

ABOPD: Antibody CDR Design via On-Policy Distillation

Zhuo Yang^{1,6}, Jiaying He^{1,2}, Jiaqing Xie³, Daolang Wang⁵, Xipeng Qiu^{1,3}, Yuxin Wang¹,
Tianfan Fu^{7,4,†}, Beilun Wang^{6,†}

¹Shanghai Innovation Institute

²Zhejiang University

³Fudan University

⁴Shanghai Artificial Intelligence Lab

⁵Xiamen University

⁶Southeast University

⁷Nanjing University

[†]Corresponding author.

Abstract

Antibodies are essential therapeutic molecules, and their complementarity-determining regions (CDRs) form the primary antigen-recognition interface. Recent protein generative models have demonstrated broad capabilities in biomolecular design, yet post-training strategies for downstream objectives remain limited. Standard denoising training operates on noisy states obtained by perturbing native structures, whereas recursive generation proceeds through model-generated intermediate states. For flexible antibody CDR loops such as CDR-H3, this mismatch can allow backbone deviations to accumulate along the denoising trajectory and compromise antigen-facing loop geometry. We introduce ABOPD, an Antibody design framework built on On-Policy Distillation, leveraging privileged native geometry during training to supervise states visited along the model’s own denoising trajectories. With this fine-grained structural supervision, ABOPD substantially improves structural recovery on RAbD CDR-H3 generation, reducing RMSD by 0.42 Å (2.37→1.95 Å) and outperforming supervised fine-tuning and offline distillation controls, offering a path to higher-fidelity protein design.

🔗 **GitHub:** <https://github.com/little1d/ABOPD>

🤗 **Hugging Face:** <https://huggingface.co/collections/little1d/ABOPD>

1 Introduction

Computational design of antibody complementarity-determining regions (CDRs) for antigen-specific recognition is a central goal in protein engineering. This task remains challenging because antigen recognition depends critically on the coupling between the amino-acid sequences and 3D conformations of these flexible loops. The six CDRs (H1–H3 on the heavy chain and L1–L3 on the light chain) form most of the antigen-contacting surface [1, 2]. Consequently, CDR design requires jointly generating amino-acid sequences and loop conformations while preserving framework anchoring and geometric compatibility with the antigen interface (Figure 1). Learning the joint distribution of CDR sequences and structures is further complicated by the limited availability of experimentally determined antibody–antigen complex structures [3].

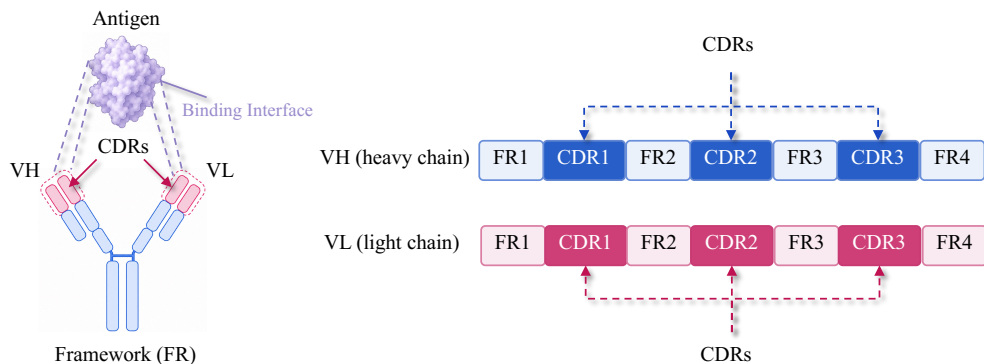


Figure 1 Antibody CDR design. The left panel locates the six CDR loops at the antigen-binding interface, while the right panel shows their organization within the variable domains of the heavy and light chains.

Recent work has advanced antigen-conditioned CDR generation by leveraging geometric representations, equivariant architectures, and diffusion-based generative modeling [4–7]. Despite these advances, reliably generating CDRs with accurate loop geometry remains challenging. Existing work has largely emphasized model architectures and structural representations, while post-training strategies for antibody CDR generation remain comparatively underexplored. Post-training therefore offers a complementary route to improve existing antibody generators and, more broadly, adapt protein generative models to downstream design objectives, such as specialization to particular protein families [8] and property optimization [9].

For diffusion-based CDR models, a key post-training challenge is the mismatch between the state distribution used for supervision and the one encountered during generation [10]. Standard denoising training constructs noisy inputs by perturbing native CDR sequences and structures, which we call reference-derived states. Supervised fine-tuning (SFT) retains the denoising objective on this distribution, while offline distillation supervises the student with teacher predictions on the same class of states. During generation, however, reverse denoising is recursive: each step conditions on states produced by earlier model predictions. Prediction errors can therefore move subsequent states away from the reference-derived distribution, so supervision confined to this distribution does not directly address errors arising along the model’s own generation trajectory. This mismatch motivates augmenting the denoising objective with supervision on states visited along the student’s own reverse trajectories.

Such a mismatch is especially consequential for antibody CDR generation because flexible loops must satisfy geometric constraints. CDR-H3 exemplifies this challenge: it is typically the most diverse and conformationally flexible CDR, yet it must remain anchored to the antibody framework at both ends and adopt a conformation compatible with the antigen interface [2, 4]. A $C\alpha$ backbone error introduced during reverse denoising can alter the loop’s placement or antigen-facing orientation. Subsequent steps then condition on the resulting geometry, allowing the error to propagate and increasing the risk of steric clashes. Because residue identities and orientations are predicted within the same evolving geometric context, backbone drift may also influence subsequent sequence and orientation predictions. Thus, the state-distribution mismatch can manifest as accumulated geometric error in CDR backbone recovery.

To address this mismatch, we introduce ABOPD, an on-policy distillation framework for post-training diffusion-based antibody CDR models. ABOPD adapts a backbone-aware teacher using native geometry available during training and distills $C\alpha$ coordinate-transition targets on states visited along the student’s reverse trajectories. The original denoising objective is retained as an offline anchor for sequence, coordinate, and orientation prediction. Our main contributions are:

- We identify a state-distribution mismatch in CDR diffusion that widens during denoising, while teacher corrections become most useful during later refinement.

- We propose ABOPD, combining on-policy $C\alpha$ coordinate-transition distillation on student rollout states with an offline denoising anchor.
- We demonstrate improved backbone recovery on RAbD CDR-H3 and simultaneous multi-CDR redesign. On RAbD, ABOPD reduces CDR-H3 RMSD from 2.37 to 1.95 Å and outperforms SFT and offline distillation.

2 Related Work

2.1 Antibody CDR Design

Antibody CDR design spans sequence modeling, energy-guided loop optimization, and antigen-conditioned sequence-structure generation. Sequence models learn antibody repertoire distributions or infill variable regions, while classical pipelines search CDR sequences and conformations through loop sampling, biophysical scoring, and Rosetta-based optimization [1, 2, 11, 12]. Neural co-design methods couple residue identities with 3D geometry to generate CDR sequences and structures using contextual information from the antibody and antigen [5–7, 13, 14]. Recent antibody and protein-binder models further incorporate all-atom representations, flow-based generation, and interface-aware conditioning [15–19]. ABOPD complements this line of work by post-training an existing CDR diffusion model [4].

2.2 Structure-Aware Protein Generative Models

More broadly, structure-aware protein generation spans sequence design under structural constraints and direct generation of 3D protein geometry. Hallucination-based methods optimize sequences through learned structure predictors [20], while inverse-folding models generate sequences conditioned on fixed backbones [21]. Diffusion and flow-based models further enable direct backbone generation through coordinate and rigid-frame representations, including diffusion and flow trajectories on SE(3) [22–26]. These geometric formulations provide the methodological foundation for modern antibody CDR generation.

2.3 Post-Training Strategies for Generative Models

Post-training strategies differ in both their feedback signals and the states on which those signals are applied. On-policy distillation applies teacher feedback to student-generated states, reducing distribution mismatch in autoregressive language models [27]; recent diffusion and flow methods extend this principle to continuous generative trajectories [28–30]. Reinforcement learning and preference optimization instead adapt generators through task-specific rewards or preferences, with applications to molecular design [31–34] and protein design [9, 35, 36]. Related but distinct approaches include conditional guidance, which steers sampling at inference [37], and privileged learning, which uses information available only during training [38, 39]. ABOPD combines on-policy distillation with training-time privileged structural supervision for antibody CDR generation.

3 Methods

ABOPD aligns structural supervision with states encountered during recursive generation while retaining the original denoising objective as an offline anchor. As shown in Figure 2, the framework comprises three stages: hybrid pretraining to obtain H-DiffAb, backbone-aware teacher adaptation using privileged native geometry, and student post-training via on-policy $C\alpha$ coordinate-transition distillation.

3.1 Problem Setup

We study antigen-conditioned generation of one or more antibody CDRs. Let $\mathcal{G}$ be the set of target residues. Each target residue $i \in \mathcal{G}$ is represented by

$$r_i = (s_i, x_i, O_i),$$

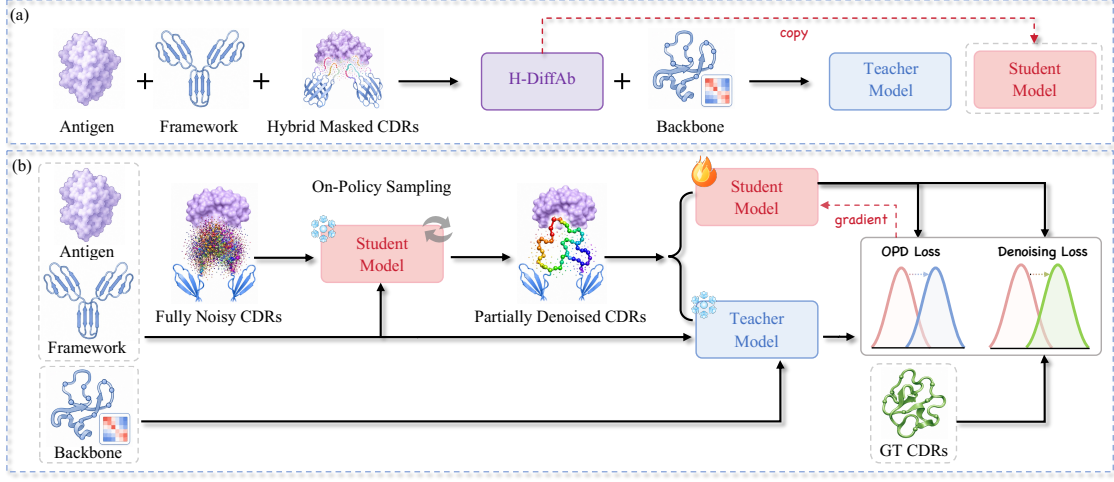


Figure 2 Overview of ABOPD. (a) Hybrid pretraining produces H-DiffAb, which initializes the student and is adapted into a backbone-aware teacher using privileged native geometry. (b) The frozen teacher supplies $C\alpha$ coordinate-transition targets on student rollout states, while the offline denoising anchor preserves the original training objective.

where s_i is its amino-acid identity, $x_i \in \mathbb{R}^3$ is its $C\alpha$ coordinate, and $O_i \in SO(3)$ is its local orientation. The context C contains the antigen and all non-target antibody residues, including the framework and any CDRs not selected for generation. Given $R_{\mathcal{G}} = \{r_i : i \in \mathcal{G}\}$, the model represents

$$p_{\theta}(R_{\mathcal{G}} | C).$$

Let R_0 denote the native target sequence and structure. Following DiffAb [4], the forward process independently perturbs amino-acid identities, $C\alpha$ coordinates, and residue orientations using categorical, Gaussian, and $SO(3)$ diffusion, respectively. We use

$$R_t = (s_t, x_t, O_t) \sim q_t(R_t | R_0, \mathcal{G})$$

to denote a *reference-derived state*, i.e., a noisy state obtained by perturbing the native CDR. Given (R_t, C, t) , the denoiser predicts a categorical reverse distribution, coordinate noise, and a denoised orientation:

$$f_{\theta}(R_t, C, t) = \left(p_{\theta}^{\text{seq}}, \hat{\epsilon}_{\theta}^{\text{pos}}, \hat{O}_{\theta}^{t-1} \right).$$

Following Luo et al. [4], the standard supervised objective is

$$\mathcal{L}_{\text{denoise}} = \mathcal{L}_{\text{seq}} + \lambda_{\text{pos}} \mathcal{L}_{\text{pos}} + \lambda_{\text{ori}} \mathcal{L}_{\text{ori}}, \quad (1)$$

where the three branch losses are

$$\mathcal{L}_{\text{seq}} = \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} D_{\text{KL}} \left(q(s_i^{t-1} | s_i^t, s_i^0) \| p_{\theta}^{\text{seq}}(s_i^{t-1} | R_t, C, t) \right), \quad (2)$$

$$\mathcal{L}_{\text{pos}} = \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} \left\| \epsilon_i^{\text{pos}} - \hat{\epsilon}_{\theta, i}^{\text{pos}}(R_t, C, t) \right\|_2^2, \quad (3)$$

$$\mathcal{L}_{\text{ori}} = \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} \left\| (O_i^0)^{\top} \hat{O}_{\theta, i}^{t-1}(R_t, C, t) - I_3 \right\|_F^2. \quad (4)$$

Here, $q(s_i^{t-1} | s_i^t, s_i^0)$ is the categorical diffusion posterior, ϵ_i^{pos} is the Gaussian noise applied to the $C\alpha$ coordinate, and $\hat{O}_{\theta,i}^{t-1}$ is the predicted denoised orientation. We use unit weights, $\lambda_{\text{pos}} = \lambda_{\text{ori}} = 1$, and retain the complete objective as the offline denoising anchor during ABOPD post-training.

The reference-derived state distribution underlying Eq. (1) differs from the distribution encountered during generation. At inference, a reverse trajectory

$$R_K^\theta \rightarrow R_{K-1}^\theta \rightarrow \dots \rightarrow R_0^\theta$$

recursively conditions each transition on previous model outputs. Consequently, even a model trained accurately on $q_t(R_t | R_0, \mathcal{G})$ may receive states outside this reference-derived distribution after its own errors accumulate. ABOPD addresses this mismatch by adding teacher supervision directly on states visited along the student’s reverse trajectory.

3.2 Hybrid Pretraining and Backbone-Aware Teacher Adaptation

Hybrid pretraining. The released DiffAb protocol trains separate models for single-CDR and multi-CDR co-design [4]. We instead pretrain a single H-DiffAb model across single-CDR, multi-CDR, and all-six-CDR target masks. At each update, a target set $\mathcal{G} \sim p_{\text{hyb}}(\mathcal{G})$ is sampled; only residues in $\mathcal{G}$ are noised and reconstructed, while the remaining antibody and antigen residues provide structural context. The objective is

$$\mathcal{L}_{\text{hyb}} = \mathbb{E}_{\mathcal{G}, t, R_t \sim q_t(\cdot | R_0, \mathcal{G})} [\mathcal{L}_{\text{denoise}}(f_\theta(R_t, C, t), R_0; \mathcal{G})]. \quad (5)$$

We retain the same target-mask distribution during teacher adaptation and student post-training. The H-DiffAb pretraining budget is approximately 8× that of a default DiffAb checkpoint; this larger budget should be considered when interpreting its improvement over DiffAb. The resulting parameters θ_0 initialize the teacher, the ABOPD student, and all post-training controls, so comparisons among the post-training methods share the same starting point. The sampling probabilities and training configuration are reported in Appendix B and summarized in Table 8.

Backbone-aware teacher adaptation. The teacher augments H-DiffAb with a privileged descriptor B constructed from native target geometry. For each target residue, B contains $C\alpha$ -centered Cartesian coordinates and validity masks for the N, $C\alpha$, C, O, and observed $C\beta$ atoms, together with target-to-complex $C\alpha$ distance features. Two lightweight residual MLP adapters encode the residue- and pair-level descriptors and add them to the corresponding H-DiffAb encoder embeddings. The enriched embeddings and the unchanged noisy state R_t are then processed by the original EpsilonNet, so the privileged geometry conditions the denoising prediction without replacing the diffusion state. Explicit target residue labels and atoms beyond $C\beta$ are excluded, although the observed $C\beta$ geometry and validity mask can carry limited sequence-correlated information. Exact descriptor encodings and adapter architectures are provided in Appendix B.

We initialize the H-DiffAb backbone from θ_0 , randomly initialize both adapters, and jointly optimize the full teacher on reference-derived states:

$$\mathcal{L}_{\text{teacher}} = \mathbb{E}_{\mathcal{G}, t, R_t \sim q_t(\cdot | R_0, \mathcal{G})} [\mathcal{L}_{\text{denoise}}(f_\phi(R_t, C, B, t), R_0; \mathcal{G})]. \quad (6)$$

After adaptation, the resulting parameters ϕ^* are frozen, and B is used only when querying the teacher during student post-training.

3.3 ABOPD: On-Policy Distillation

Student-visited rollout states. The student f_θ is initialized with θ_0 and receives only (R_t, C, t) . For each training batch, the current student generates a reverse-denoising trajectory using the inference-time sampler:

$$R_K^S \rightarrow R_{K-1}^S \rightarrow \dots \rightarrow R_0^S, \quad R_{0:K}^S \sim p_\theta^{\text{rollout}}(\cdot | C, \mathcal{G}).$$

Each rollout starts at schedule index $K = 80$, with the generated residues sampled from the factorized initialization distribution used by DiffAb [4]: amino-acid types are sampled uniformly over 20 categories, normalized $C\alpha$ coordinates are sampled from $\mathcal{N}(0, I)$, and residue orientations are sampled uniformly over $SO(3)$. The rollout follows $80 \rightarrow 79 \rightarrow \dots \rightarrow 0$ on the original 100-step schedule, while the antibody–antigen context remains fixed. The rollout is performed without gradients, and the cached states are detached before distillation. We cache the states at $\mathcal{T} = \{80, 70, 60, 50, 40, 30, 20, 10, 5\}$; this training timestep set is distinct from the eight-timestep grid used only for the mechanism analysis in Section 4.4. Thus, “on-policy” refers only to the fact that distillation states are drawn from the current student’s own inference process. It does not imply policy-gradient optimization.

At every cached state, the student and frozen teacher are evaluated on exactly the same R_t^S . For compactness, denote their coordinate predictions by

$$\hat{e}_{S,t}^{\text{pos}} := \hat{e}_{\theta}^{\text{pos}}(R_t^S, C, t), \quad \hat{e}_{T,t}^{\text{pos}} := \hat{e}_{\phi^*}^{\text{pos}}(R_t^S, C, B, t). \quad (7)$$

The teacher remains frozen throughout student optimization, and its predictions are treated as fixed structural targets; gradients propagate only through the student branch. Although f_{ϕ^*} is fixed, its target is state-dependent because it is recomputed for each student-visited R_t^S .

From reverse KL to coordinate transition matching. Let $P_{S,t}$ and $P_{T,t}$ denote the student and teacher one-step reverse kernels evaluated at the same student-visited state. At each selected rollout timestep $t \in \mathcal{T}$, the local OPD target follows the reverse-KL direction used for student-generated diffusion trajectories [30]:

$$\ell_t^{\text{OPD}} = D_{\text{KL}}(P_{S,t}(\cdot | R_t^S, C, \mathcal{G}) \| P_{T,t}(\cdot | R_t^S, C, B, \mathcal{G})). \quad (8)$$

The student kernel appears first because distillation is evaluated on states visited by the student, following the reverse-KL OPD convention.

We instantiate Eq. (8) using the DDPM coordinate transition adopted by DiffAb [4, 40]. This Gaussian transition is inherited from the coordinate sampler rather than introduced by ABOPD. Conditioned on a shared student-visited state, model $j \in \{S, T\}$ induces the one-step reverse kernel

$$P_{j,t}^{\text{pos}} = \mathcal{N}(\mu_{j,t}^{\text{pos}}, \sigma_t^2 I), \quad \mu_{j,t}^{\text{pos}} = a_t x_t^S + b_t \hat{e}_{j,t}^{\text{pos}}. \quad (9)$$

Given x_t^S , the model prediction is deterministic and the only transition randomness is the sampler’s Gaussian noise. The coefficients a_t , b_t , and σ_t^2 are determined by the shared DiffAb schedule. Student and Teacher therefore have the same covariance and differ only through their coordinate-noise predictions. In particular, their mean difference is

$$\begin{aligned} \mu_{S,t}^{\text{pos}} - \mu_{T,t}^{\text{pos}} &= (a_t x_t^S + b_t \hat{e}_{S,t}^{\text{pos}}) - (a_t x_t^S + b_t \hat{e}_{T,t}^{\text{pos}}) \\ &= b_t (\hat{e}_{S,t}^{\text{pos}} - \hat{e}_{T,t}^{\text{pos}}). \end{aligned}$$

Substituting this difference into the KL between the shared-covariance Gaussian kernels in Eq. (9) reduces the coordinate component of Eq. (8) to

$$D_{\text{KL}}(P_{S,t}^{\text{pos}} \| P_{T,t}^{\text{pos}}) = \frac{b_t^2}{2\sigma_t^2} \left\| \hat{e}_{S,t}^{\text{pos}} - \hat{e}_{T,t}^{\text{pos}} \right\|_2^2. \quad (10)$$

To remove the explicit schedule-dependent KL prefactor and simplify implementation, we drop the factor $b_t^2/(2\sigma_t^2)$ and assign equal weight to all selected timesteps, Let $\mathcal{D}$ denote the training-data distribution over

(C, R_0, B) :

$$\mathcal{L}_{\text{pos}}^{\text{OPD}} = \mathbb{E}_{\substack{(C, R_0, B) \sim \mathcal{D}, \mathcal{G} \sim p_{\text{hyb}}, \\ R_{0:K}^S \sim p_{\theta}^{\text{rollout}}(\cdot | C, \mathcal{G})}} \left[\frac{1}{|\mathcal{T}| |\mathcal{G}|} \sum_{t \in \mathcal{T}} \sum_{i \in \mathcal{G}} \left\| \hat{e}_{S,t,i}^{\text{pos}} - \hat{e}_{T,t,i}^{\text{pos}} \right\|_2^2 \right]. \quad (11)$$

Equation (11) is a uniformly reweighted coordinate-transition regression objective. It retains the per-state teacher-matching target induced by Eq. (10), while omitting the schedule-dependent KL weights. Because rollout states are detached, gradients pass only through the student predictions at the sampled states, not through the rollout distribution. Thus, reverse KL motivates the state distribution and distillation direction, whereas Eq. (11) defines the practical ABOPD update.

We distill only the coordinate branch because it provides the most direct on-policy correction to CDR geometry. The $C\alpha$ trace controls framework anchoring and loop placement at the antigen interface; through Eq. (9), coordinate errors directly alter the next loop state and thereby the geometric context of subsequent denoising steps. Sequence predictions do not themselves specify a spatial correction, whereas residue orientations describe local frames on the current coordinate scaffold and cannot alone correct translational drift in loop placement. The Backbone-aware Teacher improves coordinate denoising and backbone recovery without improving sequence recovery (Table 1). Coordinate matching also provides the best trade-off in the distillation-target ablation (Fig. 5). We therefore apply OPD only to $C\alpha$ coordinates, while sequence and orientation remain supervised by the denoising anchor below.

Offline denoising anchor. On-policy coordinate supervision alone does not preserve the joint sequence, coordinate, and orientation objective learned during pretraining. We therefore retain the original denoising loss on an independently sampled reference-derived state:

$$\mathcal{L}_{\text{anchor}} = \mathbb{E}_{\substack{(C, R_0) \sim \mathcal{D}, \mathcal{G} \sim p_{\text{hyb}}, \\ t \sim \mathcal{U}\{1, \dots, K\}, R_t \sim q_t(\cdot | R_0, \mathcal{G})}} \left[\mathcal{L}_{\text{denoise}}(f_{\theta}(R_t, C, t), R_0; \mathcal{G}) \right]. \quad (12)$$

The final objective is

$$\boxed{\mathcal{L}_{\text{ABOPD}} = \mathcal{L}_{\text{anchor}} + \beta \mathcal{L}_{\text{pos}}^{\text{OPD}}}. \quad (13)$$

Here, β controls the relative strength of on-policy coordinate supervision, and we set $\beta = 0.6$. ABOPD is therefore a mixed-state-distribution objective: $\mathcal{L}_{\text{anchor}}$ is evaluated on $q_t(R_t | R_0, \mathcal{G})$, whereas $\mathcal{L}_{\text{pos}}^{\text{OPD}}$ is evaluated on $p_{\theta}^{\text{rollout}}(\cdot | C, \mathcal{G})$. We maintain an exponential moving average (EMA) copy of the student for validation and generation, while the current non-EMA student produces all training rollouts and the teacher remains frozen. Unless otherwise stated, reported ABOPD results use the EMA student $\hat{f}_{\bar{\theta}}$, whereas SFT and offline-distillation controls use their directly optimized parameters. The EMA analysis is reported in Appendix A.5. The complete post-training procedure is summarized in Algorithm 1.

Post-training controls. SFT continues the reference-state denoising objective without using the teacher or student rollout states:

$$\mathcal{L}_{\text{SFT}} = \mathbb{E}_{\substack{(C, R_0) \sim \mathcal{D}, \mathcal{G} \sim p_{\text{hyb}}, \\ t \sim \mathcal{U}\{1, \dots, K\}, R_t \sim q_t(\cdot | R_0, \mathcal{G})}} \left[\mathcal{L}_{\text{denoise}}(f_{\theta}(R_t, C, t), R_0; \mathcal{G}) \right]. \quad (14)$$

Thus, SFT is identical in form to $\mathcal{L}_{\text{anchor}}$ but is optimized without an OPD term. Offline distillation instead applies the same frozen teacher’s coordinate target to reference-derived states. Below, $\hat{e}_S^{\text{pos}}(R_t)$ and $\hat{e}_T^{\text{pos}}(R_t)$ abbreviate the student and teacher predictions under conditioning C and (C, B) , respectively:

$$\mathcal{L}_{\text{pos}}^{\text{offline}} = \mathbb{E}_{\substack{(C, R_0, B) \sim \mathcal{D}, \mathcal{G} \sim p_{\text{hyb}}, \\ \{R_t^{\text{ref}} \sim q_t(\cdot | R_0, \mathcal{G})\}_{t \in \mathcal{T}}}} \left[\frac{1}{|\mathcal{T}| |\mathcal{G}|} \sum_{t \in \mathcal{T}} \sum_{i \in \mathcal{G}} \left\| \hat{e}_{S,t,i}^{\text{pos}}(R_t^{\text{ref}}) - \hat{e}_{T,t,i}^{\text{pos}}(R_t^{\text{ref}}) \right\|_2^2 \right]. \quad (15)$$

Algorithm 1: ABOPD with On-Policy Coordinate Transition Matching

Require: f_{θ_0} ; frozen teacher f_{ϕ^*} ; data $\mathcal{D}; \mathcal{T}; \beta$; EMA decay ρ

- 1: $\theta \leftarrow \theta_0, \bar{\theta} \leftarrow \theta$
- 2: **for** each optimization step **do**
- 3: $(C, R_0, B) \sim \mathcal{D}, \mathcal{G} \sim p_{\text{hyb}}(\mathcal{G})$
- 4: $t_a \sim \mathcal{U}\{1, \dots, K\}, R_{t_a}^{\text{ref}} \sim q_{t_a}(\cdot | R_0, \mathcal{G})$
- 5: $\mathcal{L}_{\text{anchor}} \leftarrow \mathcal{L}_{\text{denoise}}(f_{\theta}(R_{t_a}^{\text{ref}}, C, t_a), R_0; \mathcal{G})$
- 6: With gradients disabled, $\{R_{\tau}^S\}_{\tau \in \mathcal{T}} \leftarrow \text{Rollout}(f_{\theta}, C, \mathcal{G})$
- 7: **for all** $\tau \in \mathcal{T}$ **do**
- 8: $\hat{\epsilon}_{S, \tau}^{\text{pos}} \leftarrow f_{\theta}^{\text{pos}}(R_{\tau}^S, C, \tau)$
- 9: $\hat{\epsilon}_{T, \tau}^{\text{pos}} \leftarrow f_{\phi^*}^{\text{pos}}(R_{\tau}^S, C, B, \tau)$
- 10: **end for**
- 11: Compute $\mathcal{L}_{\text{pos}}^{\text{OPD}}$ using Eq. (11) and set $\mathcal{L} \leftarrow \mathcal{L}_{\text{anchor}} + \beta \mathcal{L}_{\text{pos}}^{\text{OPD}}$
- 12: $\theta \leftarrow \text{Update}(\theta, \nabla_{\theta} \mathcal{L}); \bar{\theta} \leftarrow \text{EMA}_{\rho}(\bar{\theta}, \theta)$
- 13: **end for**
- 14: **return** $f_{\bar{\theta}}$

Offline distillation uses the same Teacher target, selected timestep set, and uniform timestep weighting as ABOPD, but applies the target to reference-derived states. It does not include $\mathcal{L}_{\text{anchor}}$. All post-training controls use the same H-DiffAb initialization, training data, and number of optimizer updates; they differ in their supervision and state distributions.

4 Experiments

The experiments address three questions: whether the Backbone-aware Teacher provides more accurate geometric targets, whether ABOPD improves CDR generation across two complementary benchmarks, and whether on-policy supervision provides gains beyond SFT and offline distillation. All post-training controls start from the same H-DiffAb initialization.

4.1 Setup

Data. Following Luo et al. [4], we construct the data split from the SAbDab database [41]. We exclude structures with resolution worse than 4 Å and antibodies targeting non-protein antigens, then cluster the remaining antibody–antigen complexes by CDR-H3 sequence at 50% identity. Five clusters are held out for testing, and the remaining clusters are used for training.

Benchmarks. We evaluate two complementary generation settings. Simultaneous multi-CDR redesign jointly generates the sequences and structures of all six CDRs for 19 held-out SAbDab antibody–antigen complexes. RAbD CDR-H3 generation evaluates the most diverse and conformationally flexible heavy-chain CDR loop across 60 target cases [2].

Metrics. **AAR (%)** measures amino-acid recovery over generated CDR residues; we report it as a conventional metric but interpret it cautiously because it does not necessarily reflect overall design quality [42]. Structural fidelity is evaluated using Cα **RMSD** in the fixed complex coordinate frame without Kabsch alignment, **IDDT**, and heavy-chain **TM-Score**. Following Wu et al. [43], we compute antibody-global IDDT on PyRosetta-packed all-atom structures. Interface quality is assessed using **ddG** and **DockQ**. We define ddG as $dG_{\text{gen}} - dG_{\text{ref}}$, where lower values indicate a more favorable interface-energy change [44, 45]. In simultaneous multi-CDR redesign, ddG is computed over the complete antibody–antigen interface.¹ DockQ measures agreement with the native protein–protein interface [46] and is computed between the antigen and complete antibody heavy chain following Wu et al. [43]. Geometric realism is evaluated using **Clash_{in}**, **Clash_{out}**, and **JSD_{bb}**. The

¹Before aggregating ddG, we exclude catastrophic Rosetta relaxation or energy-evaluation failures, affecting 0.433% of simultaneous multi-CDR samples.

clash metrics measure $C\alpha$ -level steric conflicts within the generated CDRs and between generated CDRs and the antigen, respectively, using a 3.6574 Å cutoff and excluding covalently adjacent CDR residues [47]. JSD_{bb} is the Jensen–Shannon divergence between generated and reference backbone dihedral-angle distributions. All three metrics are computed on raw generated structures before relaxation.

Sampling protocol. For each target and sampling seed, we generate 100 designs and aggregate results for our evaluated variants over three fixed seeds. Baseline implementation and evaluation details are provided in Appendix C.

4.2 Backbone-Aware Teacher Validation

Because the Teacher supplies training-time distillation targets rather than serving as the final generator, we evaluate H-DiffAb, Teacher w/o Backbone, and the Backbone-aware Teacher on identical reference-derived noisy states (R_t, C, t) . The two adapted variants follow the same procedure and differ only in access to privileged backbone descriptors. Table 1 shows that the Backbone-aware Teacher reduces coordinate MSE from 0.2019 to 0.1503 and orientation loss from 0.1445 to 0.1324 relative to H-DiffAb, while its sequence accuracy remains lower, localizing the benefit to geometric prediction. The no-backbone continuation control improves sampled CDR RMSD despite exhibiting poorer one-step denoising accuracy, showing that sampled-generation quality alone is insufficient for selecting a distillation teacher. We therefore freeze the Backbone-aware Teacher, which provides the most accurate geometric denoising targets, for subsequent student distillation. Additional adaptation variants and timestep-resolved results are reported in Appendix A.1.

Table 1 Backbone-aware Teacher validation. Coord. MSE, Ori. loss, and Seq. Acc. evaluate one-step denoising on shared reference-derived states, whereas AAR and RMSDs characterize samples generated by the same variants. CDR BB RMSD is computed over the backbone atoms N, $C\alpha$, and C in the fixed complex coordinate frame. Best and second-best values are shown in **bold** and underlined, respectively.

Variant	Coord. MSE ↓	Ori. loss ↓	Seq. Acc. ↑	AAR (%) ↑	CDR $C\alpha$ RMSD (Å) ↓	CDR BB RMSD (Å) ↓
H-DiffAb	<u>0.2019</u>	<u>0.1445</u>	0.7382	60.56	1.4863	1.5298
Teacher w/o Backbone	0.2212	0.1662	0.7105	57.92	<u>1.0520</u>	<u>1.1428</u>
Backbone-aware Teacher	0.1503	0.1324	<u>0.7246</u>	<u>59.26</u>	0.4855	0.6642

4.3 Main Results

Simultaneous multi-CDR redesign. Table 2 shows that, relative to H-DiffAb, ABOPD reduces RMSD across all six CDRs, with the largest improvement on CDR-H3, from 2.94 Å to 2.48 Å. It also reduces intra-CDR and CDR–antigen clash rates from 0.93% and 1.58% to 0.42% and 0.67%, respectively, while maintaining a similar JSD_{bb} . Together, these results indicate more accurate multi-loop geometry while preserving backbone-dihedral realism. Single-CDR redesign results are reported in Appendix A.4.

Table 2 Simultaneous multi-CDR redesign. H3 is highlighted as the most flexible and challenging CDR. For entries containing $\pm$, we report the mean and 95% confidence interval over three sampling seeds. Best and second-best results are shown in **bold** and underlined, respectively.

Method	AAR (%) ↑	IDDT ↑	ddG ↓	Per-CDR RMSD (Å) ↓						Geometric Realism ↓		
				H1	H2	H3	L1	L2	L3	Clash _{in} (%)	Clash _{out} (%)	JSD_{bb}
dyMEAN [5]	60.07	0.8029	60.03	1.65	1.47	6.15	1.58	1.23	1.59	–	–	–
AbX [48]	26.52	0.8365	22.65	1.46	1.49	6.01	3.19	2.78	1.30	–	–	–
PPIFlow [19]	–	0.3075	–	2.17	1.68	5.41	3.64	2.73	4.89	–	–	–
IgGM [16]	28.49	0.4234	23.93	1.73	1.55	4.37	1.62	1.51	1.71	25.63	<u>1.45</u>	0.2900
DiffAb	52.06±0.29	0.8525±0.0006	27.727±25.835	0.79±0.05	0.77±0.26	3.41±0.18	1.15±0.16	1.11±0.32	1.12±0.24	1.40	1.87	0.2345
H-DiffAb	<u>60.27±0.11</u>	0.8670±0.0004	<u>22.431±12.142</u>	<u>0.68±0.02</u>	<u>0.66±0.18</u>	<u>2.94±0.22</u>	<u>0.95±0.20</u>	<u>1.04±0.23</u>	<u>1.01±0.23</u>	0.93	1.58	0.2308
ABOPD	60.73±0.10	0.8710±0.0004	20.574±10.627	0.63±0.02	0.64±0.20	2.48±0.16	0.90±0.21	1.00±0.17	0.97±0.20	0.42	<u>0.67</u>	<u>0.2325</u>

Table 3 RAbD CDR-H3 generation. For entries containing $\pm$, we report the mean and 95% confidence interval over three sampling seeds. Methods marked with * use the pipeline IgFold $\Rightarrow$ HDock $\Rightarrow$ Rosetta side-chain packing [44, 49, 50]. Best and second-best results are shown in **bold** and underlined, respectively.

Method	AAR (%) $\uparrow$	IDDT $\uparrow$	TM-score $\uparrow$	RMSD ($\text{\AA}$) $\downarrow$	DockQ $\uparrow$	Clash _{in} (%) $\downarrow$	Clash _{out} (%) $\downarrow$	JSD _{bb} $\downarrow$
RosettaAb* [2]	32.31	0.8272	0.9717	17.70	0.14	–	–	–
MEAN* [13]	37.38	0.8252	0.9688	17.30	0.16	–	–	0.5290
GeoAB* [6]	40.02	0.8367	0.9695	15.43	0.19	–	–	–
dyMEAN [5]	<u>41.84</u>	0.8392	0.9718	8.10	0.41	–	–	0.5420
DGENet [51]	42.67	0.8551	0.9747	7.19	0.43	–	–	–
BoltzGen [52]	39.07	0.8372	0.9675	2.69	0.47	–	–	–
DiffAb	32.30 $\pm$ 0.63	0.9105 $\pm$ 0.0002	0.9772	2.79 $\pm$ 0.02	0.79	3.18	<u>3.42</u>	0.2622
H-DiffAb	36.35 $\pm$ 0.59	<u>0.9158$\pm$0.0001</u>	<u>0.9809</u>	<u>2.37$\pm$0.02</u>	<u>0.81</u>	<u>3.46</u>	<u>3.69</u>	0.2382
ABOPD	37.31 $\pm$ 0.21	0.9193$\pm$0.0002	0.9845	1.95$\pm$0.02	0.83	0.94	2.24	<u>0.2418</u>

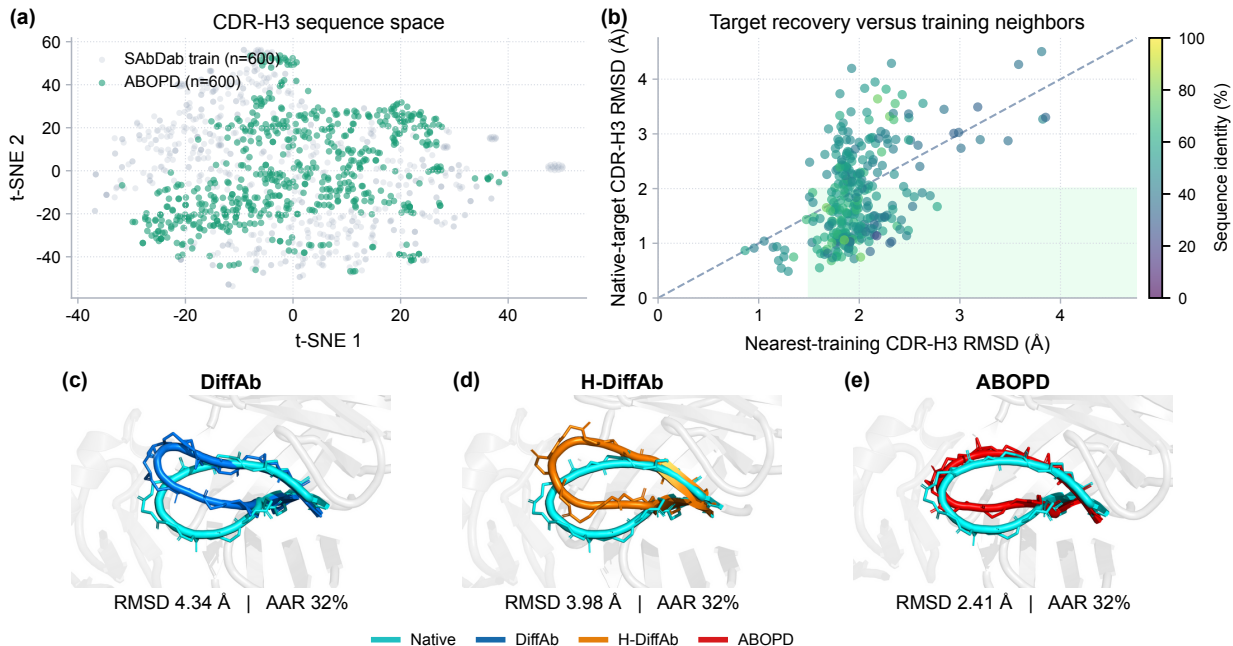


Figure 3 (a) t-SNE comparison of ABOPD and SabDab training sequences. (b) Native-target RMSD versus nearest-training CDR-H3 RMSD. (c–e) Qualitative structural comparison with the native loop.

RAbD CDR-H3 evaluation. On the external RAbD benchmark, ABOPD reduces CDR-H3 RMSD from 2.37 Å to 1.95 Å, while AAR increases only modestly from 36.35% to 37.31% (Table 3), identifying improved backbone recovery as its primary benefit. The accompanying gains in IDDT, TM-score, and DockQ, together with lower intra-CDR and CDR–antigen clash rates, indicate consistent improvements in local geometry, global heavy-chain structure, and interface placement. Additional interface and side-chain analyses are reported in Appendix A.6.

Sequence and structure analysis of generated CDR-H3 loops. Figures 3(a) and 3(b) show that ABOPD sequences broadly overlap with SABDab training CDRs in the t-SNE embedding, while the generated structures are closer on average to their native targets than to their nearest training neighbors. Together, these observations suggest that the structural gains are not explained by simple nearest-neighbor retrieval. In a representative case where all three methods attain the same AAR, ABOPD better preserves framework anchoring and antigen-facing loop placement than DiffAb and H-DiffAb (Fig. 3(c)–(e)).

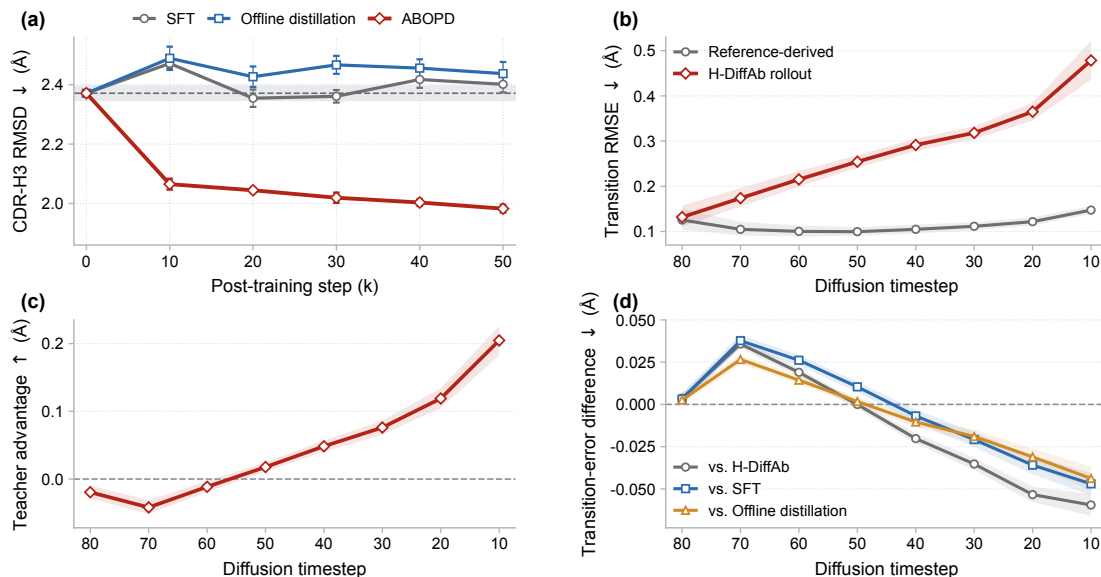


Figure 4 (a) CDR-H3 RMSD over post-training steps; the dashed line and gray band denote the H-DiffAb initialization. The ABOPD curve uses EMA checkpoints, whereas the control curves use directly optimized parameters. (b) H-DiffAb $C\alpha$ transition error on reference-derived and rollout states. (c) Teacher advantage on rollout states, computed as paired H-DiffAb error minus Teacher error; positive values favor the Teacher. (d) Paired ABOPD error minus control error on the shared H-DiffAb rollout-state bank; negative values favor ABOPD. Curves in (b–d) are complex-level means after seed averaging; error bars in (a) and shaded bands in (b–d) denote 95% confidence intervals computed across the three sampling seeds and by complex-level bootstrap, respectively. Timesteps decrease from left to right, corresponding to early-to-late reverse denoising.

4.4 On-Policy Distillation: Controls and Mechanism

Post-training controls. Starting from the same H-DiffAb initialization, we compare ABOPD with SFT, which continues supervised denoising, and offline distillation, which applies the same frozen Teacher and coordinate-transition target to reference-derived states. Fig. 4(a) shows that both controls remain close to H-DiffAb, whereas ABOPD steadily reduces CDR-H3 RMSD. Thus, neither continued denoising nor Teacher matching on reference-derived states reproduces ABOPD’s structural gain, consistent with the benefit of combining the denoising anchor with Teacher supervision on rollout states. Although the main ABOPD curve uses EMA checkpoints, the non-EMA student retains most of the gain (Appendix A.5, Table 6), indicating that EMA alone does not explain the improvement.

Timestep-level mechanism analysis. Using all 60 RAbD CDR-H3 complexes and three sampling seeds, we compare reference-derived states from forward noising with states cached from H-DiffAb rollouts at $t \in \{80, 70, 60, 50, 40, 30, 20, 10\}$. At each timestep, H-DiffAb and the frozen Teacher are evaluated on the same state, with the Teacher additionally receiving the privileged descriptor B . Transition error is the one-step $C\alpha$ RMSE to a native-directed transition-mean target, which serves as a local reference rather than the exact reverse posterior on rollout states. Teacher advantage is computed as H-DiffAb error minus Teacher error, whereas the ABOPD–control difference is ABOPD error minus the corresponding control error; positive and negative values favor the Teacher and ABOPD, respectively.

H-DiffAb has higher transition error on rollout states than on reference-derived states throughout denoising, with the gap widening toward later timesteps (Fig. 4(b)). Teacher advantage becomes positive at $t = 50$, whereas ABOPD attains lower transition error than all controls at $t = 40$ – 10 (Fig. 4(c)–(d)). Together, these temporal patterns support a stage-dependent mechanism in which rollout-state supervision becomes most useful during later geometric refinement (Fig. 4(b)–(d)).

4.5 Distillation Target Ablation

We compare sequence, coordinate, orientation, and combined distillation targets under the same H-DiffAb initialization, Teacher, rollout budget, and sampling protocol. Sequence-only matching decreases AAR without improving RMSD, whereas orientation-only matching yields only a limited structural gain (Fig. 5). All variants containing the coordinate target substantially reduce RMSD, but coordinate-only matching achieves the largest RMSD reduction and AAR gain; adding sequence or orientation provides no further improvement. Together with the Teacher’s advantage in geometric denoising but not sequence accuracy (Table 1), these results support the coordinate-only distillation target adopted by ABOPD.

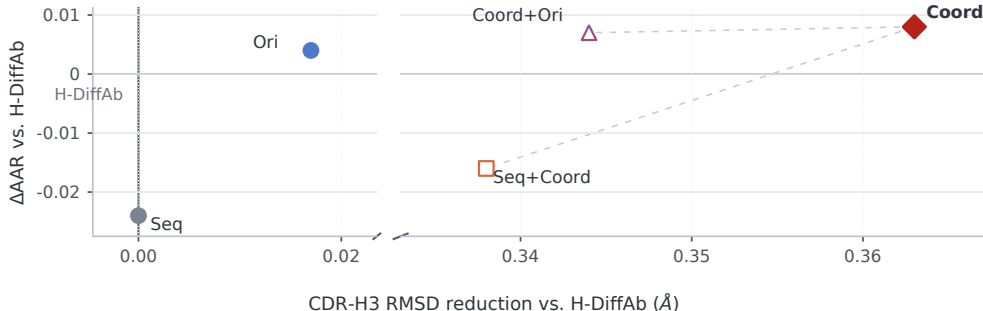


Figure 5 Distillation target ablation on RAbD CDR-H3. Each point reports the AAR gain and CDR-H3 RMSD reduction relative to H-DiffAb; the x-axis is broken for readability.

5 Conclusion

We introduce ABOPD, an on-policy distillation framework that improves the structural accuracy of antibody CDR design by transferring privileged geometric supervision to states visited along the student’s own trajectories. ABOPD reduces RAbD CDR-H3 RMSD from 2.37 Å to 1.95 Å and improves RMSD across all six CDRs in simultaneous redesign, outperforming SFT and offline distillation controls. Timestep-level analysis reveals a stage-dependent correspondence: the gap between rollout and reference-derived states widens during denoising, while Teacher corrections become most useful during later geometric refinement. These findings highlight that effective structural post-training depends not only on the supervision signal, but also on the generated states to which it is applied. ABOPD therefore provides a complementary direction to larger-scale pretraining and new generative architectures for more reliable antibody and protein design.

References

- [1] R. J. Pantazes and C. D. Maranas. OptCDR: a general computational method for the design of antibody complementarity determining regions for targeted epitope binding. *Protein Engineering Design and Selection*, 23(11):849–858, September 2010. ISSN 1741-0134. doi: 10.1093/protein/gzq061. URL <http://dx.doi.org/10.1093/protein/gzq061>.
- [2] Jared Adolf-Bryfogle, Oleks Kalyuzhnyi, Michael Kubitz, et al. RosettaAntibodyDesign (RABD): A general framework for computational antibody design. *PLOS Computational Biology*, 14(4):e1006112, April 2018. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006112. URL <http://dx.doi.org/10.1371/journal.pcbi.1006112>.
- [3] 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.
- [4] 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 *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.
- [5] Xiangzhe Kong, Wenbing Huang, and Yang Liu. End-to-end full-atom antibody design. In *Proceedings of the 40th International Conference on Machine Learning, ICML'23*. JMLR.org, 2023.
- [6] Haitao Lin, Lirong Wu, Yufei Huang, et al. GeoAB: towards realistic antibody design and reliable affinity maturation. In *Proceedings of the 41st International Conference on Machine Learning, ICML'24*. JMLR.org, 2024.
- [7] Karolis Martinkus, Jan Ludwiczak, Kyunghyun Cho, et al. AbDiffuser: full-atom generation of in vitro functioning antibodies. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, Red Hook, NY, USA, 2023. Curran Associates Inc.
- [8] Ali Madani, Ben Krause, Eric R. Greene, et al. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41:1099–1106, 2023. doi: 10.1038/s41587-022-01618-2.
- [9] Xiangxin Zhou, Dongyu Xue, Ruizhe Chen, Zaixiang Zheng, Liang Wang, and Quanquan Gu. Antigen-specific antibody design via direct energy-based preference optimization. In *Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS '24*, Red Hook, NY, USA, 2024. ISBN 9798331314385.
- [10] Mang Ning, Enver Sangineto, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Input perturbation reduces exposure bias in diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 26245–26265. PMLR, 2023.
- [11] Richard W. Shuai, Jeffrey A. Ruffolo, and Jeffrey J. Gray. IgLM: Infilling language modeling for antibody sequence design. *Cell Systems*, 14(11):979–989.e4, November 2023. ISSN 2405-4712. doi: 10.1016/j.cels.2023.10.001.
- [12] Rosalie Lipsh-Sokolik, Dina Listov, and Sarel J. Fleishman. The AbDesign computational pipeline for modular backbone assembly and design of binders and enzymes. *Protein Science*, 30(1):151–159, October 2020. ISSN 1469-896X. doi: 10.1002/pro.3970. URL <http://dx.doi.org/10.1002/pro.3970>.
- [13] Xiangzhe Kong, Wenbing Huang, and Yang Liu. Conditional antibody design as 3D equivariant graph translation. In *The Eleventh International Conference on Learning Representations*, 2023.
- [14] Wengong Jin, Jeremy Wohlwend, Regina Barzilay, and Tommi S. Jaakkola. Iterative refinement graph neural network for antibody sequence-structure co-design. In *International Conference on Learning Representations*, 2022.
- [15] Nathaniel R. Bennett, Joseph L. Watson, Robert J. Ragotte, et al. Atomically accurate de novo design of antibodies with RFdiffusion. *Nature*, 649(8095):183–193, November 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09721-5. URL <http://dx.doi.org/10.1038/s41586-025-09721-5>.
- [16] Rubo Wang, Fandi Wu, Xingyu Gao, Jiaxiang Wu, Peilin Zhao, and Jianhua Yao. IgGM: A generative model for functional antibody and nanobody design. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [17] Wenda Wang, Yang Zhang, Zhewei Wei, and Wenbing Huang. AbFlow: End-to-end paratope-centric antibody design by interaction enhanced flow matching. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery*

and Data Mining V.1, KDD '26, page 1471–1482, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722585. doi: 10.1145/3770854.3780296. URL <https://doi.org/10.1145/3770854.3780296>.

- [18] Odin Zhang, Xujun Zhang, Haitao Lin, et al. ODesign: A world model for biomolecular interaction design, 2025.
- [19] Qilin Yu, Liangyue Guo, Xiayan Qin, et al. High-affinity protein binder design via flow matching and in silico maturation. *bioRxiv*, January 2026. doi: 10.64898/2026.01.19.700484. URL <http://dx.doi.org/10.64898/2026.01.19.700484>.
- [20] Ivan Anishchenko, Samuel J. Pellock, Tamuka M. Chidyausiku, et al. De novo protein design by deep network hallucination. *Nature*, 600(7889):547–552, December 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-04184-w. URL <http://dx.doi.org/10.1038/s41586-021-04184-w>.
- [21] J. Dauparas, I. Anishchenko, N. Bennett, et al. Robust deep learning-based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, October 2022. ISSN 1095-9203. doi: 10.1126/science.add2187. URL <http://dx.doi.org/10.1126/science.add2187>.
- [22] Joseph L. Watson, David Juergens, Nathaniel R. Bennett, et al. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, July 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06415-8. URL <http://dx.doi.org/10.1038/s41586-023-06415-8>.
- [23] John B. Ingraham, Max Baranov, Zak Costello, et al. Illuminating protein space with a programmable generative model. *Nature*, 623(7989):1070–1078, November 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06728-8. URL <http://dx.doi.org/10.1038/s41586-023-06728-8>.
- [24] Jason Yim, Brian L. Trippe, Valentin De Bortoli, et al. SE(3) diffusion model with application to protein backbone generation. In *Proceedings of the 40th International Conference on Machine Learning, ICML'23*. JMLR.org, 2023.
- [25] Jason Yim, Andrew Campbell, Andrew Y. K. Foong, et al. Fast protein backbone generation with SE(3) flow matching. In *NeurIPS 2023 Workshop on Machine Learning in Structural Biology*, 2023.
- [26] Joey Bose, Tara Akhound-Sadegh, Guillaume Huguet, et al. SE(3)-stochastic flow matching for protein backbone generation. In *The Twelfth International Conference on Learning Representations*, 2024.
- [27] Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, et al. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*, 2024.
- [28] Dengyang Jiang, Xin Jin, Dongyang Liu, et al. D-OPSD: On-policy self-distillation for continuously tuning step-distilled diffusion models, 2026.
- [29] Zhen Fang, Wenxuan Huang, Yu Zeng, et al. Flow-OPD: On-policy distillation for flow matching models, 2026.
- [30] Quanhao Li, Junqiu Yu, Kaixun Jiang, et al. DiffusionOPD: A unified perspective of on-policy distillation in diffusion models, 2026.
- [31] Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics*, 9(1):48, 2017. doi: 10.1186/s13321-017-0235-x. URL <https://doi.org/10.1186/s13321-017-0235-x>.
- [32] Jiaxuan You, Bowen Liu, Rex Ying, Vijay Pande, and Jure Leskovec. Graph convolutional policy network for goal-directed molecular graph generation. In *Advances in Neural Information Processing Systems*, volume 31, 2018.
- [33] Zhenpeng Zhou, Steven Kearnes, Li Li, Richard N. Zare, and Patrick Riley. Optimization of molecules via deep reinforcement learning. *Scientific Reports*, 9(1):10752, 2019. doi: 10.1038/s41598-019-47148-x. URL <https://doi.org/10.1038/s41598-019-47148-x>.
- [34] Zhuo Yang, Yeyun Chen, Jiaqing Xie, et al. MolAct: An agentic RL framework for molecular editing and property optimization, 2025.
- [35] Matt Sternke and Joel Karpiak. ProteinRL: Reinforcement learning with generative protein language models for property-directed sequence design. In *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*, 2023.
- [36] Yasha Ektefaie, Olivia Viessmann, Siddharth Narayanan, Drew Dresser, J. Mark Kim, and Armen Mkrtchyan. Reinforcement learning on structure-conditioned categorical diffusion for protein inverse folding, 2024.

- [37] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL <https://openreview.net/forum?id=qw8AKxfYbI>.
- [38] Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: similarity control and knowledge transfer. *J. Mach. Learn. Res.*, 16(1):2023–2049, January 2015. ISSN 1532-4435.
- [39] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations*, 2016.
- [40] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.
- [41] James Dunbar, Konrad Krawczyk, Jinwoo Leem, et al. SABDab: the structural antibody database. *Nucleic Acids Research*, 42(D1):D1140–D1146, 2014. ISSN 1362-4962. doi: 10.1093/nar/gkt1043.
- [42] Xiangzhe Kong, Yinjun Jia, Wenbing Huang, and Yang Liu. Full-atom peptide design with geometric latent diffusion. In *Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS '24*, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.
- [43] Fang Wu, Weihao Xuan, Heli Qi, et al. Proteo-R1: Reasoning foundation models for de novo protein design. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.
- [44] Andrew Leaver-Fay, Michael Tyka, Steven M. Lewis, et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods in Enzymology*, 487:545–574, 2011. doi: 10.1016/B978-0-12-381270-4.00019-6.
- [45] Rebecca F. Alford, Andrew Leaver-Fay, Jeliasko R. Jeliaskov, et al. The Rosetta all-atom energy function for macromolecular modeling and design. *Journal of Chemical Theory and Computation*, 13(6):3031–3048, 2017. ISSN 1549-9626. doi: 10.1021/acs.jctc.7b00125.
- [46] Sankar Basu and Björn Wallner. DockQ: A quality measure for protein-protein docking models. *PLOS ONE*, 11(8): e0161879, 2016. doi: 10.1371/journal.pone.0161879.
- [47] Xiangzhe Kong, Zishen Zhang, Ziting Zhang, et al. UniMoMo: Unified generative modeling of 3D molecules for de novo binder design. In *Forty-second International Conference on Machine Learning*, 2025.
- [48] Tian Zhu, Milong Ren, and Haicang Zhang. Antibody design using a score-based diffusion model guided by evolutionary, physical and geometric constraints. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 62531–62548, 21–27 Jul 2024.
- [49] Jeffrey A. Ruffolo, Lee-Shin Chu, Sayan P. Mahajan, and Jeffrey J. Gray. IgFold: Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nature Communications*, 14:2389, 2023. doi: 10.1038/s41467-023-38063-x.
- [50] Yumeng Yan, Huanyu Tao, Jiahua He, and Sheng-You Huang. The HDock server for integrated protein-protein docking. *Nature Protocols*, 15:1829–1852, 2020. doi: 10.1038/s41596-020-0312-x.
- [51] Weihong Huang, Feng Yang, Qiang Zhang, and Juan Liu. Dynamic geometric equivariant network for full-atom antibody design. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(1):444–452, March 2026. ISSN 2159-5399. doi: 10.1609/aaai.v40i1.37007. URL <http://dx.doi.org/10.1609/aaai.v40i1.37007>.
- [52] Hannes Stark, Felix Faltings, MinGyu Choi, et al. BoltzGen: Toward universal binder design. *bioRxiv*, November 2025. doi: 10.1101/2025.11.20.689494. URL <http://dx.doi.org/10.1101/2025.11.20.689494>.
- [53] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. In *Advances in Neural Information Processing Systems*, volume 37, 2024.
- [54] Javad Rajabi, Soroush Mehraban, Seyedmorteza Sadat, and Babak Taati. Token perturbation guidance for diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.
- [55] Dongxu Zhang, Zhichao Yang, Sepehr Janghorbani, et al. Fast and effective on-policy distillation from reasoning prefixes. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 25553–25569. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1276.

A Supplementary Experimental Results

A.1 Additional Teacher Evaluation

Table 4 Teacher training variants. The adapter-only Teacher achieves the highest AAR, whereas the backbone-aware Teacher minimizes both CDR geometry errors and provides the frozen structural target used by ABOPD. Both RMSDs evaluate generated CDR geometry rather than the fixed framework. Best and second-best results are shown in **bold** and underlined, respectively.

Variant	AAR (%) $\uparrow$	CDR C α RMSD ($\text{\AA}$) $\downarrow$	CDR BB RMSD ($\text{\AA}$) $\downarrow$
H-DiffAb	<u>60.56</u>	1.4863	1.5298
Teacher w/o Backbone	57.92	1.0520	1.1428
Adapter-only Teacher	61.48	<u>0.9133</u>	<u>0.9962</u>
Backbone-aware Teacher	59.26	0.4855	0.6642

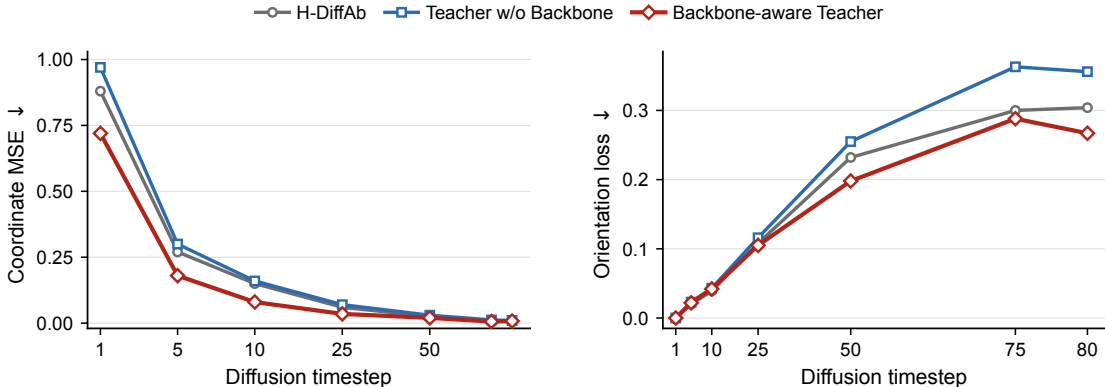


Figure 6 Timestep-resolved Teacher evaluation. On shared reference-derived noisy states, backbone-aware conditioning most clearly reduces coordinate MSE at low-index timesteps and orientation loss at higher-index timesteps. Orientation loss is diagnostic only and is not distilled.

A.2 Additional Post-Training Trajectories

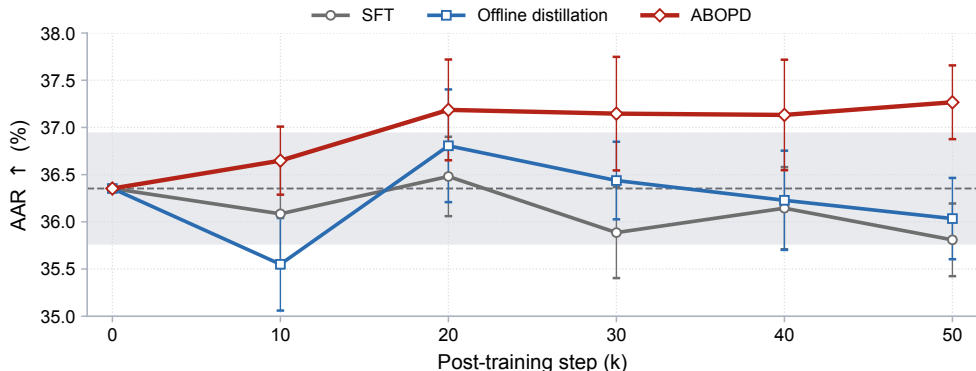


Figure 7 Post-training AAR trajectories on RAbD CDR-H3. ABOPD attains higher mean AAR than SFT, offline distillation, and the H-DiffAb initialization at every evaluated checkpoint after initialization. Markers and error bars show the mean and 95% confidence interval over three sampling seeds; the dashed line and gray band denote the H-DiffAb initialization.

A.3 Teacher-Correction Geometry Across Timesteps

The scalar Teacher advantage in Fig. 4(c) reveals whether the Teacher is closer than H-DiffAb to the native-conditioned one-step target, but not how the Teacher changes H-DiffAb’s transition. We therefore treat the difference between their predicted transition means as an additive correction. This view is motivated by diffusion-guidance analyses that interpret differences between denoiser predictions as corrections to a base vector field and characterize such corrections through their directional relation to a reference denoising signal and their norm across timesteps [53, 54]. Accordingly, we ask whether the Teacher correction points toward the native-conditioned transition and how large that correction is. This analogy is diagnostic only: the correction is not applied as inference-time guidance.

For a current coordinate $x_{t,i}$ and its corresponding native coordinate $x_{0,i}$, we construct the native-directed transition mean as

$$\hat{e}_{t,i}^* = \frac{x_{t,i} - \sqrt{\bar{\alpha}_t} x_{0,i}}{\sqrt{1 - \bar{\alpha}_t}}, \quad \mu_{t,i}^* = a_t x_{t,i} + b_t \hat{e}_{t,i}^*.$$

Here, $\bar{\alpha}_t$ denotes the cumulative signal-retention coefficient of the shared diffusion schedule. The term “reference-derived” describes the source of the state rather than a model prediction. A reference-derived state is obtained by forward-noising the native coordinates as $x_{t,i} = \sqrt{\bar{\alpha}_t} x_{0,i} + \sqrt{1 - \bar{\alpha}_t} \epsilon_{t,i}$. Consequently, $\hat{e}_{t,i}^*$ recovers the sampled forward noise and $\mu_{t,i}^*$ equals the mean of the native-conditioned posterior $q(x_{t-1,i} | x_{t,i}, x_{0,i})$ under the standard DDPM parameterization [40]. A rollout state is instead produced by H-DiffAb’s reverse trajectory and need not be a forward-noised realization of the paired native structure. On such a state, the same mapping defines a counterfactual, schedule-consistent native reference rather than the posterior mean of the rollout-generating process. Reference-derived states therefore serve as a forward-process control for determining whether the observed correction geometry is specific to model-visited states.

For each generated CDR residue $i \in \mathcal{G}$, let

$$\delta_{t,i} = \mu_{T,t,i}^{\text{pos}} - \mu_{H,t,i}^{\text{pos}}, \quad d_{t,i}^* = \mu_{t,i}^* - \mu_{H,t,i}^{\text{pos}}.$$

Viewed from the H-DiffAb transition mean, $\delta_{t,i}$ is the displacement to the Teacher prediction, whereas $d_{t,i}^*$ is the remaining displacement to the native-conditioned target. We measure their cosine alignment and the RMS magnitude of the Teacher correction as

$$\text{Align}_t = \frac{\sum_{i \in \mathcal{G}} \delta_{t,i}^\top d_{t,i}^*}{\sqrt{\sum_{i \in \mathcal{G}} \|\delta_{t,i}\|_2^2} \sqrt{\sum_{i \in \mathcal{G}} \|d_{t,i}^*\|_2^2}}, \quad \text{Mag}_t = \sqrt{\frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} \|\delta_{t,i}\|_2^2}. \quad (16)$$

An alignment near one indicates that the Teacher changes H-DiffAb in the same direction as the native-conditioned displacement; zero indicates an orthogonal change, and negative values indicate a change away from that direction. Magnitude reports the RMS transition-mean displacement per generated residue in Å. Alignment is directional and magnitude is scale-only; neither alone establishes improvement, which is assessed separately by Teacher advantage in Fig. 4(c). For each complex-seed state, alignment is computed after concatenating the generated-residue Ca vectors; results are then averaged over seeds and macro-averaged across complexes. Coordinates remain in the fixed complex frame, and the coordinate scale is restored before reporting magnitude.

Beginning at $t = 50$, where Teacher advantage becomes positive, Align_t on H-DiffAb rollout states increases from 0.347 to 0.855 at $t = 10$, while Mag_t increases from 0.115 to 0.272 Å. The corresponding reference-derived curves remain comparatively stable over the same interval (Fig. 8). Thus, the late-timestep Teacher advantage is accompanied by corrections that become both larger and increasingly native-directed specifically on rollout states. This provides a geometric interpretation of the stage-dependent transition behavior rather than an independent measure of final generation quality.

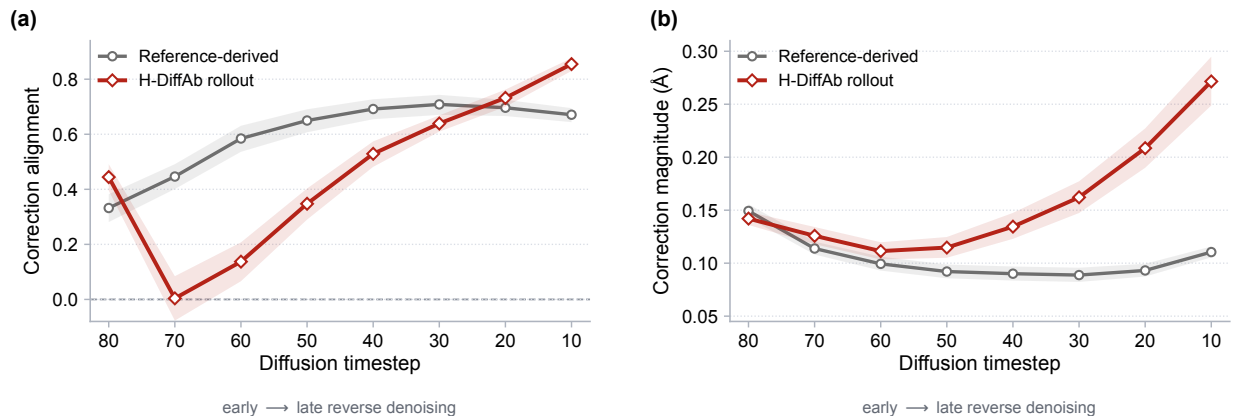


Figure 8 Teacher-correction geometry on RAbD CDR-H3. (a) Cosine alignment between the Teacher correction relative to H-DiffAb and the native-directed correction; higher values indicate a more native-directed correction. (b) RMS correction magnitude in Å, which reports correction size rather than quality alone. Gray curves show reference-derived (forward-noised native) states as a forward-process control; red curves show H-DiffAb rollout states. Curves show macro means over 60 complexes after averaging three fixed seeds; bands denote 95% complex-level bootstrap confidence intervals.

A.4 Per-CDR redesign

Table 5 Per-CDR redesign on the 19 SABDab test complexes used for simultaneous multi-CDR redesign. Each CDR is redesigned independently. AAR is averaged over all target cases, whereas RMSD and ddG are reported separately for each redesigned CDR. ABOPD achieves the highest overall AAR and the lowest H3 RMSD and ddG, while H-DiffAb performs better on several shorter or light-chain CDR metrics. Best and second-best results are shown in **bold** and underlined, respectively.

Method	AAR (%) ↑	Per-CDR RMSD (Å) ↓						Per-CDR ddG ↓					
		H1	H2	H3	L1	L2	L3	H1	H2	H3	L1	L2	L3
DiffAb	50.89	0.880	0.754	3.359	1.212	1.050	1.070	4.45	5.17	31.17	4.64	3.70	0.02
H-DiffAb	61.72	0.778	0.662	<u>3.049</u>	0.954	0.991	0.987	<u>2.56</u>	0.28	<u>20.31</u>	3.01	-0.20	1.89
ABOPD	62.29	0.727	<u>0.688</u>	2.547	0.945	<u>1.032</u>	<u>0.996</u>	1.14	<u>0.59</u>	11.52	2.86	<u>0.89</u>	3.17

A.5 EMA analysis

Table 6 EMA ablation across evaluation settings. The upper block compares ABOPD with and without EMA, while the lower reports paired per-CDR differences computed as EMA minus non-EMA; negative Δ RMSD and positive Δ AAR favor EMA. EMA improves all aggregate metrics in the upper block and per-CDR RMSD for five of six CDRs, with a +0.56 trade-off in aggregated per-CDR ddG. Best and second-best results are **bold** and underlined, respectively.

Method	Simultaneous multi-CDR redesign			RAbD CDR-H3	
	AAR (%) ↑	RMSD (Å) ↓	ddG ↓	AAR (%) ↑	RMSD (Å) ↓
ABOPD w/o EMA	<u>60.53</u>	<u>1.5357</u>	<u>26.32</u>	<u>36.25</u>	<u>1.9954</u>
ABOPD w/ EMA	60.73	1.4972	20.57	37.31	1.9500

Paired effect of EMA on per-CDR redesign								
Difference	Δ AAR (pp) ↑	Δ RMSD (Å) ↓						Δ ddG ↓
		H1	H2	H3	L1	L2	L3	
w/ EMA – w/o EMA	+0.66	-0.007	-0.005	<u>-0.055</u>	-0.033	+0.002	-0.003	+0.56

A.6 CDR-H3 interface and side-chain diagnostics

We additionally report interface-improvement percentage (IMP) and side-chain dihedral realism to distinguish backbone-geometry gains from thresholded interface behavior and side-chain packing. IMP is the percentage of designs whose Rosetta InterfaceAnalyzer energy improves over the native reference after the common relaxation pipeline; JSD_{sc} is computed after side-chain packing and relaxation, with lower values indicating closer side-chain dihedral distributions. Table 7 reports these additional diagnostics.

Table 7 CDR-H3 interface-improvement and side-chain-dihedral diagnostics. IMP is reported as a percentage; lower JSD_{sc} is better. Best results are **bold** and second-best are underlined.

Method	IMP (%) $\uparrow$	JSD_{sc} $\downarrow$
DiffAb	15.84	0.2368
H-DiffAb	24.84	<u>0.2451</u>
ABOPD	<u>23.32</u>	0.2566

ABOPD attains the best CDR-H3 RMSD in Table 3 and lowers mean ddG in the simultaneous and per-CDR evaluations in Tables 2 and 5, but its IMP remains slightly below H-DiffAb and its JSD_{sc} is higher. Thus, lower mean backbone RMSD and ddG do not imply improved side-chain realism or a larger fraction of designs crossing the $ddG < 0$ threshold.

A.7 Additional qualitative CDR-H3 cases

Figures 9–13 show five additional RAbD cases, where only CDR-H3 is redesigned. They span short and difficult long-loop conformations; ABOPD generally follows the native trace more closely, although the hardest case remains challenging.

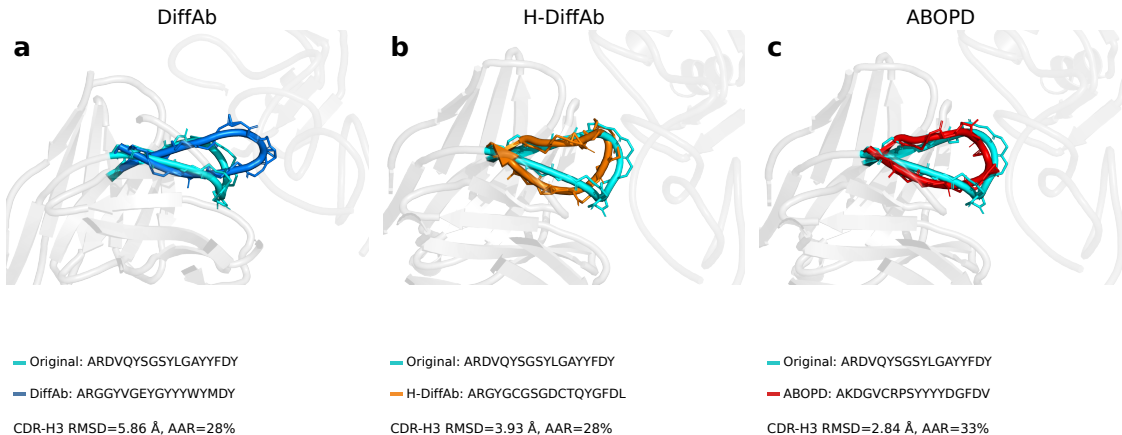


Figure 9 Qualitative CDR-H3 comparison for RAbD target 4fqj.

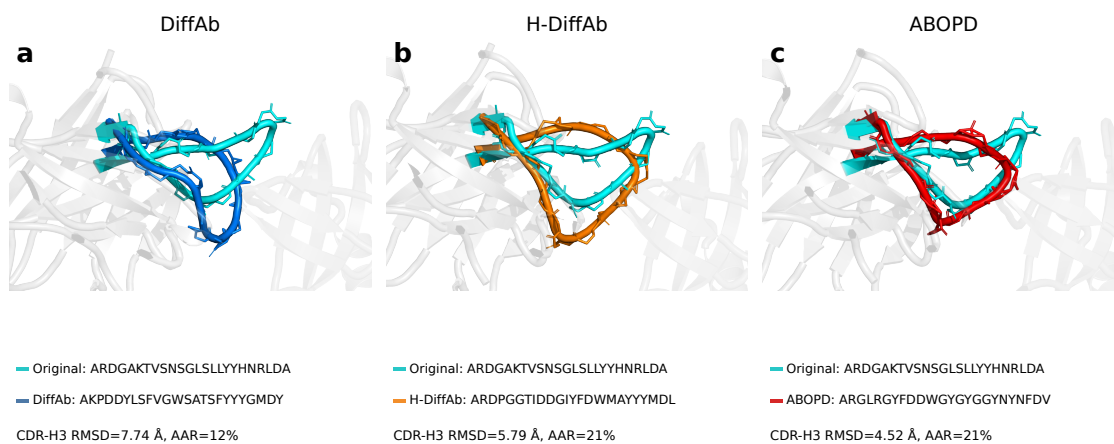


Figure 10 Qualitative CDR-H3 comparison for RAbD target 4ot1.

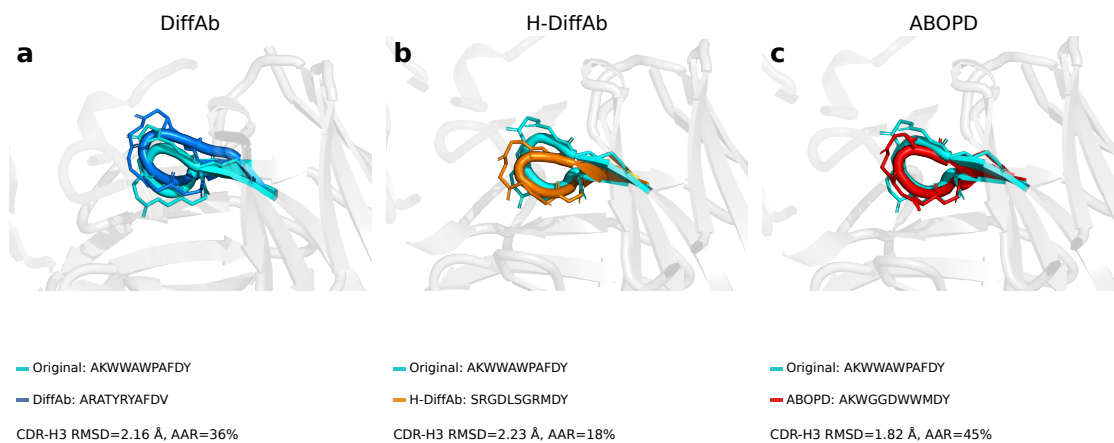


Figure 11 Qualitative CDR-H3 comparison for RAbD target 3k2u.

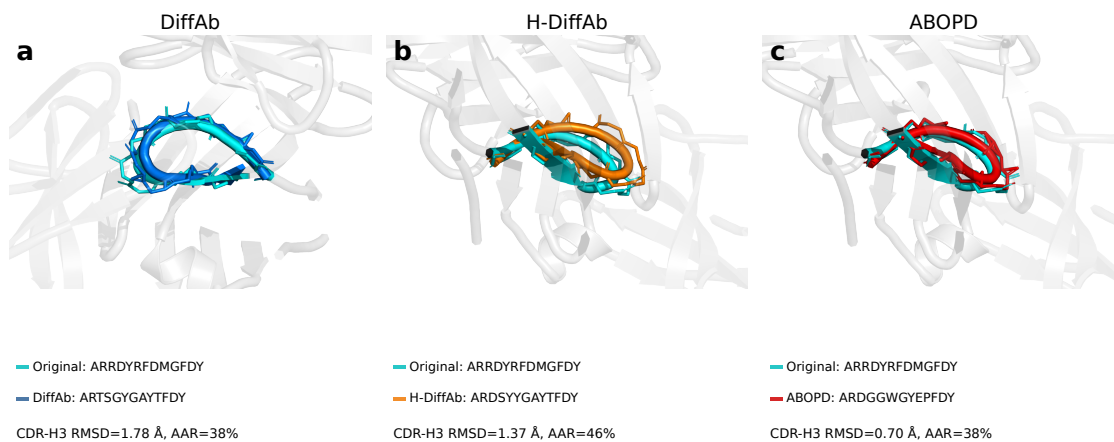


Figure 12 Qualitative CDR-H3 comparison for RAbD target 5ggs.

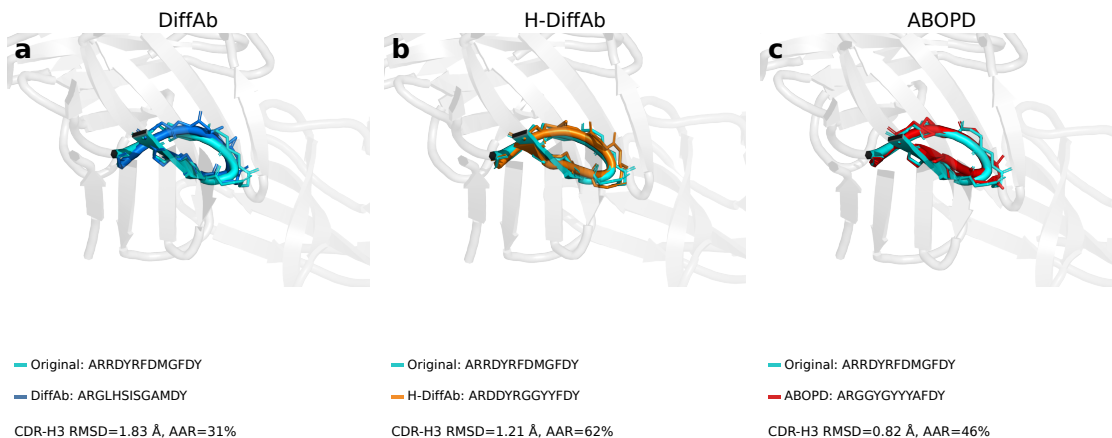


Figure 13 Qualitative CDR-H3 comparison for RAbD target 5b8c.

B Training and Implementation Details

Unless otherwise stated, coordinate preprocessing and diffusion scheduling follow the public DiffAb implementation [4]. All stages use Adam, unit weights for the three denoising terms, and the same target-mask distribution. Table 8 summarizes the stage-specific configuration. For the post-training controls, SFT uses the denoising anchor’s timestep sampling, whereas offline distillation uses the same fixed teacher–student matching timesteps as ABOPD; both are evaluated without EMA. ABOPD uses its EMA student unless otherwise specified. H-DiffAb pretraining processes approximately 8× as many examples as the default DiffAb configuration.

Table 8 Unified training configuration for H-DiffAb pretraining, backbone-aware teacher adaptation, and ABOPD post-training.

Component	H-DiffAb pretraining	Teacher adaptation	ABOPD student
Target-mask distribution		$p_{\text{hyb}}(\text{single}, \text{multi}, \text{all}) = (0.25, 0.40, 0.35)$	
Timestep sampling	$\mathcal{U}\{1, \dots, 100\}$	$\mathcal{U}\{1, \dots, 100\}$	Anchor: $\mathcal{U}\{1, \dots, 80\}$
Rollout start index	–	–	$K = 80$ on the original 100-step schedule
Rollout timesteps	–	–	$\{80, 70, 60, 50, 40, 30, 20, 10, 5\}$
Distillation weights	–	–	Coordinate: 0.6; sequence/orientation: 0
EMA usage	–	–	$\rho = 0.9995$; updated after each student step; evaluation and sampling only
Learning rate	10^{-4}	10^{-4}	10^{-5}
Training steps	200k	30k	50k
Global batch size	128	128	32

Privileged descriptor and Teacher adaptation. The privileged descriptor B is injected through two residual adapters between the standard H-DiffAb encoder and EpsilonNet. It enriches the encoder representations without replacing or modifying the noisy diffusion state R_t ; the student and Teacher therefore receive the same R_t , but only the Teacher receives B .

For each target residue, the residue descriptor concatenates the native N, C α , C, O, and C β coordinates, five corresponding atom-validity masks, and a target-region indicator, yielding 21 input dimensions. Coordinates are centered at the residue’s native C α position and divided by 10 Å, while retaining the original Cartesian axes of the complex rather than constructing a rotation-normalized residue frame. A $21 \rightarrow 128 \rightarrow 128$ ReLU MLP maps this descriptor to a masked residual that is added to the base residue embedding.

For each target–complex residue pair with C α distance d_{ij} , the pair descriptor contains $\exp[-(d_{ij}/(10 \text{ Å}))^2]$,

$\min(d_{ij}/(100 \text{ \AA}), 1)$, and the target and context $C\alpha$ -validity indicators. A $4 \rightarrow 64 \rightarrow 64$ ReLU MLP encodes the masked directed features. Its output is averaged with its transpose and added to the base pair embedding. Output masking ensures zero adapter contributions for non-target residues and invalid pairs.

Explicit target amino-acid labels and atoms beyond $C\beta$ are excluded, although the observed $C\beta$ coordinate and validity mask may carry limited sequence-correlated information. In the main configuration, the H-DiffAb backbone and both adapters are jointly optimized under the Teacher-adaptation objective and frozen afterward. The two adapters add only 23.8K parameters in total; adapter-only optimization is evaluated only as an ablation.

C Baseline and Evaluation Protocols

We distinguish locally reproduced baselines from published baseline results. In the simultaneous multi-CDR redesign benchmark (Table 2), dyMEAN, AbX, PPIFlow, and IgGM were run by us on the same SAbDab split whenever an official implementation was available, using default hyperparameters unless a format bridge was required. DiffAb, H-DiffAb, and ABOPD are evaluated with the same sampling, relaxation, and diagnostic pipeline. H-DiffAb is our hybrid-pretrained DiffAb variant, and its parameters θ_0 initialize SFT, offline distillation, and ABOPD. In the RAbD CDR-H3 benchmark (Table 3), the RosettaAb, MEAN, GeoAB, dyMEAN, DGENet, and BoltzGen rows are copied from Wu et al. [43] and Kong et al. [47].

For a protocol-matched comparison with DiffAb, H-DiffAb, and ABOPD, we disabled PPIFlow’s strict post-generation acceptance filter, which otherwise retains a sample only when the framework RMSD is below 1 Å and no backbone clash is detected. In our SAbDab simultaneous multi-CDR run, this filter accepted only 64 of 1900 raw samples (3.37%). We therefore evaluate all raw PPIFlow samples, so that PPIFlow is compared as a generator under the same sample budget rather than as a filtered candidate-selection pipeline.

D Limitations and Future Directions

Structural fidelity and design diversity. ABOPD is designed to improve $C\alpha$ backbone trajectories, and its gains do not extend uniformly to side-chain packing or interface improvement, as reflected by JSD_{sc} and IMP in Table 7. The lower RMSD and clash rates, together with the more concentrated CDR-H3 sequence-space pattern in Fig. 3(a), are consistent with a fidelity–diversity trade-off: dense trajectory supervision may concentrate generation on structurally reliable modes while reducing exploration and limiting gains in IMP. Future work could extend trajectory-level supervision from $C\alpha$ backbone transitions to all-atom geometry, while developing diversity-aware objectives or sampling strategies that better balance structural fidelity with exploration of the sequence–structure space.

Task-aware trajectory supervision. ABOPD uniformly weights the selected rollout timesteps, whereas our timestep-level transition analysis on shared H-DiffAb rollout states indicates that the Teacher’s local advantage and ABOPD’s lower transition errors are concentrated toward the end of reverse denoising (Fig. 4(c)–(d)). By contrast, on-policy distillation for long-form mathematical reasoning exhibits larger token-level reverse-KL losses in early output prefixes [55]. Although autoregressive token positions and reverse-diffusion stages are not directly equivalent, these differing patterns suggest that supervision allocation may need to reflect task-specific generation dynamics. Because our temporal evidence is limited to CDR-H3 generation with the current DiffAb-based model, future work should examine adaptive timestep selection or weighting across antibody CDR design, protein-binder design, and broader protein-design settings.