S2S2FUN: DECODING PROTEIN FUNCTION FROM LATENT STRUCTURAL REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting mutational effects on protein function from sequences alone remains an unsolved challenge, despite its importance for protein engineering. Protein functions such as enzymatic activity are highly sensitive to mutations in a structure-dependent manner. Recent advances in structure prediction including AlphaFold3 and its open-source counterparts have enabled atomic-level modeling of biomolecular complexes. We hypothesize that AlphaFold3's latent structural features of protein–ligand complexes can be harnessed for decoding functional differences of sequence variants. Focusing on the optical properties of light-sensitive proteins, we demonstrate that AlphaFold3 $pair$ and $single$ representations can effectively predict absorption peaks, fluorescence brightness, and protein stability of natural and *de novo* designed proteins. Our "sequence-to-structure-to-function (S2S2Fun)" approach offers an effective method for ranking protein function and provides an *in silico* metric for metagenomic protein discovery and protein engineering applications.

1 INTRODUCTION

The central paradigm of structural biology posits that a protein's primary sequence dictates its three-dimensional structure, which in turn determines its biological function. Following this principle, deep learning models such as AlphaFold2 (AF2) (Jumper et al., 2021) and AlphaFold3 (AF3) (Abramson et al., 2024) have revolutionized biomedical science by predicting highly accurate protein structures from sequences alone. However, the intricate relationship between structure and function remains beyond the predictive power of current models. Although these tools were developed to bridge sequence to function via structure, accumulating evidence suggests that structure prediction models like AF2 and AF3 are often insensitive to sequence variation.

To predict functional changes for sequence variants, researchers have turned to Protein Language Models (PLMs) represented by ESM2 and ESM3 (Lin et al., 2023; Hayes et al., 2025). The success of PLMs is grounded in the notion that they implicitly encode evolutionary fitness from sequence statistics, and that this fitness landscape serves as a proxy for molecular function (Meier et al., 2021). This PLM-based "sequence-to-fitness-to-function" approach, however, has inherent limitations: (1) atomic-level structural information is omitted or reduced to residue-level representations; (2) functionally essential non-protein components, such as small molecules, RNA, or DNA, are excluded entirely; and (3) the function of *de novo* designed proteins with out-of-distribution sequences, which lack evolutionary context, cannot be accurately inferred.

To address these limitations, we return to the structural biology paradigm of "sequence-to-structure-to-function" and propose a structure-centric method called **S2S2Fun**. We hypothesize that the internal representations learned by biomolecular structure prediction models can better capture function-relevant features. Specifically, models such as AF3 encode ligand-aware structural information within their $pair$ and $single$ representations that can be harnessed for function prediction. This is particularly relevant when protein function depends on interactions with small-molecule, protein, RNA or DNA ligands, where even subtle sequence changes at the binding interface can lead to significant functional divergence.

To test this hypothesis, we focus on the optical properties of light-sensitive proteins that are highly sensitive to interactions between the protein and its chromophore ligand. Our methodological contributions are threefold:

- We introduce function prediction tasks for light-sensitive proteins of multiple protein folds covering both natural and *de novo* designed sequences. Our results demonstrate that S2S2Fun effectively predicts key optical properties, including absorption peaks and fluorescence brightness.
- We systematically evaluate multiple learning objectives and prioritize S2S2Fun's training with ranking relative order rather than predicting absolute values.
- For protein families that share similar functions but differ in their overall folds, we implement the Domain-Adversarial Neural Network (DANN) framework to discourage the model from learning structural bias and to focus on extracting the common functional features. This allows function-specific knowledge to transfer across protein families.

2 METHODS

2.1 FEATURE EXTRACTION FROM ALPHAFOLD3 LATENT REPRESENTATION

Our model extracts latent structural features from AF3's internal representations of protein-ligand complexes, and uses those features to decode their functional difference(Figure 1A). For each protein sequence and its ligand, AF3's pairformer module generates unique $single$ and $pair$ representations. Although their final three-dimensional structures are almost identical, sequence variants often produce drastically different $single$ and $pair$ representations (see detailed examples in Appendix D).

For a protein with L residues, its final AF3 $pair$ representation p_{ij} has the shape of $(L, L, 128)$ and the $single$ representation q_i has the shape of $(L, 384)$. We utilize the predicted complex structures and select the p_{ij} and q_i corresponding to the ligand-binding sites, which are considered to be the most relevant features for functional prediction. The cropped representations are pooled into a fixed-dimensional feature vector (see detailed information in Appendix E.3).

2.2 LEARNING OBJECTIVES FOR FUNCTIONAL PREDICTION

Protein function prediction is typically framed as a regression problem that estimates an absolute functional value y. Given the noisy nature of biological experiments, we argue that the relative rank order of functional changes are more reliable than the absolute values. Instead of treating it as a regression problem, we frame the function prediction as a ranking problem and systematically compare the influence of pointwise, pairwise and listwise ranking objectives on prediction accuracy.

In the pointwise setting, the model predicts an absolute score $\hat{y}_i = M(\mathbf{f}_i)$ for each protein variant x_i with feature representation $\mathbf{f}_i$, optimized using Mean Squared Error (MSE):

$$\mathcal{L}_{\text{Pointwise}} = \sum_i (y_i - \hat{y}_i)^2.$$

The pairwise objective reformulates function prediction as a ranking task, encouraging correct relative ordering between variants. For any pair (x_i, x_j) with $y_i > y_j$, the model is trained to satisfy $\hat{y}_i > \hat{y}_j$. We adopt the RankNet loss from Burges et al. (2005) and Burges (2010):

$$\mathcal{L}_{\text{Pairwise}} = -\sum_{(i,j):y_i>y_j} \log \sigma(\hat{y}_i - \hat{y}_j) - \sum_{(i,j):y_i<y_j} \log\big(1 - \sigma(\hat{y}_i - \hat{y}_j)\big),$$

where $\sigma(z) = (1 + e^{-z})^{-1}$.

The listwise objective operates on sets of variants jointly by aligning the predicted and true score distributions. We employ a ListNet-style loss as in Bruch et al. (2019):

$$\mathcal{L}_{\text{Listwise}} = -\sum_{l \in \mathcal{L}} \sum_i P_l(y_i) \log P_l(\hat{y}_i),$$

where $\mathcal{L}$ denotes the collection of lists and $P_l(\cdot)$ is approximated via a softmax over scores within each list.

2.3 DOMAIN-ADVERSARIAL LEARNING FOR KNOWLEDGE TRANSFER ACROSS PROTEIN FAMILIES

Divergent protein folds can perform similar functions under shared physicochemical constraints (Gherardini et al., 2007). We aim to reduce fold-specific structural bias and improve generalization of protein function prediction. To this end, we adopt domain-adversarial neural networks (DANN) (Ganin et al., 2016) to transfer functional knowledge across protein families. As illustrated in Figure 1B, for proteins performing the same function, we split training data into source and target domains by their fold category. All sequences are processed through the AF3 feature pipeline (see Methods 2.1). The DANN module includes a domain-confusion encoder Gc, a task predictor Gy, and a domain classifier Gd. Given an input x, the encoder produces $c = Gc(x)$, which goes to both Gy and Gd. Training minimizes the functional loss Ly and maximizes the domain loss Ld with respect to Gc through a Gradient Reversal Layer (GRL). This encourages the model to learn representations that are predictive of function yet invariant to fold-specific structural signals. The overall training objective is:

$$\min_{G_c,G_y} \max_{G_d} \mathcal{L}_{\text{DANN}} = \mathcal{L}_y(\mathbf{y}, G_y(G_c(\mathbf{x}))) - \lambda \mathcal{L}_d(\mathbf{d}, G_d(G_c(\mathbf{x}))).$$

Here, $\mathcal{L}_y$ denotes the functional loss, $\mathcal{L}_d$ is a cross-entropy loss over domains, and λ controls the strength of adversarial training (Appendix E.2).

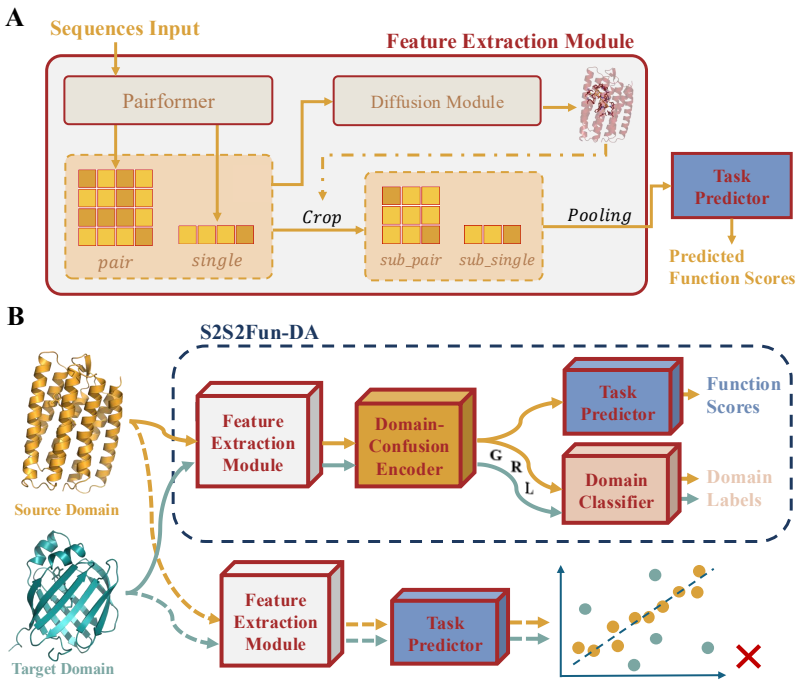

Figure 1: Overview of the S2S2Fun framework and its domain-adversarial variant. (**A**)S2S2Fun inference pipeline. A protein sequence (optionally with ligand) is processed by AlphaFold3 Pairformer to obtain structural representations. *pair* and *single* features are cropped based on the predicted structure (dashed line) to obtain function-site *sub_pair* and *sub_single* features, which are pooled and fed into a trainable MLP to predict function scores. (**B**) S2S2Fun-DA workflow. Yellow and cyan solid lines indicate source and target domain information flows respectively. Dashed lines indicate the workflow that adds target-domain samples directly to the training set. S2S2Fun-DA follows the same inference pipeline as S2S2Fun, except that representations are first transformed by Gc before going to the task predictor.

3 RESULTS

To test whether S2S2Fun can distinguish subtle functional changes, we focus on the optical properties of light-sensitive proteins including microbial rhodopsin proteins (MRPs), GFP-like fluorescent proteins, human cellular retinal binding proteins (CRBPs) and the deep mutation scanning (DMS) dataset of a *de novo* designed fluorescence-activating protein (HBI dataset). Our datasets consists of sequence variants paired with their functional labels of light absorption peaks, relative fluorescence brightness or proteolysis stability (detailed information in Appendix B). For each supervised function-prediction task, we use sequence embedding information from EMS-1v and ESM3 as the baseline.

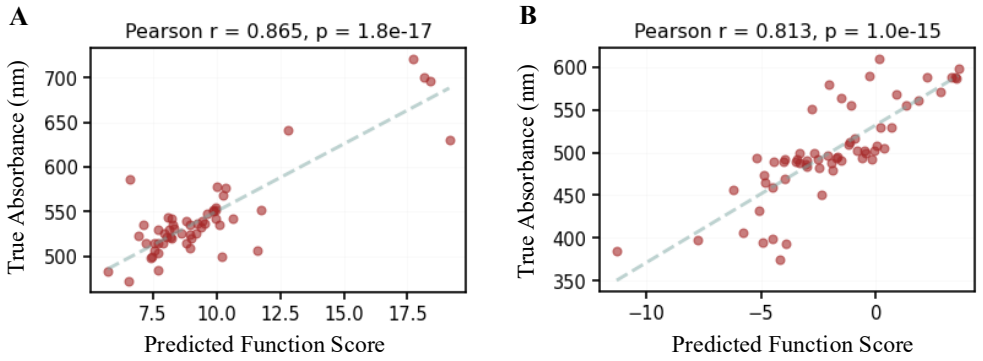

Figure 2: S2S2Fun performance in predicting absorption peaks of MRPs and GFP-like proteins. Model predictions on the test sets for MRPs (**A**) and GFP-like proteins (**B**) using S2S2Fun trained with the pairwise loss. The dashed line represents the linear regression fitted to the data.

Table 1: Performance of models on the MRP and GFP datasets, all trained with pairwise loss.

	Pair Accuracy↑	**Average Precision↑**	**Spearman↑**	**Pearson↑**
MRP				
ESM-1v	.669	.644	.462	.280
ESM3	.745	.709	.645	.405
S2S2Fun	**.772**	**.768**	**.669**	**.865**
GFP				
ESM-1v	.794	.787	.722	.665
ESM3	.781	.800	.730	.643
S2S2Fun	**.831**	**.826**	**.846**	**.813**

3.1 S2S2FUN PREDICTS ABSORPTION PEAKS OF LIGHT-SENSITIVE PROTEINS

We first benchmarked our S2S2Fun workflow for predicting absorption peaks using different learning objectives, evaluating performance on separate datasets of 891 MRPs and 754 GFP-like protein samples. As summarized in Table 5, models trained with a pairwise ranking objective consistently outperformed other objectives across all four correlation metrics. This result confirms that learning relative functional order is more robust than predicting absolute values from noisy biological data. Notably, the pairwise S2S2Fun achieved Pearson correlations of 0.865 for MRPs and 0.813 for GFP-like proteins (Figure 2), surpassing existing PLM-based approaches (Table 1).

Next, we assessed S2S2Fun's generalization capability. CRBP and MRP proteins share the same retinal-binding function with distinct structures (Figure 1B, Appendix B). S2S2Fun trained on MRP samples performed poorly on CRBP data (Table 4), indicating that the model is strongly biased by the structural background even though the training features were focused their shared ligand. Adding

the CRBP samples directly to training set was ineffective(Table 2, Figure 1B). A similar, perhaps more pronounced, generalization gap was observed for language model–based methods (Table 2).

To mitigate this structural bias, we applied DANN strategy (see Methods 2.3), treating MRP samples as the source domain and CRBP samples as the target. In this framework, the task predictor (G_y) is trained exclusively on MRP data, while the domain classifier (G_d) is trained to distinguish between domains. The feature encoder is optimized to rank the absorption peaks regardless of the background structural difference. As a result, S2S2Fun with DANN (S2S2Fun-DA) maintains its original performance on the MRP test set while gaining the ability to generalize to CRBP samples (Table 2). This results demonstrate that domain adversarial training effectively transfers functional knowledge across distinct protein scaffolds.

Table 2: Performance of models on RET-binding proteins. Test metrics are reported separately for MRP and CRBP as MRP_test/CRBP_test. Training data include both MRP and CRBP.

	Pair Accuracy↑	**Average Precision↑**	**Spearman↑**	**Pearson↑**
ESM-1v	.724/.291	.723/.517	.637/-.582	.620/-.614
ESM3	.738/.455	.712/.554	.643/-.182	.542/-.377
S2S2Fun	**.777**/.717	**.786**/.689	**.687**/.615	**.832**/.569
S2S2Fun–Da	.767/**.775**	.772/**.732**	.671/**.747**	.764/**.717**

3.2 S2S2Fun Predicts Fluorescence Brightness and Stability of de novo sequences

To further assess generalizability, we evaluated our S2S2Fun workflow on the DMS dataset of *de novo* designed HBI-binding protein (Dou et al., 2018), predicting fluorescence brightness and stability for its single-point mutations. As shown in Table 3, S2S2Fun outperformed ESM-1v and achieved comparable accuracy with ESM3 on brightness prediction. However, for stability prediction, S2S2Fun underperformed both ESM-1v and ESM3. This divergence likely arises because fluorescence brightness depends critically on protein-ligand interactions, which S2S2Fun captures by its ligand-aware latent structural features. In contrast, protein stability is determined by the protein sequence alone, a context where protein language models excel. This result underscores S2S2Fun's unique advantage in predicting ligand-dependent functions, effectively complementing the general-purpose applications of PLMs.

Table 3: Performance of models on the HBI dataset across multiple functional labels. Values are formatted Stability/Brightness.

	Pair Accuracy↑	**Average Precision↑**	**Spearman↑**	**Pearson↑**
ESM-1v	.790/.728	.712/.687	.764/.651	.797/.636
ESM3	**.796/.749**	**.719/.706**	**.783/.693**	**.842/.668**
S2S2Fun	.759/.745	.699/**.706**	.711/.683	.756/.664

4 Discussion

Our work presents S2S2Fun, a structure-centric framework for protein function prediction. By combining AF3 internal representations with pairwise learning, S2S2Fun effectively ranks functional properties for natural and *de novo* light-sensitive proteins. Incorporating a Domain-Adversarial Neural Network (DANN) reduces structural bias and enables knowledge transfer across distinct protein folds. Building on its strong performance across MRPs, CRBPs, and GFP-like proteins, our next step is experimental validation of sequences variants predicted to show red-shifted absorbance. Despite these results, S2S2Fun is currently limited to single-function prediction and does not generalize well to multi-property tasks. Addressing these limitations will be the focus of future work.

REFERENCES

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.

Claire N Bedbrook, Kevin K Yang, J Elliott Robinson, Elisha D Mackey, Viviana Gradinaru, and Frances H Arnold. Machine learning-guided channelrhodopsin engineering enables minimally invasive optogenetics. *Nature methods*, 16(11):1176–1184, 2019.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 2000.

Surojit Biswas, Grigory Khimulya, Ethan C Alley, Kevin M Esvelt, and George M Church. Low-n protein engineering with data-efficient deep learning. *Nature methods*, 18(4):389–396, 2021.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008 (10):P10008, 2008.

Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.

Matthias Broser, Wayne Busse, Anika Spreen, Maila Reh, Yinth Andrea Bernal Sierra, Songhwan Hwang, Tillmann Utesch, Han Sun, and Peter Hegemann. Diversity of rhodopsin cyclases in zoospore-forming fungi. *Proceedings of the National Academy of Sciences*, 120(44): e2310600120, 2023.

Sebastian Bruch, Xuanhui Wang, Michael Bendersky, and Marc Najork. An analysis of the softmax cross entropy loss for learning-to-rank with binary relevance. In *Proceedings of the 2019 ACM SIGIR international conference on theory of information retrieval*, pp. 75–78, 2019.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pp. 89–96, 2005.

Christopher JC Burges. From ranknet to lambdarank to lambdamart: An overview. *Learning*, 11 (23-581):81, 2010.

Jiayi Dou, Anastassia A Vorobieva, William Sheffler, Lindsey A Doyle, Hahnbeom Park, Matthew J Bick, Binchen Mao, Glenna W Foight, Min Yen Lee, Lauren A Gagnon, et al. De novo design of a fluorescence-activating β-barrel. *Nature*, 561(7724):485–491, 2018.

Janani Durairaj, Yusuf Adeshina, Zhonglin Cao, Xuejin Zhang, Vladas Oleinikovas, Thomas Duignan, Zachary McClure, Xavier Robin, Gabriel Studer, Daniel Kovtun, et al. Plinder: The protein-ligand interactions dataset and evaluation resource. *BioRxiv*, pp. 2024–07, 2024.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: towards cracking the language of life's code through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:7112–7127, 2021.

Seth A Frazer, Mahdi Baghbanzadeh, Ali Rahnavard, Keith A Crandall, and Todd H Oakley. Discovering genotype–phenotype relationships with machine learning and the visual physiology opsin database (vpod). *GigaScience*, 13:giae073, 2024.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Pier Federico Gherardini, Mark N Wass, Manuela Helmer-Citterich, and Michael JE Sternberg. Convergent evolution of enzyme active sites is not a rare phenomenon. *Journal of molecular biology*, 372(3):817–845, 2007.

Peter Mørch Groth, Mads Kerrn, Lars Olsen, Jesper Salomon, and Wouter Boomsma. Kermut: Composite kernel regression for protein variant effects. *Advances in Neural Information Processing Systems*, 37:29514–29565, 2024.

Alex Hawkins-Hooker, Shikha Surana, Jack Simons, Jakub Kmec, Oliver Bent, and Paul Duckworth. Likelihood-based fine-tuning of protein language models for few-shot fitness prediction and design. *bioRxiv*, pp. 2024–05, 2024.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

Xin Hong, Bowen Gao, Yinjun Jia, Wenyu Zhu, Qixuan Chen, Xiaohe Tian, Zhenyi Zhong, Jianhui Wang, and Yanyan Lan. How good is alphafold3 at ranking drug binding affinities? *bioRxiv*, pp. 2025–05, 2025.

Chloe Hsu, Hunter Nisonoff, Clara Fannjiang, and Jennifer Listgarten. Learning protein fitness models from evolutionary and assay-labeled data. *Nature biotechnology*, 40(7):1114–1122, 2022.

Keiichi Inoue, Masayuki Karasuyama, Ryoko Nakamura, Masae Konno, Daichi Yamada, Kentaro Mannen, Takashi Nagata, Yu Inatsu, Hiromu Yawo, Kei Yura, et al. Exploration of natural red-shifted rhodopsins using a machine learning-based bayesian experimental design. *Communications biology*, 4(1):362, 2021.

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.

Masayuki Karasuyama, Keiichi Inoue, Ryoko Nakamura, Hideki Kandori, and Ichiro Takeuchi. Understanding colour tuning rules and predicting absorption wavelengths of microbial rhodopsins by data-driven machine-learning approach. *Scientific reports*, 8(1):15580, 2018.

Talley J Lambert. Fpbase: a community-editable fluorescent protein database. *Nature methods*, 16 (4):277–278, 2019.

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.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in neural information processing systems*, 34:29287–29303, 2021.

Pascal Notin, Ruben Weitzman, Debora Marks, and Yarin Gal. Proteinnpt: Improving protein property prediction and design with non-parametric transformers. *Advances in Neural Information Processing Systems*, 36:33529–33563, 2023.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

Philip A Romero, Andreas Krause, and Frances H Arnold. Navigating the protein fitness landscape with gaussian processes. *Proceedings of the National Academy of Sciences*, 110(3):E193–E201, 2013.

Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

Shunki Takaramoto, Chenxiang Zhao, Masayuki Karasuyama, Frederik Schulz, Masaya Watanabe, Yuma Kawasaki, Takashi Nagata, Masae Konno, Naoya Morimoto, Masahiro Fukuda, et al. Functional characterization of red-shifted rhodopsin channels from giant viruses explored by a machine-learning model for long-wavelength optogenetics. *bioRxiv*, pp. 2025–09, 2025.

Felix Teufel, Aaron W Kollasch, Yining Huang, Ole Winther, Kevin K Yang, Pascal Notin, and Debora S Marks. Few-shot protein fitness prediction via in-context learning and test-time training. *arXiv preprint arXiv:2512.02315*, 2025.

Wenjing Wang. *Protein Design: Reengineering of Cellular Retinol Binding Protein II (CRBPII) Into a Rhodopsin Mimic, Functionalization of CRBPII Into a Fluorescent Protein Tag and Design of a Photoswitchable Protein Tag*. Michigan State University. Chemistry, 2012.

Yang Zhang and Jeffrey Skolnick. Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic acids research*, 33(7):2302–2309, 2005.

A RELATED WORK

Protein Language Models for Downstream Prediction. Recent advances in protein language models (pLMs) have demonstrated that sequence-only pretraining on large-scale protein databases can yield embeddings that capture rich biochemical and evolutionary information. Models such as ESM-1v, ProtBERT (Brandes et al., 2022), and ESM-2 have been widely adopted for downstream tasks including stability prediction, mutational effect prediction, and protein engineering, often via simple linear probes or shallow neural networks (Elnaggar et al., 2021; Hsu et al., 2022; Rives et al., 2021).

Structure-Based Representations from AlphaFold. Recent studies have explored AlphaFold 3 representations for ranking drug binding affinities by exploiting its protein–ligand co-folding capability (Hong et al., 2025). The results indicate that AF3-derived structural features contain informative interaction signals that enable relative ranking of ligand-binding affinities, highlighting the potential of AF3 representations for downstream ranking-based tasks.

Single-Mutation Fitness Prediction and Ranking Objectives. Large-scale DMS experiments have motivated extensive research on predicting the effects of single amino acid substitutions (Groth et al., 2024; Hawkins-Hooker et al., 2024; Notin et al., 2023; Teufel et al., 2025). The reference-based loss used in (Hawkins-Hooker et al., 2024) is conceptually similar to our pairwise ranking objective, as both optimize relative ordering between samples rather than absolute functional values.

Absorbance Functional Prediction. Predicting optical properties, such as absorption maxima in GFP and rhodopsins, has long been a benchmark problem in protein engineering. Multiple machine learning-based methods have been developed (Bedbrook et al., 2019; Inoue et al., 2021; Frazer et al., 2024; Takaramoto et al., 2025) for rhodopsin protein absorbance prediction. Early studies combined mutational scanning with statistical and machine learning models to characterize GFP fitness landscapes (Biswas et al., 2021; Romero et al., 2013).

B THREE FUNCTIONAL DATASETS FOCUSED ON LIGHT-SENSITIVE PROTEINS

We curate three datasets with experimental light absorption peaks and fluorescence brightness, including sequences of both natural and *de novo* designed proteins and their corresponding functional labels:

- **RET dataset** consists of *all-trans* retinal-binding proteins (RET-binding proteins) that execute diverse functions upon light absorbance. Our RET-binding proteins mainly come from two protein families: (1) We obtained 68 microbial rhodopsin proteins (MRP) from the PDB (Berman et al., 2000), 823 MRP data(from paper Karasuyama et al. (2018); Broser et al. (2023))that are membrane proteins characterized by their GPCR-like fold. and (2) 182 mutants of human Cellular Retinoid Binding Protein 2 (CRBP) (Wang, 2012), we obtained 14 homologous protein families from PDB. In total, our RET-binding dataset contains 1087 unique sequences paired with their absorption maximum. (data from PDB full list in Appendix G)
- **GFP dataset** contains 754 GFP-like fluorescent proteins from the FPbase resource (Lambert, 2019). We collect fluorescent proteins that generate their own chromophores via autocatalysis (i.e., without exogenous cofactor), which come from a single protein family with

a typical 11-strand beta-barrel fold. Their sequence similarity ranges from 0.2 to 0.99 and absorption peaks range from 343nm to 634nm.

- **HBI dataset** is a Deep Mutational Scanning (DMS) dataset containing single mutations of a *de novo* designed fluorescence-activating protein (Dou et al., 2018). The functional labels include the protein variants' relative fluorescence brightness and proteolysis stability measured on the surface of yeast cells.

To ensure reliable model performance in small datasets, we set up rigorous criteria for training/test data splitting. Protein samples are clustered and split based on both sequence and structure similarities: sequence similarity is defined by MMseqs2 (Steinegger & Söding, 2017) and pocket similarity follows the definition of the PLINDER database (Durairaj et al., 2024). These carefully examined criteria ensure a controlled data partitioning process to prevent data leakage and to reduce model overfitting (details in Appendix C).

C Data Splitting Strategy Based on Sequence and Pocket Similarity

To ensure a strict evaluation of model generalization, train–test splits were constructed by jointly controlling global sequence similarity and local binding pocket similarity. Protein sequences were clustered using MMseqs2 at an 0.8 identity threshold. Proteins within the same sequence cluster were treated as homologous and kept in the same data split. Structural alignments between protein pairs were computed using TM-align (Zhang & Skolnick, 2005) . Let (P_A) and (P_B) denote the sets of annotated pocket residues in proteins A and B, respectively. Using the residue mapping induced by structural alignment, pocket similarity is defined as

$$S_{\text{pocket}}(A, B) = \frac{1}{|P_A|} \sum_{i \in P_A} \mathbb{I}\Big(\text{align}(i) \in P_B \ \wedge \ \text{AA}_i = \text{AA}_{\text{align}(i)} \Big)$$

where $align(i)$ is the aligned residue in protein B (if not a gap), and AA denotes amino acid identity. The score is symmetrized to obtain a pairwise similarity matrix.

Based on the similarity definition, all the samples were connected in a graph (threshold is 0.8), and Louvain community (Blondel et al., 2008) detection was applied to obtain pocket similarity clusters. Each protein was assigned a merged cluster label: merged_cluster = (sequence cluster, pocket cluster). Clusters containing only one protein were grouped into an "other" category. Train–test splitting was then performed at the cluster level using stratified sampling based on the majority class label of each cluster.

We considered both sequence similarity and pocket similarity when splitting the RET-binding data. In the end, 66 data points were divided into the test set, of which 11 are CRBP (most of the collected CRBP data are single mutation data, and to prevent data leakage, only a small portion was selected as the test set). The splits for GFP and HBI data were randomly divided based on the types of motifs that form the chromophore and the location information of the mutations. Ultimately, there are 62 in the GFP test and 222 in the HBI test (the number of tests for stability and brightness is consistent). All model training in this paper maintains a consistent train-test split.

D An Example Illustrating Differences in AF3 Representations

Using single-mutation data from the HBI dataset, we examined a case where the wild-type (WT) sequence contains lysine (K) at position 40, and mutation to proline (P) leads to a substantial experimental decrease in binding affinity. However, AlphaFold3 complex structure predictions (protein + ligand input) show nearly identical structures between WT and mutant (RMSD = 0.31; WT pLDDT: 96.35, mutant pLDDT: 96.09), with indistinguishable ligand conformations (Figure 3). Thus, the final predicted structures alone do not explain the functional change.

We therefore compared internal AF3 representations, specifically the *pair* and *single* embeddings. Similarity was defined as the mean squared error (MSE) along the last feature dimension. Both

representations exhibit clear differences between WT and mutant, as visualized in the heatmaps (Figure 3, right); pair diagonal excluded), suggesting AF3 latent features capture functional perturbations not reflected in global structural metrics.

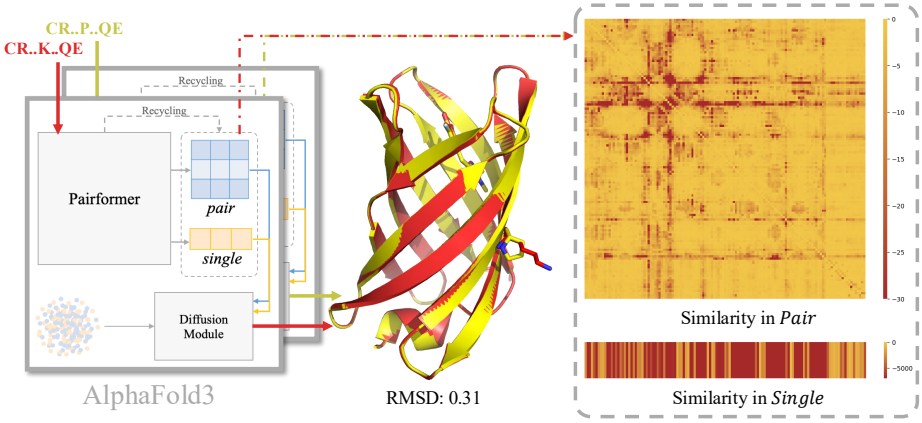

Figure 3: Divergent latent representations despite nearly identical predicted structures. Red indicates the wild-type (WT) and yellow indicates the mutant. While their AlphaFold3-predicted structures are highly similar, their internal representations differ markedly. In the heatmaps, deeper red corresponds to greater differences between representations.

E Model Setting

S2S2Fun–DA follows a domain-adversarial ranking paradigm composed of three neural modules: a domain-confusion encoder G_c, a task predictor G_y(Consistent with the task predictor module in S2S2Fun), and a domain classifier G_d. A Gradient Reversal Layer (GRL) is introduced to enable adversarial domain adaptation and encourage domain-invariant feature learning.

The GRL behaves as the identity function during the forward pass but reverses gradients during backpropagation. Formally, for an input feature vector x and scaling factor $\lambda > 0$,

$$\mathrm{GRL}(x) = x, \quad \frac{\partial \mathcal{L}}{\partial x} \leftarrow -\lambda \frac{\partial \mathcal{L}}{\partial x}. \tag{1}$$

This mechanism allows the domain-confusion encoder to learn representations that are predictive for the main task while being indistinguishable across domains.

The domain-confusion encoder G_c refines the input embedding $x \in \mathbb{R}^d$ using a residual multilayer perceptron:

$$G_c(x) = x + f(x), \tag{2}$$

where $f(\cdot)$ is a two-layer MLP with ReLU activation and dropout,

$$f(x) = W_2, \sigma\big(\mathrm{Dropout}(W_1 x)\big), \tag{3}$$

and $\sigma(\cdot)$ denotes the ReLU function. This residual design preserves the original structural representation while allowing task-specific feature refinement.

The ranking predictor G_y maps extracted features to a scalar functional score. Input features are first ℓ_2-normalized,

$$\tilde{x} = \frac{x}{|x| * 2}, \tag{4}$$

and then passed through a linear projection followed by B residual blocks:

$$h_0 = \mathrm{ReLU}(W * in\tilde{x}),\ h_{k+1} \qquad = \mathrm{ReLU}\big(h_k + F_k(h_k)\big), \quad k = 0, \ldots, B-1. \tag{5}$$

Each residual function $F_k(\cdot)$ consists of a two-layer MLP with activation and dropout,

$$F_k(h) = W_{k,2}, \sigma\big(\mathrm{Dropout}(W_{k,1} h)\big). \tag{6}$$

The final ranking score is computed by

$$\hat{y} = W_{out} h_B. \tag{7}$$

The domain classifier G_d predicts the domain label from feature representations using a three-layer MLP:

$$G_d(z) = W_3, \sigma\big(W_2, \sigma(W_1 z)\big). \tag{8}$$

During training, features are passed through the GRL before entering the domain classifier,

$$\hat{d} = G_d(\mathrm{GRL}(G_c(x))), \tag{9}$$

which encourages G_c to produce domain-invariant features.

The overall training objective combines a ranking loss $\mathcal{L} * rank$ and a domain classification loss $\mathcal{L} * domain$:

$$\mathcal{L} = \mathcal{L} * rank(G_y(G_c(x))) + \mathcal{L} * domain(G_d(\mathrm{GRL}(G_c(x))), d), \tag{10}$$

where d denotes the domain label. The GRL ensures that minimizing $\mathcal{L}_{domain}$ with respect to G_d corresponds to maximizing it with respect to G_c, thereby promoting domain-invariant representation learning while preserving ranking performance.

E.1 AlphaFold 3 Inference Settings.

All AlphaFold 3 (AF3) representations were generated using a fixed inference configuration with one random seed, five sampling runs, and ten recycling iterations. Multiple Sequence Alignment (MSA) searches were performed for all sequences, and the resulting MSA information was incorporated to produce AF3 single and pair representations. AF3 was used as a frozen feature extractor, and no model parameters were fine-tuned during downstream training. All the sequences in the HBI dataset was inferenced without MSA. The RET-bing protein had bonds set during the AF3 prediction, forming the two atoms of the *Schiff base* (N15 of retinal and NZ of lysine). Additionally, we customized the RET small molecule by removing the oxygen atom.

E.2 Scheduling of the Adversarial Weight λ.

In the Domain-Adversarial Neural Network (DANN) framework, the hyperparameter λ controls the strength of the adversarial domain loss relative to the task-specific objective. To stabilize training and avoid early over-regularization, λ is not fixed but gradually increased over the course of training. Specifically, we adopt a sigmoid-based scheduling strategy:

$$\lambda(p) = \lambda_{\max} \left(\frac{2}{1 + \exp(-10p)} - 1 \right),$$

where $p = \frac{\text{epoch}}{\text{num\_epochs}}$ denotes the normalized training progress and $\lambda_{\max}$ is the maximum adversarial weight.

E.3 Feature Pooling and Aggregation.

Given the pairwise representation $\mathbf{P} \in \mathbb{R}^{l \times l \times 128}$ and the single-residue representation $\mathbf{S} \in \mathbb{R}^{l \times 384}$, where $l = N_{\text{protein}} + N_{\text{ligand}}$, we construct a fixed-length feature vector via block-wise mean pooling. We first partition the pairwise representation into four blocks according to protein and ligand indices: protein–protein (PP), ligand–ligand (LL), protein–ligand (PL), and ligand–protein (LP). Each block is mean-pooled over both residue dimensions, resulting in four 128-dimensional vectors. These pooled vectors are then concatenated to form a 512-dimensional pairwise feature representation. For the single-residue representation, we separately apply mean pooling over protein residues and ligand residues, producing two 384-dimensional vectors. These are concatenated to obtain a 768-dimensional single-feature representation.

The N_{protein} of all datasets is 30. The RET-binding dataset $N_{\text{ligand}} = 20$. For the GFP dataset, we will consider three amino acids that form the chromophore as small molecules, with $N_{\text{ligand}} = 3$. For the HBI dataset, for stability $N_{\text{ligand}} = 0$; for brightness $N_{\text{ligand}} = 18$.

Table 4: Train S2S2Fun only use MRP, the performance in MRP_test and CRBP data.

	Pair Accuracy↑	Average Precision↑	Spearman↑	Pearson↑
pointwise				
MRP	.677	.705	.519	.687
CRBP	.453	.503	-.136	-.106
pairwise				
MRP	.731	.739	.638	.753
CRBP	.389	.430	-.320	-.283

Table 5: ESM3 perfomance with different learning objectives in MRP data and 2S2Fun trained with different learning objectives on MRP and GFP. *Listwise (all)* uses the entire training set as a single list, while *listwise (random)* constructs lists by randomly partitioning the training data.

	Pair Accuracy↑	Average Precision↑	Spearman↑	Pearson↑
ESM3				
Pointwise	.706	.733	.568	.673
Pairwise	**.745**	**.709**	**.645**	**.405**
Listwise(all)	.513	.625	.069	.703
Listwise(random)	.501	.622	.047	.649
S2S2Fun				
MRP_Pointwise	.646	.692	.419	.778
MRP_Pairwise	**.772**	**.768**	**.669**	**.865**
MRP_Listwise (all)	.668	.711	.457	.696
MRP_Listwise (random)	.610	.675	.311	.755
GFP_Pointwise	.699	.690	.560	.543
GFP_Pairwise	**.831**	**.826**	**.846**	**.813**
GFP_Listwise (all)	.671	.690	.499	.539
GFP_Listwise (random)	.687	.693	.532	.552

E.4 Crop Ablation

Without applying cropping, S2S2Fun directly uses the full $pair$ and $single$ representations from AlphaFold3. Under identical training settings and with the same pairwise loss, this results in consistently worse performance on both datasets (Table 6). This indicates that cropping is beneficial when utilizing AF3 representations, potentially by alleviating redundancy in the $pair$ and $single$ features.

Table 6: Performance of S2S2Fun on the MRP and GFP datasets with and without crop.

	Pair Accuracy↑	Average Precision↑	Spearman↑	Pearson↑
MRP	.772	.768	.669	.865
MRP w/o crop	.700	.716	.553	.726
GFP	.831	.826	.846	.813
GFP w/o crop	.784	.765	.737	.679

F Evaluation Metrics

To evaluate the performance of our models, we report four complementary metrics that assess ranking consistency, threshold-based precision, and correlation between predicted and true scores: *Pair Accuracy*, *Average Precision*, *Pearson Correlation*, and *Spearman Correlation*. In this parper, the best epoch for all models is selected based on the average value of these four metrics.

Pair Accuracy measures the fraction of correctly ordered sample pairs based on their predicted scores. Given a set of n samples with predicted scores $\{s_i\}_{i=1}^n$ and ground-truth labels $\{y_i\}_{i=1}^n$, we consider all unordered pairs (i, j) with $i < j$. Pairs with identical labels ($y_i = y_j$) are optionally excluded. For each remaining pair, a prediction is considered correct if the relative ordering of predicted scores matches the ordering of true labels:

$$\mathbb{I}\left[(s_i > s_j) = (y_i > y_j)\right].$$

The Pair Accuracy is computed as

$$\text{PairAcc} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \mathbb{I}\left[(s_i > s_j) = (y_i > y_j)\right],$$

where $\mathcal{P}$ denotes the set of valid sample pairs.

Average Precision as a threshold-based precision metric computed by jointly sweeping prediction and label thresholds. For each sample k, we use its predicted score s_k as a prediction threshold and its true label y_k as a ground-truth threshold. Binary predictions and labels are defined as

$$\hat{p}_i^{(k)} = \mathbb{I}(s_i \geq s_k), \quad p_i^{(k)} = \mathbb{I}(y_i \geq y_k).$$

The precision at threshold k is then

$$\text{Prec}_k = \begin{cases} \frac{\text{TP}_k}{\text{TP}_k + \text{FP}_k}, & \text{if } \text{TP}_k + \text{FP}_k > 0, \\ 0, & \text{otherwise,} \end{cases}$$

where TP_k and FP_k denote the number of true positives and false positives under threshold k. The final Average Precision is computed by averaging over all samples:

$$\text{AP} = \frac{1}{n} \sum_{k=1}^{n} \text{Prec}_k.$$

This metric emphasizes how well high-scoring predictions align with high ground-truth values across multiple thresholds.

Pearson Correlation measures the linear relationship between predicted scores and true labels. It is defined as

$$\rho_p = \frac{\sum_{i=1}^n (y_i - \bar{y})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^n (s_i - \bar{s})^2}},$$

where $\bar{y}$ and $\bar{s}$ denote the mean of true labels and predicted scores, respectively. If the denominator is zero, the correlation is defined as zero.

Spearman Correlation evaluates the monotonic relationship between predictions and labels by computing the Pearson correlation between their ranks. Let $\text{rank}(y_i)$ and $\text{rank}(s_i)$ denote the ranks of the true labels and predicted scores. The Spearman correlation is then given by

$$\rho_s = \rho_p\left(\text{rank}(\mathbf{y}), \text{rank}(\mathbf{s})\right).$$

G COLLECTION DATA FROM PDB

We retrieved RET-binding proteins from the PDB and determined the corresponding absorbance values of these proteins by reviewing the literature, creating a new list of 82 items regarding protein structure IDs and their functional absorbance values. The ones highlighted in red belong to the CRBP family of RET-binding proteins.

PDBID	$\lambda_{max}(nm)$	PDBID	$\lambda_{max}(nm)$	PDBID	$\lambda_{max}(nm)$	PDBID	$\lambda_{max}(nm)$
6xyt	530	8yel	525	6g7i	570	6csm	515
4kly	550	6wp8	522	8zan	470	6cso	483
7crj	550	5jje	510	6eig	485	1fbk	560
5b2n	550	8h86	521	4xxj	570	6csn	483
5g2d	562	3t45	554	5zim	570	4fpd	560
2l6x	540	4yzi	455	1o0a	570	6mor	461
5uk6	543	1mgy	574	1brr	570	7lhn	613
5g2c	523	7w9w	520	4pxk	560	4zr2	496
7o8z	550	1xji	570	8jh0	530	4ydb	530
6ybz	530	8h87	486	6rmk	570	6mr0	457
6rf7	530	7cj3	492	7zbc	500	6mqj	522
8qqz	535	7clj	542	7d7q	492	4yfr	564
6lm0	550	6jo0	509	1jgj	498	4i9s	556
6lm1	550	5dys	380	1kg9	570	4i9r	610
4tl3	549	8cqc	525	8yek	505	4m7m	616
7aky	500	6su4	552	3am6	532	4m6s	619
5h2l	570	6s63	555	3oax	500	4ykm	557
7zmy	558	7pl9	661	6xl3	540	4ruu	508
7zov	536	1p8u	560	5j7a	570	4yko	545
7zng	558	6gyh	535	6i9k	535		
1c3w	570	8zao	521	7xjd	570		