SAMPLING CDR-H3 ENSEMBLES WITH ANTIBODY LANGUAGE MODELS AND DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Antibodies derive their binding specificity primarily from the third complementarity-determining region of the heavy chain (CDR-H3), a hypervariable loop characterized by extreme sequence diversity and conformational flexibility. Despite its functional importance, accurate structural prediction of CDR-H3 remains an open problem. Here we present H3Fold, a fast generative framework for CDR-H3 structure prediction that couples antibody-specific language models with a denoising diffusion model. Sequence context is provided by two pre-trained encoders: H3BERTa, trained on over 18M CDR-H3 sequences, and Ab-RoBERTa, trained on over 57M full-length variable regions. To address the scarcity of high-resolution antibody structures, we train the diffusion trunk under a progressive curriculum that transitions from general protein backbone loops to CDR-H3 loops. Systematic ablations isolate the contribution of each architectural choice to structural quality. On benchmark splits filtered for strict sequence identity to the training set, H3Fold generates geometrically plausible low-RMSD conformations, competitive with state-of-the-art models, at a fraction of the inference cost of existing predictors. Since the diffusion trunk samples structural ensembles, it can integrate directly into iterative antibody design pipelines where speed and conformational diversity are decisive.

1 INTRODUCTION

Antibodies are the main effectors of humoral immunity and have emerged as a leading class of therapeutic molecules for cancer, autoimmune diseases and infectious diseases (Wilkinson & Hale, 2022). Their binding specificity is encoded in three hypervariable loops on the heavy chain and three on the light chain, the complementarity-determining regions (CDRs) (Janeway et al., 2001). While five of the six CDR loops adopt canonical conformations predictable from sequence homology (Chothia & Lesk, 1987; North et al., 2011), CDR-H3 lacks a structural classification. V(D)J recombination generates extreme variation in loop length and sequence diversity that defeats the homology-based classification applicable to the other five loops (Regep et al., 2017). As a result, CDR-H3 remains a large source of structural error in computational antibody modeling. (Chen et al., 2024; Valdés-Tresanco et al., 2023).

Early approaches modeled CDR-H3 in isolation using fragment-matching databases or physics-based loop sampling anchored at the stem (Jacobson et al., 2004; Choi & Deane, 2010; Weitzner et al., 2017), but failed to generalize to long or novel loop sequences (Regep et al., 2017). Dedicated deep learning methods improved upon classical approaches yet remained limited by the small number of experimentally resolved structures (Dunbar et al., 2014). The development of end-to-end structure prediction shifted the broader paradigm by predicting complete V_H/V_L structures from sequence, subsuming CDR-H3 within a global folding problem (Jumper et al., 2021; Ruffolo et al., 2023; Abanades et al., 2023); CDR-H3 nonetheless remains the largest source of error across all full-structure models, and any modification to the loop sequence requires complete re-prediction of the entire antibody structure (Kenlay et al., 2024a).

In this work, we present H3Fold, a denoising diffusion probabilistic model for CDR-H3 loop structure prediction conditioned on loop sequence and V_H framework context, trained through a four-stage progressive curriculum and evaluated against a systematic ablation study and a temporal holdout. Our main contributions are:

- A diffusion model for CDR-H3 loop structure prediction conditioned on loop sequence and V_H framework context via two complementary protein language models (H3BERTa and Ab-RoBERTa), running at $0.70\,\mathrm{s}$ per Fv (Table 11).
- A four-stage progressive curriculum bootstrapping from over 180,000 general protein loop structures to overcome CDR-H3 data scarcity.
- Systematic ablations establishing that loop pre-training and V_H conditioning are each necessary for generalization to novel sequences, with V_H conditioning contributing most for longer ($L \geq 16$) loops.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2 Related Work

2.1 Antibody Structure Prediction

Early computational approaches to antibody structure prediction relied on homology modeling of the framework regions combined with rule-based or energy-guided sampling of the CDR loops (Weitzner et al., 2017; Marcatili et al., 2008). These methods were accurate for frameworks but degraded sharply on CDR-H3 due to its structural diversity.

End-to-end deep learning architectures substantially improved over homology modeling by predicting full V_H/V_L structure from sequence (Jumper et al., 2021; Ruffolo et al., 2022; 2023; Abanades et al., 2023; Abramson et al., 2024; Dreyer et al., 2025). Replacing MSA-based co-evolutionary features with antibody-specific language models was particularly consequential for the CDR loops, where MSA coverage is sparse and co-evolutionary signal weak, motivating antibody-specific architectures trained directly on immune repertoire data (Ruffolo et al., 2023; Abanades et al., 2023).

Despite these advances, CDR-H3 remains consistently the most challenging loop in independent benchmarks (Chen et al., 2024).

2.2 CDR Loop Modeling

Dedicated CDR loop modeling methods isolate loop generation from the global folding problem, enabling faster inference and explicit focus on loop conformation.

Ab initio conformational sampling was first shown to be tractable at all-atom resolution through a hierarchical loop-closure algorithm (Jacobson et al., 2004). Clustering loops into discrete structural clusters (North et al., 2011) is put to direct use by database retrieval methods such as FREAD (Choi & Deane, 2010) and by fragment-threading methods such as DaReUS-Loop (Karami et al., 2018). Complementarily, fragment-based samplers such as RosettaAntibody enumerate backbone conformations through fragment insertion combined with knowledge-based energy minimization (Weitzner et al., 2017), a pipeline subsequently repurposed for design (Adolf-Bryfogle et al., 2018). All of these approaches depend on proximity to known structural or fragment space and consequently degrade on long or sequentially novel CDR-H3 loops.

Removing this dependence motivated a shift toward learned, loop-centric generalization. ABlooper learns CDR-H3 conformation directly through an equivariant graph neural network ensemble conditioned on the antibody stem geometry (Abanades et al., 2022). This formulation was extended to jointly resolve side-chain conformations alongside the backbone (Akpinaroglu et al., 2022) and to end-to-end design and docking by dyMEAN (Kong et al., 2023).

Due to CDR-H3's vast conformational space, attention has shifted away from deterministic models.

2.3 Diffusion Models for Protein Structure

Denoising diffusion probabilistic models have emerged as a dominant paradigm for protein structure generation. RFdiffusion generates novel protein backbones by diffusing over rigid-body frames and denoising with a fine-tuned RoseTTAFold network (Watson et al., 2023). Frame-based diffusion over $SE(3)$ (Yim et al., 2023b;a), coordinate-diffusion with arbitrary conditioning (Ingraham et al., 2023), and dihedral-angle diffusion (Wu et al., 2024b) have each been explored as alternatives; Proteina subsequently scales flow-matching backbone generation to millions of general protein structures (Geffner et al., 2025).

Diffusion-based generative models bring explicit probabilistic sampling to the antibody loop design problem. DiffAb jointly diffuses CDR sequence and structure conditioned on the antigen epitope (Luo et al., 2022). AbDiffuser extends antibody-specific diffusion to full-atom, physics-informed joint sequence–structure generation (Martinkus et al., 2023). Generalizing beyond antibody training data, a diffusion model pretrained on general protein backbones (Watson et al., 2023) can be fine-tuned to generate CDR loops conditioned on framework context (Bennett et al., 2026). These models are limited by the scarcity of antibody data (Wu & Li, 2023); protein language models allow advancements by emphasizing sequence data (Ruffolo et al., 2022).

2.4 Protein Language Models for Antibodies

Protein language models pre-trained on large sequence corpora learn transferable representations of protein sequence, structure, and function (Rives et al., 2021; Lin et al., 2023). General-purpose models such as ESM-2 capture evolu-

tionary co-variation across homologous protein families (Lin et al., 2023), but fall short on CDR-H3 loops due to the diversity introduced by V(D)J recombination.

Antibody-specific language models address this by pre-training on immune repertoire data. AntiBERTy learns representations that correlate with binding affinity and maturation (Ruffolo et al., 2021). AbLang pre-trains separate masked language models on heavy and light chain sequences, performing strongly on antibody engineering tasks (Olsen et al., 2022). IgLM conditions language model generation on chain type and species (Shuai et al., 2023). Paired heavy/light training data has also been used directly. IgBERT and IgT5 pre-train on paired OAS sequences, learning cross-chain representations that single-chain models cannot access (Kenlay et al., 2024b). Ab-RoBERTa improves upon IgBERT by utilizing a RoBERTa-based architecture with dynamic masking and a smaller, more efficient parameter size (Huh & Lee, 2025). Beyond sequence-only pretraining, AntiFold conditions an inverse-folding model on antibody structure to score and generate CDR sequences consistent with a given backbone (Høie et al., 2025).

Although these models train on complete antibody sequences, framework regions dominate the training signal, since they are longer and more conserved. H3BERTa addresses this gap by pre-training exclusively on CDR-H3 sequences from over 18 million immune repertoire records (Rodella & Lemmin, 2026).

3 METHOD

3.1 DATASET AND CANONICAL FRAME

We source all structural data from the ALL-conformations dataset (Spoendlin et al., 2025), which aggregates crystallographic protein loop structures from the PDB including antibody CDR-H3, CDR-L3, and TCR CDRβ3 loops. Stage 0 pre-training uses 181,013 general protein loops (178,219 PDB entries, 1,771 antibody entries, and 1,023 TCR CDRβ3; 29,415 unique sequences). For Stages 1–3, we focus exclusively on 1,919 unique sequences (yielding 4,648 entries in Stage 1/2 and 4784 entries after Stage 3 re-extraction), partitioned by 60% sequence identity clustering into training (4,648 entries), validation (461 entries), and test (430 entries) splits with 23% singletons (breakdown in Appendix B Table 6 & Figure 1). A harder evaluation subset retains only the 58 test entries with $<$40% identity to any training sequence. For evaluation, validation and test sets are deduplicated to one structure per unique CDR-H3 sequence, yielding 192 and 162 unique sequences respectively (28 for the hard test subset). All coordinates are re-extracted from the original mmCIF files and expressed in a structurally grounded canonical frame defined by the conserved anchor residues flanking the CDR-H3 loop (Cys at IMGT position 104 and Trp at position 118).

This choice eliminates five of the six rigid-body degrees of freedom shared across all loops. The remaining degree of freedom, roll about the $+x$-axis, is resolved by Gram-Schmidt orthogonalization against $\mathbf{z} = (0, 0, 1)^\top$ in Stages 0–2, where no structurally grounded roll reference is available.

In Stage 3, we replace this convention with a *grounded* frame whose roll is fixed by the backbone-plane normal derived from the anchor residue peptide planes, following Dauparas et al. (2022) (derivation in Appendix C). This grounded frame reduces the same-sequence pair RMSD from $11.7\,\text{Å}$ to $0.6\,\text{Å}$ and enables absolute loop placement without superposition at inference, making the canonical frame a tractable alternative to SE(3)-equivariant architectures for this geometrically constrained problem.

3.2 MODEL ARCHITECTURE

We parameterize the denoising network ϵ_θ as a Transformer with $L = 6$ layers, hidden dimension $d = 256$, 8 attention heads, and feed-forward dimension $1{,}024$ ($\approx$10M parameters in the denoising network; total trainable parameters including LoRA adapters are reported in Figure 2, Appendix D). Each residue's noisy Cα coordinate $\mathbf{x}_{t,i} \in \mathbb{R}^3$ is linearly projected to $\mathbb{R}^d$ and summed with a learned positional embedding. The diffusion timestep t is encoded via sinusoidal frequencies through a two-layer MLP ($d_t = 128$) and injected into every transformer layer via Adaptive Layer Normalization (Peebles & Xie, 2023) (AdaLN) with zero-initialized affine modulation, ensuring identity at initialization.

Each transformer block contains four AdaLN-preconditioned sub-layers: (i) masked self-attention over loop residues; (ii) H3BERTa (Rodella & Lemmin, 2026) cross-attention, providing per-token CDR-H3 sequence context in all stages; (iii) Ab-RoBERTa (Huh & Lee, 2025) cross-attention, providing V_H framework context, gated by $\lambda_\text{Ab} \in \{0, 1\}$; and (iv) a position-wise FFN. PLM hidden states (768-dim) are mapped to d by a two-layer projection (Linear 768$\to$256, GELU, Linear 256$\to$256) before cross-attention. A final linear head predicts the noise $\hat{\boldsymbol{\epsilon}}_i \in \mathbb{R}^3$ per residue.

H3BERTa is active in all stages, but its adaptation alternates: frozen embeddings at Stages 0 and 2, and LoRA adapters (Hu et al., 2022) ($r = 8$, $\alpha = 16$) trained at Stages 1 and 3. Ab-RoBERTa is gated off ($\lambda_\text{Ab} = 0$) in

Stages 0–1 and activated with LoRA adapters from Stage 2 onward. In both cases, base PLM weights remain frozen throughout training.

3.3 Diffusion Process

We adopt a discrete-time DDPM (Ho et al., 2020) with $T = 1{,}000$ steps and a cosine noise schedule (Nichol & Dhariwal, 2021) ($s = 0.008$). The forward process corrupts Cα coordinates as $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. At inference we use DDIM (Song et al., 2021) with 50 steps, generating 5 candidate structures per sequence in approximately 3.5 s on an A40 GPU (Table 11 Appendix H).

3.4 Training Objective

All coordinates are normalized to zero mean and unit variance (per-dataset statistics μ_α, σ_α) before loss computation. The training loss combines noise-prediction with two geometric auxiliary terms:

$$\mathcal{L} = \mathbb{E}_{t,\boldsymbol{\epsilon}}\Big[\sum_i w_i \|\hat{\boldsymbol{\epsilon}}_i - \boldsymbol{\epsilon}_i\|^2 + 0.5\,\lambda(t)\,\mathcal{L}_{\text{anc}} + 0.05 \cdot \mathbf{1}[t < 750]\,\mathcal{L}_{\text{bond}}\Big], \tag{1}$$

where $w_i = 5$ for anchor residues and $w_i = 1$ otherwise, and

$$\lambda(t) = e^{-t/300} \cdot \mathbf{1}[t < 750]. \tag{2}$$

Anchor loss $\mathcal{L}_{\text{anc}}$. Jointly penalizes both stem residues in normalized space: the N-terminal anchor (residue 0) is penalized toward the canonical frame origin across all three coordinates, and the C-terminal anchor (residue $L-1$) is penalized toward the positive x-axis by driving its y and z coordinates toward zero. The timestep weighting $\lambda(t)$ concentrates this supervision on low-noise timesteps ($t \ll 750$), where the model's predictions approach the final structure, and suppresses it entirely at high noise ($t \geq 750$), where enforcing endpoint geometry carries no meaningful signal.

Bond-length loss $\mathcal{L}_{\text{bond}}$. penalizes C$\alpha$–C$\alpha$ distances outside $[3.5, 4.2]$ Å. The indicator $\mathbf{1}[t < 750]$ restricts this term to the low-to-mid noise regime, where predicted coordinates are sufficiently close to the data manifold for local geometry regularization to be effective.

3.5 Training Curriculum

Training proceeds in four sequential stages, each initializing from the previous checkpoint (Table 1).

Table 1: Training curriculum. *Frozen*: PLM weights active, no LoRA adaptation. *LoRA*: LoRA adapters active ($r = 8$, $\alpha = 16$), base weights frozen. *Off*: gated out ($\lambda_{\text{Ab}} = 0$); no embeddings consumed.

Stage	Data	H3BERTa	Ab-RoBERTa
0	PDB loops + CDR-L3 + TCR CDR3 (∼181k)	Frozen	Off
1	CDR-H3 + CDR-L3 + TCR CDR3	LoRA	Frozen
2	CDR-H3 only (4,648)	Frozen	LoRA
3	CDR-H3, grounded frame	LoRA	LoRA

Stage 0 trains the denoising network on diverse loop geometry using frozen H3BERTa embeddings, establishing general stem-constrained structural priors before any CDR-H3-specific supervision. **Stage 1** activates H3BERTa LoRA adapters and utilizes a frozen Ab-RoBERTa on a mixed dataset of CDR-H3, CDR-L3, and TCR CDR3 loops, progressively specializing the sequence representation toward immune loops. **Stage 2** freezes H3BERTa and activates Ab-RoBERTa LoRA adapters to fine-tune exclusively on CDR-H3 structures, isolating the incorporation of V_{H} framework context. **Stage 3** re-extracts CDR-H3 coordinates in the grounded canonical frame and fine-tunes from the Stage 2 checkpoint, enabling absolute loop placement at inference.

4 Experiments

4.1 Ablation Study

We systematically evaluate the contribution of each design choice by training a suite of ablated models that each modify a single component of the curriculum, architecture, or loss function, and advancing each variant through all

four stages to obtain a Stage 3 checkpoint. All main results are reported as median Kabsch Cα RMSD on the 192-sequence validation set (Table 2). Full ablation results are reported in Appendix E.3 Table 7. The robustness of the primary effects is confirmed by a joint ordinary-least-squares regression of Kabsch RMSD on all ablation-axis flags simultaneously: Stage 0 pretraining, H3BERTa, Ab-RoBERTa conditioning, and immune loop data each remain significant after mutual adjustment (Appendix E, Figure 3).

Dual PLM architecture. Our central hypothesis is that CDR-H3 structure is co-determined by two distinct sequence signals: the loop's own residue identity, encoded by H3BERTa, and the surrounding V_{H} framework, encoded by Ab-RoBERTa. Each signal should be necessary, and their relative importance should scale with loop length, since longer CDR-H3 loops are more loosely constrained by stem geometry alone.

Removing Ab-RoBERTa entirely by skipping Stage 2 produces the largest degradation in our pLM ablations ($+0.56\mathring{A}$), confirming that framework context is the dominant conditioning signal. To assess whether Ab-RoBERTa can be replaced by a lighter-weight alternative, we substitute it with HeavyBERTa (Brönnimann et al., 2025), a shallower 4-layer antibody language model. Performance degrades by $+0.54$ Å even with this substitute, indicating that the benefit is not explained by architectural capacity alone but requires the sequence representations learned by Ab-RoBERTa. On the sequence side, replacing the frozen H3BERTa embeddings at Stage 0 with a null token degrades performance by $+0.52$ Å, confirming that CDR-H3-specific sequence context is essential even in the earliest training stage. As hypothesized, the contribution of Ab-RoBERTa is strongly length-dependent: the gap between models with and without Stage 2 conditioning widens sharply beyond $L = 16$ (slope $= +0.087$ Å/residue, $p = 7.6 \times 10^{-7}$; Appendix E.2 Figure 4), consistent with V_{H} framework context providing geometric anchoring precisely when the loop has the most conformational freedom. In contrast, H3BERTa's benefit is flat across loop lengths (slope ≈ 0 Å/residue, $p = 0.97$; Appendix E.2 Figure 4), consistent with its role as a CDR-H3-local sequence model rather than a cross-chain context encoder.

Training curriculum. The four-stage curriculum was designed on the hypothesis that a model asked to learn CDR-H3 geometry from scratch would overfit to the small available dataset (4,648 structures), and that a progressively specialized prior from generic PDB loops through immune loops to CDR-H3 would generalize better.

Removing all pretraining and initializing Stage 3 directly from random weights leads to near-total failure ($+2.46$ Å Kabsch), validating the premise. The origin of this collapse is not simply the absence of sequence conditioning: the Stage 0 pretraining on generic loops is what establishes the geometric prior for stem-constrained denoising, and without it the model cannot learn a coherent coordinate space. Within the curriculum, the immune loop data at Stage 1 is the second most important factor ($+0.58$ Å when removed). CDR-L3 and TCR CDR3 loops share structural context, immune-system origin, conserved stem geometries, with CDR-H3 but provide far greater sequence diversity; their inclusion forces the model to build representations that generalize across loop identity rather than memorizing CDR-H3-specific patterns. Finally, activating LoRA adapters during Stage 0 rather than using frozen PLM embeddings degrades performance by $+0.26$ Å. We attribute this to premature specialization: fine-tuning the PLM on the broad and heterogeneous Stage 0 distribution biases its representations away from the CDR-H3 sequence grammar encountered in later stages, erasing part of the benefit that frozen pre-trained features would otherwise carry forward.

Loss function. We originally included a differentiable Ramachandran pseudo-dihedral penalty under the assumption that regularizing backbone geometry would improve structural quality. Contrary to this expectation, enabling the Ramachandran loss *hurts* performance by $+0.16$ Å Kabsch. A pseudo-Ramachandran analysis (Appendix F.3, Table 8) reveals that 75.2% of CDR-H3 residues in the ground-truth dataset fall outside the canonical α/β regions, compared to 50.3% for PDB structures in general. The Ramachandran penalty draws predicted coordinates toward the general loop distribution, which is systematically wrong for CDR-H3. Removing it, our model's predictions converge to 73.7% non-canonical residues, closely matching the ground-truth distribution (75.2%) and substantially closer than any general-loop prior ($\mathrm{KL}_{\mathrm{pred}\to\mathrm{GT}} = 0.38$ versus $\mathrm{KL}_{\mathrm{PDB70}\to\mathrm{GT}} = 1.15$; sliced Wasserstein-1: 9.3 versus 24.7). Notably, the model trained *with* the Ramachandran loss achieves the same non-canonical fraction (75.2%) but a *worse* distributional match (KL$= 0.47$, $W = 12.5$): the loss does not reduce non-canonical occupancy but distorts the angular geometry within non-canonical regions, explaining its detrimental effect on RMSD. This detrimental effect is moreover length-dependent (slope $= -0.029$ Å/residue, $p = 0.016$; Appendix E.2 Figure 4), consistent with the Ramachandran penalty over-constraining an increasingly complex dihedral space in longer loops.

The anchor loss, which enforces stem endpoint constraints during denoising, is beneficial ($+0.34$ Å when removed) but secondary to the dominant conditioning and curriculum effects. Its main role appears to be training stability: without it, the model must infer endpoint constraints implicitly from the canonical frame definition, which it can do but less reliably in early training.

Table 2: Ablation study on the 192-sequence validation set. We report median Cα RMSD under three superposition regimes (Direct: 0 DOF; Roll-aligned: 1 DOF (Appendix F.1); Kabsch: 6 DOF). Δ is the change in Kabsch RMSD relative to the best model (*noRama*). All variants are evaluated at Stage 3; each modifies a single component of the curriculum, architecture, or loss.

	Variant	Direct (Å)	Roll (Å)	Kabsch (Å)	Δ (Å)
	Ours (*noRama*)	**2.04**	**1.89**	**1.45**	—
Loss	+ Ramachandran loss	2.34	2.08	1.61	+0.16
Loss	− Anchor loss	2.53	2.25	1.79	+0.34
Architecture	− Ab-RoBERTa (no Stage 2)	2.85	2.68	2.01	+0.56
Architecture	Ab-RoBERTa → HeavyBERTa	2.70	2.55	2.00	+0.54
Architecture	H3BERTa → null (Stage 0)	2.76	2.55	1.97	+0.52
Curriculum	− Immune loops (Stage 1)	2.91	2.69	2.03	+0.58
Curriculum	LoRA at Stage 0 (vs. frozen)	2.39	2.22	1.71	+0.26
Curriculum	− All pre-training (Stages 0–2)	7.95	7.50	3.92	+2.46

4.2 Evaluation on The Test Set

To confirm that our design choices generalize beyond the validation set on which they were selected, we evaluate *noRama_v3* on the held-out test split and the hard test split which vary in sequence-identity stringency (Table 3). Median Kabsch RMSD is 1.452 Å on the validation set and 1.448 Å on the test set, a gap of under 0.01 Å, confirming that model selection introduced no measurable overfitting and that all design decisions transfer cleanly to unseen sequences. The hard-test set (40% sequence identity threshold) yields a lower median Kabsch RMSD of 1.181 Å, due to the hard-test split containing no CDR-H3 loops of length $L \geq 20$ (standard test set has 22). Restricting both splits to the L5–L19 length range yields comparable Kabsch RMSD, consistent with genuine generalization at low sequence identity rather than a distribution artifact.

The grounded canonical frame introduced at Stage 3 allows loop placement to be assessed without superposition. On the test set, the median roll correction angle is $|\theta| = 4.51°$, and 95.7% of predictions fall within $15°$ of the ground-truth roll orientation. This indicates that the backbone-plane grounding (Appendix C) resolves the roll ambiguity reliably for the large majority of loops, validating the use of Direct RMSD as a meaningful absolute placement metric. For short loops ($L5$–8), direct placement reaches $1.24\mathring{A}$ median without any superposition on the test set, a result that has direct relevance for antibody engineering applications where loop geometry must be predicted in the context of a fixed framework.

Performance degrades monotonically with loop length (Table 3): Kabsch RMSD rises from 0.66 Å at $L5$–8 to 2.76 Å at $L20+$ on the test set. This length dependence is expected, since longer CDR-H3 loops have greater conformational freedom and a weaker geometric relationship to the stem anchors.

4.3 Temporal Holdout and Baseline Comparison

To benchmark our model against existing methods, we curated a temporally held-out set of 71 CDR-H3 loops deposited after 2021, the AlphaFold3 training cutoff. This cutoff postdates the training data of all evaluated methods, ruling out exact-structure leakage. However, we can only guarantee the absence of sequence-level leakage for our model, whose training set was explicitly filtered by sequence identity against the holdout set. The training corpora of the baselines, ColabFold (Mirdita et al., 2022), ESMFold (Lin et al., 2023), tFold (Wu et al., 2024a), IgFold (Ruffolo et al., 2023), and Chai-1 (Chai Discovery team et al., 2024), were not filtered in the same way. All comparisons are therefore performed on the 69-structure subset processable by every method.

Baseline comparison. On a single random sample (*Ours (rand)*), our model reaches 2.95 Å Kabsch on the holdout, behind ColabFold (2.20 Å), ESMFold (2.19 Å), and tFold (2.29 Å). The oracle, best of five samples selected post hoc, reaches 2.31 Å, comparable to the oracles of the deterministic baselines (ESMFold 2.19 Å, tFold 2.29 Å, IgFold 2.28 Å) and to Chai-1's oracle (2.11 Å). ColabFold's oracle (1.80 Å) is the lowest, benefiting from its multi-seed diversity (Table 4).

Length dependence is present in all methods. At short loops (L5–8), our single sample already outperforms ColabFold, ESMFold, IgFold, and Chai-1 (0.89 Å vs 1.49 Å, 1.49 Å, 1.16 Å, 1.35 Å), and our oracle (0.63 Å) is the lowest in this bin. At L12–15, our oracle (1.62 Å) matches ColabFold (1.64 Å). The gap widens at L16–19 and L20+, where greater conformational freedom weakens our single-sample baseline (Table 9, Appendix G.1).

Split	Bucket	n	**Direct** (Å)	**Kabsch** (Å)	**Window** (Å)	$\lvert\theta\rvert$ (°)	**W15** (%)
Validation	**all**	**192**	$\mathbf{2.04 \pm 1.85}$	$\mathbf{1.45 \pm 1.12}$	$\mathbf{1.02 \pm 0.44}$	**3.56**	**95.8**
	L5–8	14	1.21 ± 0.48	0.60 ± 0.42	0.55 ± 0.39		
	L9–11	40	1.25 ± 0.67	0.81 ± 0.39	0.62 ± 0.28		
	L12–15	79	1.93 ± 1.31	1.43 ± 0.65	1.05 ± 0.36		
	L16–19	40	3.64 ± 1.89	2.57 ± 0.90	1.47 ± 0.32		
	L20+	19	5.24 ± 1.59	3.56 ± 1.16	1.55 ± 0.26		
Test	**all**	**162**	$\mathbf{2.29 \pm 1.93}$	$\mathbf{1.45 \pm 0.99}$	$\mathbf{1.05 \pm 0.43}$	**4.51**	**95.7**
	L5–8	13	1.24 ± 0.43	0.66 ± 0.20	0.54 ± 0.23		
	L9–11	41	1.36 ± 0.88	0.93 ± 0.45	0.71 ± 0.33		
	L12–15	60	2.37 ± 1.02	1.62 ± 0.60	1.14 ± 0.33		
	L16–19	26	3.63 ± 2.99	2.13 ± 0.98	1.36 ± 0.42		
	L20+	22	4.82 ± 1.68	2.76 ± 1.09	1.46 ± 0.39		
Hard test	**all**	**28**	$\mathbf{1.78 \pm 2.29}$	$\mathbf{1.18 \pm 0.93}$	$\mathbf{0.95 \pm 0.39}$	**4.65**	**85.7**
	L5–8	4	0.89 ± 0.24	0.45 ± 0.06	0.35 ± 0.05		
	L9–11	9	1.56 ± 2.56	0.88 ± 0.40	0.64 ± 0.28		
	L12–15	9	1.72 ± 1.28	1.27 ± 0.75	1.06 ± 0.32		
	L16–19	6	3.89 ± 2.20	2.37 ± 0.80	1.30 ± 0.13		
	L20+	0					

Table 3: Per-length performance of *noRama_v3* across internal evaluation splits. Metrics are median $\pm$ std Cα RMSD (Å). *Direct*: no superposition (0 DOF). *Kabsch*: 6-DOF superposition. *Window*: mean 5-residue Kabsch RMSD, measuring local backbone geometry. $|\theta|$: median absolute roll-correction angle; *W15*: fraction of loops with $|\theta| < 15°$. Test uses 60% sequence identity clustering; hard test uses 40%. The hard-test split contains no loops of length $L \geq 20$, whereas the test set has 22 such sequences.

Method	**n**	**Kabsch (Å)** median (IQR); mean±std	**Roll (Å)** median (IQR); mean±std	$\lvert\theta\rvert$ (°)	**Conf.**
ColabFold	69	2.20 (1.28–3.00); 2.23±1.18	3.14 (2.09–4.58); 3.47±2.00	6.4	0.76
↪ oracle	69	1.80 (1.08–2.66); 1.93±1.13	2.56 (1.52–4.48); 3.04±1.85	6.1	–
ESMFold†	69	2.19 (1.18–3.31); 2.38±1.49	3.39 (1.77–4.70); 3.57±2.34	5.9	0.69
tFold†	69	2.29 (1.39–3.20); 2.42±1.29	2.97 (2.11–5.11); 3.59±2.09	6.9	0.73
IgFold	69	2.52 (1.48–3.54); 3.22±2.25	3.77 (2.38–6.05); 5.91±5.43	8.0	–
↪ oracle	69	2.28 (1.37–3.54); 2.70±1.67	3.52 (2.01–5.52); 4.68±3.77	8.0	–
Chai	69	2.52 (1.80–3.67); 2.73±1.40	4.26 (2.70–5.70); 4.46±2.74	9.8	0.92
↪ oracle	69	2.11 (1.13–3.24); 2.28±1.35	3.39 (1.81–4.69); 3.76±2.66	7.1	–
Ours (rand)	69	2.95 (1.81–3.87); 3.00±1.47	4.31 (2.47–6.52); 4.83±2.80	8.8	–
direct (no align)	69	5.05 (2.94–8.07); 5.68±3.16	–	–	–
Ours (ens)	69	2.82 (1.72–3.67); 2.86±1.29	3.96 (2.45–4.97); 4.30±2.35	6.3	–
Ours (oracle)	69	2.31 (1.48–3.05); 2.31±1.12	3.36 (2.15–4.53); 3.76±2.36	10.2	–

Table 4: Bake-off comparison on the SAbDab2 n=69 external holdout. Each cell reports median (IQR) and mean±std (semicolon-separated) across the n Fv listed. *oracle* rows select the best-of-5 sample per loop by ground-truth Kabsch RMSD; *Ours (rand)* draws one random sample per loop and *Ours (ens)* selects the sample closest to the centroid of the 5-sample distribution. †ESMFold and tFold are deterministic (single structure per input), so their oracle and random-sample statistics are identical by construction and are shown once. Conf. is each method's own confidence metric, not comparable across rows: mean CDR-H3 pLDDT$/100$ for ColabFold and ESMFold, mean CDR-H3 confidence for tFold, and chain PTM for Chai; IgFold's is omitted as its scale is non-standard. Appendix G.1 Table 9 gives the per-length breakdown of Kabsch RMSD for all methods.

The $+0.94\mathring{A}$ gap between holdout and test performance is systematic. The generalization to the internal test set, which contains sequentially novel CDR-H3 loops, would argue against simple train–test contamination (Section 4.2).

4.3.1 Systematic analysis of the test–holdout gap.

Since the gap is present at *every* length bin, including the shortest (L5–8: $+0.21\mathring{A}$; L9–11: $+0.59\mathring{A}$), this indicates a systematic drift. To identify whether any known structural or distributional factor accounts for it, we test five hypotheses in turn; the full numerical breakdown is in Table 10 and Figure 6 (Appendix G.2).

Holo/apo status. The holdout is predominantly antigen-bound (83% holo vs 42% in the test set; 49% of test structures carry no annotation). However, splitting both sets by holo/apo status yields nearly identical gaps in both subgroups (holo $\Delta = +0.91$ Å, apo $\Delta = +1.01$ Å), ruling out antigen contact as an explanation.

Crystal contacts. Within both sets, loops with close symmetry-mate contacts show higher RMSD than loops without (packing artifacts distort the crystallized conformation). The test set contains *more* contacts than the holdout (67% vs 42%), so if crystal packing drove the gap we would expect the test set to be harder, the opposite of what we observe. The gap persists even in structures where no symmetry mate lies within 15 Å (holdout 2.67 vs test 1.99 Å, $\Delta = +0.68$ Å). Crystal contacts are ruled out.

Sequence identity to training. Stratifying by identity quartile exposes a confound: Q3 and Q4 *test* loops have mean lengths of 11.0 and 12.0 residues (only 2% and 15% with $L \geq 16$), whereas the same holdout quartiles contain substantially longer loops (15.4 and 14.8 residues, 48% and 39% with $L \geq 16$). The apparent increase in Δ at high identity is a length artifact; cross-set identity at matched length bins confirms holdout structures are uniformly more distant from the test set at all lengths. Sequence identity does not explain the gap.

Cysteine-containing loops. The holdout contains substantially more cysteine-bearing CDR-H3 loops than the test set (12.7% vs 2.5%; 9 vs 4 structures), and these loops are harder to predict ($\Delta = +3.06$ Å for Cys loops vs $+0.76$ Å for the remainder). CDR-H3 loops bearing cysteine might form intra-loop disulfide bonds that rigidly constrain the backbone. This is the single factor that fully accounts for the disproportionate difficulty at long lengths: after removing cysteine loops, the L20+ gap is *eliminated*, holdout (3.55 Å) is marginally easier than the test set (3.74 Å, $\Delta = -0.19$ Å), because five of the sixteen holdout L20+ structures bear cysteine (31%), compared with one of twenty-two on the test set (5%).

Residual gap and oracle analysis. After excluding cysteine loops, a gap of $+0.76$ Å persists across all length bins after ruling out all examined confounders. Oracle analysis distinguishes model failures (the correct conformation is not generated) from selection failures (it is generated but not ranked first). On the test set, oracle reduces Kabsch RMSD by 25.3% overall and by 21.3% at L12–15. On the holdout, L12–15 shows only 2.5% oracle gain ($2.19 \rightarrow 2.14$ Å). The model does not generate correct conformations for specific L12–15 holdout structures. This failure is not easily attributed to data coverage: CDR-H3 loops of length 12–15 are well represented in training, and their sequence identities to training are comparable to those in the test set. At L16–19 and L20+, oracle gains remain comparable across test and holdout (10–27%), confirming adequate ensemble diversity for longer loops.

Summary. The test–holdout gap is systematic across all loop lengths and is not explained by holo/apo status, crystal contacts, or sequence identity. A higher prevalence of cysteine-containing loops in the holdout (12.7% vs 2.5%) fully accounts for the L20+ component of the gap. A residual gap of $+0.76$ Å in cysteine-free structures remains after this correction, and oracle analysis identifies a specific model-level failure at L12–15 that cannot be attributed to training coverage or sequence novelty. We present this as a concrete open problem: CDR-H3 prediction on genuinely post-cutoff structures is systematically harder than internal evaluation predicts, and the source of this difficulty has not been identified.

5 Conclusion

We presented H3Fold, a four-stage diffusion model for CDR-H3 loop structure prediction that grounds predictions in a canonical stem frame, conditions generation on two complementary protein language model representations, and runs at 0.70 s per Fv at inference (Table 11, Appendix H). A systematic ablation study identified the principal drivers of performance and clarified the distinct role of each component.

The most consequential finding is the irreplaceable value of Stage 0 pretraining on generic loop geometries: without it, the model fails almost entirely ($+2.46$ Å Kabsch on the validation set), confirming that a coherent geometric prior for stem-constrained denoising cannot be acquired from the CDR-H3 dataset alone (Table 2). Within the dual PLM architecture, H3BERTa and Ab-RoBERTa contribute orthogonal signals: H3BERTa provides CDR-H3-local sequence context that is equally important at all loop lengths, while Ab-RoBERTa encodes V_{H} framework context whose benefit grows with length as loops gain conformational freedom. Immune loop data at Stage 1 ($+0.58$ Å when removed) is the second-largest contributor overall, forcing the model to generalize across loop identity rather than memorize CDR-H3-specific patterns. These structural gains are confirmed to reflect learned geometry rather than encoder confidence: prediction error correlates negligibly with PLM perplexity ($\rho = 0.14$ for H3BERTa, $\rho \approx 0$ for Ab-RoBERTa), while loop length dominates ($\rho = 0.72$; Figure 5, Appendix F.2).

Temporal holdout evaluation reveals a systematic performance drop that internal benchmarks do not predict, and that exhaustive investigation could only partially explain. Cysteine-containing loops, more prevalent in the holdout

Hypothesis / Subgroup	Test set			SAbDab2 holdout			Δ rand
	n	Rand (Å)	Oracle (Å)	n	Rand (Å)	Oracle (Å)	(Å)
Overall	162	1.937	1.448	71	2.873	2.448	+0.936
Per loop length:							
L5–8	13	0.683	0.663	4	0.890	0.627	+0.207
L9–11	41	1.069	0.931	10	1.659	1.535	+0.591
L12–15	60	2.064	1.624	23	2.193	2.139	+0.129
L16–19	26	2.376	2.133	18	3.427	2.753	+1.051
L20+	22	3.760	2.761	16	4.740	3.776	+0.980
Cys loops:							
Has Cys (2.5% / 12.7%)	4	1.900	1.826	9	4.965	3.871	+3.064
No Cys	158	1.937	1.442	62	2.698	2.394	+0.761
No-Cys per length:							
L5–8	13	0.683	0.663	4	0.890	0.627	+0.207
L9–11	41	1.069	0.931	10	1.659	1.535	+0.591
L12–15	58	2.064	1.624	21	2.193	2.139	+0.129
L16–19	25	2.402	2.164	16	3.277	2.692	+0.875
L20+	21	3.737	2.605	11	3.547	3.220	−0.190
Oracle improvement (rand → oracle):							
Overall	162	1.937 (25.3%)	1.448	71	2.873 (14.8%)	2.448	+0.936
L5–8	13	0.683 (2.9%)	0.663	4	0.890 (29.6%)	0.627	+0.207
L9–11	41	1.069 (12.8%)	0.931	10	1.659 (7.5%)	1.535	+0.591
L12–15	60	2.064 (21.3%)	1.624	23	2.193 (2.5%)	2.139	+0.129
L16–19	26	2.376 (10.2%)	2.133	18	3.427 (19.7%)	2.753	+1.051
L20+	22	3.760 (26.6%)	2.761	16	4.740 (20.3%)	3.776	+0.980

Table 5: Test–holdout performance gap (*noRama_v3*, $N = 5$ samples, Kabsch RMSD Å, median). *Rand*: single sample. *Oracle*: best of 5. Δ: holdout−test rand; positive = holdout harder. Holdout $n = 71$ for our model; the common benchmark (Table 4) uses the $n = 69$ subset processable by all methods. Cys percentages are test / holdout. Full analysis including holo/apo, crystal contacts, and sequence identity is in Table 10 (Appendix G.2).

(12.7% vs 2.5%), account for the disproportionate long-loop component through intra-loop disulfide bonds invisible to Cα-only modeling; removing them eliminates the L20+ deficit entirely. A residual gap of $+0.76$ Å in cysteine-free structures, present at all loop lengths, remains without structural explanation. We present this as an open problem: it establishes that CDR-H3 prediction on genuinely novel structures is harder than internal benchmarks suggest and that identifying this difficulty is a prerequisite for reliable use in antibody engineering.

Two limitations bound this work. First, although the model is trained on both apo (unbound) and holo (antigen-bound) structures, it does not receive antigen information at inference and so cannot account for the antigen that drives CDR-H3 conformational selection. Second, conditioning on V_{H} alone omits the paired V_{L} chain, which shapes CDR-H3 geometry through interfacial packing in the complete Fv.

Paired-chain context, side-chain heavy atoms, and disulfide bond topology are the clearest paths toward closing the gap between internal and temporal performance.

In conclusion, by pairing loop-local diffusion with dual protein language model conditioning, H3Fold achieves competitive accuracy at a fraction of the cost of full-structure prediction, laying a foundation for the future development of antibody engineering.

Reproducibility Statement

We ensure that our dataset construction, model architecture, training curriculum, inference-time sampling, and baseline comparisons are reproducible. The appendices document each of these in full.

Beyond Section 3.1, Appendix B describes how the datasets are created, extracted, normalized, and annotated. Split sizes are broken down per length bin in Table 6, and Appendix B.2 characterizes the resulting distributions. The canonical frame in which all coordinates are expressed is derived step by step in Appendix C, so the coordinate convention can be reconstructed exactly. Next, the denoising architecture is specified in Section 3.2, with a per-module parameter inventory and the per-stage adaptation state of both encoders in Figure 2 (Appendix D). The diffusion formulation and DDIM sampling settings are given in Section 3.3, the full training objective with both auxiliary terms in Section 3.4, and the four-stage curriculum in Section 3.5 and Table 1. Every reported metric is defined explicitly: Direct, Kabsch, and Window RMSD in Table 3, and the closed-form roll-aligned RMSD in Appendix F.1. Baseline sample counts, selection rules, and timing methodology are given in Section 4.3 and Appendix H, so the external comparison can be repeated as run.

For model weights, training and inference code, and the exact split definitions, see our GitHub repository.

References

Brennan Abanades, Guy Georges, Alexander Bujotzek, and Charlotte M. Deane. ABlooper: fast accurate antibody CDR loop structure prediction with accuracy estimation. *Bioinformatics*, 38(7):1877–1880, 2022. doi: 10.1093/bioinformatics/btac016.

Brennan Abanades, Wing Ki Wong, Fergus Boyles, Guy Georges, Alexander Bujotzek, and Charlotte M. Deane. ImmuneBuilder: Deep-learning models for predicting the structures of immune proteins. *Communications Biology*, 6(1):575, 2023. doi: 10.1038/s42003-023-04927-7.

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, 2024. doi: 10.1038/s41586-024-07487-w.

Jared Adolf-Bryfogle, Oleksandr Kalyuzhniy, Michael Kubitz, Brian D. Weitzner, Xiaozhen Hu, Yumiko Adachi, William R. Schief, and Roland L. Dunbrack. RosettaAntibodyDesign (RAbD): A general framework for computational antibody design. *PLOS Computational Biology*, 14(4):e1006112, 2018. doi: 10.1371/journal.pcbi.1006112.

Deniz Akpinaroglu, Jeffrey A. Ruffolo, Sai Pooja Mahajan, and Jeffrey J. Gray. Simultaneous prediction of antibody backbone and side-chain conformations with deep learning. *PLOS ONE*, 17(6):e0258173, 2022. doi: 10.1371/journal.pone.0258173.

Nathaniel R. Bennett, Joseph L. Watson, Robert J. Ragotte, Andrew J. Borst, DéJenaë L. See, Connor Weidle, Riti Biswas, Yutong Yu, Ellen L. Shrock, Russell Ault, Philip J. Y. Leung, Buwei Huang, Inna Goreshnik, John Tam, Kenneth D. Carr, Benedikt Singer, Cameron Criswell, Basile I. M. Wicky, Dionne Vafeados, Mariana Garcia Sanchez, Ho Min Kim, Susana Vázquez Torres, Sidney Chan, Shirley M. Sun, Timothy T. Spear, Yi Sun, Keelan O'Reilly, John M. Maris, Nikolaos G. Sgourakis, Roman A. Melnyk, Chang C. Liu, and David Baker. Atomically accurate de novo design of antibodies with RFdiffusion. *Nature*, 649(8095):183–193, 2026. doi: 10.1038/s41586-025-09721-5. Originally posted as bioRxiv 2024.03.14.585103.

Lea Brönnimann, Thomas Lemmin, and Chiara Rodella. Deep generative modeling captures maturation-dependent pairing patterns in human antibodies. *iScience*, 2025. doi: 10.1016/j.isci.2025.114447.

Chai Discovery team, Jacques Boitreaud, Jack Dent, Matthew McPartlon, Joshua Meier, Vinicius Reis, Alex Rogozhonikov, and Kevin Wu. Chai-1: Decoding the molecular interactions of life. *bioRxiv*, 2024. doi: 10.1101/2024.10.10.615955. Preprint.

Hedi Chen, Xiaoyu Fan, Shuqian Zhu, Yuchan Pei, Xiaochun Zhang, Xiaonan Zhang, Lihang Liu, Feng Qian, and Boxue Tian. Accurate prediction of CDR-H3 loop structures of antibodies with deep learning. *eLife*, 12:RP91512, 2024. doi: 10.7554/eLife.91512.

Yoonjoo Choi and Charlotte M. Deane. FREAD revisited: Accurate loop structure prediction using a database search algorithm. *Proteins: Structure, Function, and Bioinformatics*, 78(6):1431–1440, 2010. doi: 10.1002/prot.22658.

Cyrus Chothia and Arthur M. Lesk. Canonical structures for the hypervariable regions of immunoglobulins. *Journal of Molecular Biology*, 196(4):901–917, 1987. doi: 10.1016/0022-2836(87)90412-8.

Justas Dauparas, Ivan Anishchenko, Nathaniel R. Bennett, Hua Bai, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Alexis Courbet, Robbert J. de Haas, Neville P. Bethel, Philip J. Y. Leung, Timothy F. Huddy, Samuel J. Pellock, Doug Tischer, Frederick Chan, Brian Koepnick, Hannah Nguyen, Alex Kang, Banumathi Sankaran, Asim K. Bera, Neil P. King, and David Baker. Robust deep learning-based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, 2022. doi: 10.1126/science.add2187.

Frédéric A. Dreyer, Jan Ludwiczak, Karolis Martinkus, Brennan Abanades, Robert G. Alberstein, Pan Kessel, Pranav Rao, Jae Hyeon Lee, Richard Bonneau, Andrew M. Watkins, and Franziska Seeger. Conformation-aware structure prediction of antigen-recognizing immune proteins. *arXiv preprint arXiv:2507.09054*, 2025.

James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M. Deane. SAbDab: the structural antibody database. *Nucleic Acids Research*, 42(D1):D1140–D1146, 2014. doi: 10.1093/nar/gkt1043.

Tomas Geffner, Kieran Didi, Zhonglin Cao, Danny Reidenbach, Zuobai Zhang, Christian Dallago, Emine Kucukbenli, Karsten Kreis, and Arash Vahdat. Proteina: Scaling flow-based protein structure generative models. In *International Conference on Learning Representations (ICLR)*, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. doi: 10.48550/arXiv.2006.11239.

Magnus Haraldson Høie, Alissa M. Hummer, Tobias H. Olsen, Broncio Aguilar-Sanjuan, Morten Nielsen, and Charlotte M. Deane. AntiFold: improved structure-based antibody design using inverse folding. *Bioinformatics Advances*, 5(1):vbae202, 2025. doi: 10.1093/bioadv/vbae202.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Eunna Huh and Hyeonsu Lee. Antibody foundational model: Ab-RoBERTa. *arXiv preprint arXiv:2506.13006*, 2025.

John B. Ingraham, Max Baranov, Zak Costello, Karl W. Barber, Wujie Wang, Ahmed Ismail, Vincent Frappier, Dana M. Lord, Christopher Ng-Thow-Hing, Erik R. Van Vlack, Shan Tie, Vincent Xue, Sarah C. Cowles, Alan Leung, João V. Rodrigues, Claudio L. Morales-Perez, Alex M. Ayoub, Robin Green, Katherine Puentes, Frank Oplinger, Nishant V. Panwar, Fritz Obermeyer, Andrew R. Root, Andrew L. Beam, Frank J. Poelwijk, and Gevorg Grigoryan. Illuminating protein space with a programmable generative model. *Nature*, 623(7989):1070–1078, 2023. doi: 10.1038/s41586-023-06728-8.

Matthew P. Jacobson, David L. Pincus, Chaya S. Rapp, Tyler J. F. Day, Barry Honig, David E. Shaw, and Richard A. Friesner. A hierarchical approach to all-atom protein loop prediction. *Proteins: Structure, Function, and Bioinformatics*, 55(2):351–367, 2004. doi: 10.1002/prot.10613.

Charles A. Janeway, Paul Travers, Mark Walport, and Mark J. Shlomchik. *Immunobiology: The Immune System in Health and Disease*. Garland Science, New York, 5 edition, 2001. ISBN 978-0815336426.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Yasaman Karami, Frédéric Guyon, Sjoerd de Vries, and Pierre Tufféry. DaReUS-Loop: accurate loop modeling using fragments from remote or unrelated proteins. *Scientific Reports*, 8(1):13673, 2018. doi: 10.1038/s41598-018-32079-w.

Henry Kenlay, Frédéric A. Dreyer, Daniel Cutting, Daniel A. Nissley, and Charlotte M. Deane. ABodyBuilder3: improved and scalable antibody structure predictions. *Bioinformatics*, 40(10):btae576, 2024a. doi: 10.1093/bioinformatics/btae576.

Henry Kenlay, Frédéric A. Dreyer, Aleksandr Kovaltsuk, Dom Miketa, Douglas Pires, and Charlotte M. Deane. Large scale paired antibody language models. *PLOS Computational Biology*, 20(12):e1012646, 2024b. doi: 10.1371/journal.pcbi.1012646.

Xiangzhe Kong, Wenbing Huang, and Yang Liu. End-to-end full-atom antibody design. *arXiv preprint arXiv:2302.00203*, 2023. doi: 10.48550/arXiv.2302.00203.

Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006. doi: 10.1093/bioinformatics/btl158.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Shitong Luo, Yufeng Su, Xingang Peng, Sheng Wang, Jian Peng, and Jianzhu Ma. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 9754–9767, 2022.

Paolo Marcatili, Alessandra Rosi, and Anna Tramontano. PIGS: automatic prediction of antibody structures. *Bioinformatics*, 24(17):1953–1954, 2008. doi: 10.1093/bioinformatics/btn341.

Karolis Martinkus, Jan Ludwiczak, Wei-Ching Liang, Julien Lafrance-Vanasse, Isidro Hotzel, Arvind Rajpal, Yan Wu, Kyunghyun Cho, Richard Bonneau, Vladimir Gligorijevic, and Andreas Loukas. AbDiffuser: Full-atom generation of in vitro functioning antibodies. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 40729–40759, 2023.

Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. ColabFold: making protein folding accessible to all. *Nature Methods*, 19(6):679–682, 2022. doi: 10.1038/s41592-022-01488-1.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171, 2021.

Benjamin North, Andreas Lehmann, and Roland L. Dunbrack. A new clustering of antibody CDR loop conformations. *Journal of Molecular Biology*, 406(2):228–256, 2011. doi: 10.1016/j.jmb.2010.10.030.

Tobias H. Olsen, Iain H. Moal, and Charlotte M. Deane. AbLang: an antibody language model for completing antibody sequences. *Bioinformatics Advances*, 2(1):vbac046, 2022. doi: 10.1093/bioadv/vbac046.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

Cristian Regep, Guy Georges, Jiye Shi, Bojana Popovic, and Charlotte M. Deane. The H3 loop of antibodies shows unique structural characteristics compared to other loops. *Proteins: Structure, Function, and Bioinformatics*, 85(7):1311–1318, 2017. doi: 10.1002/prot.25291.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. 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. doi: 10.1073/pnas.2016239118.

Chiara Rodella and Thomas Lemmin. H3BERTa: A CDR-H3-specific language model for antibody repertoire analysis. *Patterns*, 7(7):101561, 2026. doi: 10.1016/j.patter.2026.101561.

Jeffrey A. Ruffolo, Jeffrey J. Gray, and Jeremias Sulam. Deciphering antibody affinity maturation with language models and weakly supervised learning. *arXiv preprint arXiv:2112.07782*, 2021.

Jeffrey A. Ruffolo, Jeremias Sulam, and Jeffrey J. Gray. Antibody structure prediction using interpretable deep learning. *Patterns*, 3(2):100406, 2022. doi: 10.1016/j.patter.2021.100406.

Jeffrey A. Ruffolo, Lee-Shin Chu, Sai Pooja Mahajan, and Jeffrey J. Gray. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nature Communications*, 14(1):2389, 2023. doi: 10.1038/s41467-023-38063-x.

Richard W. Shuai, Jeffrey A. Ruffolo, and Jeffrey J. Gray. IgLM: Infilling language modeling for antibody sequence design. *Cell Systems*, 14(11):979–989, 2023. doi: 10.1016/j.cels.2023.10.001.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

Fabian C. Spoendlin, Monica L. Fernández-Quintero, Sai Sundar Rajan Raghavan, Hannah L. Turner, Anant Gharpure, Johannes R. Loeffler, Wing Ki Wong, Alexander Bujotzek, Guy Georges, Andrew B. Ward, and Charlotte M. Deane. Predicting the conformational flexibility of antibody and T cell receptor complementarity-determining regions. *Nature Machine Intelligence*, 7(10):1755–1767, 2025. doi: 10.1038/s42256-025-01131-6.

Martin Steinegger, Markus Meier, Milot Mirdita, Harald Vöhringer, Stephan J. Haunsberger, and Johannes Söding. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*, 20(1):473, 2019. doi: 10.1186/s12859-019-3019-7.

Mario S. Valdés-Tresanco, Mario E. Valdés-Tresanco, Daiver E. Jiménez-Gutiérrez, and Ernesto Moreno. Structural modeling of nanobodies: A benchmark of state-of-the-art artificial intelligence programs. *Molecules*, 28(10):3991, 2023. doi: 10.3390/molecules28103991.

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Émile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with RFdiffusion. *Nature*, 620 (7976):1089–1100, 2023. doi: 10.1038/s41586-023-06415-8.

Brian D. Weitzner, Jeliazko R. Jeliazkov, Sergey Lyskov, Nicholas Marze, Daisuke Kuroda, Rahel Frick, Jared Adolf-Bryfogle, Naireeta Biswas, Roland L. Dunbrack, and Jeffrey J. Gray. Modeling and docking of antibody structures with Rosetta. *Nature Protocols*, 12(2):401–416, 2017. doi: 10.1038/nprot.2016.180.

Ian Wilkinson and Geoff Hale. Systematic analysis of the varied designs of 819 therapeutic antibodies and Fc fusion proteins assigned international nonproprietary names. *mAbs*, 14(1):2123299, 2022. doi: 10.1080/19420862.2022.2123299.

Fandi Wu, Yu Zhao, Jiaxiang Wu, Biaobin Jiang, Bing He, Longkai Huang, Chenchen Qin, Fan Yang, Ningqiao Huang, Yang Xiao, et al. Fast and accurate modeling and design of antibody-antigen complex using tfold. *bioRxiv*, pp. 2024.02.05.578892, 2024a. doi: 10.1101/2024.02.05.578892.

Fang Wu and Stan Z. Li. A hierarchical training paradigm for antibody structure-sequence co-design. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://openreview.net/forum?id=hV52oj0Sik.

Kevin E. Wu, Kevin K. Yang, Rianne van den Berg, Sarah Alamdari, James Y. Zou, Alex X. Lu, and Ava P. Amini. Protein structure generation via folding diffusion. *Nature Communications*, 15:1059, 2024b. doi: 10.1038/s41467-024-45051-2.

Jason Yim, Andrew Campbell, Andrew Y. K. Foong, Michael Gastegger, José Jiménez-Luna, Sarah Lewis, Victor Garcia Satorras, Bastiaan S. Veeling, Regina Barzilay, Tommi Jaakkola, and Frank Noé. Fast protein backbone generation with SE(3) flow matching. *arXiv preprint arXiv:2310.05297*, 2023a.

Jason Yim, Brian L. Trippe, Valentin De Bortoli, Émile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. SE(3) diffusion model with application to protein backbone generation. In *International Conference on Machine Learning (ICML)*, 2023b. doi: 10.48550/arXiv.2302.02277.

A Appendix

B Dataset Details

B.1 Structural Sources

We construct our training dataset from the ALL-conformations benchmark (Spoendlin et al., 2025), comprising: (i) 6,401 antibody VH/VL structures with annotated CDR-H3 and CDR-L3 loops; (ii) 691 TCR structures with annotated CDR-α3 and CDR-β3 loops; and (iii) approximately 1.2M general protein loop instances.

Table 6: Dataset statistics for all training, validation, evaluation, and holdout splits, stratified by CDR-H3 loop length. For each split, the first row counts all entries including duplicate CDR-H3 sequences from different PDB structures, and the *unique* row counts distinct CDR-H3 amino acid sequences.

Split	**L5–8**	**L9–11**	**L12–15**	**L16–19**	**L20+**	**Total**
Stage 0	50,977	69,939	40,233	12,699	7,165	181,013
↪ unique	6,698	11,783	6,953	2,403	1,578	29,415
Stage 1/2 train (CDR-H3)	326	1,069	2,082	718	453	4,648
↪ unique	98	372	640	305	150	1,565
Stage 1/2 val (CDR-H3)	28	96	185	114	38	461
↪ unique	14	40	79	40	19	192
Stage 3 train (CDR-H3)	326	1,082	2,139	739	498	4,784
↪ unique	98	372	640	305	150	1,565
Stage 3 val (CDR-H3)	28	99	185	120	38	470
↪ unique	14	40	79	40	19	192
Test set (Stage 3 join)	72	98	141	78	41	430
↪ unique	13	41	60	26	22	162
Hard test set (<40% id.)	10	19	14	15	0	58
↪ unique	4	9	9	6	0	28
SAbDab2 holdout (<70% id.)	4	10	23	18	16	71
↪ unique	4	10	23	18	16	71

B.2 Dataset Characterization

Figure 1 characterizes the CDR-H3 splits used in Stages 1–3.

B.3 Coordinate Extraction

For each loop entry we extract the full backbone (N, Cα, C, O, Cβ) from the corresponding mmCIF file. Loops are represented by their Cα trace and padded to $L_{\max} = 100$ residues. Loops are excluded if they exhibit any of the following: consecutive Cα–Cα distances outside $[3.3, 4.5]$ Å (chain breaks); non-bonded Cα clashes (< 3.0 Å); missing backbone atoms at any loop residue; non-standard amino acids; inter-anchor distance outside $[4.0, 8.0]$ Å; or degenerate alignment geometry.

B.4 Coordinate Normalization

Cα coordinate statistics $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ are computed from the CDR-H3 training set in the anchor frame. All coordinates are normalized channel-wise as $\tilde{\mathbf{x}}_i = (\mathbf{x}_i - \boldsymbol{\mu})/\boldsymbol{\sigma}$ prior to diffusion. The same statistics are used across all training stages and at inference.

B.5 Sequence Clustering

CDR-H3 sequences are clustered using CD-HIT (Li & Godzik, 2006) at 60% sequence identity and length-stratified into 80/10/10 train/validation/test splits; a hard test set uses 40% identity. Stage 0 loops are clustered at 50% (general PDB) or 60% (CDR-L3, TCR) identity with a 90/10 split. Any sequence with $\geq$60% identity to a CDR-H3 test sequence is excluded from pre-training.

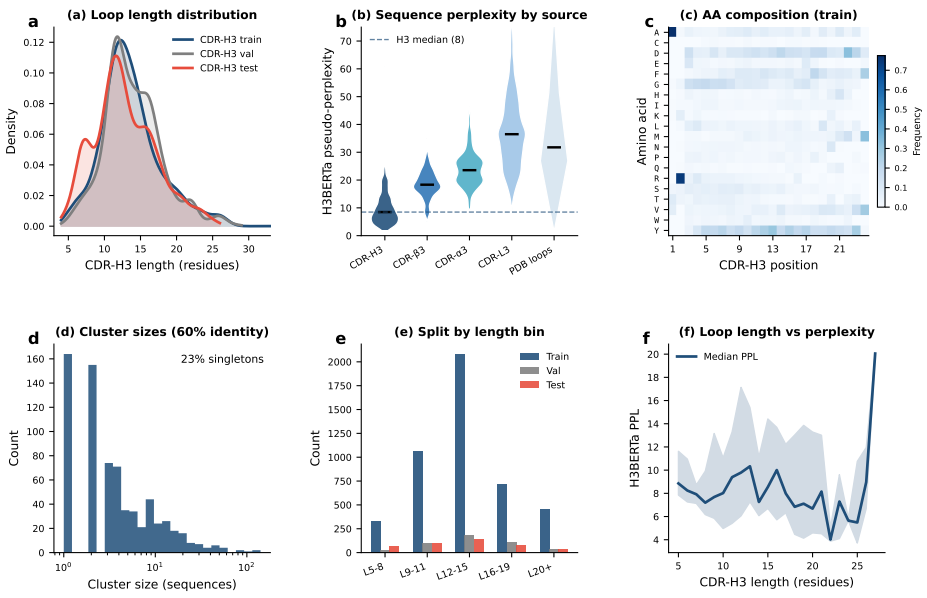


Figure 1: Characterization of the CDR-H3 dataset. **(a)** Loop length distributions for the train, validation, and test splits. **(b)** H3BERTa pseudo-perplexity by loop source; CDR-H3 loops have the lowest perplexity (median 8), with CDR-β3, CDR-α3, CDR-L3, and general PDB loops progressively higher, reflecting the encoder's CDR-H3-specific pre-training. **(c)** Position-wise amino acid composition of the training split. **(d)** CD-HIT cluster sizes at 60% sequence identity; 23% of clusters are singletons. **(e)** Split composition by loop length bin. **(f)** Median H3BERTa pseudo-perplexity as a function of loop length, with interquartile range shaded.

C Grounded Canonical Frame Derivation

The canonical frame places anchor$_N$ at the origin and anchor$_C$ on the positive x-axis, constraining five of six rigid-body degrees of freedom. To fix the sixth (roll about $+x$), we use the backbone geometry of the anchor residues.

For each anchor residue $r \in \{r_N, r_C\}$, the peptide-plane normal is:

$$\mathbf{n}_r = \frac{\mathbf{b}_r \times \mathbf{c}_r}{\|\mathbf{b}_r \times \mathbf{c}_r\|}, \qquad \mathbf{b}_r = \mathbf{p}_r^{C\alpha} - \mathbf{p}_r^{N}, \quad \mathbf{c}_r = \mathbf{p}_r^{C} - \mathbf{p}_r^{C\alpha}. \tag{3}$$

The sign is consistent for L-amino acids by chirality convention. We average over both anchors to reduce sensitivity to coordinate noise:

$$\bar{\mathbf{n}} = \frac{\mathbf{n}_{r_N} + \mathbf{n}_{r_C}}{\|\mathbf{n}_{r_N} + \mathbf{n}_{r_C}\|}. \tag{4}$$

Since r_N and r_C reside in the same β-strand framework, their peptide planes are nearly parallel, making the average geometrically stable. The three orthonormal frame axes are:

$$\mathbf{e}_1 = \mathbf{u} = \frac{\mathbf{p}_{r_C}^{C\alpha} - \mathbf{p}_{r_N}^{C\alpha}}{\|\mathbf{p}_{r_C}^{C\alpha} - \mathbf{p}_{r_N}^{C\alpha}\|}, \tag{5}$$

$$\mathbf{e}_2 = \frac{\bar{\mathbf{n}} - (\bar{\mathbf{n}} \cdot \mathbf{u})\,\mathbf{u}}{\|\bar{\mathbf{n}} - (\bar{\mathbf{n}} \cdot \mathbf{u})\,\mathbf{u}\|}, \tag{6}$$

$$\mathbf{e}_3 = \mathbf{e}_1 \times \mathbf{e}_2. \tag{7}$$

The rotation $\mathbf{R} = [\mathbf{e}_1\ \mathbf{e}_2\ \mathbf{e}_3] \in \mathrm{SO}(3)$ and translation $\mathbf{t} = \mathbf{p}_{r_N}^{C\alpha}$ define the transform; loop coordinates are expressed as $\tilde{\mathbf{p}}_i = \mathbf{R}^\top(\mathbf{p}_i - \mathbf{t})$.

This grounded frame reduces the same-sequence pair Cα RMSD from 11.7 Å (Gram-Schmidt against $[0, 0, 1]^\top$) to 0.6 Å, confirming that the roll degree of freedom is structurally determined and learnable.

Hyperparameter	Pre-training Stage 0	Fine-tuning Stage 1	Fine-tuning Stage 2	Fine-tuning Stage 3
H3Fold Denoiser Architecture				
Coordinate dim	3			
Max loop length	100			
Time enc. dim	128			
Coord. proj.	1,024			
Positional emb.	25,600			
Time MLP	33,024			
Denoiser width / heads / head dim / FFN	256 / 8 / 32 / 1024			
Transformer blocks / dropout	6 / 0.1			
H3 cross-attn proj. (proj_h3; 768→256→256)	262,656			
Ab cross-attn proj. (proj_ab; 768→256→256)	262,656			
Ab-RoBERTa null/frozen embedding	768	768	—	—
H3BERTa frozen embedding	768	—	768	—
AdaLN (LN no affine + Lin zero-init), 4 per block, 6 blocks total	1,585,152			
Self-attn MHA (Q=K=V=h, mask=loop_pad), per block	263,168			
H3 cross-attn MHA (Q=h, K=V=h3_kv), per block	263,168			
Ab cross-attn MHA (Q=h, K=V=ab_kv) — sublayer skipped entirely at scale 0	263,168			
FFN, per block	525,568			
Per block total / six blocks	1,579,264 / 9,475,584			
Output head (epsilon-prediction, not x0-prediction)	771			
Ghost *BackboneFrameEncoder* (zero-init, unused in forward)	136,448			
Module inventory total (trunk, all modules, excl. ghost)	**10,062,083**			
H3BERTa				
Hidden / layers / heads / FFN / vocab / max_pos	768 / 12 / 12 / 3072 / 25 / 102			
Backbone params, `AutoModel` incl. pooler / excl. pooler	85,744,896 / **85,154,304**			
Conditioning state (`h3_mode`)	frozen	LoRA	frozen	LoRA
LoRA rank / scale (Q,V only, 12 layers), params	—	8/16, 294,912	—	8/16, 294,912
Ab-RoBERTa				
Hidden / layers / heads / vocab / max_pos	768 / 12 / 12 / 25 / 152			
Backbone params, `AutoModel` incl. pooler / excl. pooler	**85,783,296** / 85,192,704			
Extra vs. H3 (excl. pooler)	38,400			
`ab_mode` (forward-pass mode)	null (no PLM call)	frozen, no_grad	LoRA	LoRA
Ab cross-attn gate (`phase2_scale`) — skips sublayer, not a zeroed matmul	0	0	1	1
LoRA rank / scale (Q,V; 12 layers), params	—	—	8/16, 294,912	8/16, 294,912
Summary				
Module inventory total (trunk + active LoRA, excl. ghost/frozen backbones)	10,062,083	10,356,995	10,356,995	10,651,907
Optimizer trainable params (AdamW; incl. Ab MHA/AdaLN gating, active LoRA, H3/Ab null embeddings)	**8,224,771**	**10,494,979**	**10,494,979**	**10,789,891**
Total params	**≈182.2M**			

Figure 2: Hyperparameters for H3Fold model training. Values are grouped by module: the denoiser trunk (transformer blocks, projections, and output head), the H3BERTa and Ab-RoBERTa conditioning backbones (frozen vs. LoRA-adapted, with active-adapter parameter counts), and a per-stage summary of trainable and total resident parameters. The final row reports the fixed total parameter count (trunk + both frozen pLM backbones + both LoRA adapters) resident in memory at any stage.

D Model Parameters

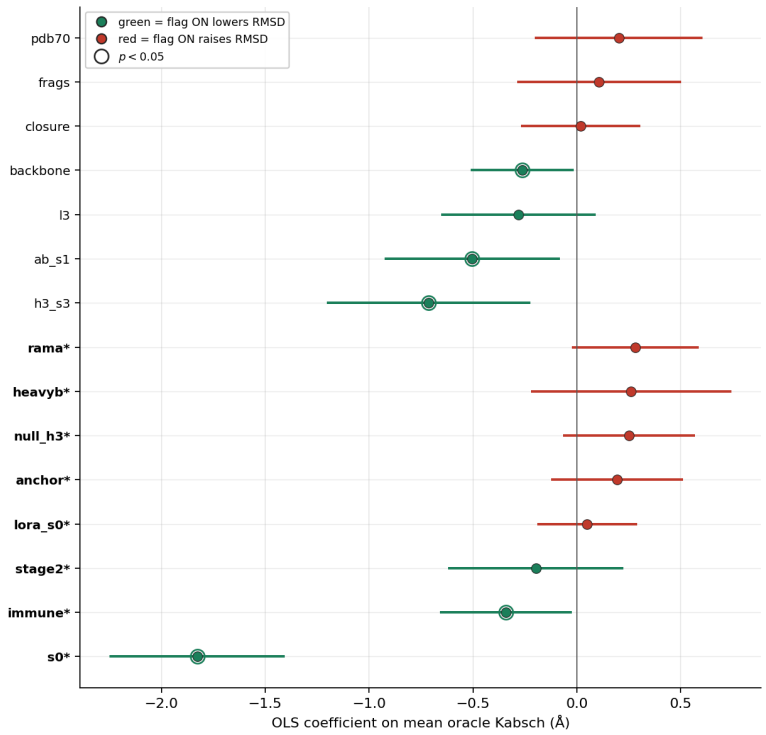
Figure 2 reports total hyperparameter counts for the noRama_v3 model training. This model contained the standard training regimen with H3BERTa frozen in s0 and s2 and fine-tuned with LORA in s1 and s3, and the Ab-RoBERTa embeddings null, frozen, then LORA fine tuned. The final row reports the fixed total parameter count (trunk + both frozen pLM backbones + both LoRA adapters) resident in memory at any stage.

E Ablation Study Details

E.1 Joint Regression Analysis

To verify that the univariate ablation findings are not artifacts of correlated design choices, we fit an ordinary least-squares (OLS) model that regresses mean oracle Kabsch Cα RMSD on 15 binary ablation-axis flags simultaneously, evaluated on the 192-sequence validation set. Confidence intervals on each coefficient are constructed by bootstrapping loops within the validation set. Figure 3 plots the resulting coefficients and gives the full regression output.

Controlling for all other axes, Stage-0 pretraining (*s0*), the H3BERTa representation (*h3_s3*), Ab-RoBERTa conditioning (*ab_s1*), and immune loop data (*immune*) each remain significant at $p < 0.05$, consistent with their rank in the univariate ablation table. The Ramachandran (*rama*) and anchor (*anchor*) losses do not reach significance in the joint model, reflecting their modest univariate effects. Notably, the HeavyBERTa substitution (*heavyb*) loses significance in the joint model ($\beta = 0.262$, $p = 0.260$), indicating that its apparent univariate effect ($+0.54$ Å) is partially confounded



Flag	β	SE	CI low	CI high	p
pdb70	+0.202	0.187	−0.195	+0.599	0.296
frags	+0.108	0.183	−0.281	+0.496	0.565
closure	+0.018	0.132	−0.262	+0.298	0.892
backbone	**−0.262**	**0.114**	**−0.504**	**−0.020**	**0.036**
l3	−0.281	0.173	−0.647	+0.085	0.123
ab_s1	**−0.504**	**0.196**	**−0.919**	**−0.089**	**0.020**
h3_s3	**−0.714**	**0.228**	**−1.198**	**−0.230**	**0.0065**
rama*	+0.282	0.141	−0.016	+0.580	0.062
heavyb*	+0.262	0.224	−0.214	+0.738	0.260
null_h3*	+0.251	0.147	−0.060	+0.562	0.107
anchor*	+0.193	0.147	−0.118	+0.504	0.206
lora_s0*	+0.050	0.110	−0.184	+0.284	0.657
stage2*	−0.197	0.196	−0.612	+0.218	0.330
immune*	**−0.342**	**0.147**	**−0.653**	**−0.031**	**0.033**
s0*	**−1.827**	**0.196**	**−2.243**	**−1.412**	7.2×10^{-8}

Figure 3: OLS coefficients (with bootstrapped 95% CIs) of mean oracle Kabsch Cα RMSD regressed jointly on all 15 binary ablation-axis flags, validation set ($n = 192$). Green markers indicate the flag lowers RMSD when active; red markers indicate it raises RMSD. Ringed markers denote $p < 0.05$. Controlling for all other axes, Stage-0 pretraining, the H3BERTa swap, Ab-RoBERTa conditioning, immune data, and the backbone-frame token remain significant; the Ramachandran and anchor losses do not. Variants included in the main ablation table (Table 2) are marked with *.

with correlated PLM and curriculum flags; we therefore treat the univariate comparison as the more controlled estimate for that ablation axis.

E.2 Length-Stratified Ablation Effects

Figure 4 plots the paired per-loop Δ Kabsch RMSD (ablated model minus reference) as a function of CDR-H3 loop length for three ablation axes.

E.3 Full Ablation Grid

Table 7 reports all Stage 3 ablation variants evaluated on the 192-sequence validation set, including runs not discussed in the main text. Each entry reports median Cα RMSD under three superposition regimes (Direct, Roll-aligned, Kabsch).

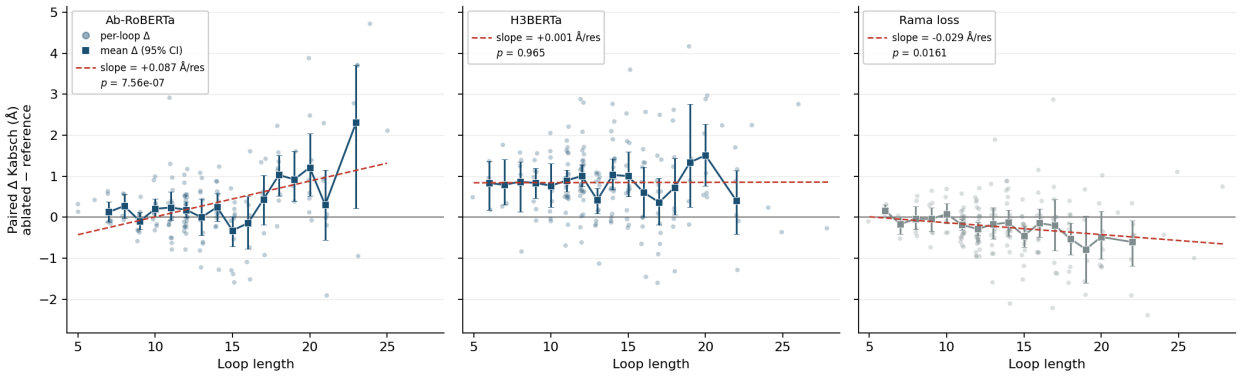


Figure 4: Paired ablation-impact Δ (ablated $-$ reference) in Kabsch Cα RMSD as a function of loop length, for the Ab-RoBERTa (left), H3BERTa (middle), and Ramachandran-loss (right) ablation axes. Light points are per-loop deltas; dark markers and error bars show the mean Δ ($\pm$95% CI) within each length bin; the dashed line is an OLS fit of Δ on loop length, with slope and p-value in each panel legend. Ab-RoBERTa's benefit grows significantly with length (slope $= +0.087$ Å/residue, $p = 7.6 \times 10^{-7}$), reflecting the increasing informational value of V_{H} framework context for longer, more conformationally free loops. H3BERTa's benefit is length-invariant (slope ≈ 0, $p = 0.97$), consistent with its role as a CDR-H3-local sequence model. The Ramachandran loss becomes *more* harmful with length (slope $= -0.029$ Å/residue, $p = 0.016$), consistent with over-constraining an increasingly complex dihedral space.

Table 7: Full ablation grid on the 192-sequence validation set (Stage 3 checkpoints, 5 samples per sequence). Variants included in the main ablation table (Table 2) are marked with $\star$.

Variant	Direct (Å)	Roll (Å)	Kabsch (Å)	Notes
noRama_v3*	2.04	1.89	1.45	Best model
full_v3*	2.34	2.08	1.61	+Ramachandran loss
noAnchor_v3*	2.53	2.25	1.79	−Anchor loss
noStage2_v1*	2.85	2.68	2.01	−Ab-RoBERTa
noRama_heavyB_v1*	2.70	2.55	2.00	HeavyBERTa sub.
nullS0_v3*	2.76	2.55	1.97	H3BERTa→null
noImmune_v3*	2.91	2.69	2.03	−Immune loops
loraS0both_v1*	2.39	2.22	1.71	LoRA at Stage 0
noPretrain_v1*	7.95	7.50	3.92	−All pretraining
allImmune_v1	2.39	2.05	1.61	CDR-L3+TCR at Stage 3
withL3_v1	2.51	2.24	1.81	
withL3_pdb70_v1	2.65	2.39	1.86	
noL3_v1	2.81	2.63	1.95	−CDR-L3 from Stage 1
noL3_pdb70_v1	2.95	2.63	1.98	
loraS0_v1	2.69	2.43	1.92	LoRA S0 (H3BERTa only)
closure1_v1	2.36	2.15	1.71	Closure loss variant
anchor05_v1	2.68	2.46	1.90	Anchor weight $\times 0.5$
anchor0_v1	2.28	1.93	1.57	No anchor weight bonus
bb_scratch_v1	2.28	2.06	1.54	Backbone-atom pred., random init
bb_abs_v1	2.28	2.04	1.56	Backbone-atom pred., abs. frame
bb_v1	2.24	2.05	1.64	Backbone-atom pred.
lora4_v1	2.56	2.32	1.91	LoRA $r = 4$
deep_v1	2.78	2.43	1.97	Deeper network
frags_bb_v1	2.57	2.40	1.85	Fragment + backbone pred.
frags_v1	2.77	2.53	1.91	Fragment conditioning
ramaLen_v1	2.83	2.55	2.04	Length-scaled Rama loss
noAbS1_v1	3.11	2.88	2.32	−Ab-RoBERTa at Stage 1
noImmune_v1	3.16	2.96	2.32	Earlier immune ablation
nullS0_v1	3.03	2.74	2.22	Earlier null ablation
noH3berta_v1	4.31	3.29	2.25	*Confounded: also skips Stage 2*

F Evaluation Details

F.1 Roll-Aligned RMSD

The roll-aligned RMSD removes only the roll degree of freedom via the closed-form optimal rotation angle:

$$\phi^* = \operatorname{atan2}(B,\, A), \qquad A = \sum_i (p_{iy}q_{iy} + p_{iz}q_{iz}), \quad B = \sum_i (p_{iy}q_{iz} - p_{iz}q_{iy}), \tag{8}$$

where $\mathbf{p}_i$ and $\mathbf{q}_i$ are predicted and ground-truth Cα coordinates. The roll correction angle $|\phi^*|$ is reported as a diagnostic for frame learning quality; W15% denotes the fraction of loops with $|\phi^*| < 15°$.

F.2 Evaluating Sources of Error

Figure 5 correlates per-loop CDR-H3 Kabsch Cα RMSD on the test set (n=162) against six candidate sources of error: sequence-level uncertainty under the two pre-trained encoders (H3BERTa perplexity; Ab-RoBERTa loop and VH perplexity), loop length, redundancy in the training set (CD-HIT cluster size), and species of origin.

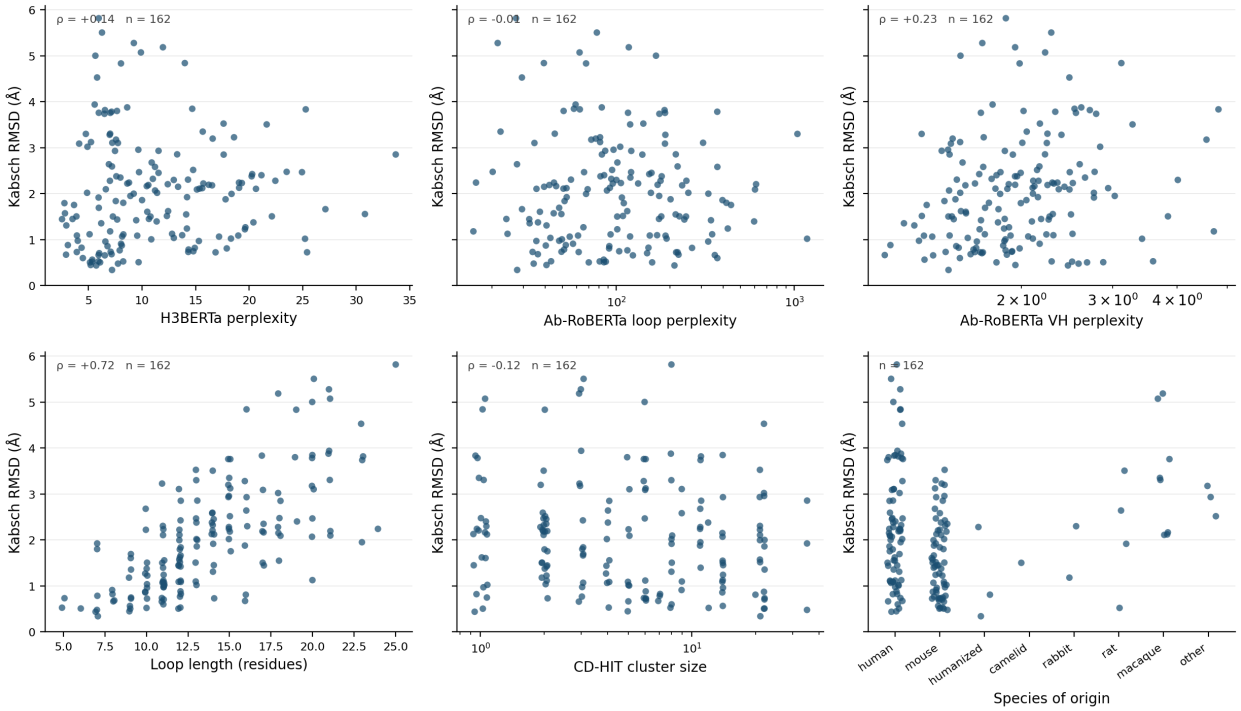


Figure 5: Sources of error in CDR-H3 Kabsch Cα RMSD on the test set (n=162). Each panel plots per-loop RMSD against a candidate covariate: H3BERTa perplexity, Ab-RoBERTa loop perplexity, Ab-RoBERTa VH perplexity, loop length, CD-HIT cluster size, and species of origin. Pearson ρ is reported for the five continuous covariates.

F.3 Pseudo-Ramachandran Analysis

Because CDR-H3 structures are Cα-only (side-chain and carbonyl atoms are not predicted), we use Cα-only pseudo-dihedral angles as a proxy for backbone geometry. For each interior residue i we compute two torsions from overlapping quadruples of Cα atoms. We get ϕ_i from $(i{-}2, i{-}1, i, i{+}1)$ and ψ_i from $(i{-}1, i, i{+}1, i{+}2)$, and treat the pair as a (ϕ, ψ)-like coordinate, requiring $L \geq 5$. We form a 2-D histogram over these pairs. We define *non-canonical* residues as those falling outside the canonical α-helical ($\phi \in [-100, -30]$, $\psi \in [-67, -7]$) and β-sheet ($\phi \in [-170, -60]$, $\psi \in [90, 180]$) regions.

We compare four populations (Table 8): ground-truth CDR-H3 loops from SAbDab (n = 4,784 pseudo-dihedral observations over all CDR-H3 loops before sequence-level deduplication); a reference set of general PDB loops drawn from PDB70 (Steinegger et al., 2019) (n = 5,000 loops, randomly sampled); and predictions from our best model

(*noRama*) and the ablated model trained *with* the Ramachandran loss (*full_v3*). Both predicted sets comprise 5 samples for each of the 192 validation sequences ($n = 960$ pseudo-dihedral observations). Distributional distance is quantified by $D_{\mathrm{KL}}(\mathrm{row}\|\mathrm{col})$ and by sliced Wasserstein-1 distance (degrees), averaged over 200 random projection directions.

			KL divergence ($D_{\mathrm{KL}}(\mathrm{row}\|\mathrm{col})$)				Sliced Wasserstein-1 (degrees)			
Population	n	**Other (%)**	**→GT**	**→PDB70**	**→noRama**	**→withRama**	**→GT**	**→PDB70**	**→noRama**	**→withRama**
GT CDR-H3	4,784	75.2	—	0.766	0.415	0.470	—	24.7	9.3	12.5
PDB70 gen. loops	5,000	50.3	1.151	—	1.237	1.213	24.7	—	25.3	22.8
Pred. (noRama)	960	73.7	0.377	0.678	—	0.278	9.3	25.3	—	13.2
Pred. (withRama)	960	75.2	0.473	0.864	0.343	—	12.5	22.8	13.2	—

Table 8: Pseudo-Ramachandran analysis of CDR-H3 loop populations using Cα-only pseudo-dihedral angles. *Other*: fraction of residues outside canonical α/β regions. GT observations cover all CDR-H3 loops in SAbDab before sequence-level deduplication. Prediction sets use the 192-sequence validation split, 5 samples each.

G Holdout Evaluation Details

G.1 Per-Length Baseline Comparison

Table 9 gives the per-length breakdown of Kabsch RMSD for all methods on the common 69-structure holdout.

Method	**L5–8**	**L9–11**	**L12–15**	**L16–19**	**L20+**
ColabFold	1.49 ±0.52	0.92 ±1.09	1.89 ±0.94	2.89 ±0.93	2.69 ±1.20
↪ oracle	0.90 ±0.78	0.84 ±1.06	1.64 ±0.92	1.94 ±0.87	2.58 ±1.21
ESMFold	1.49 ±0.47	0.83 ±1.13	1.73 ±1.21	2.38 ±1.06	3.32 ±1.73
↪ oracle	1.49 ±0.47	0.83 ±1.13	1.73 ±1.21	2.38 ±1.06	3.32 ±1.73
tFold	0.72 ±0.61	1.32 ±0.46	1.77 ±0.76	2.79 ±0.78	4.25 ±1.07
↪ oracle	0.72 ±0.61	1.32 ±0.46	1.77 ±0.76	2.79 ±0.78	4.25 ±1.07
IgFold	1.16 ±0.56	1.42 ±2.58	1.77 ±2.71	2.58 ±1.54	3.38 ±1.90
↪ oracle	0.90 ±0.67	1.23 ±1.58	1.77 ±1.85	2.36 ±1.28	3.16 ±1.53
Chai	1.35 ±0.73	2.06 ±1.00	2.52 ±1.13	2.76 ±1.23	3.69 ±1.54
↪ oracle	1.30 ±0.63	0.91 ±1.05	1.89 ±1.14	2.50 ±1.11	2.95 ±1.43
Ours (rand)	0.89 ±0.33	1.66 ±0.78	2.22 ±0.89	3.43 ±1.05	4.74 ±1.21
Ours (oracle)	0.63 ±0.15	1.38 ±0.58	1.62 ±0.74	2.50 ±0.52	3.42 ±0.86
n	4	10	21	18	16

Table 9: Per-length Kabsch RMSD (median $\pm$ std, Å) on the common 69-structure SAbDab2 holdout. For each method, *rand* is the top-ranked prediction and *oracle* is the best of up to 5 samples, computed consistently across all methods. ESMFold and tFold are deterministic (oracle = rand). n: structures per length bin.

G.2 Full Holdout Gap Analysis

Table 10 extends the summary in Table 5 with the complete breakdown across all five hypothesis-testing sections: per loop length, holo/apo status, crystal contacts (with per-distance-bin detail), cysteine-containing loops, sequence identity to training (with per-quartile length metadata), cross-set sequence identity, and oracle improvement. Figure 6 properly visualizes these hypothesis-testing sections and compares the SAbDab2 temporal holdout against the internal test set along the axes examined in Section 4.3

Hypothesis / Subgroup	Test set			SAbDab2 holdout			Δ rand
	n	Rand (Å)	Oracle (Å)	n	Rand (Å)	Oracle (Å)	(Å)
Overall	162	1.937	1.448	71	2.873	2.448	+0.936
Per loop length: does the gap scale with length?							
L5–8	13	0.683	0.663	4	0.890	0.627	+0.207
L9–11	41	1.069	0.931	10	1.659	1.535	+0.591
L12–15	60	2.064	1.624	23	2.193	2.139	+0.129
L16–19	26	2.376	2.133	18	3.427	2.753	+1.051
L20+	22	3.760	2.761	16	4.740	3.776	+0.980
Holo/apo: does antigen contact explain the gap?							
Holo (42% / 83%)	68	1.962	1.417	59	2.873	2.448	+0.911
Apo (9% / 17%)	15	1.794	1.403	12	2.805	2.351	+1.011
Crystal contacts: do packing artifacts explain the gap?							
Has any contact (67% / 42%)	108	1.997	1.498	30	2.984	2.691	+0.987
No contact	53	1.754	1.245	41	2.727	2.235	+0.973
Min dist <5 Å	50	2.198	1.676	14	3.899	3.110	+1.701
Min dist 5–10 Å	48	1.504	1.144	14	2.377	2.395	+0.873
Min dist 10–15 Å	39	1.606	1.215	18	2.966	2.685	+1.361
Min dist >15 Å (mean: 9.3 / 12.7 Å)	24	1.988	2.047	23	2.669	2.235	+0.682
Cys loops: do disulfide bonds explain the gap?							
Has Cys (2.5% / 12.7%)	4	1.900	1.826	9	4.965	3.871	+3.064
No Cys	158	1.937	1.442	62	2.698	2.394	+0.761
No-Cys per length:							
L5–8	13	0.683	0.663	4	0.890	0.627	+0.207
L9–11	41	1.069	0.931	10	1.659	1.535	+0.591
L12–15	58	2.064	1.624	21	2.193	2.139	+0.129
L16–19	25	2.402	2.164	16	3.277	2.692	+0.875
L20+	21	3.737	2.605	11	3.547	3.220	−0.190
Has-Cys per length:							
L12–15	2	1.900	1.754	2	2.059	1.920	+0.158
L16–19	1	–	–	2	5.377	3.238	–
L20+	1	–	–	5	5.533	4.096	–
Seq. identity to training: does novelty explain the gap?							
Q1 <0.47 (mean L: 17.1 / 19.0; L≥16: 64% / 71%)	50	2.362	2.192	7	3.138	3.382	+0.776
Q2 0.47–0.55 (mean L: 14.0 / 15.2; L≥16: 28% / 50%)	32	2.364	1.981	18	3.085	2.577	+0.721
Q3 0.55–0.62 (mean L: 11.0 / 15.4; L≥16: 2% / 48%)	41	1.309	1.003	23	2.505	2.343	+1.196
Q4 >0.62 (mean L: 12.0 / 14.8; L≥16: 15% / 39%)	39	1.396	1.215	23	2.763	2.563	+1.367
Cross-set sequence identity (best match, T→Tr / H→T / H→Tr):							
L5–8	13	0.571	—	4	0.412	0.417	−0.159
L9–11	41	0.600	—	10	0.477	0.551	−0.123
L12–15	60	0.571	—	23	0.385	0.533	−0.187
L16–19	26	0.438	—	18	0.314	0.441	−0.123
L20+	22	0.386	—	16	0.244	0.386	−0.142
Oracle improvement (rand → oracle): sampling or model problem?							
Overall	162	1.937 (25.3%)	1.448	71	2.873 (14.8%)	2.448	+0.936
L5–8	13	0.683 (2.9%)	0.663	4	0.890 (29.6%)	0.627	+0.207
L9–11	41	1.069 (12.8%)	0.931	10	1.659 (7.5%)	1.535	+0.591
L12–15	60	2.064 (21.3%)	1.624	23	2.193 (2.5%)	2.139	+0.129
L16–19	26	2.376 (10.2%)	2.133	18	3.427 (19.7%)	2.753	+1.051
L20+	22	3.760 (26.6%)	2.761	16	4.740 (20.3%)	3.776	+0.980

Table 10: Systematic analysis of the test–holdout performance gap (*noRama_v3*, $N = 5$ samples, Kabsch RMSD Å, median). Each section tests one hypothesis by splitting both sets into subgroups. *Rand*: single sample. *Oracle*: best of 5 (dashes where not computed). Δ: holdout−test rand; positive = holdout harder. Subgroup label percentages are test / holdout fractions. For Has-Cys per length, test subgroups with $n = 1$ have no median (reported as –) and Δ is not reported. Cross-set identity columns report best-match sequence identity: T→Tr (test to training), H→T (holdout to test), H→Tr (holdout to training).

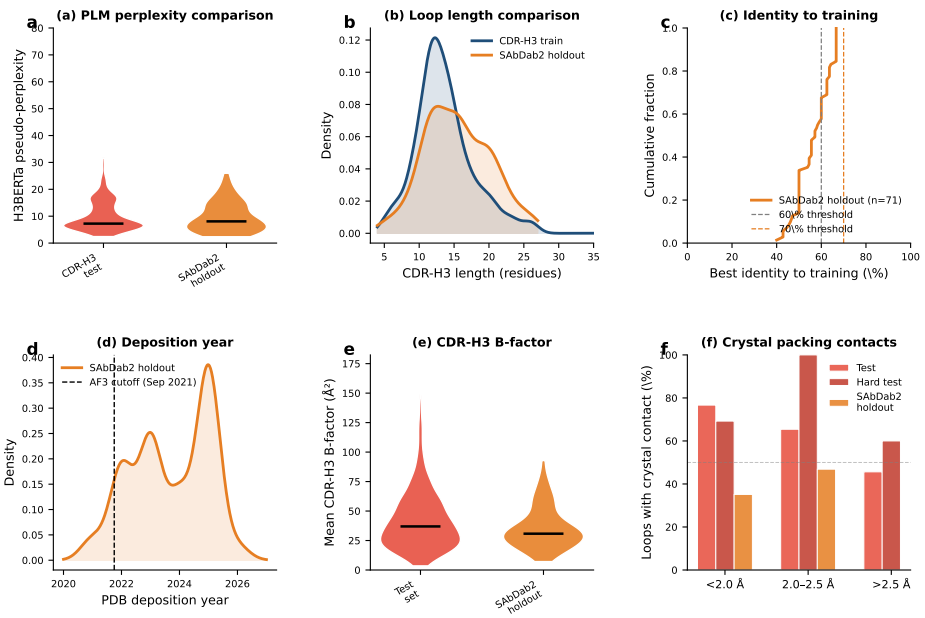


Figure 6: Characterization of the SAbDab2 temporal holdout relative to the internal test set. **(a)** H3BERTa pseudo-perplexity is comparable between the two sets, indicating the holdout is not sequence-anomalous under the encoder. **(b)** The holdout is shifted toward longer loops than the training distribution. **(c)** Cumulative best sequence identity to training for all 71 holdout loops, with the 60% and 70% thresholds marked. **(d)** PDB deposition year, with the AlphaFold3 training cutoff (September 2021) marked. **(e)** Mean CDR-H3 B-factor. **(f)** Fraction of loops with a crystal packing contact, stratified by structure resolution.

H Inference Compute

Table 11 reports wall-clock inference time on the SAbDab2 n=69 holdout for five external baselines and for H3Fold, measured on the NCSA Delta cluster using a single NVIDIA A40 GPU. All timings include model load and I/O, are computed as chunk-wall-clock / n_{chunk}, and are reported as the mean over chunks. External baselines were run in four chunks of 20/20/20/9 structures over n=69 unique Fv, each with 5 diffusion samples per Fv where applicable.

Model	Wall-clock time (s/Fv)
IgFold	4.8 on a single NVIDIA A40 GPU
tFold-Ab	48.6 on a single NVIDIA A40 GPU
ESMFold v1	125.6 on a single NVIDIA A40 GPU
Chai-1	155.5 on a single NVIDIA A40 GPU
LocalColabFold (MSA + fold)	186.7 on a single NVIDIA A40 GPU
H3Fold (ours)	0.70 on a single NVIDIA A40 GPU

Table 11: Inference wall-clock time per antibody Fv. All diffusion models ran 5 samples per loop and ran on the NCSA Delta HPC.

I AI Disclosure Statement

In this work, we used generative AI tools to aid or polish writing, write and execute code, and for discovery (e.g., finding related work). We have not used generative AI tools to generate synthetic data, propose or refine hypotheses, design research methodology or experiments, clean and reformat real datasets, or interpret results. The rest of the required disclosure tasks, such as formulate mathematical claims, prove mathematical claims, assist in the writing of proofs, etc. are not applicable to this work. Additionally, we used generative AI tools to create and edit software, draft portions of the manuscript, analyze existing literature, and format references. Both authors have reviewed all AI-assisted work extensively via manual code reviews, code tracing, manual literature surveys of all cited sources.

We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.