ADAPROTEINA: AN ADAPTER IS ALL YOU NEED FOR DEVELOPABLE ANTIBODY GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern flow-matching generators for antibody design sample from an unconditional prior and offer no control over the physicochemical properties (binding, humanness, aggregation) that determine therapeutic viability. Retraining the 160M-parameter denoiser under a conditional objective is expensive and must be repeated per property specification. Classifier-free guidance trades sample diversity for conditioning strength and doubles per-step inference cost. We propose AdaProteina, a lightweight noise-space adapter that leaves the denoiser and VAE decoder of La-Proteina Geffner et al. (2026) entirely frozen and instead learns a refiner g_ϕ mapping base Gaussian noise and a class label to a developability-aware initial latent, with $|\phi| << |\theta|$ learnable parameters and no change in per-sample inference cost. Training combines cross-entropy against a frozen developability classifier with a per-class MMD loss that anchors refined latents to the data manifold and prevents the mode collapse, e.g., tyrosine/tryptophan-dominated sequences induced by naïve property optimization. To compose multiple developability properties without training a combinatorial number of joint classifiers, we introduce a Kernel Perron–Frobenius Huang et al. (2022); Klus et al. (2020) transfer operator that maps between per-property noise spaces in closed form, enabling training-free generation of, e.g., human-like low non-specific binders. AdaProteina improves substantially over La-Proteina Geffner et al. (2026) on sequence diversity, developability separability, and every structural-validity metric evaluated and augmenting downstream training. To the best of our knowledge, this is the first large scale exploration of noise-space refinement for developability-aware antibody generation.

1 INTRODUCTION

Generative models for antibody structure and sequence design have converged on a two-stage paradigm inherited from general protein design: specify a backbone conformation, then decorate it with side chains and sequence identity Watson et al. (2023); Geffner et al. (2026). This decomposition is computationally convenient, i.e., shape and chemistry are optimized by separate, specialized machinery and it is biologically motivated for antigen-specific design, where backbone geometry constrains the paratope's spatial complementarity to a defined epitope. However, developability liabilities such as non-specific (polyreactive) binding do not arise from complementarity to any single target; they are governed by more diffuse surface properties, e.g., aromatic and charged-residue content, hydrophobic patch area, and CDR loop flexibility Chen et al. (2024); Makowski et al. (2022) that are only partially, and non-obviously, encoded in backbone shape. Whether a backbone-then-side-chain generative pipeline can meaningfully control this class of property, rather than simply reproducing whatever propensity is already latent in the input scaffold, has not been directly tested.

We take this question as central motivation for our fixed-backbone conditioning setup. If backbone geometry alone strongly predicts non-specific binding class, a side-chain generator conditioned on a class label risks "cheating" — learning to read the label off the backbone rather than performing genuine conditional design — which would make claims of controllable generation illusory precisely in the regime practitioners care about most: redesigning an existing scaffold to tune developability without altering its structural identity. We show this concern is not hypothetical: a ridge regressor on C_α coordinates alone predicts non-specific binding class at $\sim$70% accuracy, and a naïve-ly trained side-chain generator asked to produce the off-class label from a same-class backbone performs no better than chance. This motivates a label-swapping training scheme that decouples side-chain-level

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.

property control from backbone-encoded propensity, and raises the structural question we return to later: *for a given backbone, how biologically achievable is it to realize either a high or a low non-specific binding surface?* Before presenting the method, we briefly situate it among guidance techniques for conditional generation.

Noise optimization over guidance techniques. Guidance techniques were the *de-facto* technique for conditional generation Ho & Salimans (2022b); Chung et al. (2025); Ahn et al. (2024); Berrada et al. (2026). However guidance techniques often lead to off-manifold/out-of-distribution sample generation by forcing the trajectory off the desired path Park et al. (2025). Course-correction of the off-manifold trajectory due to the guidance vector often involves projection Roy et al. (2026) and second order techniques Roy et al. that are computationally expensive and require approximation of the unknown latent manifold. These prior works also highlight the necessity to respect the natural denoising trajectory to reach the data manifold while ensuring conditional generation and maintaining sample diversity that are not possible with guidance techniques. *Thus optimizing the starting point, i.e., the noise space, without changing the natural path, i.e., the denoising trajectory, has emerged as the most effective way.* Prior works explore the controllability of the generation process through initial noise Eyring et al. (2024); Guo et al. (2024); Samuel et al. (2024); Sundaram et al. (2025). It has also been observed that global behaviour is controlled by specific noise seeds Xu et al. (2025). However, the most popular approach is to utilize best-of-n sampling approaches Karthik et al. (2023); Kirstain et al. (2023); Ma et al. (2025); Ramesh et al. (2021); Razavi et al. (2019) or direct noise optimization approaches. Recent works Harrington et al. (2026); Kim et al. (2024) have highlighted that simple noise optimization is a better and highly effective approach in terms of computation for generating high quality samples. We follow the direction of noise optimization to introduce the developability property in the noise space for efficient and computationally effective antibody sequence generation. *To the best of our knowledge, this is the first work to explore noise space refinement for developability-property aware antibody sequence generation.*

2 METHOD

2.1 MOTIVATION AND PROBLEM SETUP

La-Proteina Geffner et al. (2026) factorizes the protein distribution as

$$p(x_{\mathrm{Ca}}, x_{\overline{\mathrm{Ca}}}, s, z) \;=\; \underbrace{p_\theta(x_{\mathrm{Ca}}, z)}_{\text{partially latent flow}} \cdot \underbrace{q_\vartheta(x_{\overline{\mathrm{Ca}}}, s \mid x_{\mathrm{Ca}}, z)}_{\text{VAE decoder}}, \tag{1}$$

where $x_{\mathrm{Ca}} \in \mathbf{R}^{L\times 3}$ are the α-carbon coordinates, $z \in \mathbf{R}^{L\times 8}$ is a per-residue latent encoding sequence and side-chain detail, $x_{\overline{\mathrm{Ca}}} \in \mathbf{R}^{L\times 36\times 3}$ collects the remaining heavy-atom coordinates, and $s \in \{0,\dots,19\}^L$ is the amino-acid sequence. The sampling step proceeds in two stages: (i) (x_{Ca}, z) is drawn from the partially latent flow p_θ (ii) $(x_{\overline{\mathrm{Ca}}}, s)$ is decoded by q_ϑ conditioned on that drawn sample. The flow itself is realized by a denoiser v_θ trained under the conditional flow matching (CFM) Tong et al. (2023) objective $\min_\theta \; \mathbb{E}\Big[\big\|v_\theta^x - (x_{\mathrm{Ca}} - x_{\mathrm{Ca}}^0)\big\|^2 + \big\|v_\theta^z - (z - z^0)\big\|^2\Big]$, where $(v_\theta^x, v_\theta^z) = v_\theta(x_{\mathrm{Ca}}^{t_x}, z^{t_z}, t_x, t_z)$, (x_{Ca}, z) is the data pair, $(x_{\mathrm{Ca}}^0, z^0) \sim \mathcal{N}(0, I)$ is the initial noise and the expectation is taken over data and independent time variables (t_x, t_z). Decoupling t_x and t_z is a deliberate design choice that lets the backbone and the latents be integrated on different schedules; empirically, generating the backbone faster than the latents ($t_x > t_z$) yields the strongest results Geffner et al. (2026). At inference, samples are drawn by numerically integrating the coupled stochastic differential equations

$$dx_{\mathrm{Ca}}^{t_x} = v_\theta^x\, dt_x + \beta_x(t_x)\,\zeta^x\, dt_x + \sqrt{2\beta_x(t_x)}\,\eta_x\, d\mathcal{W}_{t_x}, \tag{2}$$

$$dz^{t_z} = v_\theta^z\, dt_z + \beta_z(t_z)\,\zeta^z\, dt_z + \sqrt{2\beta_z(t_z)}\,\eta_z\, d\mathcal{W}_{t_z}, \tag{3}$$

where $\zeta^x \approx \nabla_{x_{\mathrm{Ca}}^{t_x}} \log p_{t_x}\big(x_{\mathrm{Ca}}^{t_x}\big)$ and $\zeta^z \approx \nabla_{z^{t_z}} \log p_{t_z}\big(z^{t_z}\big)$ are score estimates of the two (independent) marginals, obtained algebraically from v_θ via the standard velocity–score identity for linear interpolants Ma et al. (2024); Zheng et al. (2023): $\zeta^x = \frac{t_x\, v_\theta^x - x_{\mathrm{Ca}}^{t_x}}{1-t_x}$, $\zeta^z = \frac{t_z\, v_\theta^z - z^{t_z}}{1-t_z}$. Here $t_x, t_z \in [0,1]$ run from noise ($t=0$) to data ($t=1$), consistent with Eq. equation 2.1, and $\mathcal{W}_{t_x}, \mathcal{W}_{t_z}$ are independent standard Wiener processes in $\mathbf{R}^{L\times 3}$ and $\mathbf{R}^{L\times 8}$, respectively. The scaling

functions β_x, β_z modulate the Langevin correction term and are set to $\beta_x(t_x) = \frac{1}{t_x}, \qquad \beta_z(t_z) = \frac{\pi}{2}\tan\left(\frac{\pi}{2}(1-t_z)\right)$, following Geffner et al. (2026). The noise scales $\eta_x, \eta_z \in [0,1]$ tune the diversity–co-designability tradeoff: $\eta_x = \eta_z = 1$ yields unbiased sampling from the flow's true marginals for any choice of β_x, β_z Karras et al. (2022), while $\eta_x, \eta_z < 1$ performs low-temperature sampling, trading diversity for higher co-designability. The two SDEs are coupled only through the shared denoiser v_θ, which conditions each channel's drift on the current state of both.

Motivation: Eq. 1 defines an *unconditional* prior, agnostic to the physicochemical properties (binding affinity, thermostability, aggregation propensity, hydrophobic-patch score (PSH), charge asymmetry (PPC)) that jointly determine whether an antibody is viable as a therapeutic. This in turn makes the drawn samples useless for downstream tasks, as essentially we are generating samples without any control of what the samples are, i.e., *we don't know if the generated samples are human-like or if they are strong binders*. In the remainder of the paper, we will call these conditions "developability properties". This is the motivation of this work, where our goal is to draw samples from the class-conditional distribution $p(x_{\text{Ca}}, x_{\overline{\text{Ca}}}, s, z \mid y)$ that preserves La-Proteina's structural priors while shifting mass toward the desired class, at no additional inference cost.

A Naïve solution: A straightforward approach of retraining the denoiser with conditioning, then applying classifier-free guidance at inference: (i) Conditioning the denoiser on a property vector y requires retraining the full 160M-parameter network under the conditional CFM objective $\min_\theta \ \mathbb{E}_y\Big[\big\|v_\theta^x\big(x_{\text{Ca}}^{t_x}, z^{t_z}, t_x, t_z, y\big) - \big(x_{\text{Ca}} - x_{\text{Ca}}^0\big)\big\|^2 + \big\|v_\theta^z\big(x_{\text{Ca}}^{t_x}, z^{t_z}, t_x, t_z, y\big) - \big(z - z^0\big)\big\|^2\Big]$, where v_θ now takes the property vector y as an additional input (in contrast to the unconditional denoiser of Eq. 2.1), and y is dropped to a null token $\emptyset$ with some probability during training so the same network learns both the conditional $v_\theta^{x/z}(\cdot, y)$ and unconditional $v_\theta^{x/z}(\cdot, \emptyset)$ branches, following the standard classifier-free training recipe Ho & Salimans (2022a); (ii) classifier-free guidance Ho & Salimans (2022a) then replaces each velocity in Eqs. 2–3 by the guided combination $\tilde{v}_\theta^{x/z} = (1+w)\, v_\theta^{x/z}(\cdot, y) - w\, v_\theta^{x/z}(\cdot, \emptyset)$,, where $w \geq 0$ is the guidance scale ($w = 0$ recovers the plain conditional model; larger w pushes samples more aggressively toward high-y likelihood at the cost of sample diversity), and the score estimates $\zeta^{x/z}$ in Eqs. 2–3 are recomputed from the guided velocity $\tilde{v}_\theta^{x/z}$ via Eq. 2.1, so that the entire SDE — both drift and Langevin correction — is driven by the guided distribution rather than the plain conditional one.

This approach is undesirable for two reasons: (a) it requires a full retraining pass over the 160M-parameter denoiser every time the conditioning specification changes, and (b) the guidance weight w must be tuned per-property and directly trades off conditioning strength against sample diversity.

2.2 Proposed Approach

Our approach: We take a *smarter* approach. We took inspiration from NoiseRefine Kim et al. (2024), which showed in the image domain that the initial noise alone can encode enough structural information to replace guidance. Applied to Eq. 1, this suggests that conditioning need not enter through v_θ at all: it suffices to reparameterize the initial latent noise. We therefore introduce a lightweight adapter g_ϕ with learnable ϕ that maps base noise and a class label to a refined initial noise,

$$z^0_{ref} = g_\phi(z^0, y), \qquad z^0 \sim \mathcal{N}(0, I), \tag{4}$$

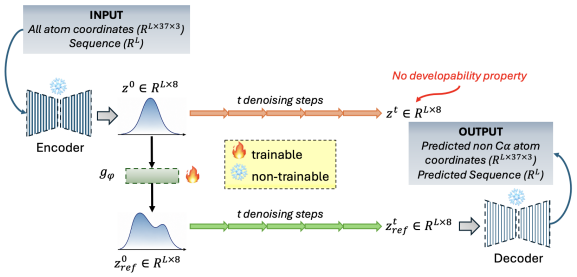


Figure 1: **Overall workflow of AdaProteina.** The encoder and decoder of Geffner et al. (2026) are frozen. g_ϕ optimizes the random noise z^0 to a developability property aware noise which is denoised by the frozen denoiser of Geffner et al. (2026). Details are provided in Sec. 2.2.

which replaces z^0 as the starting point of Eq. 3. Crucially, we target the fixed-backbone setting: x_{Ca} is a given, real backbone rather than a sample from Eq. 2, so only the side-chain and sequence detail encoded in z need be generated, conditioned on that fixed backbone. We call our method *Adapter La-Proteina* (**AdaProteina**). Consequently, Eq. 2 is never exercised at inference — x_{Ca} is held fixed at $t_x{=}1$ throughout sampling rather than integrated from noise — while the CFM training objective

(Eq. 2.1) and the VAE decoder q_ϑ remain exactly as trained in the original unconditional pipeline: v_θ is queried once per Euler step, exactly as in the unconditional pipeline, and only $|\phi| \ll |\theta|$ parameters are learned. Before going into detail about our training, we look at an ablation (Fig. 2) between La-Proteina and La-Proteina (fixed backbone, no noise-refinement); this acts as an intermediary step between *La-Proteina* and *AdaProteina*. Both metrics indicate that conditioning v_θ^z on a real, clean, fixed backbone (rather than the noisy backbone it was jointly trained to see) does not measurably degrade decoded sequence length or diversity. We will defer the structural-level analysis to the experimental section, but the *take-home message* here is that we can keep the backbone frozen and generate side-chains conditioned on the backbone and the desired developability properties.

Loss function: As mentioned, we introduce a refiner $g_\phi : \mathbf{R}^{L\times 8} \times \{0,1\} \to \mathbf{R}^{L\times 8}$ with parameters ϕ, mapping base noise and a class label to a refined noise z^0_{ref}, which replaces z^0 as the starting point of Eq. 3. The backbone SDE (Eq. 2) is left untouched: we condition on cached backbones x_{Ca} from real antibodies during training, so that g_ϕ learns to route z^0 toward regions of latent space whose rollouts decode into class-y sequences on a fixed structural scaffold. The denoiser v_θ and VAE decoder q_ϑ remain frozen throughout; only ϕ is optimized.

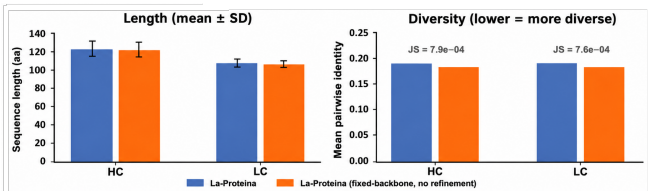


Figure 2: Sequence-level sanity comparison between La-Proteina and La-Proteina (fixed-backbone, no refinement) evaluated separately on heavy (HC) and light (LC) chains. **(Left)** Mean sequence length ± SD. **(Right)** Mean pairwise sequence identity (lower indicates greater diversity), estimated via alignment-free subsampled pairwise comparison

Training minimizes a composite objective on the amino-acid probabilities $s^{\text{soft}} \in \Delta_L^{19}$ obtained by decoding $z^{t_z=1}$ (the terminal latent of the SDE rollout from z^0_{ref}) through q_ϑ and taking the softmax over its residue-type logits:

$$\mathcal{L}(\phi) \;=\; \mathcal{L}_{\text{CE}}\big(h(s^{\text{soft}}), y\big) \;+\; \lambda\, \mathcal{L}_{\text{MMD}}\big(z^{t_z=1}, z^{t_z=1}_{\text{real}} \mid y\big), \tag{5}$$

where $\mathcal{L}_{\text{CE}}$ is a cross-entropy against a frozen pre-trained developability classifier/regressor h, evaluated on s^{soft}, and $\mathcal{L}_{\text{MMD}}$ is a per-class MMD^2 Gretton et al. (2012) between refined and cached real terminal latents that anchors $z^{t_z=1}$ to the data manifold. In the rest of the paper, we will denote $\mathcal{L}_{\text{CE}}\big(h(s^{\text{soft}}), y\big)$ as $\mathcal{L}_{dev}$, shorthand for "developability". The kernel of $\mathcal{L}_{\text{MMD}}$ is selected among RBF, Laplacian, arccos, linear, and the empirical neural tangent kernel (NTK) Jacot et al. (2018) of g_ϕ.

In order to train the above loss function, we start with a $z^0 \in \mathbf{R}^{L\times 8} \sim \mathcal{N}(0,1)$, and we concatenate the developability label, y, which passes through g_ϕ (trainable), which in turn refines the noise by $z^0_{\text{ref}} \mapsto z^0 + g_\phi(z^0, y)$. Our choice of g_ϕ is a smaller model, as one motivation of this work is to see how a smaller "adapter" works to include developability properties in antibody design.

In what follows, we discuss the experimental protocol, including a simpler baseline derived from La-Proteina. We will also discuss how to integrate multiple developability properties effectively in the adapter design.

3 Empirical Analysis

In this section, we first describe the datasets used for training and validation, followed by the details of the experiments including the hyperparameters.

Datasets:

a) *Training:* Chen et al. Chen et al. (2024): The data comprises a large paired-sequence antibody library with quantitative polyreactivity measurements relating CDR sequence/biophysical features to non-specific binding, more specifically high/low non-specific binding.

b) *Validation:* Emibetuzumab Makowski et al. (2022): A deep-sequencing dataset from yeast-display sorting of an emibetuzumab mutant library (CDR-mutated at 8 sites), labeled for high/low affinity and high/low non-specific binding: 4,000 sequences with binary labels, plus experimentally validated

subsets (125 Sanger-sequenced clones, 41 Pareto-frontier IgGs, 29 IgGs with novel CDR mutations) with measured affinity and non-specific binding values.

Experimental Protocol: We provide a description of the evaluation metrics used in this paper in Sec. A.1.1. The denoising was done for 50 denoising steps which also challenges the interpolation capability of the framework on fewer timesteps. We used the Adam optimizer with a learning rate of 0.0001 for $\mathcal{L}_{dev}$ and $\mathcal{L}_{mmd}$. 4x NVIDIA H100s were used for 3 epochs of training with a batch size of 16.

Baseline. We derive the following simpler baseline from La-Proteina: *Class-conditional shift (La-Proteina).* We draw $z^0 \sim \mathcal{N}(0, I)$ for $y = 0$ and $z^0 \sim \mathcal{N}(3, I)$ for $y = 1$. Then we pass them through the frozen La-Proteina to generate samples corresponding to $y = 0$ and $y = 1$ respectively. With abuse of notation, we will use La-Proteina to denote the class-conditioned shift version of it.

3.1 Experimental Results - Sequence level analysis

High vs. low non-specific binding: We have compared two versions of AdaProteina (with $\mathcal{L}_{dev} + \mathcal{L}_{mmd}$, $\mathcal{L}_{dev}$) with La-Proteina. We reported in Table 1 the comparative pairwise Levenshtein distances between generated antibody sequences. Each entry corresponds to the average distance between generated sequences for the low non-specific binders ($Y = 0$) and high non-specific binders ($Y = 1$) respectively.

Table 1: **Levenshtein distance and Fréchet distance (A.1.1) comparisons across methods.** Distance among the generated samples (samples of class Y=0/samples of class Y=1). "HC": Heavy chain of the antibody sequence, "LC": Light chain of the antibody sequence. "FD" indicates Fréchet Distance (samples of class Y=0/samples of class Y=1).

Method	Levenshtein Distance		FD
	HC	LC	
La-Proteina	103.3/52.8	92.9/45.3	10.74/ 47.48
AdaProteina ($\mathcal{L}_{dev}$)	85.8/28.4	78.4/23.8	28.03/61.23
AdaProteina ($\mathcal{L}_{dev} + \mathcal{L}_{mmd}$)	97.3/98.3	87.6/87.5	10.83/10.22

$\mathcal{L}_{dev}$ achieves the best separation among the generated class samples. However, it suffers from mode collapse. An important observation to note from Table 1 is that $\mathcal{L}_{dev}$ pushes the samples towards higher separability for the classes (Y). In order to do so, the majority of the sequences are the amino acids that increase binding strength. For example, a generated sequence looks like "YYYYYYYYADYYYYYYYYAYYYYYYYYYYYYYAYYAYAYYYYYAYA.....". This is also evident from the low Levenshtein distance for class Y=1. $\mathcal{L}_{mmd}$ counters this mode collapse by referring to the batch training distribution. Thus, $\mathcal{L}_{mmd}$ reduces the class separation but solves the important issue of mode collapse introduced by $\mathcal{L}_{dev}$.

Furthermore, FD quantifies the distributional similarity between the generated samples and the real samples. It can be seen in Table 1 that $\mathcal{L}_{mmd}$ is instrumental in providing a better FD which indicates sequence generation closer to the real antibodies.

Moreover, we looked at the comparative validation performance of ESM2-8M Lin et al. (2023) trained on sequences generated by each of the aforementioned three models. As mentioned earlier, we have used 96 mutated sequences from *Emibetuzumab - Makowski* as a validation set. We clearly see in Fig. 3 that AdaProteina with $\mathcal{L}_{mmd}$ significantly outperforms the other two models, specifically in terms of Recall and F1 metrics.

Mode collapse from an information-theoretic point of view: To probe whether g_ϕ preserves latent diversity or collapses onto a narrow mode, we compute the Shannon entropy of the refined noise space, averaged over the length of the sequences, i.e., number of residues. Higher entropy indicates a more uniform distribution over the g_ϕ space. In Fig. 4, we see that with just $\mathcal{L}_{dev}$, g_ϕ collapses to the lowest-entropy mode, which indicates a mode-collapse incident, as evident from Table 1. On the contrary, $\mathcal{L}_{dev} + \mathcal{L}_{mmd}$ remains highest-entropy and well above the entropy distribution of z^0.

Takeaway: La-Proteina Geffner et al. (2026) is incapable of generating samples that have the developability property. Introduction of $\mathcal{L}_{dev}$ increases the developability trait but suffers from mode collapse. $\mathcal{L}_{mmd}$ solves the "mode collapse". Our method generates samples that are closer to the training data than Geffner et al. (2026), indicating a developability-aware sequence generation.

Humanness: Now that we can generate developability-aware sequences meaningfully, a natural question comes to mind: *How human-like are these sequences?*. To answer this question we look at the proxy humanness score (Sapiens score, see Appendix) of the sequences generated using La-Proteina. As seen in Fig. 5, La-Proteina Geffner et al. (2026) suffers from generating sequences which are far from being human-like. We have also looked at the humanness of AdaProteina ($\mathcal{L}_{dev} + \mathcal{L}_{mmd}$), which is, not surprisingly, as equally poor as that of La-Proteina. Furthermore, combining $\mathcal{L}_{human}$ with $\mathcal{L}_{mmd}$ and $\mathcal{L}_{dev}$ does not lead to significant gain either. Our hypothesis is *that aggressively optimizing for binding often produces sequences dominated by tyrosine, tryptophan, and phenylalanine that score well as binders but no longer resemble natural human antibodies.* Thus we decouple this problem into three stages (i) train g_ϕ to separate high vs. low non-specific binders (ii) train g_ψ to separate human vs. non-human structures (iii) combine the above two in a training-free manner.

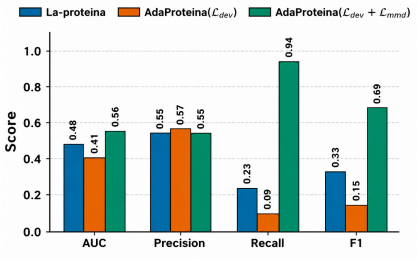


Figure 3: **Downstream classification for antibody binding** ($Y = 0/1$)**.**

We have already addressed how to achieve the first step. We can analogously train g_ψ by generating random non-human samples accompanied by the human samples from Chen et al. Chen et al. (2024). We will defer to the next part to address the mechanism for the final part to achieve both humanness and low/high non-specific binding.

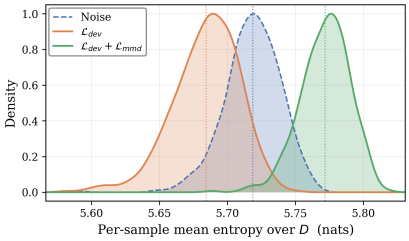


Figure 4: **Noise entropy distribution.** Plotting the entropy histogram of z^0 ("Noise") and z^0_{ref} to distinguish the contribution of $\mathcal{L}_{dev}$ and $\mathcal{L}_{mmd}$.

We further visualize the class-aware g_ψ space together with the noise space, $\mathcal{N}(0, 1)$ in Fig.7. We learned UMAP Wang et al. (2021) basis vectors from randomly drawn $\mathcal{N}(0, 1)$ noise. Then we projected the class conditioned g_ψ vectors onto the learned basis. The figure shows that g_ψ projected vectors cluster near the boundary of the noise space and have some separation between human and non-human classes.

Humanness and Binding:

Now that we can generate separate developability properties like binding and humanness independently, it is a natural question to ask whether we can combine them, i.e., *can we generate a human-like low non-specific binder?* The most natural way to do it is by training another adapter on the labeled data labeled with the multiple properties, but this is clearly not scalable. If we are asking to maintain three developable properties each with 10 choices, there are 10^3 possibilities, i.e., we need to train 1000 class classifiers, which, although not impossible, makes our adapter parameter heavy.

The next natural idea is function composition: if g_ϕ maps noise to a high/low non-spec. noise space and g_ψ maps noise to a human/non-human noise space, *can we use* $g_\phi \circ g_\psi$ *or* $g_\psi \circ g_\phi$*?* This works in principle but is hard to converge, as the separate loss terms are often conflicting.

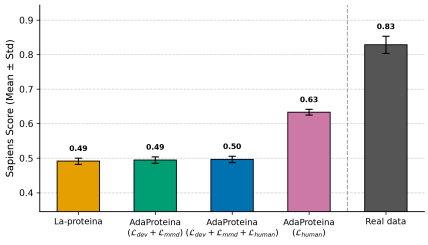


Figure 5: **Humanness comparison with real data.** Comparison of the contribution of different components towards the Sapiens score.

We therefore take a route based on transferring densities. The Perron-Frobenius operator Klus et al. (2015); Canto-Martín et al. (2014) transfers between density spaces in dynamical systems; its nonlinear kernel variant (kPF) Klus et al. (2020); Huang et al. (2022) admits a closed-form estimate. Since we can map Gaussian noise to both $\{g_\phi\}$ and $\{g_\psi\}$ space, kPF gives us a map between the binding-infused and humanness-infused noise spaces. Algorithm 1 (Appendix) gives the procedure: starting from random noise, we pass it through g_ϕ, use kPF to map it to a bimodal Gaussian $\{t(z^*)\}$ (one mode human-like, one not), then draw from the human-like mode and pass through g_ψ to preserve both bindingness and humanness.

To demonstrate the performance of the generated samples for downstream tasks, we augment the real training data with the generated samples and show performance in Table 2. Here we evaluate on

developability prediction specific to two antigens, Ovalbumin (Ova) and SMP. In both scenarios, the data augmentation helps improve the performance metrics on the validation dataset.

Beyond binary conditioning: heavy-chain germline clade. Eq. 4 places no restriction on y being binary. To test whether g_ϕ scales to a categorical axis, we retargeted it to heavy-chain germline clade, partitioning the Chen et al. repertoire into six IGHV families (clade$_h \in \{1, \ldots, 6\}$; a seventh with n=7 was dropped) and replacing h with a frozen six-way PCA+MLP germline classifier trained on HC sequences alone. Only the heavy chain is refined: light-chain latents are spliced back from the cached real pairing after the rollout, so g_ϕ solves the narrower problem of placing an HC into the correct germline family on a fixed scaffold. We drew 500 accepted samples per clade, accepting when the frozen classifier assigned the conditioned clade with confidence ≥ 0.8; acceptance ranged from 0.37 (IGHV1, the broadest family) to 1.00 (IGHV4), tracking how tightly each germline is delineated in the classifier's feature space.

Fig. 6 summarizes the accepted samples. Confidence is high and tightly concentrated across all six families (mean 0.956); IGHV4 is near-saturated (0.998 ± 0.004) while IGHV1 is most diffuse (0.906 ± 0.056). Within each clade the accepted sequences retain substantial position-level variability, e.g., mean pairwise Hamming distance 0.20 over the full heavy chain and 0.20 over CDR-H1/H2/H3 (per-clade ranges 0.15–0.27 and 0.10–0.32), so the adapter is not returning one memorized sequence per germline. This measures position-level variability rather than distributional coverage of the reference repertoire; we report the germline adapter as evidence that the noise-space mechanism transfers to categorical conditioning.

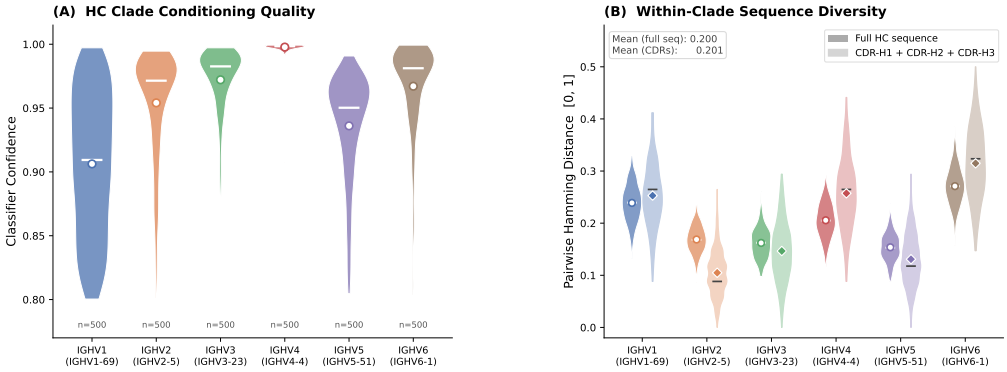


Figure 6: **Heavy-chain germline clade conditioning. (A)** Frozen six-way germline-classifier confidence for accepted samples, per IGHV family (n=500 each); white markers denote means. **(B)** Within-clade pairwise Hamming distance (scale $[0, 1]$) over the full heavy chain and over CDR-H1+H2+H3 only, computed on 500 random pairs per clade.

3.2 Experimental Results - Structure Level Analysis

Fixed backbone: Recall from Sec. 2.2, that our motivation of starting with a fixed (non-noisy) backbone is analyzed and validated in sequence-based analysis. But in order to understand the actual implications of such a design choice, it is crucial to look at structural analysis as well. We evaluated structural validity by folding sampled sequences from each variant (including fixed backbone, no refinement) with EquiFold and computing four backbone-geometry metrics on the predicted structures

(Fig. 8). AdaProteina structures showed significantly better geometry than both La-Proteina baselines across every metric (Welch's t-test West (2021), $p < 10^{-25}$ in all cases; Cohen's d-test Larner (2014) between 1.8 and 3.2, indicating large effect sizes): lower Cα–Cα deviation from ideal spacing, lower backbone bond-length deviation, a roughly 2$\times$ lower steric clash rate, and nearly double the fraction of residues in plausible Ramachandran basins Ramachandran et al. (2011).

As a sanity check, we also compared La-Proteina against its own fixed-backbone (no-refinement) variant; these did not differ significantly on three of four metrics ($p > 0.05$), consistent with the sequence-level ablation reported in Sec. 2.2, confirming that conditioning on a fixed real backbone does not itself degrade structural quality, i.e., the gains observed for AdaProteina are attributable to the noise-refinement mechanism rather than the fixed-backbone setup.

Humanness: Now, we look more deeply into our generated human-like sequences. To assess whether generated antibody heavy chains recapitulate real human repertoire statistics at the CDR level, we compared IMGT-numbered CDR-H1, -H2, and -H3 loops from 960-1000 generated sequences against a reference set of real human heavy chains, using per-position occupancy, Shannon entropy, and KL divergence, together with loop-length distributions and ANARCI-assigned V-gene germline calls Dunbar & Deane (2016).

Given sample sizes of $\sim$1000 per group, standard significance tests (a Kolmogorov-Smirnov Lilliefors (1967) test on length; a G-test analog on per-position composition) yielded p-values saturated near the limits of floating-point precision even for negligible effect sizes, and are therefore reported alongside.

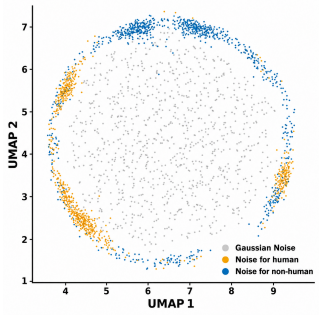


Figure 7: **UMAP visualization of learned adapter.** acts as a mapping: $\mathcal{N}(0,1) \mapsto$ *human/ non-human specific noises*

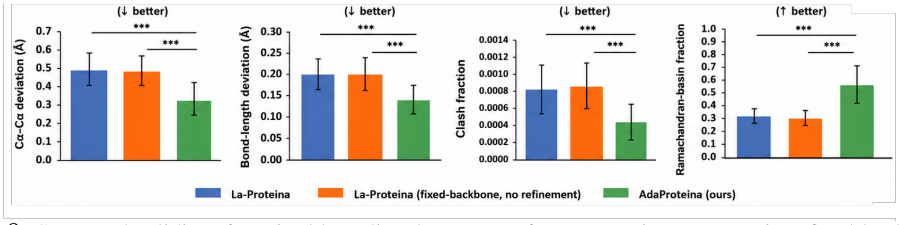


Figure 8: Structural validity of EquiFold-predicted structures for La-Proteina, La-Proteina (fixed-backbone, no refinement), and AdaProteina (ours), n=100 sequences per arm. Panels: **(i)** Cα–Cα deviation from ideal 3.80 Å spacing, **(ii)** backbone bond-length deviation, **(iii)** steric clash fraction, and **(iv)** Ramachandran-basin fraction Ramachandran et al. (2011). Arrows indicate the direction of better quality. Error bars: $\pm$1 SD; Welch's t-test, $^{***}p < 0.001$.

CDR-H1 and CDR-H2 loop lengths are collapsed, independent of germline identity (Fig. 9) Generated CDR-H1 length showed near-zero variance (mean 8.00, std. 0.06) against substantial variance in the reference set (mean 9.35, std. 0.94), and CDR-H2 showed the same pattern less severely (std ratio 0.14). Stratifying by ANARCI-assigned V-gene call ruled out simple germline collapse as the explanation: the five most frequent generated V-genes (IGHV3-33*02, IGHV3-64*04, IGHV1-46*01, IGHV3-66*01, IGHV3-30*01; n = 45-189 each) spanned two IGHV families with no single call exceeding 20% of sequences, yet CDR-H1 length was exactly 8 with zero variance *within every one of these germlines individually*. Since canonical CDR-H1 length is intrinsically germline-determined in real antibodies, this indicates the model has decoupled CDR-H1 length from V-gene identity, rather than having collapsed onto one germline. *This entails that the AdaProteina generation actually produces diversifying V-gene families.*

CDR-H3 shows a length shift with preserved variance, not a collapse (Fig. 9). In contrast to CDR-H1/H2, generated CDR-H3 loops were substantially shorter on average than reference (mean 9.61 vs. 14.71 residues; 34.7% reduction; Cohen's $d = -1.21$) but retained comparable-to-slightly-higher variance (std ratio 1.22), indicating the model continues to sample a range of CDR-H3 lengths but around a systematically shifted, shorter mean. *This again indicates that AdaProteina produces the CDR-H3 region with variability. Note that, biologically, CDR3 in the heavy chain is the area of interest in terms of drug design, so variability in this loop design is often desired.*

Table 2: **Classification performance to evaluate the generated sequences.** Performance evaluation across two datasets.

Dataset	Model	AUC	Precision	Recall	F1
Ova	w/o augmentation	51.2	**61.1**	23.4	33.9
	w augmentation	**62.5**	60.0	**83.0**	**69.6**
SMP	w/o augmentation	50.3	52.5	66.0	58.5
	w augmentation	**57.8**	**55.9**	**80.9**	**66.1**

Per-position amino acid composition diverges from reference in a position-specific rather than uniform direction (Fig. 9). Among the fifteen most divergent IMGT positions, CDR-H2 accounts for the majority of the largest divergences: position 57 shows the single largest mismatch overall (KL = 4.47 bits), and positions 62, 59, 65, 58, and 63 (all within the 57-65 range) each exceed 2.5 bits. CDR-H3 positions 116 and 117, immediately adjacent to framework 4, and CDR-H1 positions 30 and 35-38 also appear among the most divergent, indicating that composition mismatch spans all three CDRs rather than being confined to one.

Length-collapsed sequences are not distinguishable by standard humanness scoring. Sequences with CDR-H1/H2 length exactly at the dominant generated mode did not differ in Sapiens humanness score from sequences off that mode (Mann-Whitney $U = 6444$, $p = 0.43$; median 0.631 vs. 0.629), indicating this failure mode would not be flagged by humanness-based QC alone.

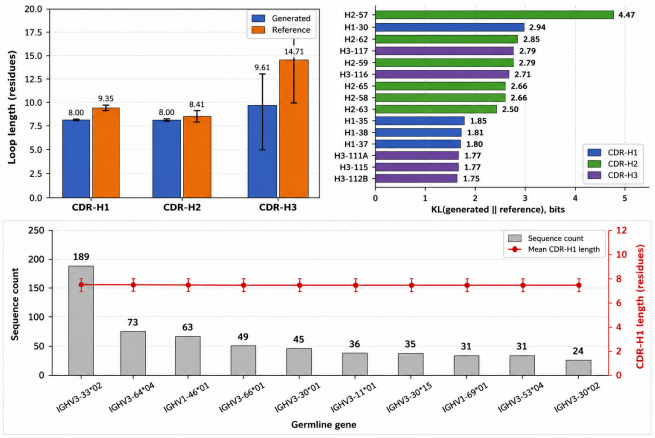


Figure 9: **CDR-level comparison of generated vs. reference heavy chains. (Left)** Mean CDR loop length ($\pm$std.). **(Middle)** Top-15 most divergent IMGT positions ($n \geq 30$ per group) by KL(generated$\|$reference) in bits. **(Right)** Top-10 ANARCI V-gene calls (bars) and mean CDR-H1 length (red line), which is exactly 8 with zero variance within every germline.

3.3 Effect of backbone in the final generation

A lingering implicit assumption is *how much does the binding property of the backbone bias that of the final structure?* We quantify this by fitting a ridge regressor on backbone atom coordinates alone to predict low/high non-specific binding: accuracy is $\sim$70%, i.e., the backbone alone is substantially predictive, so the model can "cheat" by reading the label off the scaffold. Concretely, asking for a low non-spec. binder from a high non-spec. binder backbone (and vice versa) succeeds no better than chance.

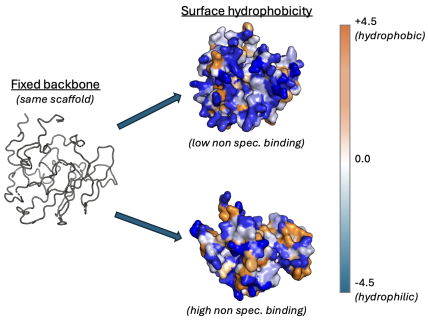


Figure 10: Generation of low and high non-spec. binders from a fixed backbone (with low non-specific binding)

This does not dilute our contribution — generating side chains from a target-biased backbone is non-trivial — and admits a simple fix: randomly swapping labels during training so the model must generate low/high non-spec. binders from high/low non-spec. binder backbones, penalizing biased generation. Generating both classes from one fixed backbone with this model (Fig. 10) shows the expected Kyte–Doolittle MetricGate (2026) signature: high non-spec. binder surfaces act as "velcro" and are visibly more hydrophobic than the low binders.

4 Conclusion

We introduced **AdaProteina**, a noise-space adapter for developability-aware antibody generation that freezes the denoiser and VAE decoder of La-Proteina (Geffner et al., 2026) and learns only a lightweight refiner g_ϕ from Gaussian noise and a class label to a developability-aware initial latent. A cross-entropy plus per-class MMD objective avoids the tyrosine/tryptophan mode collapse of naïve property optimization and the doubled per-step cost of guidance, while a Kernel Perron–Frobenius operator composes per-property noise spaces in closed form. AdaProteina yields separability La-Proteina cannot produce, better EquiFold geometry on every metric ($p < 10^{-25}$), downstream gains on Ovalbumin and SMP, and transfer from binary to six-way germline-clade conditioning.

Limitations: During geneation, heavy chains show collapsed CDR-H1/H2 length and a shortened CDR-H3, so refinement inherits length biases from the frozen decoder; the germline adapter is evaluated on classifier agreement and within-clade variability, not repertoire matching, and all structural evaluation is EquiFold-based.

REFERENCES

Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2024.

Sarah Alamdari, Nitya Thakkar, Rianne Van Den Berg, Neil Tenenholtz, Robert Strome, Alan M Moses, Alex X Lu, Nicolò Fusi, Ava P Amini, and Kevin K Yang. Protein generation with evolutionary diffusion: sequence is all you need. *BioRxiv*, pp. 2023–09, 2023.

Timothy Atkinson, Thomas D Barrett, Scott Cameron, Bora Guloglu, Matthew Greenig, Charlie B Tan, Louis Robinson, Alex Graves, Liviu Copoiu, and Alexandre Laterre. Protein sequence modelling with bayesian flow networks. *Nature Communications*, 16(1):3197, 2025.

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, et al. Atomically accurate de novo design of antibodies with rfdiffusion. *Nature*, 649(8095):183–193, 2026.

Tariq Berrada, Adriana Romero-Soriano, Michal Drozdzal, Jakob Verbeek, and Karteek Alahari. Entropy rectifying guidance for diffusion and flow models. *Advances in Neural Information Processing Systems*, 38:45750–45769, 2026.

Joey Bose, Tara Akhound-Sadegh, Guillaume Huguet, Kilian Fatras, Jarrid Rector-Brooks, Chenghao Liu, Andrei Nica, Maksym Korablyov, Michael Bronstein, and Alexander Tong. Se (3)-stochastic flow matching for protein backbone generation. In *International Conference on Learning Representations*, volume 2024, pp. 22590–22621, 2024.

Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. *arXiv preprint arXiv:2402.04997*, 2024.

Francisco Canto-Martín, Håkan Hedenmalm, and Alfonso Montes-Rodriguez. Perron-frobenius operators and the klein-gordon equation. *Journal of the European Mathematical Society (EMS Publishing)*, 16(1):31, 2014.

Can Chen, Karla-Luise Herpoldt, Chenchao Zhao, Zichen Wang, Marcus Collins, Shang Shang, and Ron Benson. Affinityflow: Guided flows for antibody affinity maturation. *arXiv preprint arXiv:2502.10365*, 2025a.

Hsin-Ting Chen, Yulei Zhang, Jie Huang, Manali Sawant, Matthew D Smith, Nandhini Rajagopal, Alec A Desai, Emily Makowski, Giuseppe Licari, Yunxuan Xie, et al. Human antibody polyreactivity is governed primarily by the heavy-chain complementarity-determining regions. *Cell reports*, 43(10), 2024.

Jiameng Chen, Xiantao Cai, Jia Wu, and Wenbin Hu. Antibody design and optimization with multi-scale equivariant graph diffusion models for accurate complex antigen binding. *arXiv preprint arXiv:2506.20957*, 2025b.

Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. Cfg++: Manifold-constrained classifier free guidance for diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 30824–30850, 2025.

Daniel Cutting, Frédéric A Dreyer, David Errington, Constantin Schneider, and Charlotte M Deane. De novo antibody design with se (3) diffusion. *Journal of Computational Biology*, 32(4):351–361, 2025.

James Dunbar and Charlotte M Deane. Anarci: antigen receptor numbering and receptor classification. *Bioinformatics*, 32(2):298–300, 2016.

Raphael R Eguchi, Christian A Choe, and Po-Ssu Huang. Ig-vae: Generative modeling of protein structure by direct 3d coordinate generation. *PLoS computational biology*, 18(6):e1010271, 2022.

Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. Reno: Enhancing one-step text-to-image models through reward-based noise optimization. *Advances in Neural Information Processing Systems*, 37:125487–125519, 2024.

Tomas Geffner, Kieran Didi, Zhonglin Cao, Danny Reidenbach, Zuobai Zhang, Christian Dallago, Emine Kucukbenli, Karsten Kreis, and Arash Vahdat. La-proteina: Atomistic protein generation via partially latent flow matching. In *International Conference on Learning Representations*, volume 2026, pp. 67350–67395, 2026.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13:723–773, 2012.

Bora Guloglu, Miguel Bragança, Alex Graves, Scott Cameron, Timothy Atkinson, Liviu Copoiu, Alexandre Laterre, and Thomas D Barrett. Abbfn2: A flexible antibody foundation model based on bayesian flow networks. *bioRxiv*, pp. 2025–04, 2025.

Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9380–9389. IEEE, 2024.

Anne Harrington, A Koepke, Shyamgopal Karthik, Trevor Darrell, and Alexei A Efros. It's never too late: Noise optimization for collapse recovery in trained diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43124–43134, 2026.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022a.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022b.

Zhichun Huang, Rudrasis Chakraborty, and Vikas Singh. Forward operator estimation in generative models with kernel transfer operators. *Proceedings of machine learning research*, 162:9148, 2022.

Guillaume Huguet, James Vuckovic, Kilian Fatras, Eric Thibodeau-Laufer, Pablo Lemos, Riashat Islam, Cheng-Hao Liu, Jarrid Rector-Brooks, Tara Akhound-Sadegh, Michael Bronstein, et al. Sequence-augmented se (3)-flow matching for conditional protein generation. *Advances in neural information processing systems*, 37:33007–33036, 2024.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. If at first you don't succeed, try, try again: Faithful diffusion-based text-to-image generation by selection. *arXiv preprint arXiv:2305.13308*, 2023.

Donghoon Kim, Gibeom Lee, et al. A noise is worth diffusion guidance. *arXiv preprint arXiv:2412.03895*, 2024.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Stefan Klus, Péter Koltai, and Christof Schütte. On the numerical approximation of the perron-frobenius and koopman operator. *arXiv preprint arXiv:1512.05997*, 2015.

Stefan Klus, Ingmar Schuster, and Krikamol Muandet. Eigendecompositions of transfer operators in reproducing kernel hilbert spaces. *Journal of Nonlinear Science*, 30(1):283–315, 2020.

Paulina Kulytė, Francisco Vargas, Simon Valentin Mathis, Yu Guang Wang, José Miguel Hernández-Lobato, and Pietro Liò. Improving antibody design with force-guided sampling in diffusion models. *arXiv preprint arXiv:2406.05832*, 2024.

Andrew J Larner. Effect size (cohen's d) of cognitive screening instruments examined in pragmatic diagnostic accuracy studies. *Dementia and geriatric cognitive disorders extra*, 4(2):236–241, 2014.

Hubert W Lilliefors. On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American statistical Association*, 62(318):399–402, 1967.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

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. *Advances in Neural Information Processing Systems*, 35:9754–9767, 2022.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision (ECCV)*, pp. 23–40. Springer, 2024.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*, 2025.

Emily K Makowski, Patrick C Kinnunen, Jie Huang, Lina Wu, Matthew D Smith, Tiexin Wang, Alec A Desai, Craig N Streu, Yulei Zhang, Jennifer M Zupancic, et al. Co-optimization of therapeutic antibody affinity and specificity using machine learning models that generalize to novel mutational space. *Nature communications*, 13(1):3788, 2022.

Karolis Martinkus, Jan Ludwiczak, Wei-Ching Liang, Julien Lafrance-Vanasse, Isidro Hotzel, Arvind Rajpal, Yan Wu, Kyunghyun Cho, Richard Bonneau, Vladimir Gligorijevic, et al. Abdiffuser: full-atom generation of in-vitro functioning antibodies. *Advances in Neural Information Processing Systems*, 36:40729–40759, 2023.

MetricGate. Kyte-Doolittle Hydropathy Calculator. `https://metricgate.com/docs/kyte-doolittle-hydropathy/`, 2026. Online statistical calculator.

Sanjay Nagaraj, Amir Shanehsazzadeh, Hyun Park, Jonathan King, and Simon Levine. Igflow: Flow matching for de novo antibody design. *Advances in Neural Information Processing Systems*, 2024.

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.

Tobias H Olsen, Iain H Moal, and Charlotte M Deane. Addressing the antibody germline bias and its effect on language models for improved antibody design. *Bioinformatics*, 40(11):btae618, 2024.

Youngrok Park, Hojung Jung, Sangmin Bae, and Se-Young Yun. Temporal alignment guidance: On-manifold sampling in diffusion models. *arXiv preprint arXiv:2510.11057*, 2025.

D Prihoda, J Maamary, A Waight, V Juan, L Fayadat-Dilman, D Svozil, and DA Bitton. Biophi: a platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. mabs. 2022; 14 (1): 2020203. *DOI*, 10(19420862.2021):2020203, 2021.

Srinivas Ramachandran, Pradeep Kota, Feng Ding, and Nikolay V Dokholyan. Automated minimization of steric clashes in protein structures. *Proteins: Structure, Function, and Bioinformatics*, 79 (1):261–270, 2011.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

Ayush Roy, Arjun Ramesh Kaushik, Vishnu Suresh Lokhande, Nalini K Ratha, and Venu Govindaraju. Ge-fm: Geometry-aware energy-based flow matching for non euclidean manifolds. *Transactions on Machine Learning Research*.

Ayush Roy, Wei-Yang Alex Lee, Rudrasis Chakraborty, and Vishnu Suresh Lokhande. Manifoldgd: Training-free hierarchical manifold guidance for diffusion-based dataset distillation. *arXiv preprint arXiv:2602.23295*, 2026.

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.

Dvir Samuel, Rami Ben-Ari, Simon Raviv, Nir Darshan, and Gal Chechik. Generating images of rare concepts using pre-trained diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4695–4703, 2024.

Justin Sanders, Luca Giancardo, Lan Guo, Yue Zhao, Kemal Sonmez, Nina Cheng, and Melih Yilmaz. Conditional generation of antibody sequences with classifier-guided germline-absorbing discrete diffusion. *arXiv preprint arXiv:2605.06720*, 2026.

Aravindan Kamatchi Sundaram, Ujjayan Pal, Abhimanyu Chauhan, Aishwarya Agarwal, and Srikrishna Karanam. Cocono: Attention contrast-and-complete for initial noise optimization in text-to-image synthesis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 9862–9870, 2025.

Cheng Tan, Yijie Zhang, Zhangyang Gao, Yufei Huang, Haitao Lin, Lirong Wu, Fandi Wu, Mathieu Blanchette, and Stan Z Li. dyab: Flow matching for flexible antibody design with alphafold-driven pre-binding antigen. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 782–790, 2025.

Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. Conditional flow matching: Simulation-free dynamic optimal transport. *arXiv preprint arXiv:2302.00482*, 2(3), 2023.

Amelia Villegas-Morcillo, Jana M Weber, and Marcel JT Reinders. Guiding diffusion models for antibody sequence and structure co-design with developability properties. *PRX Life*, 2(3):033012, 2024.

Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. Understanding how dimension reduction tools work: an empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. *Journal of Machine Learning Research*, 22(201):1–73, 2021.

Zichen Wang, Yaokun Ji, Jianing Tian, and Shuangjia Zheng. Retrieval augmented diffusion model for structure-informed antibody design and optimization. In *International Conference on Learning Representations*, volume 2025, pp. 37188–37206, 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, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Robert M West. Best practice in statistics: Use the welch t-test when testing the difference between two groups. *Annals of clinical biochemistry*, 58(4):267–269, 2021.

Katherine Xu, Lingzhi Zhang, and Jianbo Shi. Good seed makes a good crop: Discovering secret seeds in text-to-image diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3024–3034. IEEE, 2025.

Yinghua Yao, Yuangang Pan, and Xixian Chen. Generative co-design of antibody sequences and structures via black-box guidance in a shared latent space. *arXiv preprint arXiv:2508.11424*, 2025.

Jason Yim, Brian L Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. Se (3) diffusion model with application to protein backbone generation. *arXiv preprint arXiv:2302.02277*, 2023.

Siqi Zhao, Joshua Moller, Porfi Quintero-Cadena, and Lood van Niekerk. Guided generation for developable antibodies. *arXiv preprint arXiv:2507.02670*, 2025.

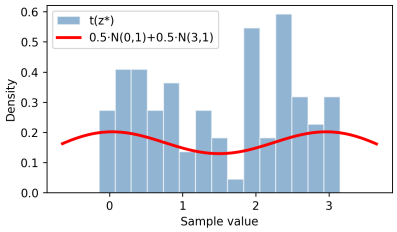
Qinqing Zheng, Matthew Le, Neta Shaul, Yaron Lipman, Aditya Grover, and Ricky T. Q. Chen. Guided flows for generative modeling and decision making. *arXiv preprint arXiv:2311.13443*, 2023.

A Appendix

A.1 Experimental Details

Algorithm 1 kPF

Input: g_ϕ, $N, \gamma, \lambda, y \in \{0, 1\}$
// Stage 1: Encoding binding properties to shared noise space.
for all $\mathbf{R}^{L\times 8} \supseteq \{z_i^0, z_j^0\} \sim \mathcal{N}(0, I)$ **do**
  $K_{ij} = \exp\left(-\frac{\|\mathrm{vec}(g_\phi(z_i^0; y)) - \mathrm{vec}(g_\phi(z_j^0; y))\|^2}{2\sigma_Z^2}\right)$
end for
for all $\mathbf{R} \supseteq \{x_i, x_j\} \sim \frac{1}{2}\mathcal{N}(0, 1) + \frac{1}{2}\mathcal{N}(3, 1)$ **do**
  $L_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma_x^2}\right)$
end for
$z^\star = g_\phi(z^0; y),\ z^0 \sim \mathcal{N}(0, I)$
$k_i^\star(z^\star) = \exp\left(-\frac{\|z^\star - z_i^0\|^2}{2\sigma_Z^2}\right)$
$\mathbf{s}(z^\star) = L\,(K + \lambda I)^{-1}\,\mathbf{k}^\star(z^\star)$
$\mathcal{I} \leftarrow \text{top-}\gamma\ \mathbf{s}(z^\star)$
$\omega_i = \max(s_i, 0) / \sum_{i' \in \mathcal{I}} \max(s_{i'}, 0)$
$t(z^\star) = \sum_{i\in\mathcal{I}} \omega_i\, x_i$ =0



// Stage 2: Projecting the shared noise space to humanness-specific noise space.
for all $\mathbf{R} \supseteq \{x_i, x_j\} \sim \mathcal{N}(0, I)$ **do**
  $M_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma_x^2}\right)$
end for
$y^\star \sim \{t(z^\star)\}$
$m_i^\star(y^\star) = \exp\left(-\frac{\|y^\star - x_i\|^2}{2\sigma_x^2}\right)$
$\mathbf{s}(y^\star) = M\,(L + \lambda I)^{-1}\,\mathbf{m}^\star(y^\star)$
$\mathcal{J} \leftarrow \text{top-}\gamma\ \mathbf{s}(y^\star)$
$\omega_i = \max(s_i, 0) / \sum_{i' \in \mathcal{J}} \max(s_{i'}, 0)$
$f(y^\star) = \sum_{i\in\mathcal{J}} \omega_i\, Z_i \in \mathbf{R}^{L\times 8}$ =0

A.1.1 Metrics

Levenshtein distance. To quantify sequence-level similarity, we use the *Levenshtein distance* (also known as the *edit distance*), which measures the minimum number of single-character edit operations required to transform one sequence into another. The permitted edit operations are substitutions, insertions, and deletions, each assigned a unit cost. Consequently, a smaller Levenshtein distance indicates greater sequence similarity, while a larger distance reflects greater divergence.

Formally, given two antibody sequences x and y, the Levenshtein distance $d_{\mathrm{Lev}}(x, y)$ is defined as the minimum number of edit operations required to transform x into y. Thus, $d_{\mathrm{Lev}}(x, y) = 0$ if and only if the two sequences are identical.

In our experiments, the Levenshtein distance is computed separately for the heavy chain (HC), light chain (LC), and the concatenated heavy–light sequence (HC+LC). For each class label, we compute the pairwise Levenshtein distance between every generated sequence from one model and every generated sequence from the comparison model, yielding a distribution of edit distances. We summarize this distribution by its mean. A lower mean Levenshtein distance indicates that the two models generate more similar antibody sequences for the corresponding class.

Compared to position-wise similarity measures such as the Hamming distance, the Levenshtein distance naturally accounts for insertions and deletions in addition to substitutions. This makes it well suited for antibody sequence analysis, where generated sequences may exhibit small length variations, particularly within the complementarity-determining regions (CDRs). Consequently, the Levenshtein distance provides a robust measure of overall sequence similarity while remaining sensitive to biologically meaningful sequence variations.

Sapiens score. To assess the human-likeness of generated antibody sequences we adopt the *Sapiens score* (a per-residue likelihood under a human antibody language model). Throughout, every generated antibody is split into its heavy chain (HC) and light chain (LC), and each chain is scored on its own. This is required rather than optional: the two chains derive from distinct germline gene families with different framework consensus, and (for Sapiens) are scored by two *separately trained* networks. We therefore report two quantities per antibody: Sapiens (HC) and Sapiens (LC), and do not collapse them into a single combined number.

Sapiens is a masked-language model of human antibody sequences, of the BERT/RoBERTa encoder family, trained on human variable-region repertoires from the Observed Antibody Space Prihoda et al. (2021). Two *distinct* models are trained independently: a heavy-chain model f_{H} on $\sim$20M VH sequences and a light-chain model f_{L} on $\sim$19M VL sequences. A chain must be scored by the model matching its type. Given a chain $s = (s_1, \ldots, s_L)$ of type $c \in \{\mathrm{H}, \mathrm{L}\}$, the model outputs a per-position probability distribution over the amino-acid alphabet,

$$f_c(s) \;=\; \big(\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_L\big), \qquad \mathbf{p}_i \in \Delta^{19}, \qquad p_i(a) \;=\; \Pr[\, s_i = a \mid s \,], \tag{6}$$

where Δ^{19} is the probability simplex over the 20 amino acids and $p_i(a)$ is the model's probability of amino acid a at position i given the (unmasked) sequence context. The Sapiens score of the chain is the mean probability the model assigns to the residue that is actually present at each position,

$$\mathrm{Sapiens}_c(s) \;=\; \frac{1}{L}\sum_{i=1}^{L} p_i(s_i) \;\in\; [0, 1]. \tag{7}$$

A high score means that, position by position, the observed residues are the ones a model of human antibodies would expect. A low score flags residues that are improbable under the human-repertoire prior. This is a more appropriate metric than OASis, as OASis asks "does the exact peptide actually occur in the human antibody repertoire?". Being a hard membership score, OASis is easily misled: *Its limitation is brittleness, i.e., a single substitution can turn a human* 9*-mer non-human, and it captures only local (*9*-residue) context, ignoring long-range dependencies.* We observe this as Geffner et al. (2026) achieves a significantly higher OASis score (10x) over AdaProteina-concatenated but suffers significantly on Sapiens score. This is due to the fact that there are repeated 9-mer sequences which bias the OASis score, and this failure mode is exactly captured by Sapiens.

Binding score. We calculate the confidence score of the generated sequences for $Y = 1$. We utilize the classifier trained for $\mathcal{L}_{dev}$, i.e., h, for calculating the confidence scores for $Y = 1$.

Downstream evaluation. We use Precision, Recall, F1 score, and AUC, which are the most standard metrics.

A.1.2 Ablations

Ablation on choice of kernel. We evaluate the choice of kernel for $\mathcal{L}_{mmd}$. We can see from Table 5 that the kernels are comparable overall in preventing the mode collapse introduced by $\mathcal{L}_{dev}$. Furthermore, without the use of any kernel, the humanness score drops significantly as seen in Fig. 11, while NTK gives the best result and is used for all experiments. All experiments for this ablation were done using the AdaProteina-shifted version.

Table 3: **Levenshtein distance (A.1.1) comparisons across kernels.** Experiments are done on AdaProteina-shifted version (see Table 1).

Method	HC	LC
AdaProteina-shifted ($\mathcal{L}_{dev} + \mathcal{L}_{mmd}$) (NTK)	99.6/99.4	89.7/88.1
AdaProteina-shifted ($\mathcal{L}_{dev} + \mathcal{L}_{mmd}$) (RBF)	93.3/88.7	85.4/78.7
AdaProteina-shifted ($\mathcal{L}_{dev} + \mathcal{L}_{mmd}$) (Laplacian)	91.0/83.0	83.8/74.3
AdaProteina-shifted ($\mathcal{L}_{dev} + \mathcal{L}_{mmd}$) (Arccos)	82.8/89.2	77.8/79.3
AdaProteina-shifted ($\mathcal{L}_{dev} + \mathcal{L}_{mmd}$) (None)	98.8/100.3	89.1/88.8

Ablation on λ (see Eq. 5). We evaluate the change in performance of AdaProteina with change in λ, i.e., change in the strength of $\mathcal{L}_{mmd}$.

$\lambda = 0.00$ clearly demonstrates the need for $\mathcal{L}_{mmd}$ (see Table 4). Furthermore, $\lambda = 0.50$ and $\lambda = 0.75$ achieve the best FD. Beyond $\lambda = 0.50$, performance plateaus, showing that AdaProteina is not sensitive to further increases in the strength of $\mathcal{L}_{mmd}$. **Ablation.**

Table 4: **Ablation on λ.** Experiments are done on AdaProteina and FD is reported (samples of class Y=0/samples of class Y=1).

λ	**FD**
0.00	19.47/41.96
0.25	11.81/11.30
0.50	10.83/10.22
0.75	10.64/10.17
1.00	10.98/10.05

Table 5: **Sapiens and binding score (A.1.1) comparison.** Experiments are done to compare KPF (Algorithm 1) and $g_\psi \circ g_\phi$ (see Table 1).

Method	**Sapiens HC**	**Sapiens LC**	**Binding** ($Y = 1$)
AdaProteina (KPF)	0.63	0.51	0.704
$g_\psi \circ g_\phi$	0.50	0.41	-
AdaProteina ($\mathcal{L}_{dev} + \mathcal{L}_{mmd} + +\mathcal{L}_{human}$)	0.43	0.35	0.58

B Related Work

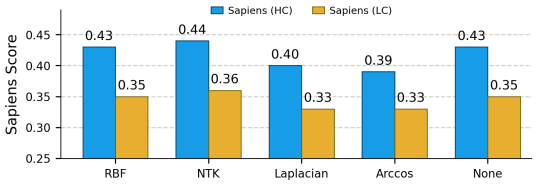


Figure 11: **Ablation study to vary the kernel for $\mathcal{L}_{mmd}$.** Experiments are done on AdaProteina-shifted version (see Table 1).

Diffusion-based antibody design. DiffAb Luo et al. (2022) pioneered antigen-conditioned CDR co-design over sequence, backbone coordinates, and residue orientations, and remains the base architecture for a large family of extensions: DIFFFORCE Kulytė et al. (2024) adds force-guided sampling for lower-energy CDRs, RADAb Wang et al. (2025) augments the reverse process with retrieval, and AbMEGD Chen et al. (2025b) adds multi-scale equivariant encoding. AbDiffuser Martinkus et al. (2023) generalizes the setting to full-atom sequence-structure co-design via an aligned protein mixer, and IgDiff Cutting et al. (2025) fine-tunes FrameDiff Yim et al. (2023) on OAS-derived variable-region structures for de novo diffusion. RFAntibody Bennett et al. (2026) similarly fine-tunes RFdiffusion Watson et al. (2023) on antibody PDB structures for epitope-conditioned VHH and scFv design.

Flow matching for proteins and antibodies. Flow-matching variants have followed the same trajectory: FoldFlow Bose et al. (2024) and FoldFlow-2 Huguet et al. (2024) established SE(3)-equivariant flow matching for protein backbones, and MultiFlow Campbell et al. (2024) extended it to joint discrete-continuous co-design. Antibody-specific instantiations include dyAb Tan et al. (2025), which combines coarse-grained interface alignment with fine-grained sequence–structure flow matching under AlphaFold2-predicted pre-binding antigens, and IgFlow Nagaraj et al. (2024), which performs masked-state discrete flow matching on antibody sequences. Most recently, La-Proteina Geffner et al. (2026), on which we build, factorizes atomistic protein generation as a partially latent flow over C_α coordinates and a per-residue latent, with sequence and side-chain heavy atoms recovered by a VAE decoder.

Sequence-only and language-model generators. Ig-VAE Eguchi et al. (2022) is an early VAE-based immunoglobulin backbone generator. On the sequence side, AntiBERTy Ruffolo et al. (2021) and AbLang Olsen et al. (2022) are BERT-style masked language models trained on OAS repertoires, later shown by AbLang-2 Olsen et al. (2024) to memorize germline patterns rather than model somatic

variation. Discrete diffusion has emerged as a competitive alternative: EvoDiff Alamdari et al. (2023) applies order-agnostic discrete diffusion to general proteins, ProtBFN and AbBFN Atkinson et al. (2025) leverage Bayesian Flow Networks for both unconditional and zero-shot conditional antibody generation, and germline-absorbing discrete diffusion Sanders et al. (2026) proposes germline sequences as the absorbing state for antibody-specific classifier guidance.

Property- and developability-conditioned generation. Several prior works have targeted developability directly. Villegas-Morcillo et al. (2024) introduce two guidance strategies-property-aware priors and sampling-by-property-on top of DiffAb for CDR design conditioned on solubility and folding stability. AbBFN2 Guloglu et al. (2025) jointly generates 31 biophysical attributes alongside sequence in a Bayesian flow framework, and LEAD Yao et al. (2025) performs property-guided sampling through black-box optimization in a shared latent space. Zhao et al. (2025) apply Soft Value-based Decoding in Diffusion (SVDD) with off-the-shelf developability predictors trained on 246 clinical-stage antibodies, on top of an ESM2-based discrete diffusion model. AffinityFlow Chen et al. (2025a) performs analogous guidance for affinity maturation in the flow-matching setting.

Across these approaches, conditioning enters either through (i) classifier or classifier-free guidance at sampling time, which doubles per-step forward passes, or (ii) multi-modal training that reintroduces the full pretraining cost. *Our contribution is orthogonal to both*: we freeze the 160M-parameter La-Proteina denoiser and its VAE decoder entirely, and introduce developability conditioning through a lightweight noise-space refiner g_ϕ with $|\phi| \ll |\theta|$ learnable parameters. This yields conditional generation at unchanged per-sample inference cost and with a training footprint governed by the adapter rather than the base model: *properties that guidance-based and jointly-trained approaches structurally cannot provide*.