Evol3D: A Large-Scale 3D Structure Dataset for Investigating Mutational Effects in Homologous Proteins

Zhiqiang Zhong*[a], **Yinghua Yao**[b], **Davide Mottin**[c], **Jun Pang**[a]

[a] *University of Luxembourg, Luxembourg* zhiqiang.zhong@uni.lu, jun.pang@uni.lu

[b] *Agency for Science, Technology and Research (A*STAR), Singapore* yao_yinghua@cfar.a-star.edu.sg

[c] *Aarhus University, Denmark* davide@cs.au.dk

* Presenting author

1. Introduction

Table 1: Data statistics.

Dataset	# Proteins	Avg. PTM	Avg. pLDDT
GB1	8,733	0.65	78.4
Fluorescence	54,025	0.92	92.4
AAV	82,583	0.72	84.9

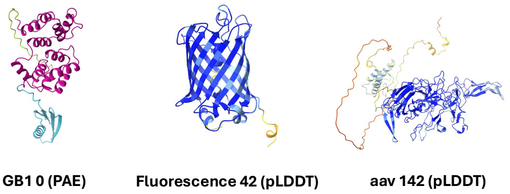


Fig. 1: Visualization of protein 3D structure, with PAE and pLDDT prediction errors.

Proteins existing today have evolved over billions of years through natural selection, undergoing continuous refinement via a vast evolutionary sieve [1]. This evolutionary process has resulted in proteins optimized to perform diverse functions essential for life, shaped by countless incremental mutations and selective pressures. Inspired by these natural evolutionary dynamics, contemporary protein engineering employs directed evolution methods that iteratively mutate a naturally occurring protein, referred to as *wild-type*, to achieve desirable properties [2]. Such mutations create families of related proteins, known as *homologous protein families*, providing valuable insights into the relationship between sequence variations and protein functionalities.

Recent work, notably EvolMPNN [3], has leveraged neural networks to effectively model evolutionary patterns and predict mutational effects within homologous protein families. However, EvolMPNN considers exclusively sequence-level data, omitting the essential 3D structural context. Protein functions are inherently determined by their 3D structures; therefore, the integration of structural information is essential for a complete understanding of mutational effects. Protein structural insights enable a deeper exploration into how mutations influence protein behaviors at the molecular level.

Latest advancements in computational protein structure prediction, particularly via deep learning models including AlphaFold [4] and ESMFold [5], have enabled researchers to generate accurate protein 3D structures at large scales. Despite these technological breakthroughs, the research community still lacks extensive and systematically annotated datasets integrating sequence information, 3D structure predictions, and functional annotations specifically tailored for homologous protein families.

To bridge this gap, we introduce Evol3D[1], a large-scale open-source dataset comprising around 150K AlphaFold 3 [6] predicted protein 3D structures derived from three extensively studied homologous protein families [7, 8]: GB1 [9], Fluorescence [10], and AAV [11]. Dataset statistics are provided in Table 1 and some example predicted 3D structures are illustrated in Figure 1. Evol3D integrates high-quality 3D structural predictions with detailed functional annotations, offering researchers the ability to systematically analyze how structural variations induced by mutations impact protein function from an evolutionary standpoint. Initial statistical assessments validate the reliability and quality of the dataset, setting the stage for future extensive analyses.

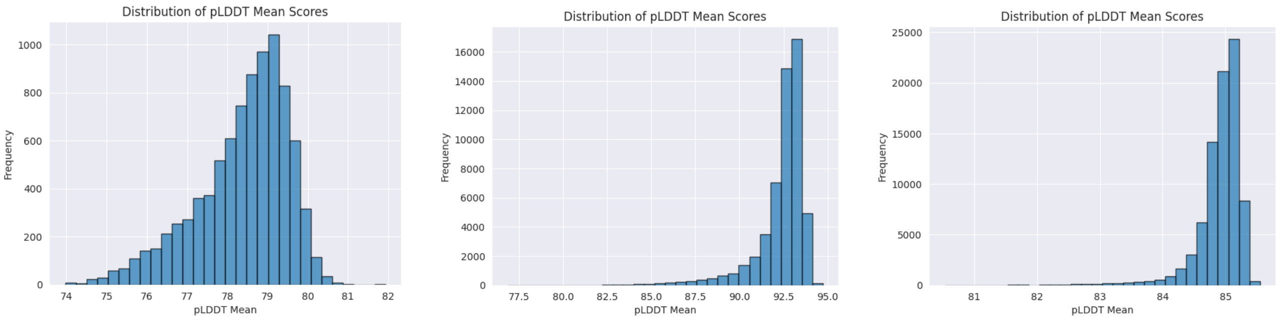
2. Evol3D Dataset Overview

The Evol3D dataset is explicitly designed to facilitate detailed investigations of the mutational effects on protein function and structure from an evolutionary perspective. Each protein structure within this dataset is systematically annotated with functional labels, allowing comprehensive analyses of the relationship between structural changes induced by mutations and corresponding protein functionalities.

The dataset comprises proteins from three well-studied homologous families:

GB1 family. The GB1 protein family [9] has emerged as a gold standard in studying epistatic interactions. GB1 refers to the binding domain of protein G, an immunoglobulin-binding protein derived from *Streptococcal* bacteria. Wu *et al.* [9] originally measured the fitness landscape for 149 protein variants, providing a foundational resource for examining epistatic interactions and adaptive protein evolution.

[1] `https://github.com/zhiqiangzhongddu/Evol3D`



(a) GB1

(b) Fluorescence

(c) AAV

Fig. 2: Distribution of mean pLDDT scores for the GB1, Fluorescence, and AAV homologous protein families, indicating structural confidence of predicted 3D structures.

Fluorescence family. The green fluorescent protein has been extensively investigated due to its importance as a fluorescent marker in biological experiments. The mutational landscape of green fluorescent protein provides insights into how specific mutations can enhance or diminish its fluorescence, offering insights into underlying mutational patterns affecting biological function [12].

Adeno-associated Virus (AAV) family. AAV capsid proteins facilitate the integration of genetic payloads into target cells and hold significant potential for gene therapy applications [13]. Bryant *et al.* [11] explored the mutational landscape of the VP-1 capsid protein, specifically mutagenizing a 28-amino acid window (positions 561–588), generating variants with between 1 and 39 mutations, and assessing their fitness. These protein variants constitute a valuable dataset for exploring mutational impacts relevant to therapeutic protein engineering.

3D Structure Prediction Procedure. To enrich the dataset with structural information, we employ AlphaFold 3 [5] [2], the state-of-the-art deep learning model for protein structure prediction. Each protein sequence is predicted using three distinct random seeds (42, 2024, and 3407), and five structures are generated per seed, resulting in a total of 15 predictions per protein. This ensemble approach increases reliability and robustness. This extensive prediction effort required approximately *36,000 A100 GPU hours*. Statistics summarizing the dataset, including the average predicted template modeling (PTM) and predicted Local Distance Difference Test (pLDDT) scores, are provided in Table 1. Representative predicted structures are visualized in Figure 1.

Preliminary Analysis. Table 1 summarizes key statistics for each protein family, indicating that the Fluorescence family achieves the highest average PTM (0.92) and pLDDT (92.4) scores, reflecting very high confidence in its predicted structures. GB1 exhibits relatively lower average scores (PTM: 0.65, pLDDT: 78.4), suggesting greater complexity or variability in structural prediction within this family. Figure 2 further highlights these distinctions, showing narrower and higher-confidence score distributions for Fluorescence and AAV families, while GB1 displays broader distributions indicating variable prediction reliability across different proteins. Additionally, we plot pLDDT scores for different regions of sampled proteins (see Appendix). The preliminary observations confirm that the majority of regions in all families are predicted with high confidence (pLDDT >90), although certain challenging regions remain, emphasizing opportunities for deeper structural investigations.

3. Future Work

While our preliminary analyses validate the quality and potential of Evol3D, several avenues for further investigation remain open:

- Evaluate and benchmark Evol3D dataset utility in enhancing mutational effect prediction through downstream tasks involving machine/deep learning models.
- Cross-validation with structures predicted by other deep learning models and traditional computational methods to benchmark and improve structural reliability.
- Collaborations with domain experts to validate the structural and functional predictions experimentally and computationally.
- Analyze detailed correlations between structural variations and specific functional outcomes, identifying critical structural motifs.
- Develop interactive visualization tools for easier exploration and analysis of mutational effects on protein structures.
- Integrate experimentally derived mutational fitness data to validate and refine computational predictions.
- Utilize Evol3D as a training dataset to develop or improve generative models for protein design.
- Benchmark protein design strategies using Evol3D to evaluate mutational pathways and structural constraints.

[2] https://github.com/google-deepmind/alphafold3

Acknowledgments

We gratefully acknowledge the EuroHPC Joint Undertaking for awarding computational GPU resources under the EuroHPC Benchmark Access and EuroHPC AI and Data-Intensive Applications Access Calls. Specifically, we thank EuroHPC for providing access to the MeluXina supercomputer hosted by LuxProvide in Luxembourg and the Vega supercomputer hosted by IZUM in Slovenia, which were essential for conducting our computational experiments.

References

[1] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[2] Roland J Siezen, William M de Vos, Jack AM Leunissen, and Bauke W Dijkstra. Homology modelling and protein engineering strategy of subtilases, the family of subtilisin-like serine proteinases. *Protein Engineering, Design and Selection*, 4(7):719–737, 1991.

[3] Zhiqiang Zhong and Davide Mottin. Efficiently predicting mutational effect on homologous proteins by evolution encoding. In *Machine Learning and Knowledge Discovery in Databases - European Conference (ECMLPKDD)*, volume 399–415, page 14947. Springer, 2024.

[4] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[5] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic level protein structure with a language model. *bioRxiv*, 2022.

[6] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

[7] Christian Dallago, Jody Mou, Kadina E. Johnston, Bruce J. Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, and Kevin Yang. FLIP: benchmark tasks in fitness landscape inference for proteins. In *Proceedings of the 2021 Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[8] Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Chang Ma, Runcheng Liu, and Jian Tang. PEER: A comprehensive and multi-task benchmark for protein sequence understanding. In *Proceedings of the 2022 Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 35156–35173, 2022.

[9] Nicholas C Wu, Lei Dai, C Anders Olson, James O Lloyd-Smith, and Ren Sun. Adaptation in protein fitness landscapes is facilitated by indirect paths. *Elife*, 5:e16965, 2016.

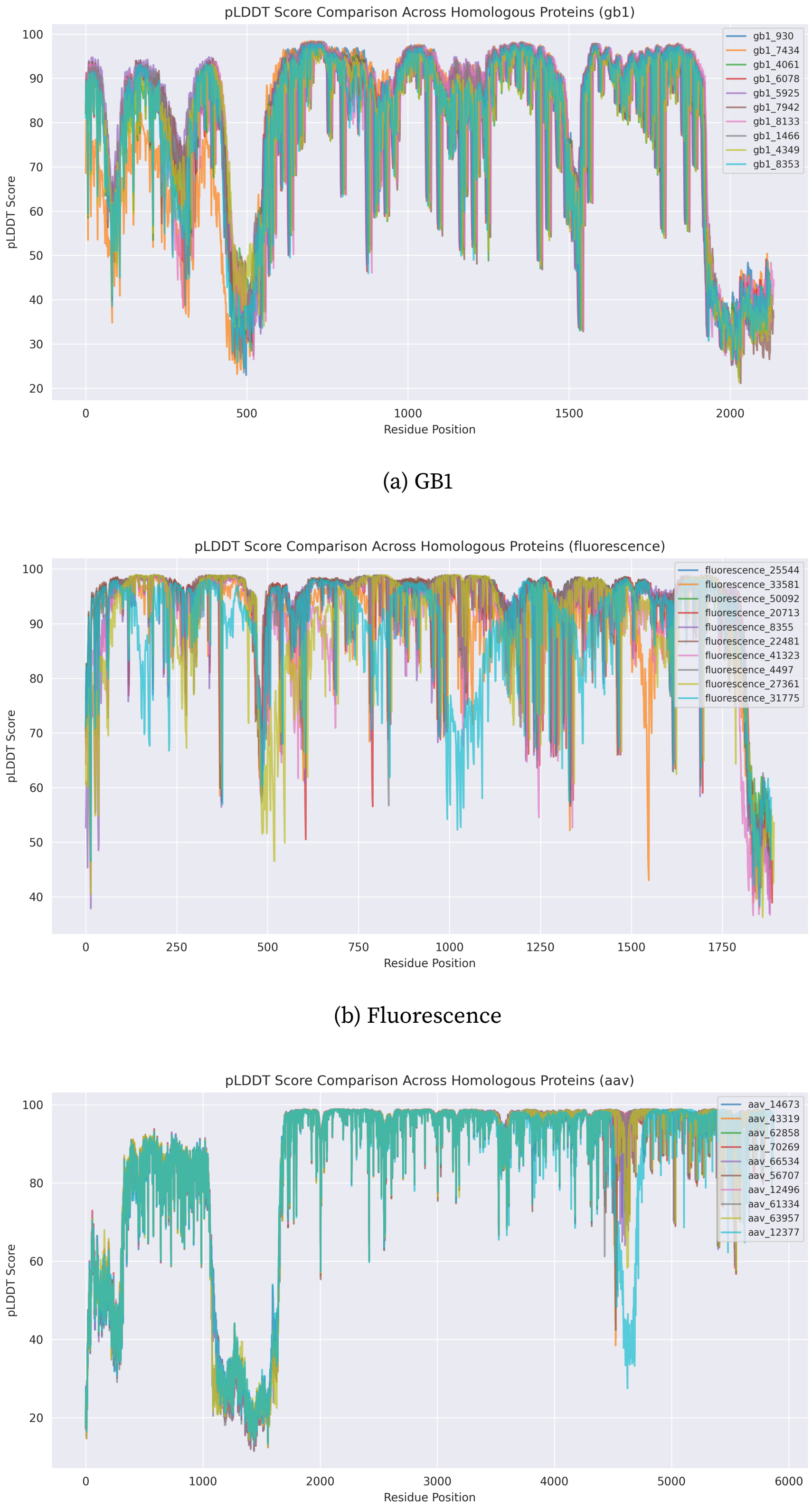
[10] Karen S Sarkisyan, Dmitry A Bolotin, Margarita V Meer, Dinara R Usmanova, Alexander S Mishin, George V Sharonov, Dmitry N Ivankov, Nina G Bozhanova, Mikhail S Baranov, Onuralp Soylemez, et al. Local fitness landscape of the green fluorescent protein. *Nature*, 533(7603):397–401, 2016.

[11] Drew H Bryant, Ali Bashir, Sam Sinai, Nina K Jain, Pierce J Ogden, Patrick F Riley, George M Church, Lucy J Colwell, and Eric D Kelsic. Deep diversification of an aav capsid protein by machine learning. *Nature Biotechnology*, 39(6):691–696, 2021.

[12] Roger Y Tsien. The green fluorescent protein. *Annual Review of Biochemistry*, 67(1):509–544, 1998.

[13] LH Vandenberghe, JM Wilson, and G Gao. Tailoring the aav vector capsid for gene therapy. *Gene Therapy*, 16(3):311–319, 2009.

Appendix A. Additional Results

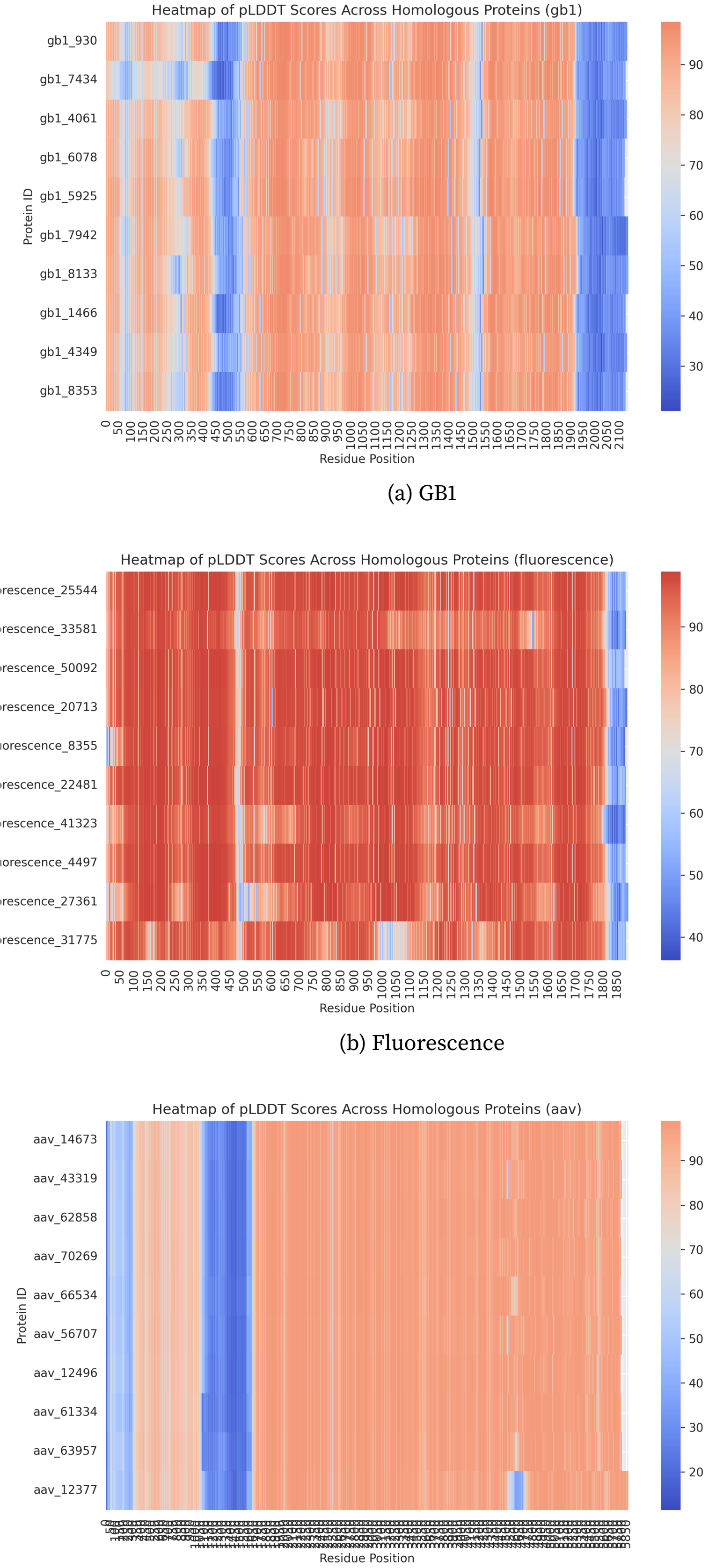


(a) GB1

(b) Fluorescence

(c) AAV

Fig. A1: Illustration of pLDDT scores of randomly selected proteins from GB1, Fluorescence, and AAV. Higher scores indicate increased confidence in predicted 3D protein structures.



(a) GB1

(b) Fluorescence

(c) AAV

Fig. A2: Illustration of pLDDT score heatmaps of randomly selected proteins from GB1, Fluorescence, and AAV. Higher scores indicate increased confidence in predicted 3D protein structures.

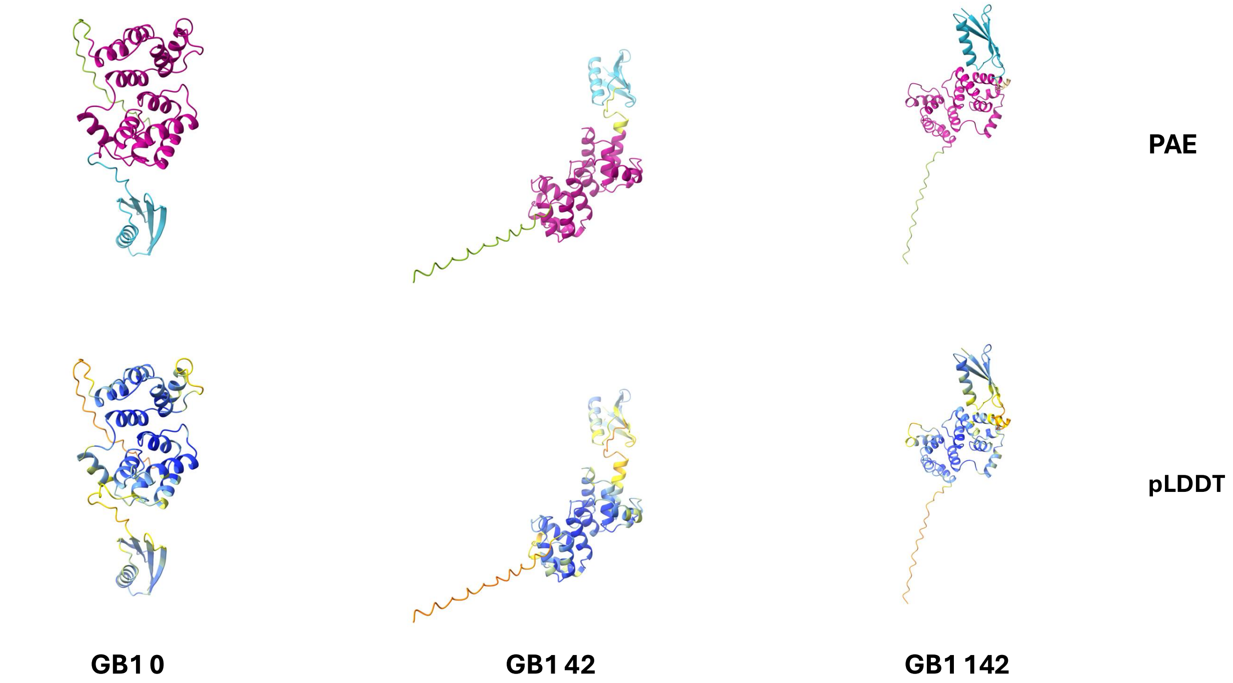


Fig. A3: Visualization of GB1 protein 3D structure, with PAE and pLDDT prediction errors.

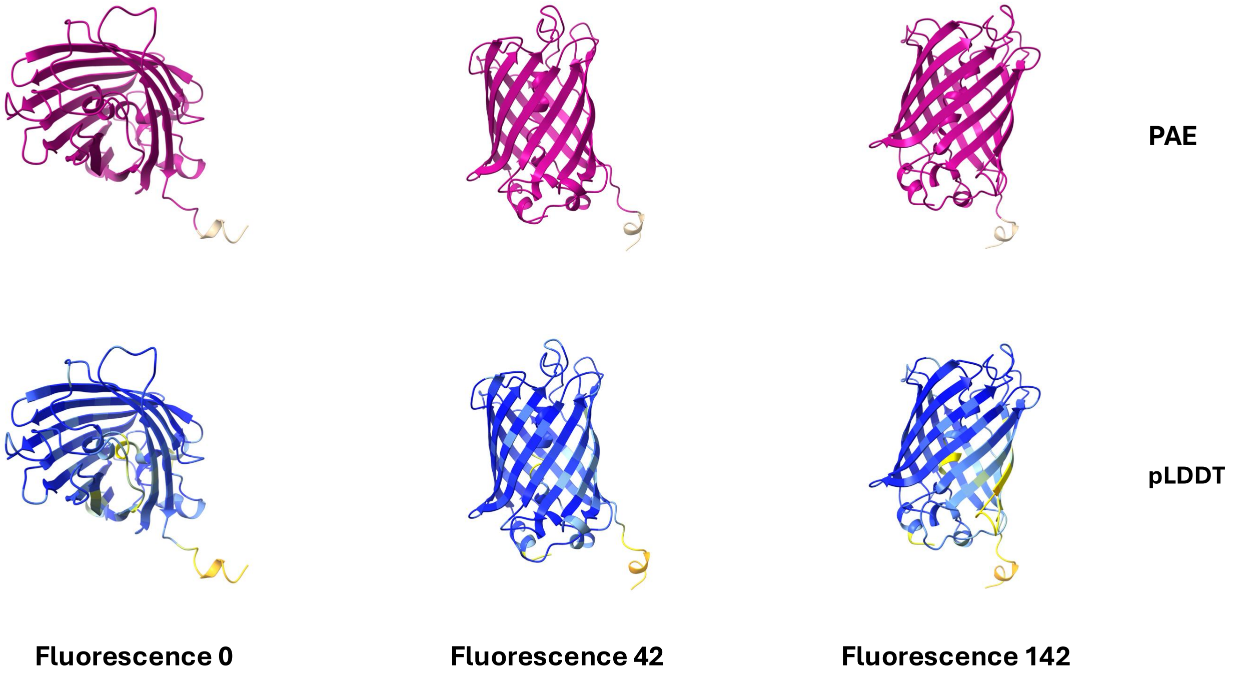


Fig. A4: Visualization of Fluorescence protein 3D structure, with PAE and pLDDT prediction errors.

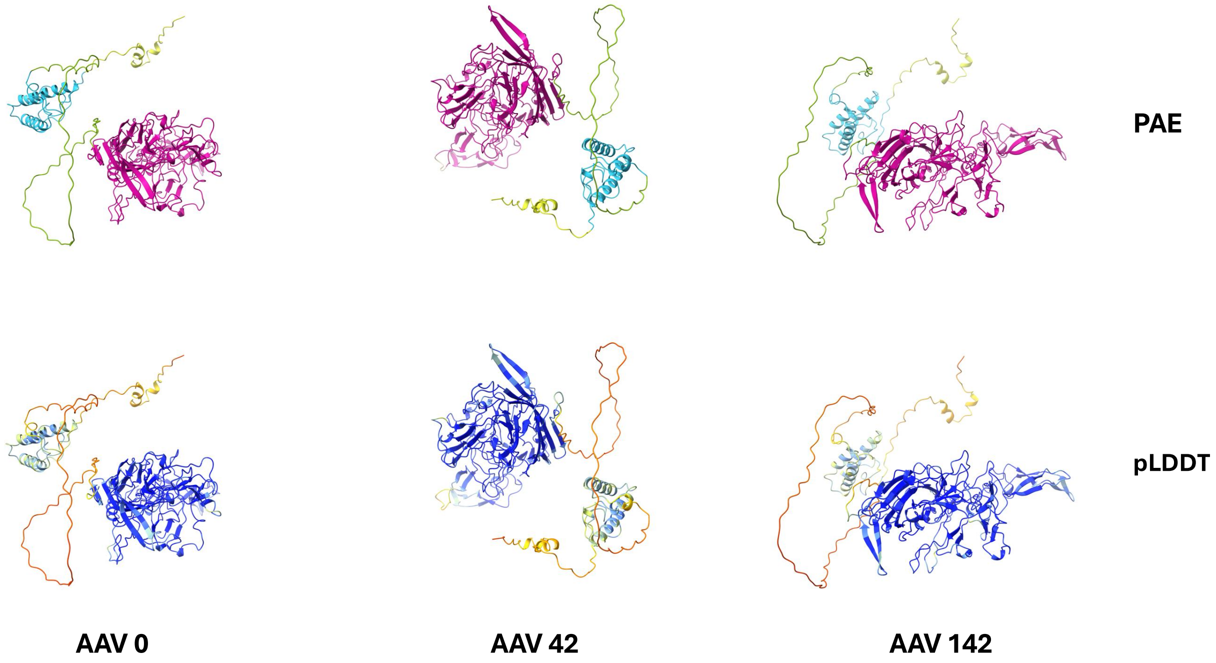


Fig. A5: Visualization of aav protein 3D structure, with PAE and pLDDT prediction errors.