FluorCode: Predicting Fluorescent Protein Photophysical Properties with LoRA-Fine-Tuned Protein Language Models

Rico Chi Kit Sou [*1] **Alicja Ziajowska** [*2]

Abstract

Predicting fluorescent protein (FP) photophysical properties from sequence remains challenging, especially for novel fluorescent proteins because of the limited diversity of existing datasets. Using FluorCode, a fine-tuned protein language model for photophysical properties, we compare alignment-based one-hot features against LoRA-fine-tuned ESM2-650M with chromophore-aware attention pooling. Under sequence-identity-clustered cross-validation at 50%, the alignment-based baseline accuracy degrades sharply, with excitation MAE increasing from 12.7 to 35.6 nm and brightness prediction even collapsing to noise ($R = 0.13$), whereas the LoRA-ESM2 MLP model remains robust (excitation MAE: 8.9 $\rightarrow$ 11.3 nm; brightness $R = 0.96$). Meanwhile, adding Pocket-3D chromophore-anchored descriptors from 913 chromophore-grafted structures yields no consistent measurable gain beyond the sequence model, indicating that explicit structural information in this hand-crafted form does not improve over LoRA-ESM2 under our current setup. Here, we demonstrate that the standard random cross-validation protocol used in prior work, previously shown to inflate performance in protein machine learning but not addressed in fluorescent protein prediction, overestimates performance by placing near-identical FP variants in both training and test folds.

* Equal contribution [1]Yusuf Hamied Department of Chemistry, University of Cambridge [2]School of Chemistry, College of Science and Engineering, University of Edinburgh. Correspondence to: Rico Chi Kit Sou <cks40@cam.ac.uk>, Alicja Ziajowska <s2713107@ed.ac.uk>.

Accepted by ICML 2026 AI for Science. *Proceedings of the 43^{rd} International Conference on Machine Learning*, Seoul, South Korea. PMLR 306, 2026.

1. Introduction

Fluorescent proteins (FPs) are widely used as genetically encoded markers for live-cell and in vivo imaging, where they enable long-term, bioorthogonal observation of protein localization, gene expression, and cellular dynamics (Tsien, 1998; Rodriguez et al., 2017). In GFP and GFP-like proteins, fluorescence originates from an autocatalytically formed chromophore derived from an internal tripeptide embedded in a β-barrel scaffold. The identity of this tripeptide and the surrounding residues jointly determine chromophore maturation, spectral position, and brightness (Tsien, 1998; Zimmer, 2002).

The design space of fluorescent proteins is correspondingly rich. Excitation and emission wavelengths span much of the visible spectrum, while quantum yield, extinction coefficient, photostability, and pK_a vary substantially across families (Rodriguez et al., 2017). This diversity underpins applications such as multicolor labeling, FRET sensing, and super-resolution microscopy, but also makes engineering difficult: small sequence changes can produce large spectral shifts or abolish fluorescence altogether. Because spectral tuning is governed by interactions between the chromophore and its local microenvironment, predictive models must capture both sequence context and structurally mediated effects rather than simple motif-level rules (Tsien, 1998; Zimmer, 2002).

Traditional protein engineering explores only a tiny fraction of this sequence space, and even dense mutational studies of GFP reveal rugged, highly constrained fitness landscapes (Sarkisyan et al., 2016). As curated datasets have grown, machine learning has become an increasingly attractive route for predicting FP properties and prioritizing candidates for synthesis. Our dataset is built from FPbase, a community-curated resource that aggregates fluorescent protein sequences and photophysical measurements across the literature (Lambert, 2019). Earlier computational tools such as FPredX (Tam et al., 2022) demonstrated the promise of alignment-based prediction, but alignment-driven methods struggle when target proteins are distantly related to the training set. Recent protein language models (PLMs), by contrast, can encode structural and functional regularities from large unlabeled sequence corpora and therefore offer a

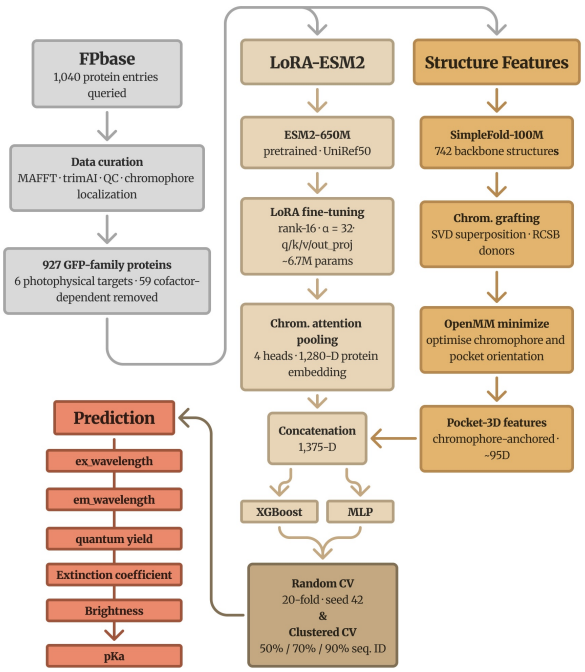

Figure 1. Overview of the FluorCode pipeline. LoRA-adapted ESM2-650M embeddings are extracted with chromophore-aware attention pooling and evaluated with both XGBoost and MLP prediction heads. A parallel structural branch extracts Pocket-3D chromophore-anchored descriptors from chromophore-grafted, energy-minimized structures for late-fusion ablation. Brightness was derived by quantum yield and extinction coefficient.

stronger basis for transfer to novel families (Lin et al., 2023). Random cross-validation sequence-similarity leakage is a common failure mode in protein ML (Rost, 1996; Dallago et al., 2021), however, it has not previously been explored for fluorescent protein property prediction.

In this work, we compare alignment-based one-hot encoding against LoRA-fine-tuned ESM2-650M representations for predicting six FP photophysical properties. Our overall workflow is summarized in Figure 1. This study demonstrates that the standard random cross-validation protocol used in prior work overestimates model performance, and that sequence-identity-clustered evaluation reveals a generalization gap of over 20 nm between one-hot and PLM-based representations. We further conduct a structural ablation study showing that Pocket-3D chromophore-anchored descriptors from 913 chromophore-grafted structures provide no consistent additional benefit when added to a strong PLM baseline.

2. Methods

2.1. Data curation

First, an API query of the FPbase, a community-editable fluorescent protein database, retrieved 1,040 protein entries with 990 entries containing amino-acid sequence (Lambert, 2019). For proteins with multiple spectral states, we retained the default state. This was followed by MAFFT sequence alignment (Katoh et al., 2002) and filtering using trimAl (Capella-Gutiérrez et al., 2009). After additional quality control we obtained a data set with 986 proteins. Each sequence was mapped to the canonical avGFP chromophore triad (Ser65–Tyr66–Gly67) using multiple sequence alignment which resulted in successful localization of the chromophore in 915 proteins. Moreover, 59 cofactor-dependent proteins were identified and excluded from GFP-family-specific analyses. This yielded a final data set containing unique 927 GFP-family proteins with six photophysical targets, including excitation maximum, emission maximum, quantum yield, extinction coefficient, pK_a, and brightness. The curated data set covers a wide range of measured properties representing variabilities within the photophysical landscape. As shown in the Appendix (Figure A1), excitation and emission values are broadly distributed after cleaning, and the data set contains significant variations across both spectral and non-spectral targets. This indicates a successful removal of noisy or incompatible entries without introducing bias toward specific wavelength ranges or only a narrow subset of fluorescent protein properties.

2.2. Sequence baselines and LoRA-ESM2 fine-tuning

Baseline: alignment-based one-hot encoding. As our baseline, we replicate the feature representation introduced by FPredX (Tam et al., 2022), a published method for fluorescent protein property prediction. All sequences are aligned using MAFFT, and each position–residue pair occurring in at least 2% of the multiple sequence alignment columns is encoded as a binary indicator, yielding 1,271 one-hot features per protein. These features are passed to an XGBoost regressor (Chen & Guestrin, 2016) with hyperparameters tuned by Optuna (Akiba et al., 2019) using 30 trials and 3-fold inner cross-validation. The search space spans the number of trees (300–1,500), maximum depth (3–9), learning rate (5×10^{-3}–0.15, log-uniform), subsample ratio (0.6–1.0), column sampling (0.3–0.9), and L1/L2 regularization (10^{-2}–10, log-uniform).

LoRA-ESM2 fine-tuning. We adapt ESM2-t33-650M (Lin et al., 2023), a protein language model pretrained on UniRef50, using Low-Rank Adaptation (LoRA) (Hu et al., 2022). Rank-16 adapters with $\alpha = 32$ are inserted into the query, key, value, and output projection matrices of the last six transformer layers (layers 27–32),

introducing approximately 983K trainable parameters, or about 0.15% of the full model. The model is trained in a multi-task setting to predict five photophysical targets simultaneously: excitation wavelength, emission wavelength, quantum yield, extinction coefficient, and pK_a. We use a masked Huber loss with $\delta = 1.0$ that ignores missing labels, together with task weights of 1.0 for excitation, emission, and quantum yield and 0.3 for extinction coefficient and pK_a. Training runs for up to 100 epochs with early stopping (patience 10), using a batch size of 8 with 2-step gradient accumulation for an effective batch size of 16. All LoRA-ESM2 training is performed on an NVIDIA H100 GPU in Google Colab. We apply differential learning rates of 5×10^{-4} for the LoRA adapters and 1×10^{-4} for the prediction heads, with 5 epochs of linear warmup followed by cosine annealing to a minimum learning rate of 10^{-6}. Gradient clipping is set to 1.0 and weight decay to 10^{-2}. All targets are Z-score normalized using statistics computed on the training fold only to avoid data leakage.

Chromophore-aware attention pooling. Standard mean pooling over per-residue embeddings treats all positions equally, even though fluorescent protein photophysics are governed primarily by the chromophore-forming triad and its immediate microenvironment. We therefore introduce a chromophore-aware attention pooling module with four independent attention heads. Each head computes a scalar attention score for every residue through a learned linear projection from 1,280 dimensions to 1, with a learned positive bias initialized to $+3.0$ on the three chromophore-triad positions to encourage attention toward the most photophysically relevant region. The four attended outputs are concatenated to form a 5,120-dimensional vector and then projected through a two-layer network ($5{,}120 \to 640 \to 1{,}280$) with GELU activation and layer normalization, yielding a 1,280-dimensional protein-level embedding.

Downstream prediction heads. We evaluate two downstream architectures on top of the LoRA-ESM2 embeddings. The **XGBoost head** takes the 1,280-dimensional embedding as input and is tuned identically to the baseline with 30 Optuna trials and 3-fold inner cross-validation. The **MLP head** consists of LayerNorm, dropout with $p = 0.1$, a hidden layer ($1{,}280 \to 256$) with GELU activation, a second dropout layer, and a linear output layer ($256 \to 1$). The MLP is trained end-to-end with the LoRA backbone and shares the same optimizer and learning-rate schedule described above.

2.3. Structure prediction and feature extraction

Structure prediction and chromophore grafting. To test whether explicit three-dimensional structural information can complement learned sequence representations, we construct a structural feature pipeline. Because experimental structures are unavailable for most fluorescent proteins, we predict backbone coordinates for 913 proteins using SimpleFold-100M (Wang et al., 2025). However, predicted structures lack the mature post-translational chromophore, a bicyclic heterocycle formed by autocatalytic cyclization of the chromophore-forming triad. To recover chromophore coordinates, we use an SVD superposition-based grafting procedure: for each target protein, we identify the highest-sequence-identity donor in a library of experimental RCSB Protein Data Bank structures with resolved chromophore HETATM records (Burley et al., 2019), superimpose donor and target using the Cα atoms within ±5 residues of the triad, and transfer the rotated chromophore coordinates. We reject grafts with local RMSD greater than 1.5 Å or more than five steric clashes under a 2.0 Å interatomic cutoff. All accepted grafted structures then undergo vacuum energy minimization with OpenMM (Eastman et al., 2017), which relaxes not only the orientation of the grafted chromophore but also the adjacent pocket residues in its local environment, yielding 913 chromophore-containing structures, as illustrated for a representative protein in Figure 2.

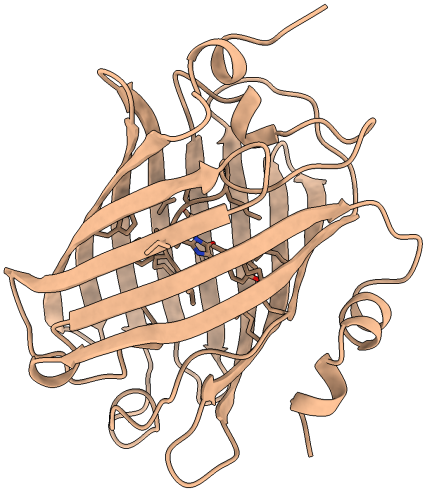

Figure 2. Energy-minimized predicted structure of crimson (FPbase ID: crimson), illustrating the chromophore-grafting and structural feature-extraction pipeline. During minimization, both the grafted chromophore orientation and the adjacent pocket residues are relaxed. The canonical β-barrel scaffold encloses the central α-helix and chromophore-forming triad, and residues in the local chromophore neighborhood are used to derive Pocket-3D chromophore-anchored descriptors.

Structural feature extraction. Our final structural representation is **Pocket-3D**, a hand-crafted, chromophore-HETATM-anchored descriptor set extracted directly from the minimized PDB structures with NumPy only. Pocket-3D was designed specifically to address a failure mode

we observed in earlier structural baselines: our initial Cα-anchored features were blind to the chemically relevant chromophore pocket because they mainly captured coarse barrel composition rather than the local ligand environment.

Pocket-3D instead anchors all spatial queries at the chromophore itself and extracts approximately 95 dimensions organized into four blocks. **Block A** encodes intrinsic chromophore chemistry and geometry, including the first residue identity of the chromophore tripeptide, phenol and imidazolinone ring planarity, the key exocyclic torsions τ and ϕ, inter-ring coplanarity, intra-chromophore distances, chromophore radius of gyration, and an atom-completeness quality flag. **Block B** is centered on the phenol hydroxyl oxygen and summarizes the local 3D environment using shell composition features at 3.5, 5.0, and 8.0 Å, nearest distances to chemically relevant side-chain atoms (for example carboxylates, histidine nitrogens, guanidinium groups, Lys NZ, Thr OG1, Ser OG, Asn/Gln donor–acceptor atoms, and Trp NE1), and a simple electrostatic proxy computed as $\sum_i q_i/d_i^2$ over nearby polar atoms. **Block C** mirrors this design around the imidazolinone-ring centroid, with shell descriptors, nearest-contact features, and an analogous electrostatic proxy. **Block D** captures broader barrel architecture through chromophore buriedness, covariance-eigenvalue ratios of the protein Cα cloud, local packing density around the phenol ring, a π-stacking candidate count, and solvent-exposure proxies.

The key design choice is that all local features are queried in the ligand-centered frame defined by the chromophore HETATM atoms rather than by sequence-adjacent backbone atoms. This allows Pocket-3D to describe the chromophore itself, the two sides of its chemical environment, and the broader β-barrel pocket in a compact representation. In downstream ablations, we use Pocket-3D to test whether explicit 3D chromophore-pocket information improves prediction beyond the LoRA-ESM2 sequence representation, rather than relying on high-dimensional pre-trained structural embeddings.

2.4. Model performance evaluation

We evaluate all models under two cross-validation schemes. **Random CV** uses standard 20-fold splits with seed 42 and no sequence-identity constraints. **Clustered CV** groups proteins using MMseqs2 (Steinegger & S"oding, 2017) at three thresholds—90% (183 clusters), 70% (82 clusters), and 50% (37 clusters)—and performs group K-fold cross-validation such that all members of a sequence cluster are assigned to the same fold. This design prevents family-level data leakage and provides a more realistic estimate of generalization to novel fluorescent protein families. We report Pearson correlation (R), mean absolute error (MAE), and root mean squared error (RMSE) as pooled metrics across all out-of-fold predictions.

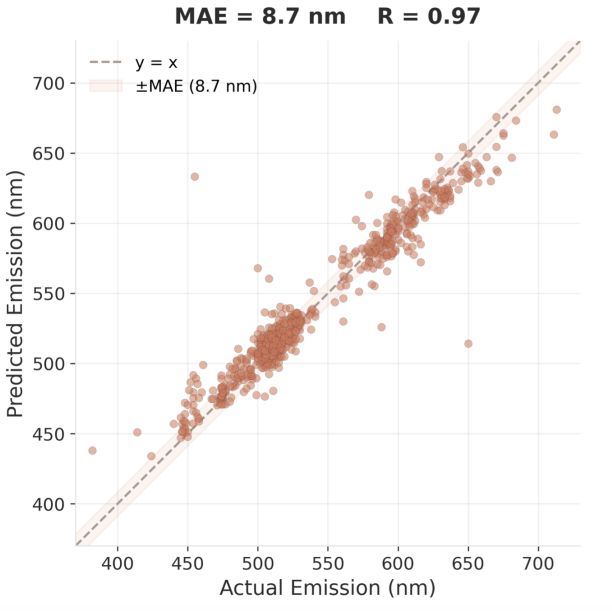

Figure 3. Scatter plot of LoRA-ESM2 (MLP) predictions under random cross-validation for emission wavelength. Predicted emission closely tracks the measured values, with MAE of 8.7 nm and Pearson correlation of $R = 0.97$.

3. Results

3.1. Random cross-validation

We first evaluate all models under standard 20-fold random cross-validation (Figures 3 and 4). Under this protocol, all three approaches—FPredX (alignment-based one-hot + XGBoost), LoRA-ESM2 with XGBoost head, and LoRA-ESM2 with MLP head—achieve apparently strong performance across all six photophysical properties. For excitation wavelength, Pearson correlations are $R = 0.89$, 0.95, and 0.97 with MAE of 12.7, 10.1, and 8.9 nm for FPredX, LoRA-ESM2 (XGBoost), and LoRA-ESM2 (MLP), respectively. Emission wavelength predictions are tighter still, with all three models achieving $R \geq 0.92$ and MAE $\leq$ 8.3 nm. For quantum yield ($R = 0.75/0.93/0.96$), extinction coefficient ($R = 0.70/0.96/0.97$), and pKa ($R = 0.48/0.88/0.91$), LoRA-ESM2 representations consistently outperform the one-hot baseline, with the MLP head yielding the highest correlations. Brightness prediction follows the same pattern ($R = 0.78/0.91/0.96$). While these results suggest that all models are competent predictors and that the performance gap between one-hot and learned representations is modest, we argue below that this evaluation is misleading.

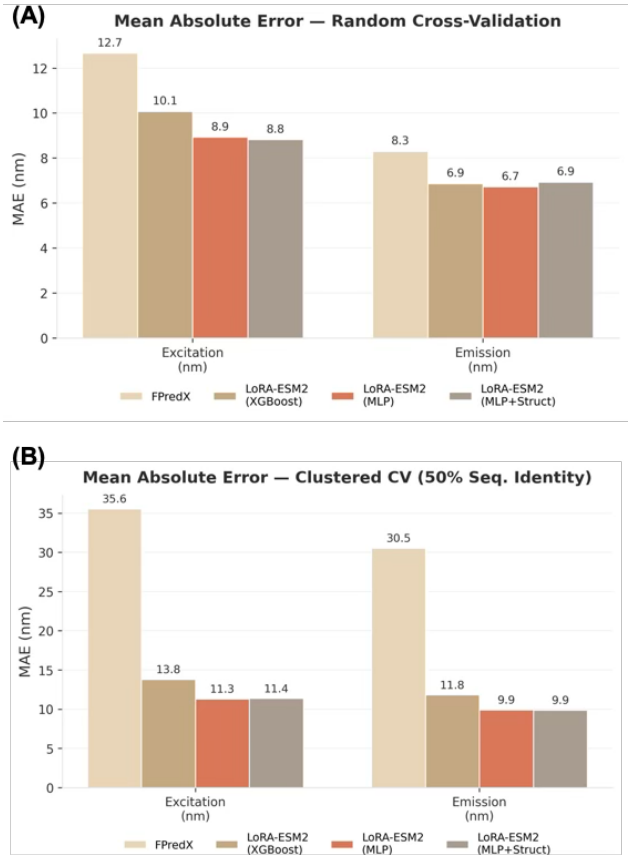

Figure 4. Comparison of model error under random and clustered cross-validation. Sequence-identity-clustered evaluation reveals a substantial increase in MAE for alignment-based features, whereas LoRA-ESM2 variants remain comparatively stable.

3.2. Clustered cross-validation

Fluorescent proteins in FPbase are heavily clustered by evolutionary family: GFP variants, mCherry derivatives, and other lineages share >90% sequence identity within each group. Standard random K-fold splitting places near-identical variants on both sides of the train/test boundary, allowing models to memorize family-specific spectral signatures rather than learning the physicochemical determinants of fluorescence. This is particularly problematic for alignment-based one-hot features, which explicitly encode positional amino acid identity and thus directly represent family membership. To obtain a more realistic estimate of generalization to *novel* FP families, we cluster all sequences using MMseqs2 at 50% sequence identity, yielding 37 clusters, and perform group K-fold cross-validation such that no two proteins sharing >50% identity appear in both training and test sets simultaneously.

3.3. Spectral property prediction under clustered CV

Under 50% identity-clustered CV (Figure 6), FPredX performance is significantly lower. Excitation MAE inflates from 12.7 to 35.6 nm (R: $0.89 \rightarrow 0.58$), and emission MAE from 8.3 to 30.5 nm (R: $0.92 \rightarrow 0.63$). In contrast, LoRA-ESM2 (XGBoost) decreases from 10.1 to 13.8 nm for excitation and 6.9 to 11.8 nm for emission, maintaining correlations of $R = 0.93$ and 0.94, respectively. Using the MLP head excitation MAE increases from 8.9 to only 11.3 nm ($R = 0.95$), and emission from 6.7 to 9.9 nm ($R = 0.95$). This reveals a generalization gap of over 20 nm between FPredX and LoRA-ESM2 for both spectral properties—a difference missed by random CV. A component ablation further shows that cross-family generalization depends on a strict interaction between LoRA fine-tuning and chromophore-aware pooling: LoRA without targeted pooling remains indistinguishable from the frozen baseline, whereas chromophore-aware pooling without LoRA is harmful; only their combination yields the full clustered-CV gain (Figure A2).

3.4. Non-spectral properties and brightness collapse

The disparity is visible especially in non-spectral targets (Figure 6 and Table 1). Under clustered CV at 50% identity, FPredX correlations collapse to near-zero for quantum yield (R: $0.75 \rightarrow 0.16$), extinction coefficient (R: $0.70 \rightarrow 0.14$), and pKa (R: $0.48 \rightarrow 0.13$), indicating that one-hot features carry no predictive signal for these properties beyond family membership. Moreover, brightness prediction—which compounds errors in both quantum yield and extinction coefficient—degrades from $R = 0.78$ to $R = 0.13$, indistinguishable from random guessing. LoRA-ESM2 representations, by contrast, retain strong predictive power across all properties: quantum yield ($R = 0.92$ XGBoost / 0.95 MLP), extinction coefficient ($R = 0.95$ / 0.96), pKa ($R = 0.87$ / 0.91), and brightness ($R = 0.90$ / 0.96). The brightness result is particularly noteworthy, while FPredX collapses entirely, both LoRA-ESM2 variants maintain near-random-CV performance, demonstrating that the fine-tuned language model has learned representations that generalize across FP families.

3.5. Degradation trajectory across clustering thresholds

To further characterize this generalization gap, we evaluate all models at intermediate clustering thresholds of 90% and 70% identity in addition to the stringent 50% threshold (Figure 5). FPredX exhibits steep, monotonic MAE inflation as the clustering threshold decreases: excitation MAE rises from 12.7 nm (random) to 23.7 nm (90%), 25.6 nm (70%), and 35.6 nm (50%). LoRA-ESM2 (XGBoost) shows a far gentler trajectory: $10.1 \rightarrow 12.3 \rightarrow 12.9 \rightarrow 13.8$ nm. The MLP variant is the most stable, with excitation MAE of $8.9 \rightarrow 9.7 \rightarrow 9.7 \rightarrow 11.3$ nm—a total inflation of only 2.4 nm even at the most stringent threshold. This progressive divergence confirms that the one-hot encoding primarily captures family membership rather than photophysical properties, as family-level information leakage is progressively

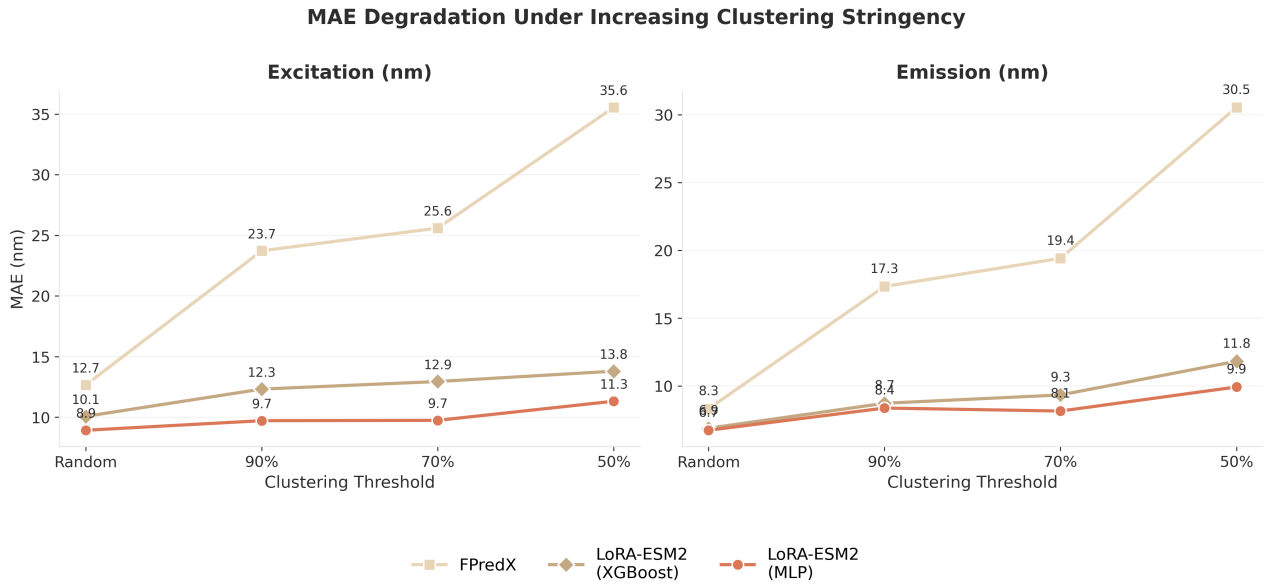

Figure 5. MAE degradation under increasing clustering stringency for excitation and emission wavelength prediction. FPredX degrades sharply as clustering becomes more stringent, whereas LoRA-ESM2 variants remain comparatively stable across thresholds.

Table 1. Pearson correlation (R) under random and 50% identity-clustered cross-validation across six photophysical properties. Clustered CV reveals that FPredX collapses to near-random performance on most targets, while LoRA-ESM2 representations remain robust. Best clustered-CV result per target in **bold**.

		Excitation	Emission	QY	Ext. Coeff.	pKa	Brightness
Random	FPredX	0.89	0.92	0.75	0.70	0.48	0.78
	LoRA-ESM2 (XGBoost)	0.95	0.97	0.93	0.96	0.88	0.91
	LoRA-ESM2 (MLP)	0.97	0.97	0.96	0.97	0.91	0.96
Clustered	FPredX	0.58	0.63	0.16	0.14	0.13	0.13
	LoRA-ESM2 (XGBoost)	0.93	0.94	0.92	0.95	0.87	0.90
	LoRA-ESM2 (MLP)	**0.95**	**0.95**	**0.95**	**0.96**	**0.91**	**0.96**

removed, one-hot predictions degrade proportionally, while LoRA-ESM2 representations—which encode transferable sequence-function relationships—remain robust.

3.6. Pocket-3D structural features

With clustered CV, adding structural features does not improve upon LoRA - ESM2 even after adapting the structural block to revolve around the chromophore ligand. Our initial Cα-anchored features mainly captured bulk barrel composition instead of the chemically relevant pocket. Pocket-3D was introduced to address this gap by anchoring all measurements at the chromophore HETATM atoms and encoding local geometry, hydrogen-bond partners, electrostatic proxies, and packing around the phenol and imidazolinone moieties.

Despite adding additional structure features, the model still provides no measurable gain beyond the sequence model under fair evaluation. Our predicted Pocket-3D descriptors do not add measurable benefit beyond LoRA-ESM2 embeddings for this dataset and evaluation setup. One possible explanation is that LoRA-ESM2 with chromophore-aware pooling already captures much of the useful local context from sequence alone; another is that the structural descriptors are still not expressive enough, or that the dataset is too small for the extra features to help reliably. PLMs pretrained on large evolutionary corpora are known to implicitly encode substantial structural information—including residue contacts, secondary structure, and solvent accessibility—from sequence alone (Rao et al., 2021); accordingly, our results suggest that, for GFP-family photophysics, adding explicit pocket-level geometry in this hand-crafted form yields little additional signal beyond a strong fine-tuned PLM.

4. Discussion

4.1. Protocol evaluation

Our results highligh the fact that evaluation protocol can be as important as model architecture. Under random cross-validation, alignment-based one-hot encoding appears to be

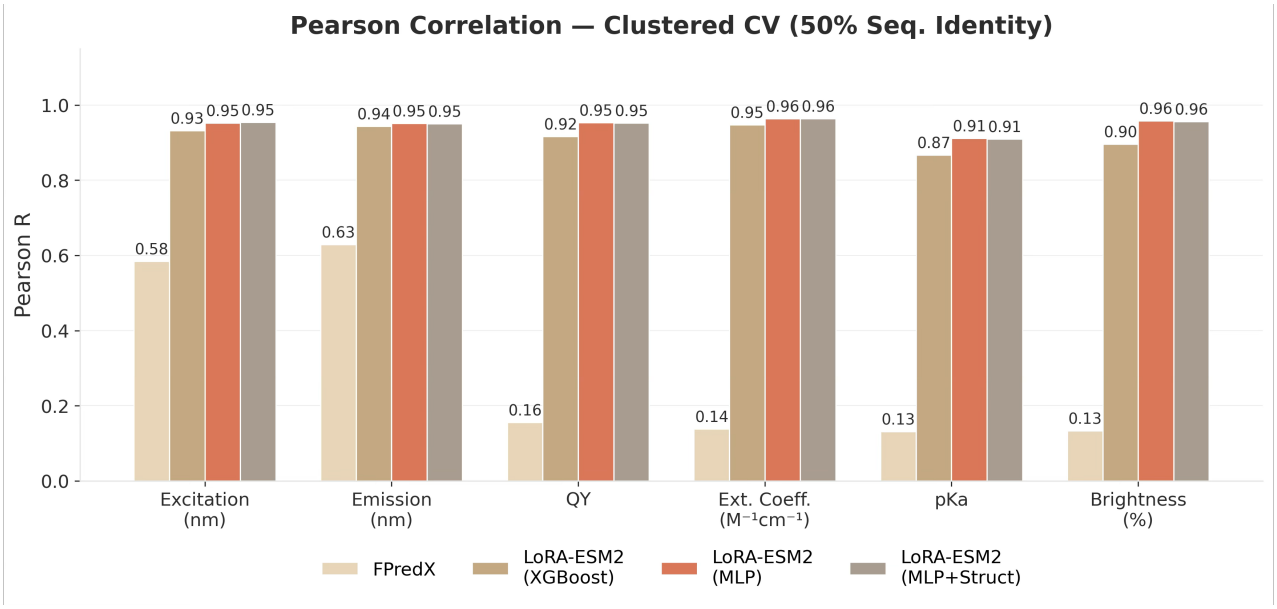

Figure 6. Pearson correlation under 50% sequence-identity-clustered cross-validation across six photophysical properties. FPredX collapses on non-spectral targets, whereas LoRA-ESM2 variants retain strong performance across excitation, emission, quantum yield, extinction coefficient, pKa, and brightness.

a competitive baseline, with performance gaps of only 2–3 nm MAE relative to PLM-based representations for spectral properties. This difference could be attributed to the additional capacity of the larger model rather than to differences in representation quality. Clustered cross-validation overturns this conclusion entirely, revealing that one-hot features encode *family membership* rather than photophysics—a form of sequence-similarity leakage (Rost, 1996; Li & Godzik, 2006; Dallago et al., 2021) that is especially severe here due to the evolutionary clustering of FPbase.

4.2. Target dependent leakage

The severity of this leakage varies by target property. Spectral wavelengths are partially predictable from family identity (GFP-derived proteins tend toward green emission regardless of mutations), so FPredX retains some residual correlation ($R \approx 0.6$) even under strict clustering. Non-spectral properties—quantum yield, extinction coefficient, pKa, and brightness—are more sensitive to individual mutations and less correlated with family membership, which is why FPredX performance on these targets collapses to noise ($R \approx 0.13$) under clustered CV while LoRA-ESM2 remains largely unaffected.

4.3. Addition of 3D structural features

The failure of Pocket-3D to improve over the sequence-only PLM deserves further consideration. We explored this question in multiple stages rather than relying on a single structural baseline. A first-generation Cα-anchored block was rejected because it described the barrel too coarsely to capture the photophysically relevant pocket. We then constructed Pocket-3D, a chromophore-HETATM-anchored structural representation that explicitly encodes chromophore torsions, ring planarity, shell composition around the phenol hydroxyl, nearest-pocket hydrogen-bond and charged contacts, electrostatic proxies, and local packing.

That even this ligand-centered structural block yields no improvement under fair evaluation suggests several possibilities. LoRA-ESM2's chromophore-aware attention pooling may already recover much of the same local microenvironment information from sequence alone; alternatively, the Pocket-3D descriptors may still be too limited, or the data may be too sparse for the structural features to provide a reliable incremental gain. This is consistent with recent findings that PLMs implicitly learn structural representations during pre-training (Rao et al., 2021; Lin et al., 2023), while also leaving open the possibility that richer structural models could still help in future work. For FP photophysics—a problem dominated by the chromophore and its immediate pocket—our results show that this particular structural representation does not outperform a properly pooled and fine-tuned sequence model.

4.4. Why the MLP head performs best

The MLP head consistently outperforms the XGBoost head, particularly under clustered CV. This could be attributed to the training. The MLP and LoRA backbone are optimized jointly, the attention pooling module can adapt its weighting to complement the downstream predictor, whereas the XGBoost head operates on frozen embeddings and cannot

backpropagate task-specific gradients into the representation. The appendix ablation clarifies that this is not just an advantage of model capacity: mean pooling suppresses the benefit of LoRA, while chromophore-aware pooling is only useful once LoRA has learned property-relevant local representations around the chromophore (Figure A2).

4.5. Current limitations

A limitation of this study is the relatively small dataset size (927 proteins), which constrains the stringency of clustering. At 50% identity, only 37 clusters remain, and fold sizes vary significantly—the largest cluster contains nearly half of all proteins. Larger and more diverse FP datasets would enable more stringent evaluation thresholds and more reliable estimates of generalization outside of family.

4.6. Conclusions and next steps

Taken together, these results show that the standard random cross-validation protocol used to benchmark fluorescent protein property prediction substantially overestimates performance, particularly for alignment-based representations. Under sequence-identity-clustered evaluation at 50% identity, LoRA-fine-tuned ESM2-650M with chromophore-aware attention pooling outperforms the FPredX baseline by over 20 nm MAE for spectral properties and maintains robust prediction across all six targets, while FPredX collapses to noise on quantum yield, extinction coefficient, pKa, and brightness. Our structural ablation further indicates that generated Pocket-3D features provide no measurable benefit beyond a properly fine-tuned PLM in the current setting, which is consistent with the possibility that the sequence model already captures much of the relevant local context.

Next possible step would be moving from forward prediction to generative fluorescent protein design. Given target excitation and emission wavelengths, quantum yield, and brightness constraints, a conditional generator could propose new sequences while using the LoRA-ESM2 predictor as a guidance signal. Future work could also explore target-conditioned attention, graph neural networks over chromophore pocket graphs, and extension of FluorCode to cofactor-dependent fluorescent proteins and larger mutational scanning datasets.

Code Availability

Code for FluorCode is available at https://github.com/ignirico/FluorCode.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2623–2631, 2019.

Burley, S. K., Berman, H. M., Bhikadiya, C., Bi, C., Chen, L., Di Costanzo, L., Christie, C., Dalenberg, K., Duarte, J. M., Dutta, S., Feng, Z., Ghosh, S., Goodsell, D. S., Green, R. K., Guranovic, V., Guzenko, D., Hudson, B. P., Kalro, T., Liang, Y., Lowe, R., Namkoong, H., Peisach, E., Persikova, I., Randle, C., Rose, A., Rose, P., Sala, R., Sekharan, M. R., Shao, C., Tan, L., Tao, Y.-P., Valasatava, Y., Voigt, M., Westbrook, J. D., Woo, J., Young, J. Y., and Zardecki, C. RCSB protein data bank: Biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Research*, 47(D1):D464–D474, 2019.

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. trimal: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25 (15):1972–1973, 2009.

Chen, T. and Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

Dallago, C., Mou, J., Johnston, K. E., Wittmann, B. J., Bhattacharya, N., Goldman, S., Bhaskara, R. M., and Yang, K. K. FLIP: Benchmark tasks in fitness landscape inference for proteins. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., Wang, L.-P., Simmonett, A. C., Harrigan, M. P., Stern, C. D., Wiewiora, R. P., Brooks, B. R., and Pande, V. S. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology*, 13(7): e1005659, 2017.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Katoh, K., Misawa, K., Kuma, K.-i., and Miyata, T. MAFFT: A novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Research*, 30(14):3059–3066, 2002.

Lambert, T. J. Fpbase: A community-editable fluorescent protein database. *Nature Methods*, 16(4):277–278, 2019.

Li, W. and Godzik, A. CD-HIT: A fast program for clus-

tering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.

Lin, Z., Akin, H., Fragoza, R., Tiang, M., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Rao, R., Bhattacharya, N., Thomas, N., et al. MSA transformer. In *Proceedings of the International Conference on Machine Learning*, 2021.

Rodriguez, E. A., Campbell, R. E., Lin, J. Y., Lin, M. Z., Miyawaki, A., Palmer, A. E., Shu, X., Zhang, J., and Tsien, R. Y. The growing and glowing toolbox of fluorescent and photoactive proteins. *Trends in Biochemical Sciences*, 42(2):111–129, 2017.

Rost, B. Prediction in 1D: Secondary and supersecondary structure. *Methods in Enzymology*, 266:525–539, 1996.

Sarkisyan, K. S., Bolotin, D. A., Meer, M. V., et al. Local fitness landscape of the green fluorescent protein. *Nature*, 533:397–401, 2016.

Steinegger, M. and S"oding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, 2017.

Tam, B. K. C., Brown, J. M. B., and Lin, M. Z. FPredX: A machine-learning framework for fluorescent protein property prediction. *Bioinformatics Advances*, 2022.

Tsien, R. Y. The green fluorescent protein. *Annual Review of Biochemistry*, 67:509–544, 1998.

Wang, Y., Lu, J., Jaitly, N., Susskind, J. M., and Bautista, M. A. Simplefold: Folding proteins is simpler than you think. *arXiv preprint arXiv:2509.18480*, 2025.

Zimmer, M. Green fluorescent protein (GFP): Applications, structure, and related photophysical behavior. *Chemical Reviews*, 102(3):759–781, 2002.

A. Appendix

A.1. Dataset Statistics

Table A1 summarizes the curated dataset. The six prediction targets have varying coverage because not all photophysical measurements are reported for every protein in FPbase.

Table A1. Dataset composition and per-target statistics for the 927 GFP-family proteins.

Target	N **available**	**Coverage**	**Unit**
Excitation (λ_{ex})	814	88%	nm
Emission (λ_{em})	775	84%	nm
Quantum yield (QY)	637	69%	0–1
Extinction coeff. (ε)	572	62%	$M^{-1}cm^{-1}$
pKa	371	40%	–
Brightness	547	59%	% of EGFP

Pipeline stage	**Count**
FPbase API query	1,040 entries
With amino-acid sequence	990
After MAFFT + trimAI + QC	986
Chromophore localized	915
Cofactor-dependent excluded	59
Final GFP-family set	**927**

MMseqs2 clustering. Table A2 shows the cluster distributions used for clustered cross-validation. At 50% identity, the largest cluster contains 497 proteins (54% of the dataset), reflecting the dominance of the GFP-like β-barrel superfamily.

Table A2. MMseqs2 cluster statistics at three identity thresholds.

Threshold	**Clusters**	**Largest cluster**	**Median size**
90%	183	285	1
70%	82	306	2
50%	37	497	2

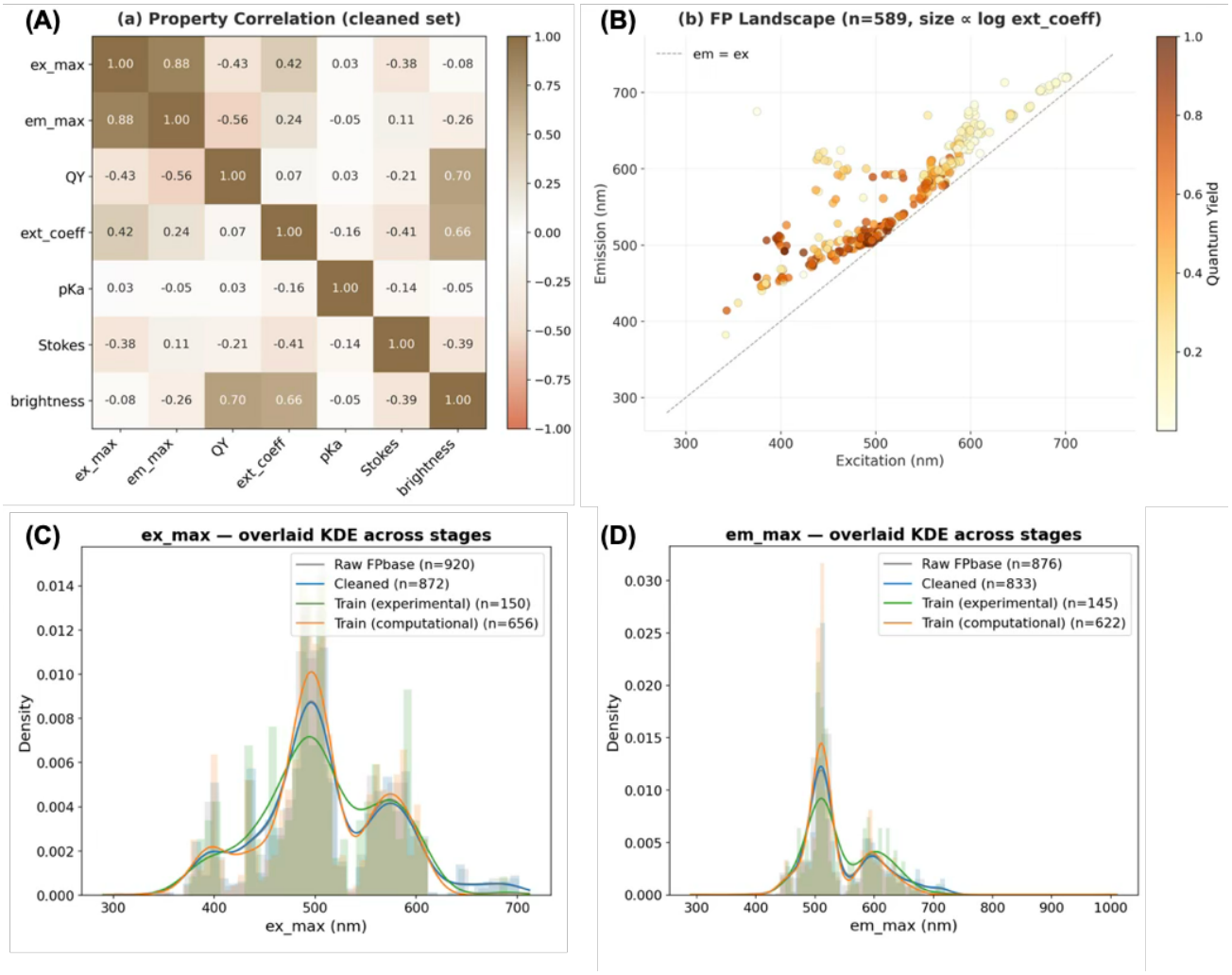

Figure A1. Overview of the curated dataset after filtering. (A) Pairwise correlations among the measured photophysical properties show the main relationships retained in the cleaned data set. (B) The target distributions remain broad across properties, indicating substantial diversity rather than collapse to a narrow range. (C) Excitation and (D) emission wavelength distributions show no obvious systematic bias after data cleaning, supporting that the filtering procedure does not preferentially retain only a restricted spectral subset.

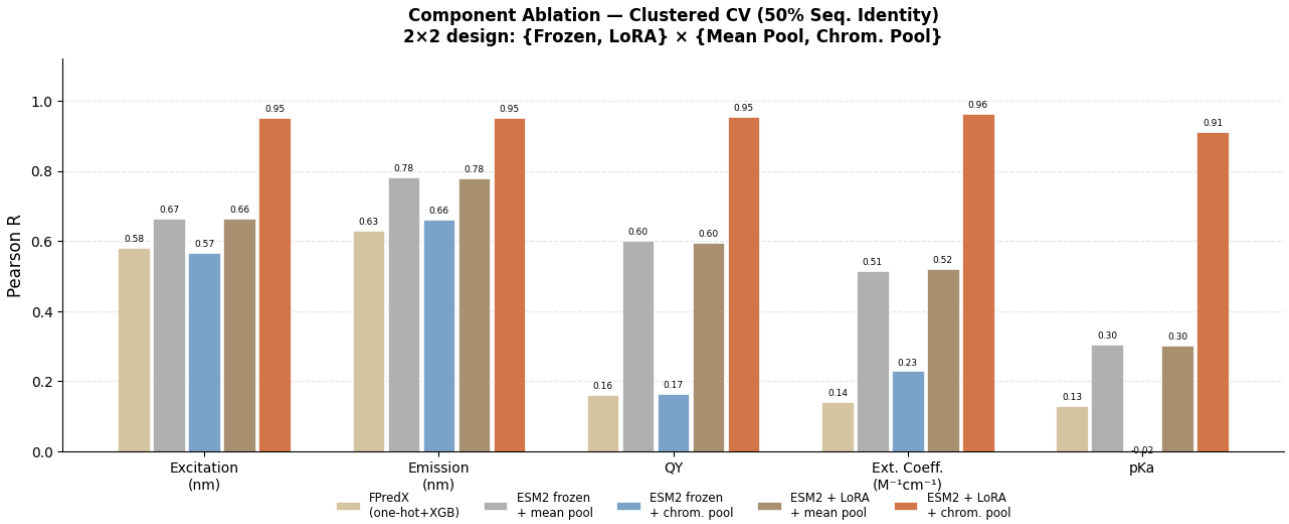

Figure A2. Component ablation reveals a strict interaction effect between LoRA fine-tuning and chromophore-aware pooling. Frozen ESM2 with mean pooling reaches a mean Pearson correlation of 0.57, whereas adding chromophore-aware pooling alone reduces performance to 0.32. LoRA fine-tuning without targeted pooling remains ineffective (0.57), indicating that mean pooling washes out the localized signal learned around the chromophore. Only the combination of LoRA and chromophore-aware pooling generalizes strongly (0.95), showing that task-adapted local representations must be paired with biologically informed feature extraction for cross-family generalization.

A.2. Training Algorithm

Algorithm 1 LoRA-ESM2 Multi-Task Training

Input: Dataset $\mathcal{D} = \{(s_i, c_i, \mathbf{y}_i)\}_{i=1}^{N}$ (sequence, chromophore positions, targets)
Hyperparams: rank r=16, α=32, layers $\{27, ..., 32\}$, LR_{LoRA}=5e-4, LR_{head}=1e-4

Load ESM2-650M; freeze all parameters
Inject LoRA adapters (r=16) into {q, k, v, out}_proj of layers 27–32
for fold $k = 1$ **to** K **do**
  Re-initialize LoRA ($\mathbf{A} \sim \mathcal{N}(0, 0.01)$, $\mathbf{B} = \mathbf{0}$), pooling, and heads
  Compute target stats (μ_t, σ_t) from train split only
  for epoch $= 1$ **to** max_epochs **do**
    for mini-batch (S, C, Y) **do**
      $\mathbf{H} \leftarrow \text{ESM2}_{\text{LoRA}}(S)$ ▷ (B, L, 1280)
      $\mathbf{z} \leftarrow \text{ChromPool}(\mathbf{H}, C)$ ▷ 4-head attention + chromophore bias
      $\hat{y}_t \leftarrow \text{MLP}_t(\mathbf{z})$ for each target t
      $\mathcal{L} \leftarrow \sum_t w_t \cdot \text{Huber}(\hat{y}_t, y_t)$ ▷ masked for missing values
      Backward with mixed-precision + gradient checkpointing
    end for
    Step AdamW (differential LR) + cosine schedule with warmup
    if val loss improves **then**
      Save checkpoint (LoRA + pooling + head weights)
    else if patience exceeded **then**
      break
    end if
  end for
end for
Output: K checkpoints, each with trainable weights + (μ_t, σ_t)

A.3. Hyperparameters

Table A3 lists all hyperparameters used for LoRA-ESM2 training. XGBoost hyperparameters are tuned per target via Optuna (30 trials, 3-fold inner CV); the search space is described in the main text.

Table A3. LoRA-ESM2 training hyperparameters.

Hyperparameter	**Value**
LoRA configuration	
Rank r / scaling α	16 / 32
Target projections	q, k, v, out_proj
Adapted layers	27–32 (last 6 of 33)
LoRA trainable params	$\sim$983K
Architecture	
Backbone	ESM2-t33-650M (frozen)
Pooling heads	4 (chromophore-aware attention)
Pooling projection	$5{,}120 \to 640$ (GELU, LN) $\to 1{,}280$
Chromophore bias init	$+3.0$
MLP head	LN $\to$ Dropout $\to$ 256 (GELU) $\to$ Dropout $\to$ 1
Dropout	0.1
Total trainable params	$\sim$6.7M (1.0% of total)
Optimization	
Loss	Huber ($\delta = 1.0$), masked for missing targets
Task weights	ex/em/qy: 1.0, ext_coeff/pKa: 0.3
Optimizer	AdamW (weight decay 10^{-2})
LR (LoRA / heads)	5×10^{-4} / 1×10^{-4}
LR schedule	5-epoch linear warmup $\to$ cosine ($\eta_{\min}$=10^{-6})
Batch size (effective)	16 (8×2 gradient accumulation)
Gradient clipping	1.0
Max epochs / patience	100 / 10
Target normalization	Z-score (train-fold μ, σ only)
Cross-validation	
Folds / seed	20 / 42
Hardware	NVIDIA H100 (Google Colab)
Time per fold	$\sim$1.5 h

A.4. Full Cross-Validation Results

Tables A4 and A5 report MAE and Pearson R for all models across all clustering thresholds. "FPredX" denotes alignment-based one-hot + XGBoost; "XGB" and "MLP" denote LoRA-ESM2 with XGBoost and MLP heads, respectively.

Table A4. Mean absolute error (MAE) under random and clustered cross-validation. Units: nm for excitation/emission/brightness; dimensionless for QY; $M^{-1}cm^{-1}$ for extinction coefficient; pH units for pKa. Best result per target and scheme in **bold**.

Target	Model	Random	90%	70%	50%
Excitation	FPredX	12.67	23.73	25.60	35.55
	LoRA-ESM2 (XGB)	10.07	12.32	12.95	13.80
	LoRA-ESM2 (MLP)	**8.93**	**9.72**	**9.74**	**11.32**
Emission	FPredX	8.30	17.33	19.41	30.54
	LoRA-ESM2 (XGB)	6.87	8.71	9.33	11.82
	LoRA-ESM2 (MLP)	**6.72**	**8.36**	**8.14**	**9.92**
QY	FPredX	0.126	0.195	0.201	0.234
	LoRA-ESM2 (XGB)	0.067	0.073	0.078	0.082
	LoRA-ESM2 (MLP)	**0.048**	**0.048**	**0.050**	**0.054**
Ext. coeff.	FPredX	18,740	30,396	38,691	41,435
	LoRA-ESM2 (XGB)	7,591	8,764	8,355	9,015
	LoRA-ESM2 (MLP)	**6,871**	**7,067**	**7,275**	**7,189**
pKa	FPredX	0.738	0.850	0.896	0.940
	LoRA-ESM2 (XGB)	0.325	0.317	0.352	0.409
	LoRA-ESM2 (MLP)	**0.266**	**0.260**	**0.271**	**0.299**
Brightness	FPredX	11.80	22.09	33.25	36.30
	LoRA-ESM2 (XGB)	8.54	9.15	9.32	9.62
	LoRA-ESM2 (MLP)	**5.11**	**5.19**	**5.59**	**5.63**

Table A5. Pearson correlation (R) under random and clustered cross-validation. Best result per target and scheme in **bold**.

Target	Model	Random	90%	70%	50%
Excitation	FPredX	0.89	0.79	0.74	0.58
	LoRA-ESM2 (XGB)	0.95	0.94	0.94	0.93
	LoRA-ESM2 (MLP)	**0.97**	**0.96**	**0.96**	**0.95**
Emission	FPredX	0.92	0.83	0.82	0.63
	LoRA-ESM2 (XGB)	0.97	0.96	0.96	0.94
	LoRA-ESM2 (MLP)	**0.97**	**0.97**	**0.97**	**0.95**
QY	FPredX	0.75	0.43	0.41	0.16
	LoRA-ESM2 (XGB)	0.93	0.93	0.92	0.92
	LoRA-ESM2 (MLP)	**0.96**	**0.96**	**0.96**	**0.95**
Ext. coeff.	FPredX	0.70	0.38	0.16	0.14
	LoRA-ESM2 (XGB)	0.96	0.95	0.95	0.95
	LoRA-ESM2 (MLP)	**0.97**	**0.97**	**0.96**	**0.96**
pKa	FPredX	0.48	0.27	0.18	0.13
	LoRA-ESM2 (XGB)	0.88	0.90	0.88	0.87
	LoRA-ESM2 (MLP)	**0.91**	**0.92**	**0.91**	**0.91**
Brightness	FPredX	0.78	0.50	0.26	0.13
	LoRA-ESM2 (XGB)	0.91	0.91	0.90	0.90
	LoRA-ESM2 (MLP)	**0.96**	**0.96**	**0.96**	**0.96**

A.5. Structural Feature Ablation

Table A6 compares LoRA-ESM2 (MLP) alone against LoRA-ESM2 + Pocket-3D ($\sim$95 dims concatenated) under all clustering thresholds. The structural block provides no consistent improvement, matching the main-text finding that Pocket-3D does not add measurable benefit beyond LoRA-ESM2 in the current setup.

Table A6. MAE comparison: LoRA-ESM2 (MLP) vs. LoRA-ESM2 (MLP) + Pocket-3D structural features. Δ is (with structure) − (without); positive values indicate worse performance with the added Pocket-3D features.

Target	Scheme	MLP only	MLP + Pocket-3D	Δ
Excitation	Random	8.93	8.82	−0.11
	90%	9.72	9.65	−0.07
	70%	9.74	9.88	+0.13
	50%	11.32	11.38	+0.06
Emission	Random	6.72	6.93	+0.21
	90%	8.36	8.22	−0.14
	70%	8.14	8.27	+0.13
	50%	9.92	9.88	−0.04
QY	Random	0.048	0.048	+0.001
	90%	0.048	0.049	+0.001
	70%	0.050	0.050	0.000
	50%	0.054	0.054	0.000
Ext. coeff.	Random	6,871	6,626	−245
	90%	7,067	6,959	−108
	70%	7,275	7,311	+36
	50%	7,189	7,298	+109
pKa	Random	0.266	0.265	−0.001
	90%	0.260	0.266	+0.006
	70%	0.271	0.272	+0.001
	50%	0.299	0.280	−0.019
Brightness	Random	5.11	5.25	+0.14
	90%	5.19	5.22	+0.03
	70%	5.59	5.51	−0.07
	50%	5.63	5.79	+0.16

A.6. Pocket-3D Feature Blocks

Table A7 describes the four blocks of the Pocket-3D descriptor, totaling ∼95 dimensions. All features are computed with NumPy from minimized PDB coordinates. Features are anchored at chromophore HETATM atoms rather than at backbone Cα.

Table A7. Pocket-3D feature blocks. All spatial queries are anchored at chromophore HETATM heavy atoms.

Block	Dims	Description
A – Chromophore chemistry	20	First-residue identity of the chromophore tripeptide (8-dim one-hot: M/Q/G/T/S/H/C/other); phenol and imidazolinone ring planarity RMSD; exocyclic torsions τ (CA2–CB2–CG2–CD1) and ϕ (N2–CA2–CB2–CG2) as $(\cos, \sin)$ pairs; inter-ring dihedral coplanarity; CZ–OH vs. phenol-plane normal angle; intra-chromophore distances (OH–N2, OH–CA3); chromophore heavy-atom radius of gyration; atom-completeness flag.
B – Phenol-OH environment	∼35	Shell composition at 3.5, 5.0, and 8.0 Å from phenol OH (counts of hydrophobic, polar, positive, negative residues); nearest-atom distances to Glu/Asp carboxylate O, His Nδ/Nε, Arg guanidinium N, Lys NZ, Thr OG1, Ser OG, Asn OD1/ND2, Gln OE1/NE2, Trp NE1; signed electrostatic proxy $\sum_i q_i/d_i^2$ over polar atoms within 5 Å.
C – Imidazolinone environment	∼20	Shell composition at 3.5, 5.0, and 8.0 Å from imidazolinone centroid; nearest-atom distances to Thr OG1, Ser OG, Trp ring centroid, Phe ring centroid (H-bond donors and π-stacking partners); nearest backbone O within 4 Å (as $1/d^2$); electrostatic proxy analogous to Block B.
D – Barrel architecture	∼10	Chromophore-to-protein-centroid distance (buriedness); eigenvalue ratios of the Cα covariance matrix (shape descriptors); local packing density (atom count within 5 Å of phenol ring); π-stacking candidate count (Phe/Tyr/Trp/His rings within 6 Å); solvent-exposure proxy (fraction of 8 Å shell that is solvent-accessible).

Structure pipeline. The 913 structures used for Pocket-3D extraction were generated as follows: (1) backbone prediction with SimpleFold-100M (742 predictions; remaining via ESMFold fallback); (2) chromophore grafting from RCSB experimental donor structures via SVD superposition on Cα atoms ± 5 residues from the triad (rejection criteria: local RMSD > 1.5 Å or > 5 steric clashes at 2.0 Å cutoff); (3) vacuum energy minimization with OpenMM (AMBER ff14SB force field). Proteins without a parseable chromophore HETATM receive a zero-vector with `has_pocket = 0`.

A.7. XGBoost Late-Fusion Pipeline

Algorithm 2 XGBoost Late Fusion with LoRA Embeddings

Input: K trained LoRA-ESM2 checkpoints, dataset $\mathcal{D}$
Optional: Pocket-3D features $\mathbf{P} \in \mathbb{R}^{N \times 95}$

// Step 1: Extract embeddings
for fold $k = 1$ **to** K **do**
  Load checkpoint k; set model to eval mode
  for each protein i in $\mathcal{D}$ **do**
    $\mathbf{z}_i^{(k)} \leftarrow \text{ChromPool}(\text{ESM2}_{\text{LoRA}}^{(k)}(s_i), c_i)$ ▷ 1,280-D
  end for
end for

// Step 2: Tune XGBoost per target
for each target t **do**
  $\theta_t^* \leftarrow \text{Optuna}(\text{XGBRegressor}, \mathbf{X}_{\text{fused}}, y_t)$ ▷ 30 trials, 3-fold inner CV
end for

// Step 3: 20-fold outer CV
for fold $k = 1$ **to** K **do**
  $\mathbf{X}_{\text{fused}} \leftarrow [\mathbf{z}^{(k)} \mid \mathbf{P}]$ ▷ concatenate 1,280-D + 95-D (if using structure)
  for each target t **do**
    Train XGBRegressor(θ_t^*) on train split of fold k
    Predict on val split; clamp to valid range
  end for
end for
Output: Out-of-fold predictions for all targets