, forming a bond with the chromophore, we decided to exclude it from the calculation for several reasons. First, it requires simulating the protein structure with the chromophore, which is currently impossible in Alpha Fold as it works with amino acids only. Additionally, the shift of the position of the sidechains is bigger for Gln-138 than for Cys-75 (Cys-75 does not move as much), thus as a larger alteration in the RMSD value might not occur, directing attention to Gln-138 proves more advantageous and practical.

The MIDDLE category 4 seems to represent the "flip" conformation of the protein [18], which can occur from both the "swing" state (an intermediate state of the path to ON) and the dark state (OFF). Since this state adds no significance for the exploration of its sequences in search of ON-prone mutations, we decided to exclude this category of clusters from further analysis.

References

[1] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, Aug. 2021.

[2] H. K. Wayment-Steele, A. Ojoawo, R. Otten, J. M. Apitz, W. Pitsawong, M. Hömberger, S. Ovchinnikov, L. Colwell, and D. Kern, "Predicting multiple conformations via sequence clustering and AlphaFold2," *Nature*, Nov. 2023.

[3] L. B. Motta-Mena, A. Reade, M. J. Mallory, S. Glantz, O. D. Weiner, K. W. Lynch, and K. H. Gardner, "An optogenetic gene expression system with rapid activation and deactivation kinetics," *Nature Chemical Biology*, vol. 10, no. 3, p. 196–202, Jan. 2014. [Online]. Available: http://dx.doi.org/10.1038/nchembio.1430

[4] C. Hsu, H. Nisonoff, C. Fannjiang, and J. Listgarten, "Learning protein fitness models from evolutionary and assay-labeled data," *Nat. Biotechnol.*, vol. 40, no. 7, pp. 1114–1122, Jul. 2022.

[5] M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, and M. Steinegger, "ColabFold: making protein folding accessible to all," *Nat. Methods*, vol. 19, no. 6, pp. 679–682, Jun. 2022.

[6] B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, and C. H. Wu, "Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches," *Bioinformatics*, vol. 31, no. 6, p. 926–932, Nov. 2014. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btu739

[7] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN revisited, revisited," *ACM Trans. Database Syst.*, vol. 42, no. 3, pp. 1–21, Sep. 2017.

[8] "Unresolved residues in the groundtruth el222 structure," https://www.rcsb.org/3d-view/3P7N.

[9] B. D. Z. et al., "Conformational switching in the fungal light sensor vivid.science316,1054-1057(2007)."

[10] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *CoRR*, vol. abs/1603.02754, 2016. [Online]. Available: http://arxiv.org/abs/1603.02754

[11] X. Zhao, X. Wang, Z. Jin, and R. Wang, "A normalized differential sequence feature encoding method based on amino acid sequences," *Mathematical Biosciences and Engineering*, vol. 20, pp. 14 734–14 755, 07 2023.

[12] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[13] "Sinusoidal positional encoding for tokenization," https://kazemnejad.com/blog/transformer_architecture_positional_encoding/.

[14] M. Almeida and R. Ranisch, "Beyond safety: mapping the ethical debate on heritable genome editing interventions," *Humanities and Social Sciences Communications*, vol. 9, no. 1, Apr 2022. [Online]. Available: https://www.nature.com/articles/s41599-022-01147-y.pdf

[15] S. P. Mann, P. V. Treit, P. E. Geyer, G. S. Omenn, and M. Mann, "Ethical principles, constraints, and opportunities in clinical proteomics," *Molecular Cellular Proteomics*, vol. 20, p. 100046, 2021.

[16] J. B. Tucker and C. Hooper, "Protein engineering: security implications," *EMBO reports*, vol. 7, no. S1, pp. S14–S17, 2006. [Online]. Available: https://www.embopress.org/doi/abs/10.1038/sj.embor.7400677

[17] J. Hadi and G. Brightwell, "Safety of alternative proteins: Technological, environmental and regulatory aspects of cultured meat, plant-based meat, insect protein and single-cell protein," *Foods*, vol. 10, no. 6, p. 1226, May 2021. [Online]. Available: https://www.mdpi.com/2304-8158/10/6/1226

[18] M. E. Kalvaitis, L. A. Johnson, R. J. Mart, P. Rizkallah, and R. K. Allemann, "A noncanonical chromophore reveals structural rearrangements of the light-oxygen-voltage domain upon photoactivation," *Biochemistry*, vol. 58, no. 22, pp. 2608–2616, Jun. 2019.

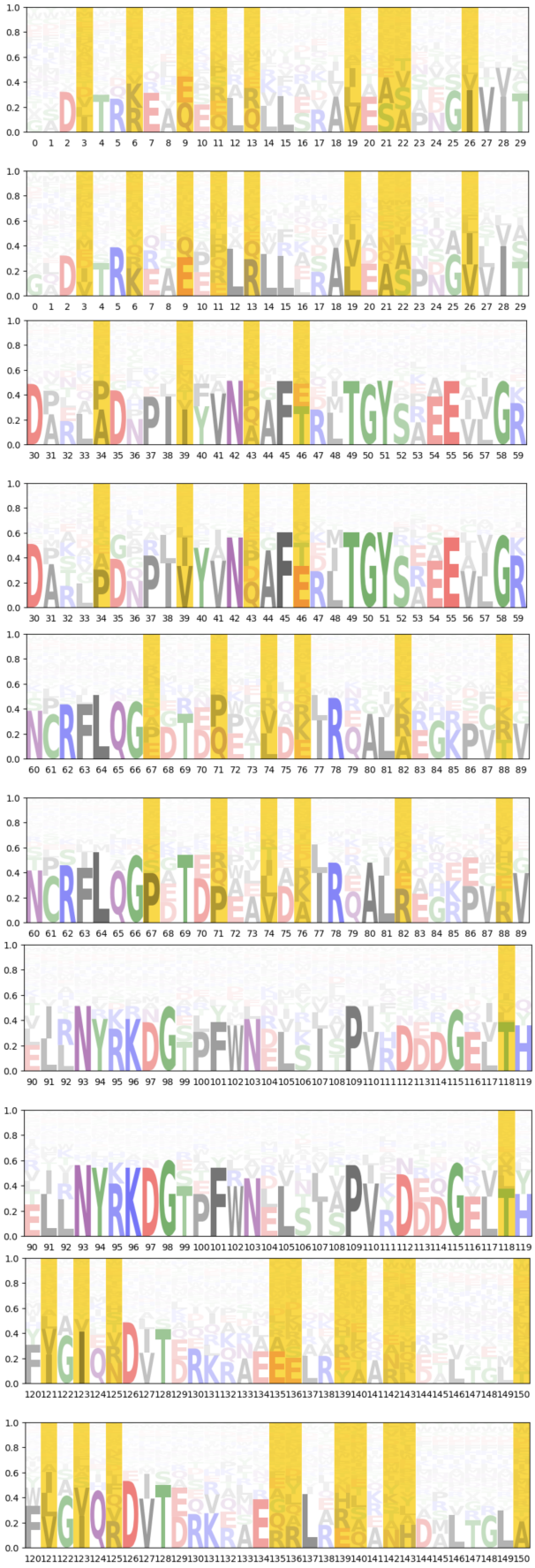


Fig. 5: Most frequent amino acids on the entire LOV domain for ON and OFF conformations