Improved Therapeutic Antibody Reformatting through Multimodal Machine Learning

Jiayi Xin[*]
University of Pennsylvania
jiayixin@seas.upenn.edu

Aniruddh Raghu
BigHat Biosciences
araghu@bighatbio.com

Nick Bhattacharya
BigHat Biosciences

Adam Carr
BigHat Biosciences

Melanie Montgomery
BigHat Biosciences

Hunter Elliott
BigHat Biosciences
helliott@bighatbio.com

Abstract

Modern therapeutic antibody design often involves composing multi-part assemblages of individual functional domains, each of which may be derived from a different source or engineered independently. While these complex formats can expand disease applicability and improve safety, they present a significant engineering challenge: the function and stability of individual domains are not guaranteed in the novel format, and the entire molecule may no longer be synthesizable. To address these challenges, we develop a machine learning framework to predict *reformatting success* – whether converting an antibody from one format to another will succeed or not. Our framework incorporates both antibody sequence and structural context, incorporating an evaluation protocol that reflects realistic deployment scenarios. In experiments on a real-world antibody reformatting dataset, we find the surprising result that large pretrained protein language models (PLMs) fail to outperform simple, domain-tailored, multimodal representations. This is particularly evident in the most difficult evaluation setting, where we test model generalization to a new starting antibody. In this challenging "new antibody, no data" scenario, our best multimodal model achieves high predictive accuracy, enabling prioritization of promising candidates and reducing wasted experimental effort.

1 Introduction

Antibodies are multi-domain proteins whose architecture underlies their diverse functional roles in the immune system as well as their success as a modern drug modality. In natural antibodies, the variable domains (VH and VL) at the tips mediate target-specific binding, while the constant domains provide structural stability and mediate effector functions [24, 28]. Due to their modularity, modern therapeutic antibodies have been designed in a variety of "formats" (Figure 1) that combine these domains in different configurations to achieve specific functional or diagnostic goals [9, 5].

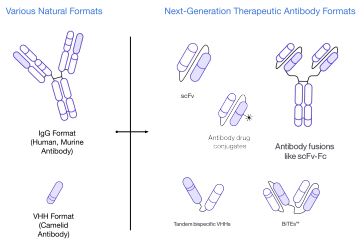


Figure 1: Examples of natural and engineered antibody formats.

"Reformatting" refers to converting an antibody from one format to another to enable a novel drug modality or for high-throughput selection, screening, or expression workflows [22]. In this work, we focus specifically on reformatting a natural multichain IgG antibody into single chain variable fragment (scFv), which can allow the binding properties of

[*]Work done while at BigHat Biosciences

Submitted to AI for Science workshop (NeurIPS 2025).

IgGs to be assayed in high throughput via cell-free expression or phage display, or ease its integration with other functions such as T-cell activation[11, 16, 25].

Yet, while antibodies are conceptually modular, reformatting is far from trivial. Properties frequently shift across formats, and binding may be lost upon conversion. The final molecule may also be prone to instability or aggregation, or may not synthesize at all. As a result, reformatting is often a trial-and-error process, with many potential final formats designed and assayed in the wet lab to find workable designs [26, 23, 3, 27, 14, 2, 12].

One promising strategy to reduce the error rate in reformatting is to use a machine learning (ML) model to predict whether a given IgG/scFv pair is likely to reformat successfully, and then reserving *in vitro* validation for designs that are predicted to perform well. However, to the best of our knowledge, no prior work has systematically modeled antibody reformatting at scale. Previous work on antibody property prediction has ranged from *developability prediction* (focusing on antibody stability, aggregation, or solubility, using physicochemical features [13]) to supervised models on IgG panels [8, 12]. While some models are tailored for specific formats (e.g., scFvs or Fabs), they are not designed to address the unique biophysical shifts introduced by *reformatting*, which can alter stability, aggregation, and even binding [7, 4].

In addition, from a machine learning perspective, a key difficulty in predicting reformatting success is the pronounced distribution shift between antibody "parental families,". Here we define a parental family as the group of reformatted antibodies derived from a single starting antibody in the source format. While antibodies within a parental family are highly similar, those in other families derived from different starting antibodies can diverge substantially in their sequence-property relationships.

Here, our contribution is to benchmark different representations for predicting reformatting success on a real-world dataset derived from multiple therapeutic development campaigns. We develop a multimodal machine learning framework that predicts IgG→scFv reformatting success by combining three types of information: sequence features, structure features, and biophysical features. We evaluate models under three real-world deployment-motivated scenarios: (1) the **scFv split**, in which the model predicts outcomes for unseen individual sequences within known families; (2) the **target-family split**, which simulates fine-tuning on a small set of data from a new family before inference; and (3) the **parental-family split**, a true zero-shot setting with no training data from a new antibody family.

Our experiments reveal two central findings. First, predicting *protein synthesis failure*—the critical gate that determines whether any downstream assays can even be run—is tractable with multimodal features that integrate sequence, structure, and biophysical representations. In the most challenging **parental-family split**, which evaluates generalization to entirely unseen antibody families, our linear multimodal model achieves AUROC values above **88%**. Second, we find that large pretrained protein language models (PLMs) such as AbLang [18], ISM [19], and DPLM2 [29] often underperform simple one-hot sequence baselines. This underscores that domain-specific, interpretable representations remain competitive in low-data structural biology settings with strong distribution shifts.

2 Method

2.1 Problem Formulation and Notation

Let each antibody be represented by a set of input modalities $\mathcal{X} = \{\mathbf{x}_{\text{seq}}, \mathbf{x}_{\text{struct}}, \mathbf{x}_{\text{bio}}\}$, where $\mathbf{x}_{\text{seq}}$ encodes the VH and VL amino acid sequences (over the 20-amino acid alphabet), $\mathbf{x}_{\text{struct}}$ encodes structure-derived descriptors from predicted 3D structures, and $\mathbf{x}_{\text{bio}}$ contains biophysical properties. The target variable $\mathbf{y}$ corresponds either to a protein synthesis success or failure, $\mathbf{y} \in \{0, 1\}$, or a continuous synthesis yield value, $\mathbf{y} \in \mathbb{R}$ (where higher is better). Given a dataset $\mathcal{D} = \{(\mathcal{X}_i, \mathbf{y}_i)\}_{i=1}^{N}$, the objective is to learn a mapping $f_\theta : \mathcal{X} \mapsto \hat{\mathbf{y}}$ that minimizes a task-appropriate loss $\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}})$ and generalizes to unseen antibody families, i.e., correctly predicting synthesis success for novel families not observed during training.

2.2 Dataset Construction

Inputs and targets. Our dataset consists of IgG→scFv reformatting experiments, where full-length immunoglobulin G (IgG) antibodies are converted into compact single-chain variable fragments (scFvs). These experiments were conducted across multiple antibody optimization campaigns. Each

scFv is represented by: (i) VH and VL amino acid sequences, (ii) the linker sequence connecting the domains, (iii) the domain ordering (VH–VL or VL–VH), and (iv) the parental family identifier from which the VH and VL are derived. We train separate models for two primary tasks: (1) *protein synthesis outcome classification*, where the target variable $y_{\mathrm{QC}} \in \{0, 1\}$ indicates whether a reformatted scFv has synthesized adequately or not, and (2) *yield regression*, where $y_{\mathrm{yield}} \in \mathbb{R}$ measures synthesis yield in ng/μL.

scFv signature and aggregation. We define the *scFv signature* of an input scFv as the tuple:

$$\textsc{Sig} = (\mathrm{VH}, \mathrm{VL}, \mathrm{linker}, \mathrm{orientation}),$$

where VH and VL are the amino acid sequences of the heavy and light chain variable regions, linker is the connecting peptide, and orientation specifies the domain order. scFvs sharing the same signature are considered equivalent: their target values are averaged. After aggregation, the dataset contains $N = 1{,}477$ unique scFv signatures drawn from 56 parental families across 7 independent antibody optimization campaigns. Summary statistics are provided in Appendix B.

Parental family generalization. Within a parental family, scFv sequences are often highly similar, differing by only a few mutations. However, across families, divergence is much greater, creating a strong distribution shift between families. This makes generalization to unseen families a central challenge for our models. This motivates our evaluation strategy, discussed next.

2.3 Evaluation Protocol

We evaluate model generalization under three disjoint data partitioning schemes, each corresponding to a realistic deployment scenario (Figure 7): (1) *Parental Family split* (NEW PARENTAL, NO DATA): entire parental families are held out from training, forcing zero-shot prediction on novel families with no prior *in vitro* data. (2) *Target Family split* (ONE BATCH FROM NEW ANTIBODY): a small batch from the target family is included in training along with data from other families; the remaining scFvs from the target family are used for validation and testing. (3) *scFv split* (LOTS OF DATA FOR ALL FAMILIES): all families appear in all splits, but individual scFv signatures are unique to a single split, preventing data leakage while maximizing same-family context.

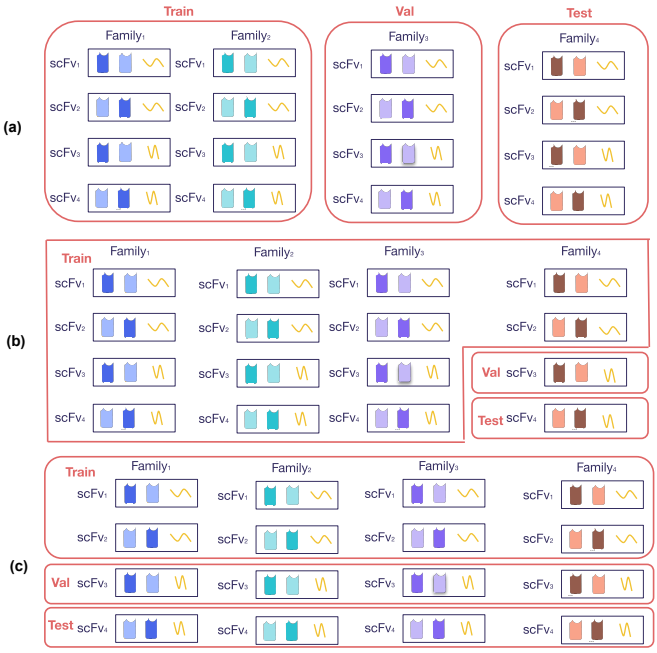


Figure 2: Three data splits illustrated. (a) Parental Family split. (b) Target Family split. (c) scFv split.

2.4 Feature Representations

Our modeling approach integrates three complementary feature modalities: sequence, structure, and biophysical properties. Each modality captures distinct aspects of the scFv–parental IgG relationship, and all features are precomputed and frozen prior to model training.

Sequence-based features. VH and VL amino acid sequences are AHo-aligned[10] to a fixed length ($L_{\mathrm{VL}} = 152$, $L_{\mathrm{VH}} = 152$) and one-hot encoded. Domain orientation (VH–VL or VL–VH) and linker peptide type are represented as categorical one-hot variables and concatenated with the sequence encodings. Given the potential of pretrained sequence models to encode rich information about an input protein, we also evaluate predicting reformatting success using frozen embeddings from two pLMs. First, we consider AbLang [17] (heavy/light encoders separately), which is an

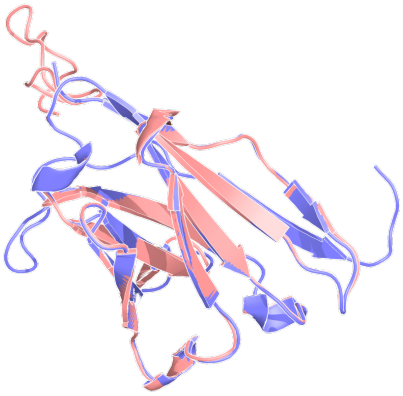


Figure 3: Overlay of predicted VH domain structure between the starting IgG (Purple) and reformatted to scFv (pink). In this example Boltz-2 predicts significant structural alteration upon reformatting.

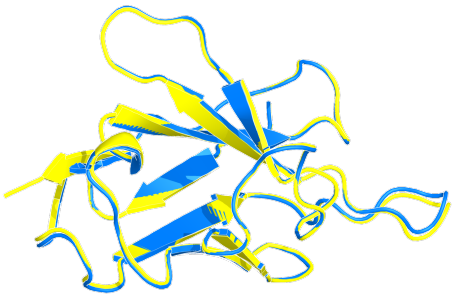


Figure 4: Overlay of predicted VL domain structure between the starting IgG (blue) and reformatted to scFv (yellow).

antibody-specific pLM trained on a large dataset of human antibody sequences. Secondly, we consider ISM [6], a fine-tuned ESM [21] model that is trained to encode structural information in addition to sequence. For both models, residue-level embeddings are computed once and then mean-pooled across the sequence to yield a fixed-length representation. These are kept fixed during downstream training. Our intuition is that such pretrained encoders capture general antibody sequence statistics (e.g., conserved CDR motifs, germline variation) that may aid generalization across families.

Structure-based features. For each unique scFv and its corresponding parental IgG, we predict full-atom 3D structures using Boltz-2 [20], a structure prediction model with accuracy comparable to AlphaFold3 across diverse proteins. Before computing RMSD or coordinate-level features, we rigid-body align the parental IgG and scFv domains to a shared reference frame. We then compute descriptors from these predictions: (i) the global RMSD of Cα atoms between VH/VL domains of the parental IgG and its reformatted scFv, and (ii) per-residue concatenation of aligned parental and scFv Cα coordinates with gap indicators, preserving spatial correspondence.

The motivation is grounded in structural biology, since the function of a protein is intimately related to function, alterations of the structure of individual domains between formats may be related to reformatting success. RMSD serves as a coarse global descriptor (A detailed analysis of RMSD distributions across families is provided in Appendix E.2). The per-residue coordinate features offer a more fine-grained view, potentially capturing subtle local rearrangements that influence reformatting outcomes. Representative overlays of VH and VL domains are shown in Figures 8 and 9.

To benchmark these domain-inspired descriptors against widely used pretrained structural models, we also extract structure-derived embeddings using two models: (i) AbMPNN [6] (an antibody inverse folding model that encodes sequences in the context of 3D backbones) and (ii) DPLM2 [29] (a structure-augmented protein language model). Both provide residue-level embeddings that are mean-pooled to fixed-length vectors for downstream tasks. Our rationale is that inverse folding and structure-augmented pLMs may capture geometric constraints and stability signals that purely sequence-based encoders miss, and thus could help with predicting reformatting success.

Biophysical features. From predicted scFv structures, we compute developability metrics using the NaturalAntibody [15] platform. Key features derived from the CDR regions include Patch Surface Hydrophobicity (PSH), Patch Negative Charge (PNC), Patch Positive Charge (PPC), and scFv Charge Separation Product (SFvCSP). These are metrics which could be expected to be associated with general antibody stability and expressability. If a score cannot be computed due to modeling failure, the dataset mean is imputed.

2.5 Model Architectures

Baseline models. We evaluate linear baselines for both protein synthesis outcome classification and yield regression. For classification, we use logistic regression with either L1 or L2 regularization (the choice of regularization is treated as one hyperparameter during tuning); for regression, we use ordinary least squares with optional regularization. Two input configurations are compared: (i) one-hot AHo-aligned VH and VL sequences concatenated with one-hot domain orientation and linker encodings, and (ii) frozen pLM embeddings. All linear models are implemented in `scikit-learn` and trained with default solvers.

Embedding-based models. We evaluate a fixed-architecture multilayer perceptron (MLP) on frozen pLM embeddings (which outperformed linear models on pLM embeddings). VH and VL embeddings are concatenated prior to the MLP, which contains two hidden layers with ReLU activations and dropout. Hyperparameters such as dropout and learning rate are tuned via grid search, while architecture depth and width remain fixed. Models are trained with AdamW and early stopping.

Multimodal ML models. To assess complementarity between modalities, we train simple linear models using combinations of sequence, structure, and biophysical features (Figure 5). Hyperparameter tuning follows the same protocol as for the embedding-based models.

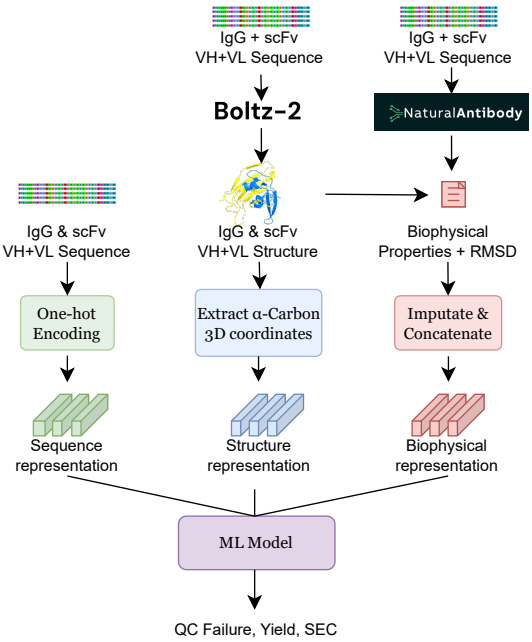


Figure 5: Multimodal ML pipeline.

3 Experiments

3.1 Experiment Setup

Unless otherwise specified, we use an 60%/10%/30% train/validation/test split by the appropriate unit (parental family or scFv signature). For synthesizability classification, we report AUROC and AUPRC in the main text, and accuracy in Appendix. For yield regression, we report Pearson correlation (r), and Spearman rank correlation (ρ) in Appendix. Each split type is repeated over 10 random folds to reduce variance from partitioning; all metrics are reported as mean $\pm$ standard deviation over folds. All experiments were conducted on a Linux server equipped with 4$\times$NVIDIA A10G GPUs (24 GB VRAM each) and 12 CPU cores, using CUDA 12.2 and NVIDIA driver version 535.247.01. Hyperparameter search details can be found in Appendix D.

3.2 Simple sequence-only baselines outperform more complex encodings

First, we investigate whether simple baselines are sufficient to predict reformatting success, and whether pretrained PLM embeddings offer an advantage over these baselines. In Table 1 we compare a logistic regression model, `LogisticReg`, trained on either one-hot encoded protein sequences or 3D structural coordinates (Section C.4), to sequence-based pLMs (`AbLang+MLP`, `ISM+MLP`), a structure-only GNN (`AbMPNN+MLP`), and a structure-augmented PLM (`DPLM2+MLP`). We highlight key findings below.

❶ Simple linear models on one-hot encodings outperform embeddings from pLMs and structure encoders. We find the surprising result that on the `scfv_signature` split, the simple one-hot `LogisticReg` significantly outperforms more complex embeddings: **+3.1** AUROC and **+5.3** AUPRC over AbLang, and nearly **+10** AUROC / **+14** AUPRC over DPLM2. Structure-only AbMPNN trails both. The same holds under `Parental_Family`, where `LogisticReg` remains best (e.g.

Table 1: **A linear model with one-hot sequence encoding outperforms pLMs on predicting reformatting synthesis failure.** We compare a logistic regression baseline using either one-hot sequence features (`vhvl_only`) or per-residue 3D coordinate features (`3D_coord`), to PLM embeddings with an MLP (`AbLang+MLP`, `ISM+MLP`), a structure GNN with an MLP (`AbMPNN+MLP`), a structure-augmented PLM with an MLP (`DPLM2+MLP`). Best results per column are **underlined in bold**.

Model	Features	scfv_signature split		Parental_Family split	
		AUROC	AUPRC	AUROC	AUPRC
AbLang+MLP	`vhvl_only`	86.35±1.39	82.15±2.26	62.58±19.37	58.75±11.70
ISM+MLP	`vhvl_only`	80.01±1.62	74.87±2.65	58.09±9.50	54.02±6.10
DPLM2+MLP	`vhvl+struct`	79.50±1.60	73.49±2.28	47.91±5.29	51.21±11.13
AbMPNN+MLP	`struct_only`	73.38±2.11	65.89±3.89	54.30±5.69	54.67±6.67
LogisticReg	`vhvl_only`	**89.46**±1.63	**87.46**±2.33	**66.35**±10.73	**59.21**±15.65
LogisticReg	`3D_coord`	77.00±1.00	71.00±2.00	52.00±12.00	48.00±8.00

Table 2: **Multimodal features consistently outperform sequence-only, with the largest gains under cross-family generalization.** Protein synthesis failure classification using *sequence+structure+biophysics* features (`multimodal`) vs. sequence-only (`seq_only`) with a linear classifier. Rows correspond to data splits; columns show head-to-head performance. Values are mean±std across runs. Best per split/metric is **underlined in bold**.

Split	AUROC		AUPRC	
	`multimodal`	`seq_only`	`multimodal`	`seq_only`
`scfv_signature`	**92.93**±3.61	89.46±1.63	**91.18**±4.59	87.46±2.33
`Parental_Family`	**88.92**±14.93	66.35±10.73	**85.68**±20.94	59.21±15.65
`Fam1`	**94.64**±2.39	87.68±3.91	**96.92**±1.53	93.23±2.15
`Fam2`	**82.96**±9.51	71.81±7.97	**66.10**±13.57	54.87±9.28
`Fam3`	**93.33**±9.33	82.71±8.87	**97.40**±3.82	92.67±4.66

+3.8 AUROC over AbLang), although performance drops and variance grows substantially. This suggests that the embeddings from these pretrained models do not adequately capture the information required to predict reformatting success. Interestingly, we see that a linear model trained on structure features alone does not outperform any model that also incorporates sequence information, but does outperform AbMPNN, the other pure-structure model. This suggests that the 3D coordinate representation might contain useful predictive signal for this task, but is not sufficient as an input representation.

❷ **Cross-family generalization remains challenging.** All unimodal models degrade substantially under `Parental_Family`, highlighting the severity of distribution shift across antibody families. We see again that pretrained pLMs and structure-based models underperform relative to the simple linear model baseline. Appendix E.2 provides additional per-family analysis, showing that the correlation between global RMSD and protein synthesis yield varies widely across parental families, often flipping sign. This heterogeneity helps explain why structure-only features do not generalize well.

3.3 Multimodal feature encodings improve generalization performance

Given the strong performance of the linear model on one-hot encoded features, and the fact that a linear model on the simple 3D coordinate representation outperformed the more complex AbMPNN encoding, we now evaluate whether combining these two modalities together with biophysical descriptors yields improved models that generalize across families.

Tables 2 and Table 3 present results for linear models trained on `multimodal` features (sequence+structure+biophysics), for the classification and regression tasks respectively. Recall that the *Target Family split* (Section 2.3) mimics a practical setting where only a small batch of measurements is available for a new antibody, and the goal is to extrapolate to the remaining scFvs. This scenario is particularly relevant for antibody engineering pipelines, where limited pilot data are collected before committing to large-scale experiments.

Table 3: **Multimodal features also improve regression, yielding higher Pearson/Spearman correlations in most splits, particularly under cross-family generalization.** Protein yield regression (ng/μL) using *sequence+structure+biophysics* features (`multimodal`) vs. sequence-only (`seq_only`) with a linear regressor. Rows correspond to data splits; columns compare head-to-head results. Values are mean$\pm$std across runs. Best per split/metric is **underlined in bold**.

Split	Pearson		Spearman	
	`multimodal`	`seq_only`	`multimodal`	`seq_only`
`scfv_signature`	**0.641**$\pm$0.031	0.531$\pm$0.044	**0.741**$\pm$0.026	0.714$\pm$0.032
`Parental_Family`	**0.625**$\pm$0.266	0.191$\pm$0.255	**0.508**$\pm$0.264	0.035$\pm$0.283
`Fam1`	**0.567**$\pm$0.035	0.550$\pm$0.061	0.630$\pm$0.039	**0.646**$\pm$0.067
`Fam2`	**0.288**$\pm$0.128	0.141$\pm$0.084	**0.207**$\pm$0.109	0.159$\pm$0.131
`Fam3`	**0.244**$\pm$0.290	0.154$\pm$0.218	0.241$\pm$0.220	**0.243**$\pm$0.202

Our key finding is that **multimodal features enable generalization across parental families, significantly boosting performance under distribution shift.** We see that adding structural and biophysical descriptors to sequence features consistently improves classification. Gains are modest in-distribution (`scfv_signature`: **+3.5** AUROC, **+3.7** AUPRC) but dramatic under cross-family generalization (`Parental_Family`: **+22.6** AUROC, **+26.5** AUPRC). Target families show similar boosts (e.g., Fam2: **+11.2** AUROC, **+11.2** AUPRC).

Importantly, this result demonstrates that: **a key factor in bridging the generalization gap is the use of multimodal features**. Despite the availability of large pretrained encoders, the strongest results come from *simple linear models on well-designed domain-specific features*. Further supporting evidence comes from our ablations (Appendix E.1), which show that the dominant synergy arises between sequence and structure features, while global RMSD alone adds little once explicit structural descriptors are included.

Generalization to SEC purity classification task. Finally, though we focus only on two tasks, synthesis outcomes and yield in the main text, we also evaluated SEC (size-exclusion chromatography) purity prediction as an orthogonal developability assay. As detailed in Appendix E.4, our multimodal models achieve strong performance across splits further underscoring the broad applicability of our feature representations.

4 Conclusion and Discussion

In this work, we studied the problem of antibody reformatting: the problem of converting one antibody format into another to enable its development as a therapeutic. Specifically, we focused on the the case of reformatting IgG antibodies into single chain variable fragments (scFvs), a key step in unlocking high-throughput antibody screening and optimization.

Typically reformatting is a process that involves significant trial-and-error, since a reformatted antibody may synthesize poorly and may have undesirable properties. Here, we proposed a multimodal machine learning framework, combining sequence, structure, and biophysical representations of antibodies, to be able to predict whether a given reformatting attempt is likely to be successful, and thus reducing the failure rate of reformatting. An important component of our framework is the evaluation methodology, where we carefully designed evaluation splits that mirror real-world use cases.

In experiments on real-world antibody reformatting data, the consistent finding is that mulitmodal **sequence**+**structure**+**biophysical** features together result in the most generalizable models. For protein synthesis failure detection, a model trained on these features achieves an AUROC of over **88%** in the most challenging evaluation scenario. Deploying this model could enable prioritization of the most promising candidates during reformatting, resulting in significant cost and time savings.

A key finding from our experiments is that contrary to recent trends, models trained on embeddings from large, pretrained protein language models (PLMs) do not result in the best performance on this task. Instead, simple domain-tailored features paired with lightweight predictors perform the best. This highlights an important takeaway: for specialized biophysical prediction problems, careful

feature design grounded in domain knowledge can outperform large pretrained models, and should remain central in data-limited settings.

Acknowledgements

We would like to thank the BigHat team, especially Ryan Henrici, Emily Delaney, Katrina Stephenson, Duc Huynh, Emily Sever, Anthony Cadena, Noelle Huskey, Taylor Skokan, Nicholas Young, and Lauren Schiff. We also thank the BigHat DS/ML team for productive discussions and insightful suggestions.

References

[1] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.

[2] M. Bailly, C. Mieczkowski, V. Juan, E. Metwally, D. Tomazela, J. Baker, M. Uchida, E. Kofman, F. Raoufi, S. Motlagh, et al. Predicting antibody developability profiles through early stage discovery screening. In *MAbs*, volume 12, page 1743053. Taylor & Francis, 2020.

[3] L. E. Boucher, E. G. Prinslow, M. Feldkamp, F. Yi, R. Nanjunda, S.-J. Wu, T. Liu, E. R. Lacy, S. Jacobs, N. Kozlyuk, et al. "stapling" scfv for multispecific biotherapeutics of superior properties. In *MAbs*, volume 15, page 2195517. Taylor & Francis, 2023.

[4] L. Chen and et al. The influence of antibody fragment format on phage display based affinity maturation. *mAbs*, 2013.

[5] S. Dickopf, G. J. Georges, and U. Brinkmann. Format and geometries matter: structure-based design defines the functionality of bispecific antibodies. *Computational and Structural Biotechnology Journal*, 18:1221–1227, 2020.

[6] F. A. Dreyer, D. Cutting, C. Schneider, H. Kenlay, and C. M. Deane. Inverse folding for antibody sequence design using deep learning. *arXiv preprint arXiv:2310.19513*, 2023.

[7] C. Gottstein and et al. Shelf-life extension of fc-fused single chain fragment variable (scfv-fc) antibodies. *Frontiers in Cellular and Infection Microbiology*, 2021.

[8] M. Hebditch and J. Warwicker. Charge and hydrophobicity are key features in sequence-trained machine learning models for predicting the biophysical properties of clinical-stage antibodies. *PeerJ*, 2019.

[9] P. Holliger and P. J. Hudson. Engineered antibody fragments and the rise of single domains. *Nature Biotechnology*, 23(9):1126–1136, 2005.

[10] A. Honegger and A. PluÈckthun. Yet another numbering scheme for immunoglobulin variable domains: an automatic modeling and analysis tool. *Journal of molecular biology*, 309(3):657–670, 2001.

[11] A. C. Hunt, B. Vögeli, A. O. Hassan, L. Guerrero, W. Kightlinger, D. J. Yoesep, A. Krüger, M. DeWinter, M. S. Diamond, A. S. Karim, and M. C. Jewett. A rapid cell-free expression and screening platform for antibody discovery. *Nature Communications*, 14:3897, 2023.

[12] T. Jain, T. Sun, S. Durand, A. Hall, N. R. Houston, J. H. Nett, B. Sharkey, B. Bobrowicz, I. Caffry, Y. Yu, Y. Cao, H. Lynaugh, M. Brown, H. Baruah, L. T. Gray, E. M. Krauland, Y. Xu, M. Vásquez, and K. D. Wittrup. Biophysical properties of the clinical-stage antibody landscape. *Proceedings of the National Academy of Sciences*, 114(5):944–949, 2017.

[13] T. M. Lauer, N. J. Agrawal, N. Chennamsetty, K. Egodage, B. Helk, and B. L. Trout. Developability index: A rapid in silico tool for the screening of antibody aggregation propensity. *Journal of Pharmaceutical Sciences*, 2012.

[14] Y. Liu, M. Gu, Y. Wu, W. Wang, R. Wang, M. Du, P. Ma, X. Zhou, Y. Wang, Y. Cao, and H. Zhang. High-throughput reformatting of phage-displayed antibody fragments to iggs by one-step emulsion pcr. *Protein Engineering, Design & Selection*, 31(11):427–436, 2018.

[15] Natural Antibody Platform. Natural antibody platform. `https://naturalantibody.com/`. Accessed: 2025-08-25.

[16] T. Ojima-Kato, S. Morishita, Y. Uchida, S. Nagai, T. Kojima, and H. Nakano. Rapid generation of monoclonal antibodies from single b cells by ecobody technology. *Antibodies*, 7(4):38, 2018.

[17] T. H. Olsen and et al. Addressing the antibody germline bias and its effect on language models for improved antibody design. *Bioinformatics*, 2024.

[18] T. H. Olsen, I. H. Moal, and C. M. Deane. Ablang: an antibody language model for completing antibody sequences. *Bioinformatics Advances*, 2(1):vbac046, 2022.

[19] J. Ouyang-Zhang, C. Gong, Y. Zhao, P. Kraehenbuehl, A. Klivans, and D. J. Diaz. Distilling structural representations into protein sequence models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[20] S. Passaro, G. Corso, J. Wohlwend, M. Reveiz, S. Thaler, V. R. Somnath, N. Getz, T. Portnoi, J. Roy, H. Stark, et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, pages 2025–06, 2025.

[21] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. 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. bioRxiv 10.1101/622803.

[22] J. V. Schaefer, A. Honegger, and A. Plückthun. Construction of scfv fragments from hybridoma or spleen cells by pcr assembly. In *Antibody Engineering*, pages 21–44. Springer, 2010.

[23] K.-T. Schneider, T. Kirmann, E. V. Wenzel, J.-H. Grosch, S. Polten, D. Meier, M. Becker, P. Matejtschuk, M. Hust, G. Russo, et al. Shelf-life extension of fc-fused single chain fragment variable antibodies by lyophilization. *Frontiers in Cellular and Infection Microbiology*, 11:717689, 2021.

[24] H. W. Schroeder and L. Cavacini. Structure and function of immunoglobulins. *Journal of Allergy and Clinical Immunology*, 125(2 Suppl 2):S41–S52, 2010.

[25] M. Stech and S. Kubick. Cell-free synthesis meets antibody production: A review. *Antibodies*, 4(1):12–33, 2015.

[26] M. Steinwand, P. Droste, A. Frenzel, M. Hust, S. Dübel, and T. Schirrmann. The influence of antibody fragment format on phage display based affinity maturation of igg. In *MAbs*, volume 6, pages 204–218. Taylor & Francis, 2014.

[27] C. Tu, V. Terraube, A. S. P. Tam, W. Stochaj, B. J. Fennell, L. Lin, M. Stahl, E. R. LaVallie, W. Somers, W. J. J. Finlay, L. Mosyak, J. Bard, and O. Cunningham. A combination of structural and empirical analyses delineates the key contacts mediating stability and affinity increases in an optimized biotherapeutic single-chain fv (scfv). *Journal of Biological Chemistry*, 291(3):1267–1276, 2016.

[28] G. Vidarsson, G. Dekkers, and T. Rispens. Igg subclasses and allotypes: from structure to effector functions. *Frontiers in Immunology*, 5:520, 2014.

[29] X. Wang, Z. Zheng, F. Ye, D. Xue, S. Huang, and Q. Gu. Dplm-2: A multimodal diffusion protein language model. In *The Thirteenth International Conference on Learning Representations*, 2025.

A Summary of Notation

Table 4: Extended notation used throughout the paper.

Symbol	Description
$\mathcal{D}$	Full dataset of N scFv and IgG antibody pairs
$(\mathcal{X}_i, \mathbf{y}_i)$	Input–label pair for the i-th scFv and IgG antibody pairs
$\mathcal{X}$	Set of input modalities for a construct
$\mathbf{x}_{\text{seq}}$	Sequence features from VH and VL domains
S_{VL}, S_{VH}	VL and VH amino acid sequences
L_{VL}, L_{VH}	Sequence lengths of VL and VH
$\mathbf{E}_{\text{PLM}}(\cdot)$	Pretrained protein language model encoder
$\mathbf{h}_{VL}, \mathbf{h}_{VH}$	Sequence embeddings for VL and VH
$\mathbf{x}_{\text{struct}}$	Structure-derived features
$\hat{\mathbf{C}}_{VL}, \hat{\mathbf{C}}_{VH}$	Predicted Cα coordinates for VL and VH
$\text{RMSD}_{VL}, \text{RMSD}_{VH}$	Root mean square deviation per domain
$\mathbf{x}_{\text{bio}}$	Biophysical property features derived from CDRs
PSH	Patch Surface Hydrophobicity
PNC	Patch Negative Charge
PPC	Patch Positive Charge
SFvCSP	scFv Charge Separation Product
$\mathbf{y}_{\text{QC}}$	Binary Protein Synthesis failure label, $\{0, 1\}$
$\mathbf{y}_{\text{yield}}$	Continuous Protein Synthesis yield label
θ	Model parameters
$f_\theta(\cdot)$	Prediction model mapping inputs to $\hat{\mathbf{y}}$
$\hat{\mathbf{y}}$	Predicted label
$\mathcal{L}(\cdot, \cdot)$	Task-appropriate loss function
$\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{val}}, \mathcal{D}_{\text{test}}$	Train/validation/test subsets
SCFV split	Random partition over unique scFv signatures
TARGET-FAMILY split	Hold-out of target family with few-shot fine-tuning
PARENTAL-FAMILY split	Zero-shot hold-out of parental family

B Dataset Statistics

Dataset Overview

Our scFv→IgG Reformatting Dataset comprises 1,477 unique scFv signatures, spanning 52 parental families. Figure 6 shows exploratory data analysis of input features.

Input Features

Sequence length is consistent over scFv→IgG pairs. VH domains average 118.9 ± 4.8 amino acids, VL domains average 108.1 ± 2.2 amino acids, and the combined sequence length averages 227.0 ± 5.6 amino acids. The dataset includes 8 linker types, 2 domain orderings, and covers 52 distinct parental families, providing diversity across scFv construct designs while maintaining tight length distributions.

Target Variables

The primary regression target, yield, has a mean of 31.33 ± 51.44 ng/μL with a wide dynamic range from 3.50 to 513.58 ng/μL. For classification tasks, 40.9% of constructs failed protein synthesis,

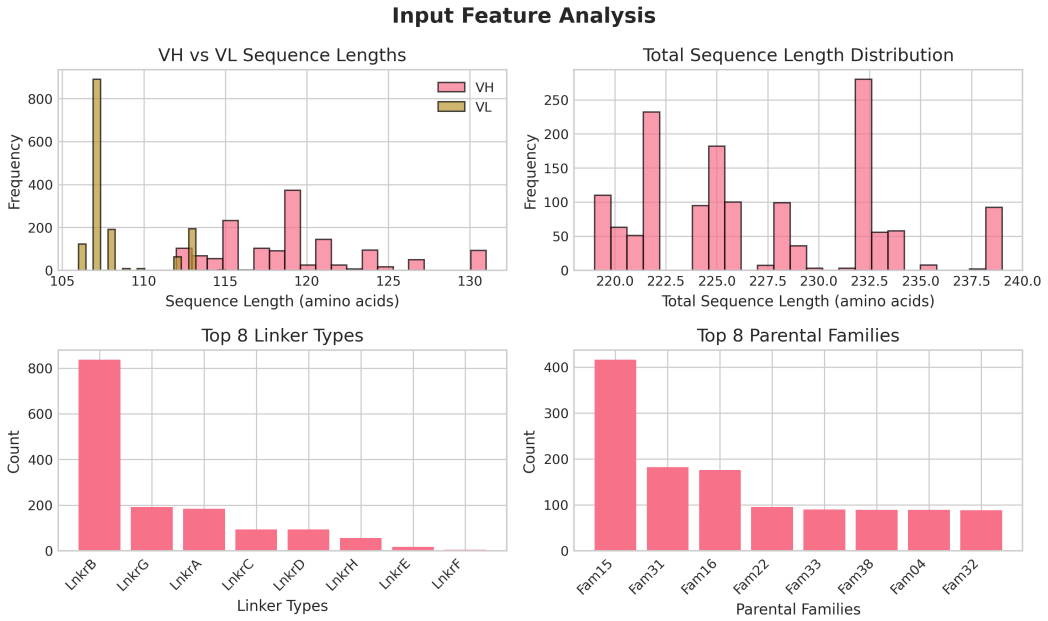


Figure 6: Exploratory data analysis of input features.

ML Considerations

With $n = 1{,}477$ observations, the dataset is limited for modern machine learning and would benefit from additional data for higher-capacity models. The protein synthesis failure outcome exhibits moderate class imbalance (40.9% failure vs. 59.1% pass), which is addressed with appropriate metrics (AUROC and AUPRC). The broad yield range is suitable for regression, while the rich sequence-based inputs (VH and VL) provide strong signal for representation learning and featurization.

C Details of Method Section

C.1 Problem Formulation and Notation

Let each antibody be represented by a set of input modalities $\mathcal{X} = \{\mathbf{x}_{\text{seq}}, \mathbf{x}_{\text{struct}}, \mathbf{x}_{\text{bio}}\}$, where $\mathbf{x}_{\text{seq}}$ encodes the VH and VL amino acid sequences (over the 20-amino acid alphabet), $\mathbf{x}_{\text{struct}}$ encodes structure-derived descriptors from predicted 3D structures, and $\mathbf{x}_{\text{bio}}$ contains biophysical properties. The target variable $\mathbf{y}$ corresponds either to a protein synthesis success or failure, $\mathbf{y} \in \{0, 1\}$, or a continuous synthesis yield value, $\mathbf{y} \in \mathbb{R}$ (where higher is better). Given a dataset $\mathcal{D} = \{(\mathcal{X}_i, \mathbf{y}_i)\}_{i=1}^{N}$, the objective is to learn a mapping $f_\theta : \mathcal{X} \mapsto \hat{\mathbf{y}}$ that minimizes a task-appropriate loss $\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}})$ and generalizes to unseen antibody families, i.e., correctly predicting synthesis success for novel families not observed during training.

C.2 Dataset Construction

Inputs and targets. Our dataset consists of IgG→scFv reformatting experiments, where full-length immunoglobulin G (IgG) antibodies are converted into compact single-chain variable fragments (scFvs). These experiments were conducted across multiple antibody optimization campaigns. Each scFv is represented by: (i) VH and VL amino acid sequences, (ii) the linker sequence connecting the domains, (iii) the domain ordering (VH–VL or VL–VH), and (iv) the parental family identifier from which the VH and VL are derived. We train separate models for two primary tasks: (1) *protein synthesis outcome classification*, where the target variable $y_{\text{QC}} \in \{0, 1\}$ indicates whether a reformatted scFv has synthesized adequately or not, and (2) *yield regression*, where $y_{\text{yield}} \in \mathbb{R}$ measures synthesis yield in ng/μL.

scFv signature and aggregation. We define the *scFv signature* of an input scFv as the tuple:

$$\textsc{Sig} = (\text{VH}, \text{VL}, \text{linker}, \text{orientation}),$$

where VH and VL are the amino acid sequences of the heavy and light chain variable regions, linker is the connecting peptide, and orientation specifies the domain order. scFvs sharing the same signature are considered equivalent: their target values are averaged. After aggregation, the dataset contains $N = 1{,}477$ unique scFv signatures drawn from 56 parental families across 7 independent antibody optimization campaigns. Summary statistics are provided in Appendix B.

Parental family generalization. Within a parental family, scFv sequences are often highly similar, differing by only a few mutations. However, across families, divergence is much greater, creating a strong distribution shift between families. This makes generalization to unseen families a central challenge for our models. This motivates our evaluation strategy, discussed next.

C.3 Evaluation Protocol

We evaluate model generalization under three disjoint data partitioning schemes, each corresponding to a realistic deployment scenario (Figure 7): (1) *Parental Family split* (NEW PARENTAL, NO DATA): entire parental families are held out from training, forcing zero-shot prediction on novel families with no prior *in vitro* data. (2) *Target Family split* (ONE BATCH FROM NEW ANTIBODY): a small batch from the target family is included in training along with data from other families; the remaining scFvs from the target family are used for validation and testing. (3) *scFv split* (LOTS OF DATA FOR ALL FAMILIES): all families appear in all splits, but individual scFv signatures are unique to a single split, preventing data leakage while maximizing same-family context.

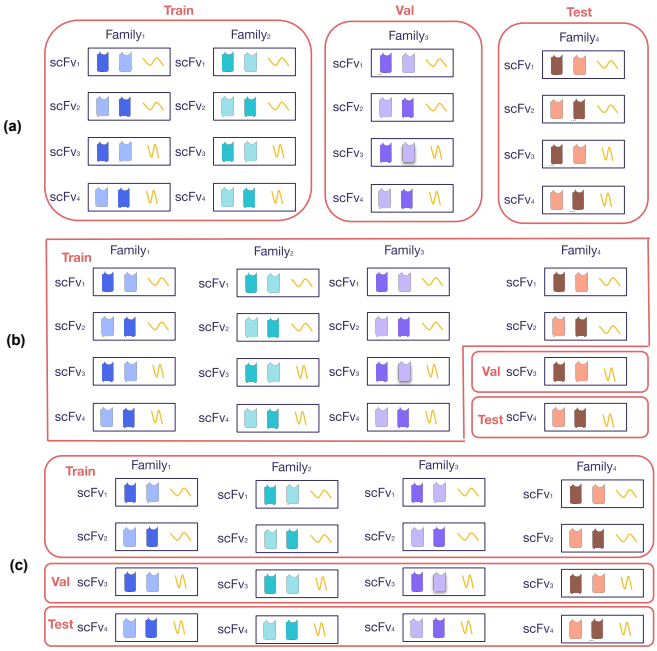


Figure 7: Three data splits illustrated. (a) Parental Family split. (b) Target Family split. (c) scFv split.

C.4 Feature Representations

Our modeling approach integrates three complementary feature modalities: sequence, structure, and biophysical properties. Each modality captures distinct aspects of the scFv–parental IgG relationship, and all features are precomputed and frozen prior to model training.

Sequence-based features. VH and VL amino acid sequences are AHo-aligned[10] to a fixed length ($L_{\text{VL}} = 152$, $L_{\text{VH}} = 152$) and one-hot encoded. Domain orientation (VH–VL or VL–VH) and linker peptide type are represented as categorical one-hot variables and concatenated with the sequence encodings. Given the potential of pretrained sequence models to encode rich information about an input protein, we also evaluate predicting reformatting success using frozen embeddings from two pLMs. First, we consider AbLang [17] (heavy/light encoders separately), which is an antibody-specific pLM trained on a large dataset of human antibody sequences. Secondly, we consider ISM [6], a fine-tuned ESM [21] model that is trained to encode structural information in addition to sequence. For both models, residue-level embeddings are computed once and then mean-pooled across the sequence to yield a fixed-length representation. These are kept fixed during downstream training. Our intuition is that such pretrained encoders capture general antibody sequence statistics (e.g., conserved CDR motifs, germline variation) that may aid generalization across families.

Structure-based features. For each unique scFv and its corresponding parental IgG, we predict full-atom 3D structures using Boltz-2 [20], a structure prediction model with accuracy comparable

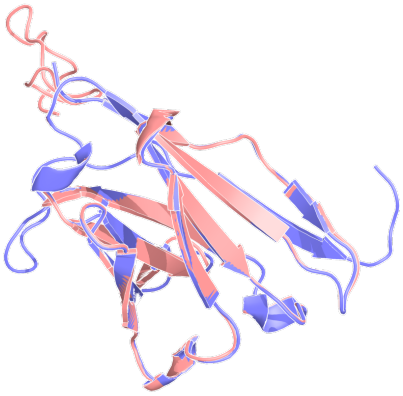


Figure 8: Overlay of predicted VH domain structure between the starting IgG (Purple) and reformatted to scFv (pink). In this example Boltz-2 predicts significant structural alteration upon reformatting.

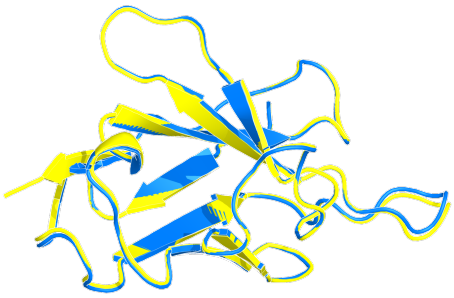


Figure 9: Overlay of predicted VL domain structure between the starting IgG (blue) and reformatted to scFv (yellow).

to AlphaFold3 across diverse proteins. Before computing RMSD or coordinate-level features, we rigid-body align the parental IgG and scFv domains to a shared reference frame. We then compute descriptors from these predictions: (i) the global RMSD of Cα atoms between VH/VL domains of the parental IgG and its reformatted scFv, and (ii) per-residue concatenation of aligned parental and scFv Cα coordinates with gap indicators, preserving spatial correspondence.

The motivation is grounded in structural biology, since the function of a protein is intimately related to function, alterations of the structure of individual domains between formats may be related to reformatting success. RMSD serves as a coarse global descriptor (A detailed analysis of RMSD distributions across families is provided in Appendix E.2). The per-residue coordinate features offer a more fine-grained view, potentially capturing subtle local rearrangements that influence reformatting outcomes. Representative overlays of VH and VL domains are shown in Figures 8 and 9.

To benchmark these domain-inspired descriptors against widely used pretrained structural models, we also extract structure-derived embeddings using two models: (i) AbMPNN [6] (an antibody inverse folding model that encodes sequences in the context of 3D backbones) and (ii) DPLM2 [29] (a structure-augmented protein language model). Both provide residue-level embeddings that are mean-pooled to fixed-length vectors for downstream tasks. Our rationale is that inverse folding and structure-augmented pLMs may capture geometric constraints and stability signals that purely sequence-based encoders miss, and thus could help with predicting reformatting success.

Biophysical features. From predicted scFv structures, we compute developability metrics using the NaturalAntibody [15] platform. Key features derived from the CDR regions include Patch Surface Hydrophobicity (PSH), Patch Negative Charge (PNC), Patch Positive Charge (PPC), and scFv Charge Separation Product (SFvCSP). These are metrics which could be expected to be associated with general antibody stability and expressability. If a score cannot be computed due to modeling failure, the dataset mean is imputed.

C.5 Model Architectures

Baseline models. We evaluate linear baselines for both protein synthesis outcome classification and yield regression. For classification, we use logistic regression with either L1 or L2 regularization (the choice of regularization is treated as one hyperparameter during tuning); for regression, we use ordinary least squares with optional regularization. Two input configurations are compared: (i) one-hot AHo-aligned VH and VL sequences concatenated with one-hot domain orientation and linker encodings, and (ii) frozen pLM embeddings. All linear models are implemented in `scikit-learn` and trained with default solvers.

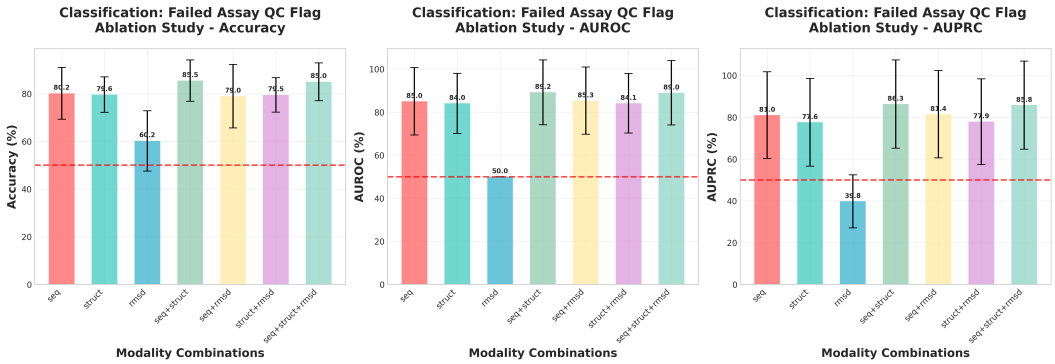


Figure 10: Ablation study on different modality combination on QC classification task.

Embedding-based models. We evaluate a fixed-architecture multilayer perceptron (MLP) on frozen pLM embeddings (which outperformed linear models on pLM embeddings). VH and VL embeddings are concatenated prior to the MLP, which contains two hidden layers with ReLU activations and dropout. Hyperparameters such as dropout and learning rate are tuned via grid search, while architecture depth and width remain fixed. Models are trained with AdamW and early stopping.

Multimodal ML models. To assess complementarity between modalities, we train simple linear models using combinations of sequence, structure, and biophysical features. Hyperparameter tuning follows the same protocol as for the embedding-based models.

D Hyperparameter Details

We summarize the hyperparameters used for all models evaluated in this work. Unless otherwise noted, all models were trained with the AdamW optimizer and early stopping on the validation loss. Hyperparameters were selected via grid or Optuna [1] search on the validation set. Below we report both the search ranges and the final values used in our experiments.

Linear models. *Search space:* regularization strength $C \in \{0.01, 0.1, 1, 10\}$, penalty $\in \{\ell_1, \ell_2\}$. *Final choice:* for linear regression, ℓ_2 penalty with $C = 0.01$; for logistic regression, ℓ_2 penalty with $C = 10$.

Pretrained embeddings + MLP. *Search space:* hidden dimension $\in \{64, 128, 256\}$, dropout $\in \{0.1, 0.2, 0.3\}$, learning rate $\in \{10^{-3}, 10^{-4}\}$, batch size $\in \{32, 64\}$, and a binary flag for using a linear head. *Final choice:* a two-layer MLP with hidden dimension 128 or 256, ReLU activations, dropout 0.2, learning rate 1×10^{-4}, and batch size 32 or 64.

1D CNNs. *Search space (Optuna):* number of dilated convolutional layers $[1, 5]$, expansion factor $[1.0, 4.0]$, representation dimension $\{16, 32, 64, 128\}$, batch normalization $\in \{\texttt{true}, \texttt{false}\}$, learning rate $[10^{-4}, 10^{-2}]$ (log-uniform), batch size $\{16, 32, 64\}$, and training epochs $[10, 50]$. *Final choice:* for classification, 5 dilated conv layers, representation dimension 32, expansion factor 1.4, no batch normalization, learning rate 3.8×10^{-4}, batch size 32, and training for 13 epochs. For regression, 1 dilated conv layer, representation dimension 16, expansion factor 2.7, no batch normalization, learning rate 1.8×10^{-4}, batch size 32, and training for 10 epochs.

E Ablation Studies and In-depth Analysis

E.1 Ablation studies of different modality combinations

We ablate seven modality configurations on the *Failed Assay QC* classifier—`seq`, `struct`, `rmsd`, all pairwise combinations, and `seq+struct+rmsd`. Figure 10 summarizes Accuracy, AUROC, and AUPRC (mean with standard-deviation bars; red dashed line denotes the random baseline).

Across all metrics, `seq+struct` is uniformly strongest, reaching **85.5 Acc**, **89.2 AUROC**, and **86.3 AUPRC**. This reflects clear complementarity: relative to the best single-modality baselines, `seq+struct` gains $+5.3/+4.2/+5.3$ points over `seq` and $+5.9/+5.2/+8.7$ over `struct` on Accuracy/AUROC/AUPRC, respectively. In contrast, `rmsd` alone carries little predictive signal (AUROC ≈ 50.0, AUPRC 39.8), and adding it to either `seq` or `struct` yields at most marginal changes (`seq+rmsd`: 85.3 AUROC, 81.4 AUPRC; `struct+rmsd`: 84.1 AUROC, 77.9 AUPRC). Notably, the full tri-modal model `seq+struct+rmsd` does not improve over `seq+struct` (89.0 vs. 89.2 AUROC; 85.8 vs. 86.3 AUPRC), suggesting that global RMSD largely overlaps with information already captured by explicit structural descriptors and can introduce redundant or noisy signal.

Takeaway. The dominant synergy is between sequence and structure: residue-level biochemical constraints from `seq` and geometric compatibility from `struct` combine to drive the best generalization, while RMSD—being a coarse, global deviation measure—adds little once structural features are explicitly modeled.

E.2 Different structure–feature distributions across parental families

To probe why RMSD features offer inconsistent gains, we quantify the within–parental-family association between structural deviation and protein yield. For each family, we compute the Pearson correlation between yield and (i) VH RMSD, (ii) VL RMSD, and (iii) their sum. Figure 11 reveals pronounced heterogeneity: several families exhibit a *positive* RMSD–yield relationship (larger deviations correlate with higher yield), whereas others show the opposite trend, and many lie near zero. These sign flips persist across VH, VL, and combined RMSD, and are not explained by sample size alone (family-wise n is indicated atop each bar).

This family-specific behavior implies non-stationarity in the mapping from global structural deviation to expression outcome. A model trained on pooled data with only global structure descriptors faces a conflicting supervision signal and will tend to regress toward a weak average effect, limiting the utility of RMSD as a standalone predictor and explaining its marginal contribution once richer structural descriptors are already present. The absence of explicit parental-family information further prevents the learner from capturing divergent RMSD–yield regimes.

Implication. Incorporating family context—e.g., parental-family identifiers, sequence backbones, or a conditional/mixture-of-experts gate—should allow the model to adapt its structural-to-yield mapping across families. This aligns with our ablation in Section E.1: the strongest performance arises when sequence (a proxy for family identity and local biophysics) is fused with structure, while global RMSD alone is insufficient.

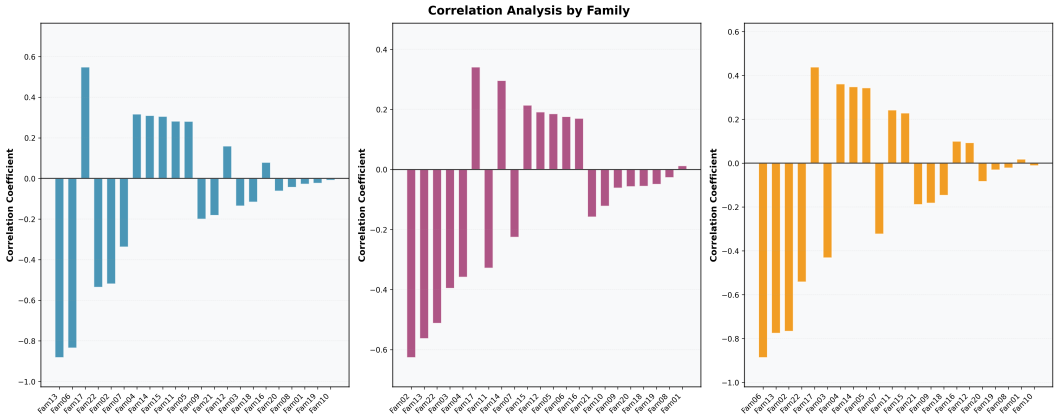
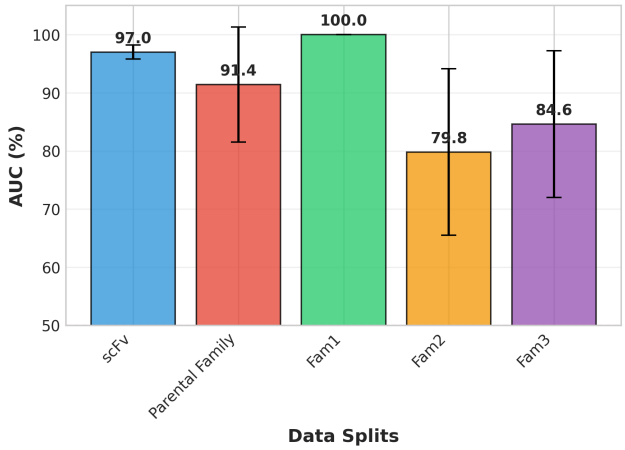


Figure 11: **Family-specific RMSD–yield relationships.** Pearson correlations between yield and VH RMSD (left), VL RMSD (middle), and VH+VL RMSD (right), computed *within* each parental family (families sorted by correlation magnitude). The wide dispersion and sign changes indicate heterogeneous, family-dependent structure–yield trends.

Figure 12: Linear model performance on multimodal feature input for SEC purity ($y = \mathbb{1}[\%$ area under the main peak 280nm $\geq 90\%]$).



E.3 Case study of the *Fam1* parental family

We analyze the model's behavior on a single parental family (*Fam1*) to understand operational impact on protein synthesis failure screening. On the held-out test set of 55 variants, the classifier attains **Recall** $= 100\%$ (no good scFv missed), **Precision** $= 87.2\%$ (five bad scFvs flagged as "pass"), and **Accuracy** $= 90.9\%$. In other words, the model screens out almost all low-quality candidates while retaining every high-quality one, eliminating wasted experiments from false negatives and confining errors to a small number of false positives.

From an operational perspective, we evaluate an actionable *screen-then-confirm* policy that only advances candidates predicted to successful protein synthesis. Under this policy, the fraction of executed experiments that yield a true pass equals the classifier's positive predictive value,

$$\text{Efficiency} \;=\; \frac{\text{TP}}{\text{TP}+\text{FP}} \;=\; \text{Precision}.$$

Relative to a trial-and-error baseline efficiency of **61.8%**, our multimodal model would raise efficiency to **87.2%**, an **absolute gain of** $+25.4$ **points** and a $1.4\times$ multiplicative improvement. Practically, this means more assays are spent confirming genuinely promising variants rather than testing poor candidates.

Takeaway. Within this family, the model provides a high-recall triage mechanism—*no good scFv is missed*—while substantially improving experimental efficiency. We note, however, that this is a single-family case study and results can vary with stochastic training and family-specific covariate shift; broader prospective evaluation and calibration across families remain important.

E.4 SEC classification

We formulate SEC (size-exclusion chromatography) purity prediction as a binary classification task with label $y = \mathbf{1}[\%$ area under main peak 280 nm $\geq 90\%]$ and evaluate generalization across several realistic data partitions that vary in sequence and family composition. The model attains strong and stable performance across splits, with mean AUCs of **97.0** on the `scFv` split, **91.4** on the `Parental Family` split, and a perfect **100.0** on `Fam1`. Performance is lower but still competitive for more distributionally shifted families, achieving **79.8** on `Fam2` and **84.6** on `Fam_102`. Error bars (standard deviation across runs) are narrow on `scFv` and `Fam1`, but widen on family-held-out splits, indicating increased variance when the model is asked to extrapolate to unseen parental backbones. Taken together, these results suggest that the learned representation captures robust determinants of SEC purity and transfers well, with degradation primarily attributable to family-specific covariate shift.

Table 5: Classification (protein synthesis failure) with sequence and/or structural features. We compare PLM embeddings with an MLP (`AbLang+MLP`, `ISM+MLP`), a structure GNN with an MLP (`AbMPNN+MLP`), a structure-augmented PLM with an MLP (`DPLM2+MLP`), and a one-hot logistic regression baseline (`LogisticReg`). Best results per column are **underlined in bold**.

Model (Acc. in %)	Features	scfv_signature split	Parental_Family split
AbLang+MLP	`vhvl_only`	80.00±1.83	59.64±16.89
ISM+MLP	`vhvl_only`	71.42±1.85	56.64±12.07
DPLM2+MLP	`vhvl+struct`	72.84±2.06	48.67±9.02
AbMPNN+MLP	`struct_only`	60.47±1.23	49.85±7.98
LogisticReg	`vhvl_only`	**83.58**±1.69	**60.98**±12.53

Table 6: Protein synthesis failure classification using per-residue 3D coordinate features (`3D_coord`). Rows correspond to data splits; columns compare `LogisticReg` vs. `1DCNN` head-to-head. Values are mean±std across runs. Best per split is **underlined in bold**.

Split	Accuracy		AUROC		AUPRC	
	LogisticReg	1DCNN	LogisticReg	1DCNN	LogisticReg	1DCNN
`scfv_signature`	72.00±2.00	**76.00**±2.00	77.00±1.00	**82.00**±2.00	71.00±2.00	**79.00**±2.00
`Parental_Family`	52.00±9.00	**54.00**±12.00	52.00±12.00	**57.00**±13.00	48.00±8.00	**52.00**±13.00
`Fam1`	**56.00**±5.00	46.00±13.00	**58.00**±6.00	49.00±22.00	**70.00**±5.00	64.00±15.00
`Fam2`	70.00±3.00	**74.00**±0.00	**59.00**±8.00	49.00±7.00	**39.00**±6.00	28.00±2.00
`Fam3`	**79.00**±5.00	38.00±8.00	**79.00**±8.00	73.00±10.00	**90.00**±5.00	89.00±4.00

Table 7: Protein synthesis failure classification using *sequence+structure+biophysics* features (`multimodal`) vs. sequence-only (`seq_only`) with a linear classifier. Rows correspond to data splits; values are mean±std across runs. Best per split is **underlined in bold**.

Split (Acc. in %)	`multimodal`	`seq_only`
`scfv_signature`	**85.24**±4.33	83.58±1.69
`Parental_Family`	**85.67**±7.60	60.98±12.53
`Fam1`	**86.55**±4.61	80.18±4.91
`Fam2`	**77.78**±4.76	75.19±3.43
`Fam3`	**84.81**±10.79	79.26±3.78

F Extended Experiment Results for Protein Synthesis Failure Classification

We present the accuracy metrics for protein synthesis failure classification in the following tables. Table 5 shows the accuracy for linear model compare to PLM embeddings with sequence features only. Table 6 shows the accuracy for structure input with Logistic Regression and 1D-CNN. Table 7 shows the accuracy for multimodal input.

G Extended Experiment Results for Protein Synthesis Yield Regression

We present the Pearson correlation and Spearman correlation metrics for protein synthesis yield regression in the following tables. Table 8 shows the Pearson correlation and Spearman correlation for linear model compare to PLM embeddings with sequence features only. Table 9 shows the Pearson correlation and Spearman correlation for structure input with Logistic Regression and 1D-CNN.

Table 8: Regression (protein yield) with sequence and/or structural features. We report mean±std across runs. Best results per column are **underlined in bold**.

Model	Features	scfv_signature split		Parental_Family split	
		Pearson	Spearman	Pearson	Spearman
AbLang+MLP	vhvl_only	**0.562**±0.018	0.610±0.012	0.014±0.161	0.052±0.331
ISM+MLP	vhvl_only	0.173±0.019	0.368±0.037	-0.110±0.142	-0.104±0.178
DPLM2+MLP	vhvl+struct	0.464±0.037	0.514±0.027	0.125±0.111	**0.132**±0.184
AbMPNN+MLP	struct_only	-0.092±0.055	0.001±0.075	0.120±0.103	0.002±0.228
LinearReg	vhvl_only	0.531±0.044	**0.714**±0.032	**0.191**±0.255	0.035±0.283

Table 9: Protein yield regression using per-residue 3D coordinate features (3D_coord). Rows correspond to data splits; columns compare LinearReg vs. 1DCNN head-to-head for Pearson and Spearman correlations. Values are mean±std across runs. Best per split/metric is **underlined in bold**.

Split	Pearson		Spearman	
	LinearReg	1DCNN	LinearReg	1DCNN
scfv_signature	0.495±0.034	**0.529**±0.034	0.512±0.037	**0.546**±0.039
Parental_Family	0.007±0.167	**0.027**±0.133	0.002±0.114	**0.088**±0.157
Fam1	**0.138**±0.106	0.070±0.252	**0.221**±0.134	0.176±0.417
Fam2	0.006±0.038	**0.020**±0.075	-0.001±0.035	**0.015**±0.071
Fam3	**0.187**±0.290	0.114±0.243	**0.022**±0.246	-0.002±0.203