# NANOIF: INTERACTION-AWARE NANOBODY INVERSE FOLDING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Nanobody inverse folding is limited by scarce VHH structures and the underuse of supervision provided by VHH–antigen complexes. We introduce NanoIF, an interaction-aware model for nanobody inverse folding that learns from VHH–antigen complexes during training and requires only the VHH backbone at inference. It first adapts to the VHH sequence–structure distribution and then incorporates antigen information through a dedicated antigen encoder, layer-wise cross-attention, and complex-derived geometric and contrastive supervision. On a SAbDab-nano-derived benchmark, NanoIF reaches state-of-the-art amino-acid recovery, improving paratope recovery by 3.2 points over the strongest baseline, with larger gains on CDR1 and CDR2 (+9.7 and +8.6 points). Controlled training shows that matched VHH-only continued training does not reproduce these gains, while the learned improvement is retained at antigen-free inference. These results establish complex-guided training as an effective strategy for fixed-backbone nanobody sequence design.

## 1 INTRODUCTION

Nanobodies, or VHHs, are the antigen-binding variable domains of camelid heavy-chain-only antibodies. Their small, single-domain architecture offers high stability, efficient microbial expression, good tissue penetration, and convenient engineering into multivalent or multispecific formats (Muyldermans, 2013; Tsumoto & Takeuchi, 2024). These advantages have already translated into therapeutic applications, with several VHH-based drugs, including caplacizumab (Scully et al., 2019), envafolimab (Markham, 2022), and ozoralizumab (Tsumoto & Takeuchi, 2024), receiving regulatory approval in different jurisdictions. Structure-based binder design provides a natural route for designing such molecules: a target-conditioned generative model first proposes a candidate backbone, after which inverse folding assigns amino-acid sequences compatible with the designed structure (Watson et al., 2023; Vázquez Torres et al., 2024; Bennett et al., 2026). The reliability of this pipeline therefore depends critically on whether inverse folding can recover sequences that faithfully realize the intended backbone.

Inverse folding is particularly challenging for nanobodies because their complementarity-determining regions (CDRs), especially CDR3, exhibit substantially greater sequence and structural variability than typical protein regions (Regep et al., 2017). The VHH paratope, however, is not a CDR3-only entity: a structural survey of 105 VHH–antigen complexes attributes only about half of the antigen contacts to CDR3, with CDR1 and CDR2 together contributing roughly a third (Zavrtanik et al., 2018). General inverse-folding models such as ProteinMPNN (Dauparas et al., 2022) and ESM-IF (Hsu et al., 2022) have therefore been specialized to antibodies and nanobodies, leading to methods such as AbMPNN (Dreyer et al., 2023), AntiFold (Høie et al., 2025), and nanoFOLD (Chomicz et al., 2025). Despite these advances, accurate design of the highly variable VHH loops remains difficult, in part because the number of experimentally resolved VHH structures available for specialized training is small (Schneider et al., 2022).

Importantly, many of these scarce structures are resolved in complex with their antigens. Such complexes contain more than VHH sequence–structure compatibility: whereas an isolated VHH primarily reflects the *intrinsic* constraints that stabilize the nanobody itself, a bound complex additionally reveals the geometry of the CDR loops in the bound state, interface packing, and the local antigen environment. Existing methods make only partial use of this additional information. Pro-



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.

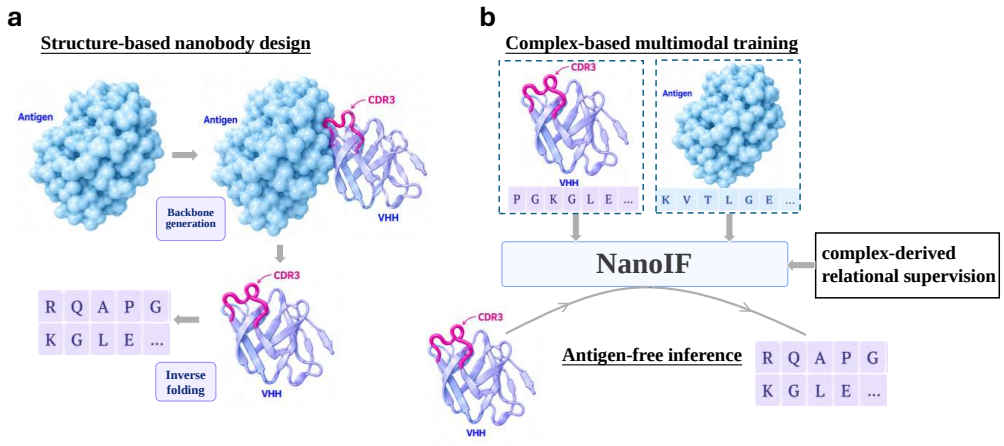


Figure 1: Motivation and overview of NanoIF. VHH-only data provide intrinsic sequence–structure supervision, whereas paired VHH–antigen complexes provide additional antigen context and complex-derived supervision. NanoIF first adapts a multimodal protein language model to the VHH sequence–structure distribution and then performs complex-guided multimodal training. The resulting model performs fixed-backbone inverse folding from the VHH backbone alone, without requiring antigen input at inference.

teinMPNN can include neighboring antigen chains as multichain context (Dauparas et al., 2022), while AbMPNN supplies epitope coordinates during antibody inverse folding (Dreyer et al., 2023). AntiFold is trained on a mixture of solved antibody–antigen complexes and predicted unbound structures (Høie et al., 2025), but does not isolate the contribution of complex supervision itself. Antigen-conditioned generative methods such as DiffAb (Luo et al., 2022) and RFdiffusion (Watson et al., 2023; Bennett et al., 2026) also learn from complexes, but focus on CDR or backbone generation rather than fixed-backbone inverse folding.

Recent benchmarks further show that simply supplying antigen context at inference yields only modest and model-dependent improvements (Li et al., 2025; Høie et al., 2025; Janusz et al., 2026). This suggests that the value of complex data may not reduce to explicit antigen conditioning at inference. A more fundamental question is what training on the bound complex itself contributes to nanobody inverse folding, separately from two simpler effects: access to the antigen identity and access to additional VHH structural examples. This motivates the question we study here: *can training on VHH–antigen complexes improve nanobody inverse folding, and if so, what aspect of complex training drives the improvement?*

To answer this question, we introduce **NanoIF**, a multimodal protein language model that learns nanobody inverse folding through complex-guided training on VHH–antigen complexes. NanoIF builds upon the pretrained sequence–structure representations of DPLM-2 (Wang et al., 2025). Because VHH–antigen complex structures are substantially scarcer than sequence or isolated-chain data, we adopt a two-stage training strategy: we first adapt DPLM-2 to the VHH sequence–structure distribution, and then perform complex-guided multimodal training on VHH–antigen complexes. During the second stage, NanoIF combines a dedicated antigen encoder and layer-wise conditional cross-attention with complex-derived geometric epitope supervision and antigen–CDR contrastive alignment. On a SAbDab-nano-derived benchmark (Dunbar et al., 2014; Schneider et al., 2022), NanoIF achieves 81.3% full-length AAR and 47.1% paratope AAR, improving over the strongest baseline by 4.0 and 3.2 points, respectively. The largest regional gains occur on CDR1 and CDR2 (+9.7 and +8.6 points), while CDR3 improves by 1.0 point to 39.4%. The same checkpoint also achieves the lowest C$\alpha$, CDR3, and paratope refolding RMSD among the compared methods (1.77/4.32/3.61 Å). Controlled training experiments further show that matched VHH-only continued training does not reproduce the Stage-2 gain and that the learned improvement persists without antigen input at inference.

Figure 1 summarizes the motivation and overall training strategy.

**Contributions.** NanoIF (i) achieves state-of-the-art nanobody inverse-folding recovery, with its largest gains on CDR1, CDR2, and the paratope, together with the lowest C$\alpha$, CDR3, and paratope refolding RMSD among the compared methods; (ii) introduces a two-stage complex-guided training framework that combines a dedicated antigen encoder and layer-wise conditional cross-attention with geometric epitope supervision and antigen–CDR contrastive alignment; and (iii) demonstrates through controlled analyses that antigen–antibody complex information provides valuable training signals beyond isolated VHH sequence–structure data, with improvements arising from multiple sources of interaction-aware information.

## 2 METHOD

### 2.1 PROBLEM FORMULATION

Let a VHH be represented by its amino-acid sequence $\mathbf{s}^V = (s_1^V, \ldots, s_L^V)$ and per-residue discrete structure tokens $\mathbf{z}^V = (z_1^V, \ldots, z_L^V)$, $z_i^V \in \mathcal{Z}$, which encode local backbone geometry and are produced by the structure tokenizer of the pretrained backbone (Sec. 2.2). For a paired antigen of $M$ residues, we denote its sequence and structure tokens by $\mathbf{s}^A$ and $\mathbf{z}^A$, and each antigen residue carries a binary epitope annotation $\mathbf{e}^A \in \{0,1\}^M$ derived from the native bound complex: an antigen residue is labeled as an epitope residue if any of its atoms lies within $5.0\,\text{Å}$ of any VHH atom; paratope annotations are obtained by the same rule on the VHH side.

**Task.** Our primary task and evaluation target is VHH inverse folding, i.e., recovering the sequence from its backbone structure:

$$p_\theta\big(\mathbf{s}^V \mid \mathbf{z}^V\big)\,. \tag{1}$$

**Two-stage training.** Rather than optimizing inverse folding in isolation, we adopt a two-stage training strategy that progressively couples sequence, structure, and antigen context. All objectives share the masked discrete-modeling paradigm of the pretrained backbone: a single network is trained to predict corrupted tokens from their uncorrupted context, so each masking pattern realizes a different distribution over $(\mathbf{s}^V, \mathbf{z}^V)$—corrupting amino-acid tokens alone yields inverse folding, corrupting structure tokens alone yields folding, corrupting both yields joint generation, and fully corrupting one modality while preventing information flow from the other modality yields the corresponding unconditional marginal distribution. No task-specific heads or separate models are required.

**Stage 1 (antigen-free adaptation)** proceeds in two sub-phases. *(1a) Sequence adaptation.* The model is trained on unconditional sequence modeling, $p_\theta(\mathbf{s}^V)$, to adapt the pretrained prior to the VHH sequence distribution. *(1b) Sequence–structure coupling.* Training extends to the joint distribution $p_\theta(\mathbf{s}^V, \mathbf{z}^V)$, whose masking patterns induce unconditional marginals $p_\theta(\mathbf{s}^V)$ and $p_\theta(\mathbf{z}^V)$ together with folding $p_\theta(\mathbf{z}^V \mid \mathbf{s}^V)$ and inverse folding $p_\theta(\mathbf{s}^V \mid \mathbf{z}^V)$. Modeling the marginals alongside the conditionals prevents the two modalities from being learned only through their mutual dependence.

**Stage 2 (complex-guided adaptation).** The multimodal objective is extended to VHH–antigen complexes by conditioning on antigen context $\mathcal{C} = (\mathbf{s}^A, \mathbf{z}^A, \mathbf{e}^A)$ :

$$p_\theta\big(\mathbf{s}^V, \mathbf{z}^V \;\big|\; \mathcal{C}\big)\,, \tag{2}$$

inducing antigen-conditioned folding $p_\theta(\mathbf{z}^V \mid \mathbf{s}^V, \mathcal{C})$, inverse folding $p_\theta(\mathbf{s}^V \mid \mathbf{z}^V, \mathcal{C})$, and joint generation $p_\theta(\mathbf{s}^V, \mathbf{z}^V \mid \mathcal{C})$.

Because unconditional inverse folding is already established in Stage 1b, the model retains the ability to inverse-fold VHH backbones without any antigen input; antigen context serves as a training-time conditioning signal rather than a required input at inference time. Moreover, the antigen encoder consumes per-residue structure tokens rather than raw complex coordinates and therefore contains no explicit representation of the VHH–antigen relative pose: bound-state interaction information is injected through the epitope annotations $\mathbf{e}^A$ and a differentiable geometric epitope objective (Sec. 2.5).

### 2.2 Model Backbone and Adaptation

**NanoIF** is built on DPLM-2 (Wang et al., 2025), a 650M-parameter multimodal diffusion protein language model that jointly represents amino-acid and discrete structure tokens and natively supports sequence generation, structure generation, and joint co-generation within a masked discrete-diffusion framework. DPLM-2's structure tokenizer maps backbone coordinates to per-residue discrete tokens via a GVP-based encoder (Jing et al., 2021) with lookup-free quantization (codebook size $|\mathcal{Z}| = 8192$), and reconstructs coordinates from tokens with an IPA-based decoder (Jumper et al., 2021); we use this tokenizer and its decoder unchanged for both the VHH and the antigen. Training follows the two-stage strategy formulated in Sec. 2.1: antigen-free adaptation to nanobody sequence–structure data (Stage 1), followed by complex-guided training on VHH–antigen complexes with complex-derived supervision (Stage 2).

To preserve the pretrained representation while enabling parameter-efficient adaptation, we freeze the DPLM-2 backbone and use rank-16 LoRA (Hu et al., 2022), $W' = W + BA$, on the query, key, value, and output projections of self-attention and on the intermediate and output projections of each feed-forward block. The token embeddings, language-model head, antigen cross-attention modules, and auxiliary projection heads remain trainable.

### 2.3 Antigen-conditioned Cross-attention

A separate, frozen DPLM-2 encoder (650M parameters; ESM architecture) processes the concatenated antigen structure and amino-acid tokens:

$$H^A = f_\phi\big(\mathbf{s}^A, \mathbf{z}^A\big) \in \mathbb{R}^{M\times d}, \tag{3}$$

where $\phi$ is initialized from the same pretrained DPLM-2 checkpoint as the VHH backbone, receives no LoRA adapters, and is held fixed during Stage 2. Freezing this tower limits overfitting to the comparatively small VHH–antigen complex dataset.

We incorporate bound-state epitope information through a trainable binary embedding table $E_{\mathrm{epi}}$. For antigen residue $j$, its epitope-aware representation is

$$\bar{h}_j^A = h_j^A + E_{\mathrm{epi}}\big(e_j^A\big)\,. \tag{4}$$

The embedding table is zero-initialized. We do not hard-mask non-epitope residues: all retained antigen residues remain visible, while the epitope bit provides a soft additive bias.

At each VHH transformer layer, the current VHH representations attend to the epitope-aware antigen representations:

$$C_l = \mathrm{CrossAttn}_l\big(H_l^V, \bar{H}^A, \bar{H}^A\big)\,, \tag{5}$$

$$\widetilde{H}_l^V = H_l^V + C_l. \tag{6}$$

This layer-wise fusion allows VHH residues, including CDR regions, to access antigen-dependent information at multiple representation depths. When no antigen is provided, the cross-attention branch is bypassed and the model reduces to the nanobody-adapted DPLM-2 backbone.

For long antigens, we retain a window of at most 256 residues. During Stage 2, the window is chosen to maximize epitope coverage, with random tie-breaking between equally scoring windows. The complete architecture is illustrated in Figure 2.

### 2.4 Multimodal Diffusion Training

Both training stages are implemented within a unified masked discrete diffusion framework. Different masking patterns correspond to the conditional generation objectives defined in Sec. 2.1.

During Stage 2, we sample three complex-guided denoising tasks with equal probability within each batch:

$$m \in \{\text{folding}, \text{inverse folding}, \text{joint generation}\}.$$

In folding mode, amino-acid tokens are observed while structure tokens are corrupted; in inverse-folding mode, structure tokens are observed while amino-acid tokens are corrupted; in joint-generation mode, both modalities are corrupted. The resulting objectives correspond to the antigen-conditioned conditional distributions introduced in Eq. 2.

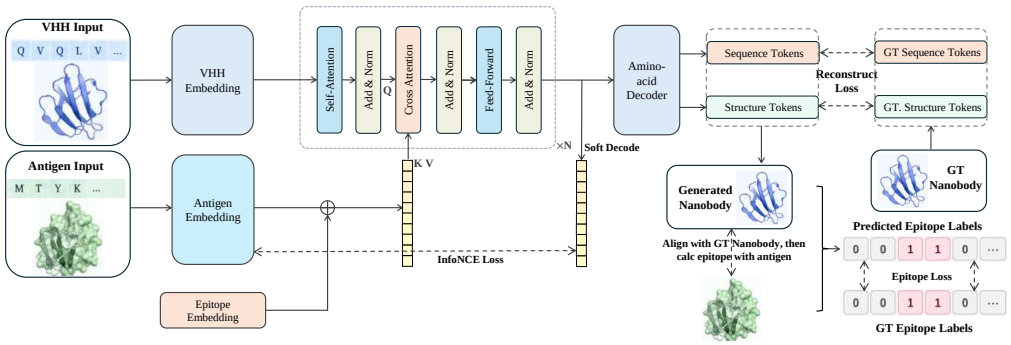


Figure 2: NanoIF architecture and training objectives. A frozen antigen encoder produces epitope-aware antigen representations that are injected into the VHH backbone through layer-wise cross-attention. Training combines multimodal denoising with geometric epitope supervision and antigen–CDR contrastive alignment.

Each modality carries its own diffusion timestep $t$, sampled uniformly from $\{1, \ldots, 500\}$; the modality kept clean under the current mode is assigned $t = 0$. A maskable position is corrupted with probability $(t/500) \cdot \text{scale}$, where scale is a corruption scale (Appendix B.1).

Let $\mathcal{M}_s$ and $\mathcal{M}_z$ denote the corrupted amino-acid and structure positions, respectively. The multimodal diffusion objective is

$$\mathcal{L}_{\text{diff}} = -\mathbb{E}_{m,t}\Bigg[\sum_{i\in\mathcal{M}_s} \log p_\theta\big(s_i^V \,\big|\, \widetilde{\mathbf{s}}_t^V, \widetilde{\mathbf{z}}_t^V, \mathcal{C}\big) + \lambda_{\text{str}} \sum_{i\in\mathcal{M}_z} \log p_\theta\big(z_i^V \,\big|\, \widetilde{\mathbf{s}}_t^V, \widetilde{\mathbf{z}}_t^V, \mathcal{C}\big)\Bigg], \tag{7}$$

where $\mathcal{C}$ denotes the available antigen context and $\lambda_{\text{str}} = 0.5$ balances the sequence and structure objectives. Stage 1 optimizes $\mathcal{L}_{\text{diff}}$ with $\mathcal{C} = \emptyset$; the epitope and contrastive objectives below are active only during Stage 2.

Beyond the task mode, we apply a hybrid masking schedule in which half of the training examples restrict maskable positions to the CDRs; details are given in Appendix B.1. An ablation suggests that this schedule does not explain the reported Stage 2 gain (Appendix B.1).

### 2.5 Geometrically Grounded Epitope Supervision

The antigen encoder consumes per-residue tokens and therefore has no explicit access to the VHH–antigen relative pose. To inject bound-state geometry, we employ a lightweight epitope predictor $g_\psi$ as a frozen geometric critic during Stage 2. The critic takes each antigen residue's identity and its distance to the VHH as input and produces a per-residue epitope logit, providing a differentiable surrogate of the 5 Å contact rule through which bound-state geometry can supervise the generative model (architecture and training details in Appendix B.3). It is fit only on the audited MFDesign training split—never on benchmark complexes—and is unused at inference.

To make the native complex frame usable as supervision, we soft-decode the predicted structure-token distribution into VHH atom37 coordinates $\widehat{X}_\theta^V$ through the expectation of the tokenizer's IPA decoder, keeping the decoding differentiable, and align the prediction to the ground-truth VHH by a Kabsch fit on framework $C_\alpha$ atoms:

$$(R^*, t^*) = \arg\min_{R\in\text{SO}(3),\, t} \sum_{i\in\mathcal{F}} \left\| R\widehat{x}_{i,\alpha}^V + t - x_{i,\alpha}^V \right\|_2^2, \tag{8}$$

where $\mathcal{F}$ denotes the framework residues. This alignment places the predicted VHH into the native complex coordinate frame while preserving differentiability through the rigid transform. The frozen

critic then produces one epitope logit $\ell_j$ per antigen residue from the aligned VHH atoms and the native antigen atom cloud. Because its minimum-distance features are differentiable in the VHH coordinates, gradients flow through the aligned prediction back to the generative model, and we optimize the residue-level binary cross-entropy against the native epitope labels:

$$\mathcal{L}_{\text{epi}} = -\frac{1}{M}\sum_{j=1}^{M}\left[e_j^A \log\sigma(\ell_j) + (1-e_j^A)\log(1-\sigma(\ell_j))\right]. \tag{9}$$

This loss is computed in all three denoising modes whenever the example carries antigen coordinates and epitope labels; in inverse-folding mode, where structure tokens are clean ($t = 0$), gradients still flow from the structure logits through the soft decoding. Unlike the antigen encoder, the critic explicitly uses the native complex coordinate frame during training; we therefore describe this component as bound-state-derived epitope supervision rather than direct relative-pose conditioning.

### 2.6 Antigen–CDR Contrastive Alignment

We further encourage the generated CDR sequence to retain information about its paired antigen. For each VHH residue $i$, we take the softmax of the predicted logits over the 20 canonical amino acids—no argmax is taken, so the pathway remains differentiable end to end—and compute a soft residue embedding as the probability-weighted average of a learnable amino-acid embedding table:

$$u_i^V = \sum_{a=1}^{20} p_\theta(s_i^V = a) E_{\text{AA}}(a), \tag{10}$$

where $E_{\text{AA}}(a) \in \mathbb{R}^H$ is a learnable embedding of amino acid $a$ and $H$ is the hidden size of the DPLM-2 backbone. The soft embedding acts as an "expected amino acid" in continuous space: it coincides with a single embedding when the model is confident and interpolates between embeddings under uncertainty. These representations are mean-pooled over the CDR positions and projected by an MLP ($\text{Linear}(H,H) \to \text{ReLU} \to \text{Linear}(H,H)$) followed by $\ell_2$ normalization, giving a VHH query $q_b$ for complex $b$. The antigen key $k_b$ is the frozen antigen encoder's last-layer hidden states—taken before the epitope bias of Eq. 4—mean-pooled over valid (non-padded) residues, detached, and $\ell_2$-normalized; query and key therefore share dimension $H$ and the key requires no learned projection.

For a minibatch of $B$ paired complexes, the contrastive objective is

$$\mathcal{L}_{\text{NCE}} = -\frac{1}{B}\sum_{b=1}^{B}\log\frac{\exp(\text{sim}(q_b,k_b)/\tau)}{\sum_{b'=1}^{B}\exp(\text{sim}(q_b,k_{b'})/\tau)}, \tag{11}$$

where sim denotes cosine similarity and $\tau = 0.07$. This loss is applied only when amino-acid tokens are corrupted, namely in inverse-folding and joint-generation modes.

### 2.7 Overall Objective and Inference

The complete Stage-2 objective is

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda_{\text{epi}}\mathcal{L}_{\text{epi}} + \lambda_{\text{NCE}}\mathcal{L}_{\text{NCE}}, \tag{12}$$

with $\lambda_{\text{epi}} = 0.2$ and $\lambda_{\text{NCE}} = 0.1$.

At inference time, inverse folding is performed by fixing the VHH structure tokens, masking the amino-acid tokens, and iteratively denoising the sequence. The model supports two inference settings. When an antigen and its epitope annotation are available, they can be supplied to the cross-attention layers. In antigen-free inference, the antigen branch is omitted and generation relies on the VHH sequence–structure prior transferred through Stage-2 adaptation. Unless otherwise specified, we use deterministic argmax decoding.

# 3 EXPERIMENTS

## 3.1 SETUP

**Benchmark.** We evaluate on nano43, a benchmark of 43 VHH–antigen complexes forming the official MFDesign test split (Yang et al., 2025), derived from SAbDab-nano (Dunbar et al., 2014; Schneider et al., 2022).

**Split integrity.** MFDesign constructs its split by clustering CDR-H3 sequences at 50% identity with MMseqs2 (Steinegger & Söding, 2017). We independently reproduced this audit and found zero complex/PDB-ID overlap and no mixed train–test CDR-H3 clusters. Nearest-neighbor checks found no nano43 target with $\geq 90\%$ full-length VHH identity to either Stage 1 or Stage 2 training data, and all Stage 2 CDR3 nearest-neighbor identities were below 70% (Appendix B.2).

**Training data and optimization.** Stage 1 adapts DPLM-2-650M in two steps. Stage 1a performs sequence-only LoRA adaptation on 9,756,679 unpaired NGS VHH training sequences from the AbNGS component of the INDI2 (INDI v2.0, the current release of the Integrated Nanobody Database) (Deszyński et al., 2022)[1] (annotated with ANARCI (Dunbar & Deane, 2016); sequences lacking complete CDR annotations were filtered out). Stage 1b continues from the step-1,000 Stage 1a checkpoint for 1,000 steps on the same corpus mixed with 1,269 MFDesign VHH structures (structure-sample ratio 0.25). Stage 2 fine-tunes on 1,231 MFDesign VHH–antigen training complexes for 1,500 steps. All stages use rank-16 LoRA, AdamW (Loshchilov & Hutter, 2019) with learning rate $10^{-4}$, and bf16 training; full optimization and masking details are given in Appendix B.1.

**Baselines.** General-purpose models (ProteinMPNN (Dauparas et al., 2022), ESM-IF (Hsu et al., 2022)), antibody-specific models (AbMPNN (Dreyer et al., 2023), AntiFold (Høie et al., 2025)), and the nanobody-specific nanoFOLD (Chomicz et al., 2025). We evaluate both authors' released checkpoints: *nanoFOLD-model*, trained on 21,276 NanoBodyBuilder2-predicted nanobody structures, and *nanoFOLD-ft*, further fine-tuned on 435 nonredundant crystallographic nanobody structures. All models are evaluated in the antigen-free setting unless noted; antigen-context variants of baselines change Global AAR by at most 1.8 points and do not alter the ranking (Table A.1). Since nanoFOLD is designed and trained antigen-free, its antigen-conditioned evaluation appends antigen coordinates through the multichain interface of its underlying ESM-IF architecture; this exercises an ESM-IF capability with nanoFOLD weights rather than a nanoFOLD-specific antigen mechanism.

**Metrics.** Amino-acid recovery (AAR) over the full sequence, framework (Fr), CDRs, and the paratope. For structural designability, designed sequences are refolded with NanoBodyBuilder2 (NBB2) (Abanades et al., 2023) and compared to the native backbone by C$\alpha$ RMSD, TM-score (Zhang & Skolnick, 2004), and per-region RMSD.

**Evaluation protocol.** Baselines are deterministic pretrained checkpoints and are evaluated once. All NanoIF results in the main paper use deterministic argmax decoding unless otherwise noted.

## 3.2 SEQUENCE RECOVERY

Table 1 reports amino-acid recovery. The **NanoIF** checkpoint achieves the best recovery in every reported sequence region: 81.3% full-length AAR, 90.1% framework AAR, and 47.1% paratope AAR. Relative to the strongest baseline, the gains are +4.0 points globally and +3.2 points on the paratope, with the largest regional improvements on CDR1 (+9.7) and CDR2 (+8.6). CDR3 improves more modestly by +1.0 point to 39.4%. Starting from the pretrained DPLM-2 backbone (68.5% Global AAR and 32.9% CDR3 AAR), the two-stage adaptation substantially improves nanobody inverse folding performance, yielding gains of +12.8 points globally and +11.7 points on paratope.

## 3.3 STRUCTURAL DESIGNABILITY AFTER REFOLDING

Sequence recovery alone does not guarantee that a designed sequence folds into the intended backbone. We therefore refold designed sequences with NBB2 and compare them with the native backbone using C$\alpha$ RMSD, TM-score, and per-region RMSD (Table 2). Designability is evaluated on a

[1] INDI2 denotes INDI v2.0; see `https://www.naturalantibody.com/indi2/`.

Table 1: Amino-acid recovery (%, ↑) on nano43 under antigen-free inference. Best in **bold**; NanoIF denotes our trained Stage 2 checkpoint, and DPLM-2 is evaluated zero-shot.

| Method | Global | Fr | CDR1 | CDR2 | CDR3 | Paratope |
|---|---|---|---|---|---|---|
| ProteinMPNN | 45.2 | 49.2 | 35.8 | 31.4 | 27.5 | 28.3 |
| ESM-IF | 47.5 | 50.0 | 41.7 | 39.8 | 35.8 | 33.9 |
| AbMPNN | 66.1 | 74.2 | 46.9 | 39.8 | 30.7 | 36.3 |
| AntiFold | 71.9 | 80.2 | 52.5 | 42.3 | 37.7 | 40.1 |
| nanoFOLD-model | 69.5 | 78.1 | 44.5 | 44.8 | 33.4 | 34.9 |
| nanoFOLD-ft | 77.3 | 86.3 | 58.0 | 46.9 | 38.4 | 43.9 |
| DPLM-2 (zero-shot backbone) | 68.5 | 76.9 | 46.0 | 41.5 | 32.9 | 35.4 |
| **NanoIF** (ours) | **81.3** | **90.1** | **67.7** | **55.5** | **39.4** | **47.1** |

Table 2: Structural designability after NBB2 refolding. RMSD in Å (↓), TM (↑). The common evaluable subset contains 41/43 targets after excluding `8r4b_B` and `8r4d_B`, for which NBB2 fails even on the native sequence. $N$ denotes targets with a valid refolded structure; † indicates additional method-specific folding failures.

| Method | C$\alpha$↓ | TM↑ | Fr↓ | CDR3↓ | Para↓ | $N$ |
|---|---|---|---|---|---|---|
| Native (refolded) | 1.82 | 0.952 | 0.80 | 4.81 | 3.82 | 41 |
| **NanoIF** (ours) | **1.77** | **0.946** | **0.82** | **4.32** | **3.61** | **41** |
| nanoFOLD-ft | 1.86 | 0.945 | 0.84 | 4.69 | 3.80 | 41 |
| AntiFold | 2.16 | 0.932 | 0.95 | 5.59 | 4.36 | 41 |
| AbMPNN | 2.22 | 0.939 | 0.85 | 5.90 | 4.68 | 34† |
| ProteinMPNN | 3.49 | 0.903 | 2.50 | 7.12 | 5.74 | 27† |
| ESM-IF | 3.99 | 0.888 | 3.50 | 5.86 | 5.35 | 21† |

fixed subset of 41/43 nano43 targets: NBB2 fails to produce a structure for `8r4b_B` and `8r4d_B` even from their native sequences, so these two targets cannot be scored by the refolding protocol and are excluded before comparing design methods. NanoIF achieves the lowest C$\alpha$ RMSD (1.77 Å), CDR3 RMSD (4.32 Å), and paratope RMSD (3.61 Å) among the compared methods, while also attaining a TM-score of 0.946. Relative to nanoFOLD-ft, these correspond to improvements of 0.09 Å, 0.37 Å, and 0.19 Å in C$\alpha$, CDR3, and paratope RMSD, respectively. NanoIF yields valid refolded structures for all 41 evaluable targets. Beyond inverse folding, the same NanoIF checkpoint supports folding and de novo CDR design without retraining; we defer these secondary generation results to Appendix A.

### 3.4 What Drives the Complex-Training Gains?

We next investigate the source of the complex-guided training gains with controlled training variants (identical architecture, recipe, and budget; antigen-free inference throughout). Table 3 reports the results. Compared with VHH-only continued training, complex-guided training provides substantial gains, motivating a further analysis of the complex-derived signals contributing to this improvement.

First, complex-guided training improves recovery across all three CDRs relative to the Stage-1 checkpoint, increasing CDR1/CDR2/CDR3 recovery from 63.9/51.5/34.4 to 67.7/55.5/39.4, respectively. In contrast, continuing training on the same VHH chains without antigen context yields only marginal changes (64.4/52.8/34.4). These results indicate that the improvement is not explained by additional optimization alone, but arises from information provided by VHH–antigen complexes beyond isolated VHH sequence–structure data.

To understand which components of the complex information contribute, component ablations reveal that the information provided by VHH–antigen complexes is multifaceted. Removing antigen cross-attention, geometric epitope supervision, or antigen–CDR contrastive alignment reduces CDR recovery relative to the full model. These results suggest that complex-guided training benefits from complementary sources of interaction-aware information, including antigen-conditioned feature interaction and complex-derived geometric and relational supervision.

Table 3: Controlled dissection of complex-guided training gains (antigen-free inference; AAR, %).

| Training variant | Global | CDR1 | CDR2 | CDR3 |
|---|---|---|---|---|
| Stage-1 only (*stage2_init*) | 79.6 | 63.9 | 51.5 | 34.4 |
| + matched VHH-only continuation, same recipe (*ab_only*) | 79.7 | 64.4 | 52.8 | 34.4 |
| **NanoIF** full complex training | **81.3** | **67.7** | **55.5** | **39.4** |
| − epitope loss ($\mathcal{L}_{\mathrm{epi}}$) | 80.5 | 66.5 | 54.4 | 37.6 |
| − contrastive loss ($\mathcal{L}_{\mathrm{NCE}}$) | 80.5 | 65.0 | 55.1 | 36.0 |
| − antigen cross-attention (*xattn_off*) | 80.5 | 65.3 | 54.1 | 37.4 |
| shuffled antigen pairing | 80.7 | 65.7 | **55.5** | 38.8 |

Next, we examine the role of correct VHH–antigen pairing. Shuffling antigen pairing preserves the VHH examples and marginal antigen distribution while disrupting pair-specific antigen identity and epitope correspondence. The shuffled variant retains most of the CDR3 improvement (38.8 vs. 39.4), suggesting that exposure to complex context itself provides substantial training signal. Nevertheless, the remaining gap suggests that correctly matched VHH–antigen pairing provides additional information beyond the marginal complex context.

Finally, although antigen representations contribute during training through cross-attention and other complex-guided objectives, the main evaluation is performed without antigen input. This indicates that interaction information learned from complexes can improve antigen-free inverse folding without requiring antigen conditioning at inference.

**Antigen context is not required at inference.** Supplying antigen context at inference yields no consistent improvement in native-sequence recovery for the compared models (Table A.1). For NanoIF, Global AAR is 81.3% without antigen and 80.2% with antigen. Thus the Stage-2 improvement is retained by the VHH model and does not require antigen input at inference; full baseline comparisons are reported in the appendix.

## 4 Discussion and Limitations

This work asks whether training on VHH–antigen complexes improves nanobody inverse folding and what drives the improvement. Complex-guided training outperforms both the Stage-1 checkpoint and matched VHH-only continued training, while the resulting improvement persists at antigen-free inference. This supports the use of solved VHH–antigen complexes as a distinct training resource rather than treating them only as sources of isolated VHH chains.

The distribution of the sequence gains is informative. Relative to the strongest baseline, NanoIF improves CDR1 by 9.7 points, CDR2 by 8.6 points, and paratope recovery by 3.2 points, whereas the CDR3 gain is more modest (+1.0 point). This pattern is consistent with the fact that the VHH paratope is not concentrated exclusively in CDR3 (Zavrtanik et al., 2018) and motivates evaluating nanobody inverse folding beyond CDR3 alone.

Several limitations remain. First, our evaluation benchmark contains only 43 complexes, so the reported point estimates should be interpreted on a relatively small benchmark. Second, structural designability is assessed through in silico refolding with a single structure predictor, which provides a useful proxy for sequence–structure consistency but does not replace experimental validation of binding, expression, or stability. Third, our antigen representation uses a window of at most 256 residues and contains no explicit cross-chain pose, limiting the interaction detail available for large antigens.

Looking forward, the data bottleneck cuts both ways: if complex-derived supervision is useful for nanobody inverse folding, additional solved complexes and experimentally validated interface measurements may provide particularly valuable supervision.

### AI use statement

In this work, we used generative AI tools for: (i) polishing and improving the clarity of the writing (grammar, style, and readability); (ii) retrieval and discovery of related work (searching for and

summarizing prior literature); (iii) research ideation and execution (refining experimental designs); (iv) drafting initial versions of sections of the paper, which we subsequently edited and rewrote; and (v) creating schematic figures and graphical illustrations. We have not used generative AI tools for authorship of the paper, peer review, critique or rebuttal of the work of others, or generating or analyzing experimental data. Generating datasets from scratch is not applicable to this work, as all datasets used are publicly available. Additionally, we used generative AI tools for coding assistance (e.g., writing, debugging, and executing experimental code). We have reviewed all AI-assisted work. Specifically, all AI-generated text was carefully verified, edited, and rewritten by the authors; AI-generated schematic figures were checked for scientific accuracy and revised by the authors; AI-suggested research ideas were validated through a manual literature survey to check for plagiarism and overlap with existing work; and all AI-assisted code was reviewed, tested, and verified for correctness by the authors. We take responsibility for the final content of this work, including text, figures, claims or artifacts produced with the aid of generative AI.

### Reproducibility statement

We have made efforts to ensure that our results are reproducible. All datasets used in our experiments are publicly available: the nano43 benchmark and MFDesign train/validation splits are derived from SAbDab-nano (Dunbar et al., 2014; Schneider et al., 2022); the Stage-1a NGS VHH sequences are from the AbNGS component of INDI2 (INDI v2.0) (Deszyński et al., 2022), which is freely available for non-commercial research. The split construction and our independent sequence-similarity audit are described in Section 3.1 and Appendix B.2. Model architecture and training objectives are specified in Section 2; all hyperparameters, the two-stage training schedule, corruption scales, and masking configurations are reported in Section 3.1 and Appendix B.1. Details of the epitope critic, including its features, training procedure, and held-out validation, are given in Appendix B.3. Baselines are evaluated with the authors' released checkpoints under the protocol of Section 3.1. Source code will be released upon publication.

## References

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.

Nathaniel R. Bennett, Joseph L. Watson, Robert J. Ragotte, Andrew J. Borst, DéJenaé 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.

Dawid Chomicz, Bartosz Janusz, Sonia Wrobel, Pawel Dudzic, Adithya Polasa, Kyle Martin, Steven Darnell, Stephen R. Comeau, and Konrad Krawczyk. nanoFOLD: sequence design of nanobodies via inverse folding. *bioRxiv*, 2025. doi: 10.1101/2025.04.29.651236.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Alexis Courbet, Rob J. de Haas, Neville Bethel, Philip J. Y. Leung, Timothy F. Huddy, Samuel Pellock, Doug Tischer, Frederick Chan, Brian Koepnick, Hannele 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.

Piotr Deszyński, Jakub Młokosiewicz, Adam Volanakis, Igor Jaszczyszyn, Natalie Castellana, Stefano Bonissone, Rajkumar Ganesan, and Konrad Krawczyk. Indi—integrated nanobody database for immunoinformatics. *Nucleic Acids Research*, 50(D1):D1273–D1281, 2022.

Frédéric A. Dreyer, Daniel Cutting, Constantin Schneider, Henry Kenlay, and Charlotte M. Deane. Inverse folding for antibody sequence design using deep learning. In *The 2023 ICML Workshop on Computational Biology*, 2023.

James Dunbar and Charlotte M. Deane. ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics*, 32(2):298–300, 2016. doi: 10.1093/bioinformatics/btv552.

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.

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.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8946–8970. PMLR, 2022.

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*, 2022.

Bartosz Janusz, Dawid Chomicz, Sonia Wrobel, Pawel Dudzic, Adithya Polasa, Kyle Martin, Steven Darnell, Stephen R. Comeau, and Konrad Krawczyk. Benchmarking antigen-aware inverse folding methods for antibody design. *Bioinformatics Advances*, 6(1):vbag081, 2026. doi: 10.1093/bioadv/vbag081.

Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael John Laurent Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. In *International Conference on Learning Representations*, 2021.

John Jumper, Richard Evans, Alexander Pritzel, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Yifan Li, Yuxiang Lang, Chenrui Xu, Yi Zhou, Ziwei Pang, and Per Jr. Greisen. Benchmarking inverse folding models for antibody CDR sequence design. *PLOS ONE*, 20(6):e0324566, 2025. doi: 10.1371/journal.pone.0324566.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

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*, volume 35, pp. 9754–9767, 2022.

Anthony Markham. Envafolimab: First approval. *Drugs*, 82(2):235–240, 2022. doi: 10.1007/s40265-022-01671-w.

Serge Muyldermans. Nanobodies: Natural single-domain antibodies. *Annual Review of Biochemistry*, 82:775–797, 2013. doi: 10.1146/annurev-biochem-063011-092449.

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

Constantin Schneider, Matthew I. J. Raybould, and Charlotte M. Deane. SAbDab in the age of biotherapeutics: updates including SAbDab-nano, the nanobody structure tracker. *Nucleic Acids Research*, 50(D1):D1368–D1372, 2022. doi: 10.1093/nar/gkab1050.

Marie Scully, Spero R. Cataland, Flora Peyvandi, Paul Coppo, Paul Knöbl, Johanna A. Kremer Hovinga, Ara Metjian, Javier de la Rubia, Katerina Pavenski, Filip Callewaert, Debjit Biswas, Hilde De Winter, and Robert K. Zeldin. Caplacizumab treatment for acquired thrombotic thrombocytopenic purpura. *New England Journal of Medicine*, 380(4):335–346, 2019. doi: 10.1056/NEJMoa1806311.

Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35:1026–1028, 2017. doi: 10.1038/nbt.3988.

Kouhei Tsumoto and Tsutomu Takeuchi. Next-generation anti-TNF$\alpha$ agents: The example of ozoralizumab. *BioDrugs*, 38(3):341–351, 2024. doi: 10.1007/s40259-024-00648-3.

Susana Vázquez Torres, Philip J. Y. Leung, Preetham Venkatesh, Isaac D. Lutz, Fabian Hink, Huu-Hien Huynh, Jessica Becker, Andy Hsien-Wei Yeh, David Juergens, Nathaniel R. Bennett, Andrew N. Hoofnagle, Eric Huang, Michael J. MacCoss, Marc Expòsit, Gyu Rie Lee, Asim K. Bera, Alex Kang, Joshmyn De La Cruz, Paul M. Levine, Xinting Li, Mila Lamb, Stacey R. Gerben, Analisa Murray, Piper Heine, Elif Nihal Korkmaz, Jeff Nivala, Lance Stewart, Joseph L. Watson, Joseph M. Rogers, and David Baker. De novo design of high-affinity binders of bioactive helical peptides. *Nature*, 626(7998):435–442, 2024. doi: 10.1038/s41586-023-06953-1.

Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. DPLM-2: A multimodal diffusion protein language model. In *International Conference on Learning Representations*, 2025.

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, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. 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.

Nianzu Yang, Songlin Jiang, Jian Ma, Huaijin Wu, Shuangjia Zheng, Wengong Jin, and Junchi Yan. Repurposing AlphaFold3-like protein folding models for antibody sequence and structure co-design. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0009.

Uroš Zavrtanik, Junoš Lukan, Remy Loris, Jurij Lah, and San Hadži. Structural basis of epitope recognition by heavy-chain camelid antibodies. *Journal of Molecular Biology*, 430(21):4369–4386, 2018. doi: 10.1016/j.jmb.2018.09.002.

Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004. doi: 10.1002/prot.20264.

# A Additional Experimental Results

## A.1 Antigen-conditioned Inference

For completeness, we compare antigen-free and antigen-conditioned inference for the inverse-folding baselines and NanoIF. As shown in Table A.1, providing antigen context changes Global AAR by at most 1.8 points for the baselines and does not alter their ranking. NanoIF likewise does not benefit from antigen input at inference: Global AAR is 81.3% without antigen and 80.2% with antigen. These results support using complex information primarily as a training-time signal rather than as a required inference-time input.

## A.2 Folding

We additionally evaluate the folding mode (sequence $\rightarrow$ structure) on the same MFDesign nano43 benchmark. Given the native VHH sequence, the model generates structure tokens that are decoded into coordinates by the IPA head; Table A.2 reports C$\alpha$ RMSD and TM-score against the native structure, overall and per region. Stage-1 VHH adaptation markedly improves folding over the pre-trained DPLM-2 backbone (C$\alpha$ 2.91 $\rightarrow$ 2.09 Å; CDR3 7.11 $\rightarrow$ 4.63 Å), i.e., adaptation strengthens rather than degrades the backbone's folding capability, and supplying the antigen at inference has only a minor effect—consistent with Section 3.4, where the complex-guided training gains are found to act on the sequence (inverse-folding) side.

Table A.1: Antigen-free vs. antigen-conditioned inference (Global AAR, %). NanoIF denotes the trained Stage 2 checkpoint.

| Model | no-Ag | with-Ag | Δ |
|---|---|---|---|
| ProteinMPNN | 45.2 | 46.4 | +1.2 |
| ESM-IF | 47.5 | 48.8 | +1.3 |
| AbMPNN | 66.1 | 67.9 | +1.8 |
| AntiFold | 71.9 | 72.7 | +0.8 |
| nanoFOLD-ft | 77.3 | 77.0 | −0.3 |
| NanoIF (ours) | **81.3** | 80.2 | −1.1 |

Table A.2: Folding (sequence → structure) on the MFDesign nano43 benchmark. Cα RMSD (Å, ↓) and TM-score (↑) against the native structure, overall and per region. Pretrained DPLM-2 is evaluated zero-shot; NanoIF denotes the trained Stage 2 checkpoint, evaluated without and with antigen conditioning at inference.

| Method | Cα | TM | Fr | CDR1 | CDR2 | CDR3 |
|---|---|---|---|---|---|---|
| Pretrained DPLM-2-650M (zero-shot) | 2.91 | 0.918 | 1.20 | 3.23 | 2.23 | 7.11 |
| NanoIF | **2.09** | **0.924** | **1.19** | **2.89** | 2.07 | **4.63** |
| + antigen at inference | 2.18 | **0.924** | **1.19** | 2.92 | **1.86** | 5.10 |

### A.3 Antigen-conditioned De Novo CDR Design

The antigen-conditioned generation mode enables de novo CDR design given an antigen. Table A.3 compares NanoIF with DiffAb (Luo et al., 2022). NanoIF improves AAR on all three CDRs, most notably on CDR-H1 (+15.9 points) and CDR-H2 (+16.1 points), while CDR-H3 recovery improves only modestly. Refolded RMSD trails DiffAb on all three CDRs, with the largest gap on CDR-H3 (7.49 vs. 4.71 Å), although the overall refolded structures remain close to native (Cα 2.29 Å, TM 0.930).

Table A.3: Antigen-conditioned de novo CDR design. AAR (%, ↑), refolded RMSD (Å, ↓). NanoIF denotes the trained Stage 2 checkpoint.

| | CDR-H1 | | CDR-H2 | | CDR-H3 | |
|---|---|---|---|---|---|---|
| Method | AAR | RMSD | AAR | RMSD | AAR | RMSD |
| DiffAb | 45.2 | **2.52** | 31.1 | **1.54** | 15.4 | **4.71** |
| NanoIF (ours) | **61.1** | 3.51 | **47.2** | 3.09 | **17.3** | 7.49 |

## B Training and Data Details

### B.1 Training Schedule, Corruption Scale, and Hybrid Masking

**Stage-wise training.** Stage 1a uses the mixed-data loader with structure sampling disabled (`struct_sample_ratio=0`), so the update stream is sequence-only and contains 9,756,679 unpaired NGS VHH training sequences (60,439 validation sequences) from the AbNGS component of INDI2 (INDI v2.0) (Deszyński et al., 2022). The Stage 1a recipe is configured for 5,000 steps; we use its step-1,000 checkpoint to initialize Stage 1b. Stage 1b mixes the same NGS corpus with 1,269 MFDesign VHH training structures (65 validation structures), giving 9,757,948 training examples in the combined source, and samples structural examples with ratio 0.25 for 1,000 steps. Stage 2 is initialized from the Stage 1b step-1,000 checkpoint and trains for 1,500 steps on 1,231 VHH–antigen complexes (64 validation complexes after requiring epitope annotations).

Across all three stages we use rank-16 LoRA with dropout 0.1, AdamW (Loshchilov & Hutter, 2019) ($\beta_1 = 0.9$, $\beta_2 = 0.98$, weight decay 0.01), gradient clipping at 0.5, and bf16 precision. The learning rate is $10^{-4}$ with polynomial decay, 200 warmup steps, and end learning rate $10^{-5}$.

Gradient accumulation is 2 for Stages 1a–1b and 8 for Stage 2. Stage 1a optimizes the single-modality objective; Stage 1b samples folding, inverse folding, joint generation, and single-modality modeling with equal weight 0.25; Stage 2 samples folding, inverse folding, and joint generation with equal probability.

**Progressive corruption and hybrid CDR masking.** For the Stage 2 recipe used by the main model, the mask probability at a maskable position is $(t/500) \cdot \mathrm{scale}$, matching Sec. 2.4, where $t \in \{1, \ldots, 500\}$ and scale increases linearly from 0.1 to 1.0 over the first 1,500 optimization steps. CDR positions follow the precomputed ANARCI–Chothia region annotations (Dunbar & Deane, 2016) used by the training pipeline. With `cdr_noise_ratio=0.5`, approximately half of the examples in each batch restrict maskable positions to the CDRs and use the current progressive scale; the remaining examples allow full-sequence masking with scale fixed to 1.0. This schedule changes only which positions can be corrupted and their Bernoulli corruption probability; the denoising objectives themselves are unchanged.

**Hybrid-mask ablation.** To test whether the CDR-focused corruption schedule explains the reported inverse-folding gains, we disable both CDR-restricted masking and the progressive mask schedule (`cdr_generation.enable=false`, `mask_schedule.enable=false`, `cdr_noise_ratio=0`) while keeping the remaining Stage 2 recipe unchanged, including $\lambda_{\mathrm{epi}} = 0.2$, $\lambda_{\mathrm{NCE}} = 0.1$, equal sampling of the three denoising modes, and 1,500 training steps. On nano43 under antigen-free inverse folding with argmax decoding, this *nocdr* variant obtains 81.1% Global AAR and 39.6% CDR3 AAR, compared with 81.3% and 39.4%, respectively, for the full model recipe. The near-identical result indicates that the hybrid CDR masking schedule does not explain the Stage 2 improvement.

Table B.1: Effect of the hybrid CDR masking schedule on nano43 (antigen-free inverse folding, argmax decoding; $N = 43$).

| Training recipe | Global | CDR1 | CDR2 | CDR3 |
|---|---|---|---|---|
| Full recipe (hybrid CDR masking) | 81.3 | 67.7 | 55.5 | 39.4 |
| No CDR-specific schedule (*nocdr*) | 81.1 | 66.4 | 55.9 | 39.6 |

### B.2 Benchmark Split Integrity and Sequence-Similarity Audit

The nano43 benchmark is the nanobody subset of the official MFDesign test partition. MFDesign constructs its split by extracting CDR-H3 sequences and clustering them with MMseqs2 (Steinegger & Söding, 2017) `easy-cluster` at 50% sequence identity; clusters containing structures released before 2021-09-30 are assigned to the training partition, and the remaining clusters are split into train, validation, and test partitions. We reproduced the same CDR-H3 clustering over the Stage 2 training set and nano43 and found no mixed train–test clusters, with zero overlapping complex/PDB identifiers.

We additionally performed nearest-neighbor sequence-identity audits. Across the 43 nano43 targets, the nearest Stage 2 full-length VHH identities had min/median/max values of 63.3/76.5/84.1%; none reached 90%, and 10/43 fell between 80% and 90%. For CDR3, nearest Stage 2 alignment identities were 28.6/41.7/63.6% (min/median/max), with every target below 70%. Against the Stage 1 NGS corpus (∼9.76M sequences), nearest full-length identities were 64.4/73.6/84.4%, again with no target at or above 90% (4/43 fell between 80% and 90%). The higher full-length similarities are consistent with the conserved VHH framework, whereas the CDR3-specific similarities remain substantially lower. Together with the cluster-level split, these audits provide no evidence of benchmark contamination at the complex, CDR-H3-cluster, or high-identity sequence level.

### B.3 Epitope Critic Details

The epitope critic $g_\psi$ (Sec. 2.5) is a small MLP over two per-residue features of each antigen residue: its identity (a 128-dimensional embedding over 21 amino-acid classes) and its minimum distance to any VHH atom, featurized by 16 fixed Gaussian radial basis functions spanning 0–20 Å; the MLP maps these features to a single logit. It is trained with plain binary cross-entropy for 3,000 steps on the MFDesign VHH–antigen training split, with labels given by the 5 Å contact rule. On a held-

out set of 64 complexes (14,286 antigen residues) the trained critic attains an F1 of $0.986$, closely matching the distance rule itself, confirming that it faithfully implements the contact criterion while remaining differentiable end to end. The critic is frozen before Stage 2 and is trained only on the audited MFDesign training partition described in Appendix B.2; it never sees benchmark complexes.

Formally, the soft-decoded VHH coordinates are $\widehat{X}_\theta^V = \mathcal{D}\big(p_\theta(\mathbf{z}^V)\big)$, where $\mathcal{D}$ denotes the expectation of the tokenizer's IPA decoder; with the Kabsch solution of Eq. 8, the aligned prediction is $\widetilde{X}_\theta^V = R^* \widehat{X}_\theta^V + t^*$, and the per-residue epitope logit is $\ell_j = g_\psi\Big(\widetilde{X}_\theta^V, X^A\Big)_j$, where $X^A$ is the native antigen atom cloud. During Stage 2, the epitope loss of each example is computed from that example's own antigen coordinates and labels, so no benchmark supervision enters the gradients.