DISAB: SELF-CONSISTENT ANTIBODY SEQUENCE–STRUCTURE CO-DESIGN WITH MULTI-CHAIN ALL-ATOM DISCRETE DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Antibody design requires amino-acid sequence and 3D structure to remain mutually compatible under antigen conditions. Existing generative methods often model sequence in a discrete space but structure in continuous geometric spaces, resulting in heterogeneous generation processes that are difficult to synchronize and reuse across different design tasks. We introduce **DisAb** (**Dis**crete Diffusion for **A**nti**b**ody), a unified framework that represents antibody sequence and multi-chain all-atom structure as residue-aligned discrete tokens and jointly generates them through a single masked discrete diffusion process. To enable this formulation, we develop **DisAb-VQ**, to our knowledge the first residue-level VQ-VAE that tokenizes multi-chain atom14 protein geometry while preserving intra-chain structure and relative inter-chain placement. We further introduce Boltz-2-based sequence–structure self-consistency as a reference-free potential for sequential Monte Carlo (SMC) guidance at inference time. We evaluate DisAb in both bound-antigen design and the more challenging unbound setting, where the complete bound antigen–VH–VL complex is generated from an isolated antigen monomer. DisAb-VQ accurately reconstructs multi-chain all-atom structures, while DisAb achieves an Aligned DockQ of 0.692 in the bound setting, improved to 0.711 with SMC guidance. In the unbound setting, SMC improves raw DockQ from 0.129 to 0.149 while also improving VH–VL geometry, CDR sequence recovery, and sequence–structure self-consistency. These results demonstrate that unified discrete modeling provides a flexible framework for antibody design.

1 INTRODUCTION

Antibodies are among the most successful molecular modalities in modern biomedicine, owing to their high target specificity, modular architecture, and broad engineering potential. Yet discovering antibodies against a desired antigen remains experimentally intensive, typically requiring large-scale screening followed by iterative optimization of affinity, specificity, and developability (Carter & Lazar, 2018; Sormanni et al., 2018). Recent advances in AI-based antibody design have begun to accelerate this process, enabling antigen-conditioned sequence generation, structure modeling, and sequence–structure co-design (Kong et al., 2023b;a; Luo et al., 2022; Wang et al., 2025b; Bennett et al., 2026).

Early antibody co-design methods such as MEAN and dyMEAN (Kong et al., 2023b;a) directly predict or refine antibody sequences and structures, but such predictive formulations provide limited ability to explore diverse solutions. Methods such as DiffAb (Luo et al., 2022), IgGM (Wang et al., 2025b), RFdiffusion for antibody design (Bennett et al., 2026), and related models introduce stochastic generation to sample diverse candidate structures and/or sequences. Yet most existing methods generate sequence and structure through heterogeneous mechanisms: sequence is modeled in a categorical space, whereas geometry is generated through continuous coordinates, rigid-body translations, and rotations. As a result, the two modalities follow different stochastic processes and objectives, making their denoising trajectories difficult to synchronize and preventing information revealed in one modality from being naturally shared with the other. This heterogeneity also complicates task unification, as sequence-to-structure prediction, structure-to-sequence design, and joint generation often require different generative procedures.

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.

A promising alternative is to represent both sequence and structure as discrete tokens. ESM3 (Hayes et al., 2025) and DPLM-2 (Wang et al., 2025c) show that aligned sequence and structural tokens can be generated within a common masked prediction or discrete diffusion framework. However, existing structural tokenizers are primarily designed for single-chain proteins and/or backbone geometry. Antibody design imposes a substantially stronger requirement: the representation must preserve not only local residue geometry, but also heavy–light chain packing and the relative pose between antibody and antigen. A unified discrete framework for multi-chain, all-atom antibody complexes therefore remains missing.

To address these limitations, we introduce **DisAb** (**Dis**crete Diffusion for **A**nti**b**ody), a unified generative framework that models antibody sequence and multi-chain all-atom structure through a *single discrete diffusion process*. The key idea is simple: if both amino-acid identities and 3D structures can be represented as residue-aligned discrete tokens, they can be corrupted, denoised, conditioned, and generated using the same stochastic mechanism. DisAb represents an antigen–VH–VL system using aligned tracks for amino-acid identity, structural tokens, and partition/conditioning information, and jointly denoises the generated tracks with masked discrete diffusion.

Enabling unified discrete generation requires a structural vocabulary that captures multi-chain all-atom geometry. We therefore develop **DisAb-VQ**, to our knowledge the first residue-level VQ-VAE that encodes a multi-chain protein complex into one structural token per residue. Its SE(3)-invariant encoder captures both intra- and inter-chain geometry, while the decoder reconstructs atom14 structures, including relative chain placement. The resulting residue-aligned sequence and structure tokens can thus be jointly modeled within the same categorical diffusion process. We further introduce **Boltz-2-guided sequential Monte Carlo (SMC)** for inference-time guidance, using reference-free sequence–structure self-consistency to reweight diffusion particles and steer generation toward mutually compatible sequence–structure pairs without retraining DisAb.

We evaluate DisAb in two antibody-design settings of increasing difficulty. **DisAb-bound** conditions on the antigen sequence, its antibody-bound conformation, specified epitope hotspots, and the antibody framework sequence. It jointly generates the CDR sequence and the VH–VL structure without using native antibody framework coordinates. **DisAb-unbound** removes the bound-antigen structural condition: given an independently available antigen monomer structure, specified epitope hotspots, and the antibody framework sequence, it jointly generates the CDR sequence and the complete bound antigen–VH–VL complex, including the antibody–antigen pose. DisAb-VQ accurately reconstructs multi-chain all-atom structures, reaching multi-chain backbone RMSDs of $0.683\,\text{Å}$ on CAMEO2024 and $1.086\,\text{Å}$ on SAbDab. In antibody design, DisAb-bound achieves an Aligned DockQ of 0.692, improved to 0.711 with SMC guidance. In the substantially harder unbound setting, SMC improves raw DockQ from 0.129 to 0.149, while also improving VH–VL geometry, CDR sequence recovery, and sequence–structure self-consistency. These results demonstrate that unified discrete modeling can support both bound-context redesign and prospective complex generation within the same sequence–structure generative framework.

In summary, our contributions are threefold:

- **Unified discrete antibody co-design.** We introduce DisAb, which represents sequence and structure as aligned categorical variables and jointly generates them through a single masked discrete diffusion process.
- **Multi-chain all-atom structural tokenization.** We develop DisAb-VQ, to our knowledge the first residue-level VQ-VAE that compresses multi-chain atom14 protein geometry—including local structure and relative inter-chain placement—into one structural token per residue, enabling antibody–antigen complexes to be modeled directly by discrete generative models.
- **Reference-free self-consistency guidance.** We introduce Boltz-2 sequence–structure self-consistency and incorporate it into an SMC-guided diffusion search procedure, improving generated sequence–structure compatibility and overall design quality at inference time without retraining DisAb.

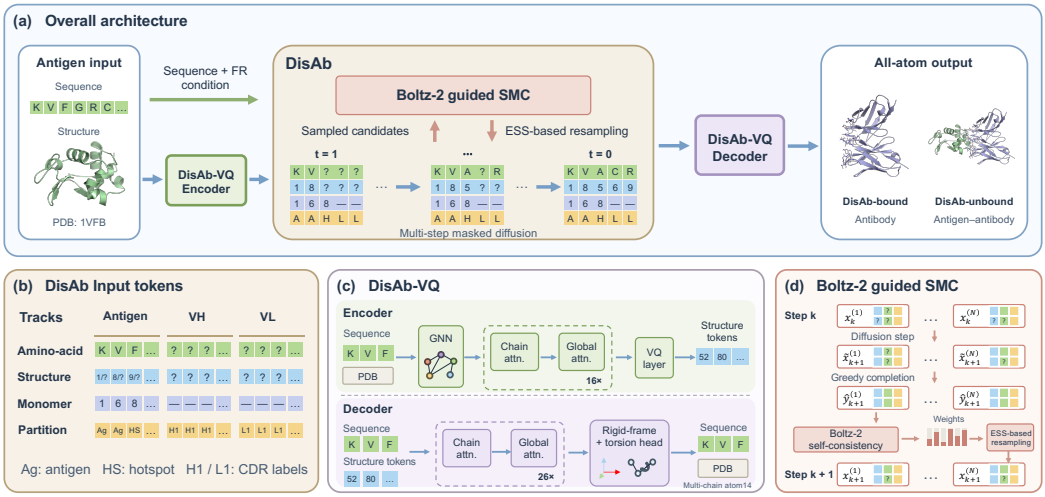


Figure 1: **Overview of DisAb.** **(a)** Overall pipeline. Given antigen conditions and an antibody framework sequence, DisAb jointly generates CDR amino-acid and discrete structure tokens through masked diffusion; DisAb-VQ decodes the latter into all-atom geometry. Boltz-2-guided SMC optionally guides generation. **(b)** Aligned residue tracks for antigen, VH, and VL, selectively observed or generated under the bound/unbound conditioning contracts. **(c)** DisAb-VQ encodes and decodes multi-chain all-atom geometry with an 8,192-entry vocabulary. **(d)** SMC scores greedy particle completions with Boltz-2 self-consistency and performs ESS-based reweighting and resampling.

2 Related Work

Structural tokenization. FoldToken, FoldToken2, GCP-VQ-VAE, ESM3, and DPLM-2 learn compact residue-level structural alphabets (Gao et al., 2024b;a; Pourmirzaei et al., 2025; Hayes et al., 2025; Wang et al., 2025c). These methods primarily target backbone geometry and are mostly designed for single-chain proteins or do not explicitly preserve multi-chain complex geometry. In contrast, DisAb-VQ discretizes multi-chain all-atom protein structures while retaining inter-chain placement in the same residue-level vocabulary.

Discrete protein generation. Categorical and masked diffusion provide a natural framework for generation and conditional completion over discrete states (Austin et al., 2021; Lou et al., 2024; Shi et al., 2024; Sahoo et al., 2024). ESM3 performs iterative generation over sequence, structure, and function tokens, while DPLM and DPLM-2 apply discrete diffusion to protein sequence and sequence–structure generation (Hayes et al., 2025; Wang et al., 2024; 2025c). DisAb extends this paradigm to multi-chain all-atom structural tokens, so sequence and structure are generated within one categorical process.

Antibody sequence–structure design. MEAN, DiffAb, dyMEAN, IgGM, AbFlow, RFdiffusion (antibody), Germinal, IgDiff, and AbDiffuser explore antibody sequence–structure design under different conditioning assumptions (Kong et al., 2023b; Luo et al., 2022; Kong et al., 2023a; Wang et al., 2025b; 2026; Bennett et al., 2026; Mille-Fragoso et al., 2026; Cutting et al., 2025; Martinkus et al., 2023). Several methods rely on native antibody framework coordinates or structural scaffolds. DisAb-bound and DisAb-unbound require no native framework coordinates; the latter further conditions only on an independently available antigen monomer rather than the native bound antigen conformation. Unlike IgGM, which combines categorical sequence diffusion with continuous coordinate/orientation diffusion, DisAb performs sequence–structure generation entirely in a shared discrete state.

Inference-time guidance. SMC-based guidance has been developed for continuous diffusion through twisted sampling and Feynman–Kac formulations, and for discrete diffusion through importance weighting and proposal design (Wu et al., 2023; Skreta et al., 2025; Kelvinius & Lindsten,

2023; Ou et al., 2026). We apply this framework to joint protein sequence–structure generation, using Boltz-2 self-consistency as a native-reference-free potential for SMC-guided antibody design.

3 NOTATIONS AND DEFINITIONS

This section introduces the aligned discrete representation and formally defines the two antibody-design tasks considered in this work.

Antigen–antibody complexes. An antigen–antibody complex contains L residues indexed by the disjoint sets $\mathcal{I}_A$, $\mathcal{I}_{VH}$, and $\mathcal{I}_{VL}$ for the antigen, heavy-chain variable domain, and light-chain variable domain, respectively. We write $\mathcal{I} = \mathcal{I}_A \cup \mathcal{I}_{VH} \cup \mathcal{I}_{VL}$. Within the antibody, $\mathcal{I}_{CDR}$ contains the six complementarity-determining regions, and $\mathcal{I}_{FR} = (\mathcal{I}_{VH} \cup \mathcal{I}_{VL}) \setminus \mathcal{I}_{CDR}$ contains the framework residues. The amino-acid sequence is $\boldsymbol{a} = (a_i)_{i=1}^{L} \in \mathcal{A}^L$, where $\mathcal{A}$ is the ESM-2 token alphabet (Lin et al., 2022). The all-atom structure is represented by atom14 coordinates $\boldsymbol{X} \in \mathbb{R}^{L\times 14\times 3}$ and a validity mask $\boldsymbol{B} \in \{0,1\}^{L\times 14}$.

Aligned discrete representation. Let E_ϕ and D_ψ denote the structural encoder and decoder, and let $\mathcal{E} = \{\boldsymbol{e}_k\}_{k=1}^{K}$ be a codebook of size $K = 8192$. The tokenizer maps $\boldsymbol{X}$ to residue-aligned structural tokens $\boldsymbol{z} = (z_i)_{i=1}^{L} \in [K]^L$ and reconstructs $\widehat{\boldsymbol{X}} = D_\psi(\boldsymbol{z}, \boldsymbol{a})$. As illustrated in Figure 1(B), every residue position is associated with an amino-acid track $\boldsymbol{a}$, a target-complex structure track $\boldsymbol{z}$, partition labels $\boldsymbol{p} \in \mathcal{P}^L$, and epitope hotspot indicators $\boldsymbol{h}_{spot} \in \{0,1\}^L$ (with hotspots defined as antigen residues whose reconstructed pseudo-Cβ is at < 8 Å from any Chothia-defined VH/VL CDR pseudo-Cβ). For the unbound task, an additional track $\boldsymbol{m}$ contains structural tokens obtained by independently encoding the isolated antigen monomer; it is populated on $\mathcal{I}_A$ and padded elsewhere.

For any residue subset $S \subseteq \mathcal{I}$, the subscript S denotes track restriction to that subset. The superscript "bound" indicates tokens encoded from the target bound complex, whereas "mono" indicates an independently available antigen-monomer structure. Partition and hotspot tracks are always observed and are not prediction targets.

Bound-antigen design. Given the antigen sequence $\boldsymbol{a}_A$, its structure in the antibody-bound conformation $\boldsymbol{z}_A^{bound}$, the antibody framework sequence $\boldsymbol{a}_{FR}$, and task metadata, the bound task is

$$p_\theta\big(\boldsymbol{a}_{CDR}, \boldsymbol{z}_{VH,VL}^{bound} \mid \boldsymbol{a}_A, \boldsymbol{z}_A^{bound}, \boldsymbol{a}_{FR}, \boldsymbol{p}, \boldsymbol{h}_{spot}\big)\,. \tag{1}$$

The model generates all VH–VL structural tokens and the amino-acid identities of the CDRs without observing native antibody framework coordinates.

Unbound-antigen design. Given an independently available antigen monomer instead of its bound conformation, the unbound task is

$$p_\theta\big(\boldsymbol{a}_{CDR}, \boldsymbol{z}_{A,VH,VL}^{bound} \mid \boldsymbol{a}_A, \boldsymbol{a}_{FR}, \boldsymbol{m}_A^{mono}, \boldsymbol{p}, \boldsymbol{h}_{spot}\big)\,. \tag{2}$$

The model jointly generates the CDR sequence and structural tokens of the full bound antigen–VH–VL complex, including the antibody–antigen pose.

Conditional discrete generation. Let $\mathcal{R}_{gen}$ denote the generated tracks. For each $r \in \mathcal{R}_{gen}$, x_0^r is the clean target and x_t^r is its noisy state at diffusion time $t \in [0,1]$. We collect the noisy generated tracks in $\boldsymbol{x}_t$ and denote all observed tracks by c. The two tasks therefore share the same conditional model $p_\theta(\boldsymbol{x}_0 \mid c)$ and differ only in which aligned track–position pairs are generated or observed.

4 METHODOLOGY

As shown in Figure 1, DisAb first converts multi-chain all-atom geometry into residue-aligned discrete codes, then jointly denoises sequence and structure tokens, and can optionally guide the reverse process with Boltz-2 self-consistency at inference time.

4.1 DISAB-VQ: MULTI-CHAIN ALL-ATOM STRUCTURE TOKENIZATION

Architecture. We develop DisAb-VQ (Figure 1(C)), a 1.67B-parameter residue-level VQ-VAE that compresses a multi-chain protein complex into one of 8,192 structural codes per residue. Unlike backbone-oriented tokenizers, it preserves local residue geometry, relative inter-chain pose, and standard protein atom14 heavy-atom structure.

The encoder removes dependence on the arbitrary global coordinate frame through an SE(3)-invariant geometric graph frontend. Local residue frames are constructed from N–Cα–C atoms, and graph edges encode invariant intra- and inter-chain relations. A scalar invariant graph encoder is augmented by a five-layer invariant graph-attention residual branch. Each residue connects to its nearest spatial neighbors separately within and across chains; edge features encode Cα distance, local-frame displacement, relative residue-frame orientation, and chain/connectivity indicators. The resulting features pass through 16 Transformer blocks alternating *chain-local* and *complex-global* attention, inspired by VGGT (Wang et al., 2025a). Chain-local attention resets positional indices within each chain, whereas global attention jointly processes all residues and chain-level registers.

The encoder produces a 256-dimensional latent $\boldsymbol{u}_i = E_\phi(\boldsymbol{X})_i$ for each residue. Normalizing both latent and code vectors, cosine quantization selects

$$z_i = \arg\max_{k\in[K]} \frac{\boldsymbol{u}_i^\top \boldsymbol{e}_k}{\|\boldsymbol{u}_i\|_2\|\boldsymbol{e}_k\|_2}, \qquad K = 8192. \tag{3}$$

The decoder embeds these codes and applies 26 alternating chain-local/complex-global Transformer blocks. A rigid-frame head reconstructs the N–Cα–C backbone, including relative chain placement, while a separate torsion head predicts ψ and $\chi_{1:4}$. Amino-acid identity is provided only to this chemical reconstruction branch, preventing sequence from bypassing the structural bottleneck. Backbone frames, torsions, and residue templates are finally converted to atom14 coordinates.

Training objective. The production DisAb-VQ is optimized with

$$\begin{aligned}\mathcal{L}_{\text{codec}} = {} & 2\mathcal{L}_{bb} + 0.5\mathcal{L}_{dist} + \mathcal{L}_{int} + 0.5\mathcal{L}_{bond} + \mathcal{L}_{vec} + \mathcal{L}_{sc-local} \\ & + 4\mathcal{L}_{\chi} + 2\mathcal{L}_{\psi} + 0.25\mathcal{L}_{spread} + \mathcal{L}_{VQ},\end{aligned} \tag{4}$$

combining backbone, pairwise/interface, local-geometry, torsional, latent-spread, and cosine-VQ commitment objectives. The codebook is initialized by K-means, updated with synchronized EMA statistics, and maintained with bounded dead-code recycling. Full loss definitions and training details are provided in Appendix B.1.

4.2 MASKED DISCRETE DIFFUSION FOR JOINT GENERATION

Architecture. Each residue occupies the same position across the tracks defined in Section 3. Their embeddings are summed into a shared hidden representation and processed by a 16-block DDiT denoiser with hidden dimension 1,024, RoPE, and adaptive timestep conditioning. Sequence inputs are represented by final hidden states from a frozen 650M tfold encoder (Wu et al., 2024; 2022). Structure inputs use the frozen DisAb-VQ codebook followed by a learned projection; the optional monomer track uses the same vocabulary but a separate input role. Partition, chain, hotspot, and timestep embeddings provide clean conditioning information. Untied two-layer SwiGLU heads predict amino-acid and structural-token logits at generated positions.

Diffusion training. For every generated track r, we use an absorbing masked-diffusion process,

$$q(x_t^r \mid x_0^r) = \mathrm{Cat}\big(\bar{\alpha}_t \mathbf{e}_{x_0^r} + (1-\bar{\alpha}_t)\mathbf{e}_{\text{MASK}}\big), \tag{5}$$

where $1-\bar{\alpha}_t$ is the masking probability and $1-\bar{\alpha}_t = t$ under our log-linear schedule. Track-position pairs designated as conditions remain clean throughout the diffusion process.

Under the SUBS parameterization (Sahoo et al., 2024), revealed tokens are copied exactly, whereas masked positions $\mathcal{M}_r(t) = \{i : x_{t,i}^r = \text{MASK}\}$ are predicted as

$$p_\theta\Big(x^r_{0,\mathcal{M}_r(t)} \mid \boldsymbol{x}_t, c\Big) = \prod_{i\in\mathcal{M}_r(t)} \mathrm{Cat}\big(x^r_{0,i}; \operatorname{softmax} \ell^r_{\theta,i}(\boldsymbol{x}_t, c, t)\big), \tag{6}$$

Writing $\mathcal{R}_{gen}$ for the generated tracks and $w(t)$ for the schedule-induced SUBS weighting, the masked-token objective is

$$\mathcal{L}_{\text{tok}} = \mathbb{E}_{\boldsymbol{x}_0,t,\boldsymbol{x}_t}\left[w(t)\sum_{r\in\mathcal{R}_{\text{gen}}}\lambda_r\frac{1}{|\mathcal{M}_r(t)|}\sum_{i\in\mathcal{M}_r(t)}-\log p_\theta(x^r_{0,i}\mid \boldsymbol{x}_t,c,t)\right], \tag{7}$$

with $\lambda_a = 0.5$ and $\lambda_z = 1.5$ for amino-acid and structure tracks. The two production objectives are

$$\mathcal{L}_{\text{bound}} = \mathcal{L}_{\text{tok}} + 0.05\mathcal{L}_{\text{dist}} + 0.02\mathcal{L}_{\text{adj}}, \tag{8}$$

$$\mathcal{L}_{\text{unbound}} = \mathcal{L}_{\text{tok}}, \tag{9}$$

where $\mathcal{L}_{\text{dist}}$ is a sampled Cα distogram cross-entropy and $\mathcal{L}_{\text{adj}}$ regularizes adjacent-residue Cα distances. Full optimization schedules and training mixtures are provided in Appendix B.3.

4.3 Boltz-2-Guided Inference

Reference-free self-consistency. For a completed candidate $y = (\boldsymbol{a}, \boldsymbol{z})$, let $X_B = F_B(\boldsymbol{a})$ be the structure independently predicted by Boltz-2 (Passaro et al., 2025) from the generated sequence, and let $X_V = D_\psi(\boldsymbol{z}, \boldsymbol{a})$ be the structure decoded from the generated structural tokens. After one joint Kabsch fit over scope $\mathcal{J}$, we define

$$\text{SC}_{BB}(y) = \text{RMSD}^{\mathcal{J}}_{N,CA,C}(X_B, X_V). \tag{10}$$

We use joint VH+VL for bound design and the full antigen+VH+VL complex for unbound design. Lower SC_{BB} indicates stronger sequence–structure agreement; it is reference-free and does not measure binding affinity.

SMC-guided search. We use SC_{BB} as a black-box potential to bias the pretrained DisAb reverse process toward

$$\pi_\beta(y \mid c) \propto p_\theta(y \mid c)\exp\{-\beta\,\text{SC}_{BB}(y)\}, \tag{11}$$

with fixed inverse temperature β. At each diffusion step, every particle first takes one stochastic reverse transition; the resulting partial state is then greedily completed only for Boltz-2 scoring, reweighted by the incremental self-consistency potential, and systematically resampled when $\text{ESS}/N < 0.70$. The greedy completion itself is not propagated.

Repeating this *diffuse–complete–score–reweight/resample* cycle allocates subsequent denoising computation toward more self-consistent trajectories without retraining DisAb. For design search, we return the lowest-SC_{BB} completed candidate encountered across all particles and diffusion levels. Because this *best-ever-seen* rule uses approximate look-ahead completions, we interpret the procedure as SMC-guided search rather than an exact sampler from Equation 11. Full weighting equations, resampling details, and analyses are provided in Appendix C.

5 Experiments

We evaluate DisAb along three main dimensions. First, we assess whether DisAb-VQ faithfully reconstructs both single-chain and multi-chain atom14 structures. Second, we evaluate antibody sequence–structure design under bound-antigen and unbound-antigen conditioning, with neither setting receiving native VH/VL framework coordinates. Finally, we analyze the inference-time guidance signal by relating reference-free self-consistency to native structural error, and characterize the diversity and training-set novelty of generated H3 sequences.

5.1 Evaluation protocol

Tokenizer comparisons use paired evaluation sets shared across methods. The SAbDab antigen benchmark is additionally structure-quality filtered to 99 common antigen chains to avoid artifacts from incomplete or discontinuous experimental structures. For antibody design, the bound benchmark contains 232 temporally held-out, antigen-homology-filtered SAbDab complexes, while the unbound benchmark contains 36 cases with paired bound and isolated-antigen structures. The bound complexes corresponding to all 36 unbound benchmark cases are explicitly excluded from

Table 1: **VQ-VAE round-trip reconstruction.** RMSD is reported in Å. The SAbDab antigen comparison uses the same structure-quality-filtered subset for all tokenizers. Whole-complex metrics use one global alignment over all chains and are available only for DisAb-VQ among the evaluated tokenizers.

Dataset	Scope	Method	N	BB RMSD ↓	TM ↑	GDT-TS ↑	All-atom RMSD ↓	Usage
CAMEO2024	monomer	DPLM-2 VQ-VAE	400	1.5413	0.9276	0.8426	–	87.99%
		GCP-VQ-VAE-large	400	**0.4128**	**0.9917**	**0.9942**	–	99.00%
		ESM3 VQ-VAE	400	0.5288	0.9853	0.9794	–	92.94%
		DisAb-VQ (Ours)	400	0.5561	0.9857	0.9841	1.1519	99.22%
SAbDab-test	VH / VL	DPLM-2 VQ-VAE	232	1.1523 / 0.8476	0.9335 / 0.9547	0.8911 / 0.9417	– / –	44.14% / 37.46%
		GCP-VQ-VAE-large	232	0.4104 / 0.3820	0.9896 / 0.9897	**0.9965** / 0.9978	– / –	64.31% / 57.18%
		ESM3 VQ-VAE	232	0.4512 / **0.3500**	0.9879 / **0.9915**	0.9934 / **0.9988**	– / –	59.99% / 54.96%
		DisAb-VQ (Ours)	232	**0.4073** / 0.3704	**0.9901** / 0.9907	0.9958 / 0.9980	0.9909 / 0.9246	40.16% / 33.57%
SAbDab-test	antigen	DPLM-2 VQ-VAE	99	11.1806	0.4812	0.3329	–	41.92%
		GCP-VQ-VAE-large	99	**0.6575**	**0.9516**	**0.9625**	–	59.64%
		ESM3 VQ-VAE	99	1.6616	0.8150	0.7706	–	64.72%
		DisAb-VQ (Ours)	99	1.1538	0.8896	0.8740	1.5365	48.79%
Whole-complex protocol (one global fit)				Multi-chain BB RMSD ↓	Multi-chain TM ↑	Multi-chain GDT-TS ↑	Multi-chain all-atom RMSD ↓	Usage
CAMEO2024	full complex	DisAb-VQ (Ours)	183	0.6825	0.9905	0.9689	1.2354	99.69%
SAbDab-test	full complex	DisAb-VQ (Ours)	180	1.0860	0.9764	0.9116	1.5170	95.98%

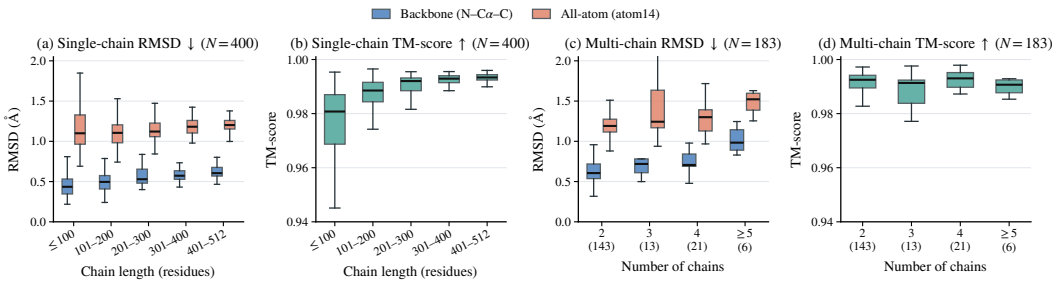


Figure 2: **Scaling behavior of DisAb-VQ round-trip reconstruction. (a–b)** Single-chain backbone/all-atom RMSD and TM-score stratified by sequence length ($N = 400$). **(c–d)** Whole-complex reconstruction stratified by chain count ($N = 183$), using one global alignment for each complex.

every stage of DisAb generator training. No downstream antibody test complex is present in the DisAb-VQ training set. For bound design, Aligned DockQ removes only the arbitrary global VH–VL rigid-body frame; for unbound design, the generated antigen–antibody pose is evaluated directly. Complete filtering criteria, alignment conventions, and metric definitions are provided in Appendix B.

5.2 Tokenizer reconstruction

Table 1 shows that DisAb-VQ performs particularly well on antibody structures, while remaining competitive with existing single-chain backbone tokenizers on general proteins. More importantly, the same discrete representation supports accurate whole-complex reconstruction under a single global fit, preserving both relative chain placement and atom14 all-atom geometry.

To examine robustness to structural complexity, we further stratify reconstruction by sequence length and chain count in Figure 2. Single-chain RMSD increases gradually with sequence length while TM-score remains high, and the same trend holds for whole-complex reconstruction as the number of chains increases. Together, these results indicate that the discrete bottleneck preserves fine-grained atomic geometry and relative chain placement across proteins of varying size and complexity.

5.3 Bound-antigen antibody design

As shown in Table 2, DisAb-bound is highly competitive among methods that do not use native VH/VL framework coordinates, and SMC guidance further strengthens its performance across every

Table 2: **Bound-antigen antibody design on 232 held-out SAbDab complexes.** "FR xyz" indicates whether native VH/VL framework coordinates are provided. Aligned DockQ removes the global VH–VL rigid-body transform before interface evaluation; all remaining alignment and metric details are given in Appendix B.5. Bold denotes the best result among methods that do not use native VH/VL framework coordinates.

Method	FR xyz	Aligned DockQ ↑	Aligned DockQ$_{BB}$ ↑	TM-score ↑	Cα ↓	iRMS ↓	H3(C-CA-N) ↓	All-atom ↓	SC$_{BB}$ ↓	CDR-AAR ↑
DiffAb	Yes	–	0.686	0.978	1.151	1.838	3.669	–	1.447	0.501
IgGM	No	–	0.615	0.952	1.397	2.303	3.587	–	2.478	0.640
dyMEAN	No	0.691	0.688	**0.961**	**1.324**	**1.674**	**3.147**	**1.642**	1.228	**0.738**
AbFlow	No	0.680	0.665	0.949	1.550	2.049	4.156	1.820	1.331	0.630
DisAb-bound (Ours)	No	0.692	0.671	0.956	1.470	1.902	3.656	1.989	1.382	0.544
DisAb-bound + SMC (Ours, N=8, K=16)	No	**0.711**	**0.703**	**0.961**	1.360	1.731	3.314	1.885	**0.916**	0.585

Table 3: **Unbound-antigen design on the 36-case paired docking benchmark.** All unguided methods are evaluated from a single generation, without best-of-N selection, reranking, or post-generation refinement; Germinal uses only its hallucination stage. RFantibody receives native VH/VL framework coordinates and is shown as an auxiliary, more strongly conditioned baseline.

Method	FR xyz	Guided Search	DockQ ↑	DockQ$_{BB}$ ↑	VHVL RMSD ↓	iRMS ↓	Complex RMSD ↓	Complex all-atom ↓	SC$_{complex,BB}$ ↓	CDR-AAR ↑
DisAb-unbound (Ours)	No	No	0.129	0.107	2.976	9.351	10.712	10.990	14.362	0.517
Germinal (hallucination only)	No	No	0.055	0.040	2.552	11.201	12.838	13.325	11.125	0.156
RFantibody	Yes	No	0.132	0.091	1.083	8.866	10.664	10.910	13.831	0.409
DisAb-unbound + SMC (Ours, N=8, K=16)	No	Yes	0.149	0.121	2.269	8.969	10.374	10.656	12.463	0.558

reported metric. DisAb-bound + SMC achieves the best Aligned DockQ (0.711), best Aligned DockQ$_{BB}$ (0.703), and lowest backbone self-consistency error (0.916 Å), while matching the best TM-score (0.961). SMC (N=8, K=16) also improves Cα, interface, H3, and all-atom geometry as well as CDR sequence recovery relative to the unguided model. These results show that DisAb provides strong joint sequence–structure design performance under the more realistic setting without native framework coordinates, while SMC further improves both structural quality and sequence–structure consistency without retraining the generator.

5.4 Unbound-antigen complex design

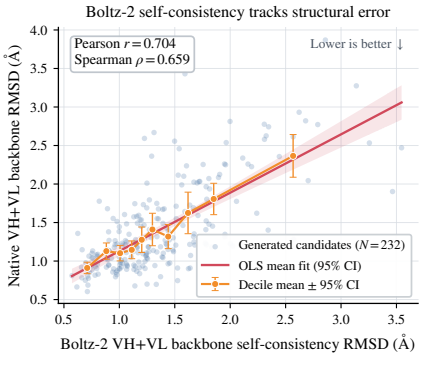


Figure 3: **Self-consistency tracks structural error.** On 232 held-out bound designs, VH+VL backbone self-consistency correlates with native VH+VL backbone RMSD. Orange markers denote decile means with 95% confidence intervals; the red line is the OLS fit with its 95% confidence band.

As shown in Table 3, the unbound setting is substantially more challenging because the model must generate both the antibody structure and its relative pose to an isolated antigen. Among methods that do not use native VH/VL framework coordinates, DisAb-unbound achieves the strongest overall docking performance, and SMC-guided search further improves DockQ, complex geometry, and CDR sequence recovery. Although RFantibody attains lower internal VH+VL RMSD, it is conditioned on native framework coordinates and is therefore not directly comparable. These results show that DisAb-unbound can generate competitive bound antigen–antibody complexes from an isolated antigen structure under a substantially weaker conditioning setting.

5.5 Self-consistency as a guidance signal

Figure 3 tests whether the native-reference-free score in Equation 10 contains information about true structural quality. Across 232 held-out bound designs, VH+VL backbone self-consistency is positively associated with native VH+VL backbone RMSD, with Pearson $r = 0.704$ and Spearman $\rho = 0.659$. The agreement between linear and rank correlations, together with the monotonic decile trend, indicates that lower self-consistency generally corresponds to lower structural error, although substantial

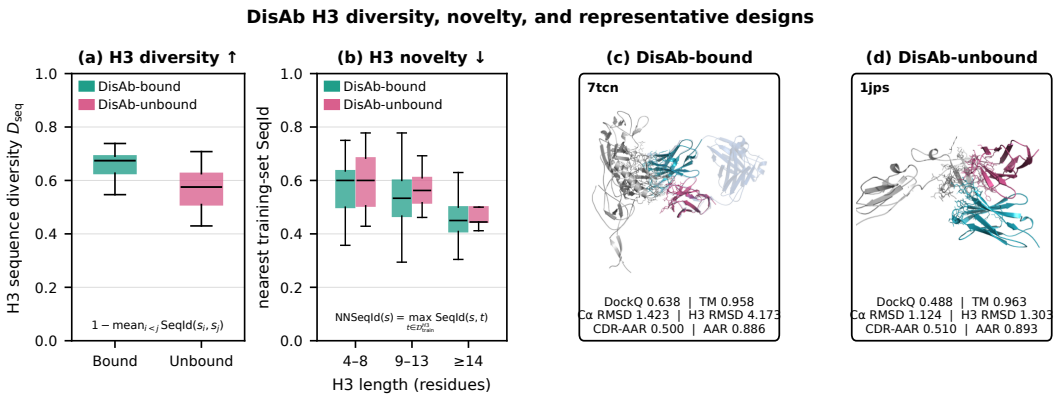


Figure 4: **H3 diversity, novelty, and representative DisAb designs. (a)** Within-target H3 sequence diversity among SMC candidates. **(b)** H3 novelty measured against the nearest H3 sequence in the SAbDab training set. **(c–d)** Representative DisAb-bound and DisAb-unbound generations.

residual variance remains. This supports its use as an inference-time potential rather than only as an evaluation statistic.

5.6 Diversity and Novelty

Figure 4 further shows that DisAb maintains substantial sequence diversity while generating H3 loops that remain distinct from the training set. Across targets, the sampled candidate pools contain diverse H3 sequences rather than collapsing to a single mode, while nearest-training sequence identity indicates that generated loops are not simply reproductions of memorized training examples. This also supports the use of SMC as a practical inference-time search procedure over a diverse generative distribution rather than merely as a deterministic post-selection mechanism. An equal-Boltz-2-budget comparison against independent best-of-16 sampling, together with additional SMC scaling analyses, is reported in Appendix C.

6 Discussion and Limitations

Our results support discrete joint modeling of antibody sequence, all-atom geometry, and multi-chain pose, but remain entirely computational; none of the reported metrics establishes experimental binding or developability. The unbound setting is further limited by scarce paired unbound/bound structural data, with evaluation covering only 36 experimentally determined docking cases. The current model assumes paired VH/VL antibodies, standard atom14 protein chemistry, and a maximum 1,024-residue context, excluding non-protein antigens, PTMs, ligands, glycans, and full Fc assemblies. In practical antibody-design settings, hotspot residues can be specified from prior knowledge or design requirements, while alternative VH/VL framework templates and CDR lengths can be explored as different conditioning choices for DisAb. Experimental validation will be necessary to assess binding specificity, developability, immunogenicity, and biosafety.

7 Conclusion

We introduced DisAb, a unified discrete framework for antibody sequence–structure co-design. DisAb-VQ provides a residue-level representation of multi-chain all-atom protein complexes, enabling sequence and structure to be modeled together through masked discrete diffusion. The same framework supports both bound-antigen and antigen-monomer conditioning without native antibody framework coordinates, while Boltz-2 self-consistency enables reference-free SMC-guided design search. Together, these components provide a general route toward discrete joint modeling of antibody sequence, structure, and multi-chain geometry. Future work will focus on improving unbound complex generation, reducing inference cost, and experimentally validating generated antibodies.

8 AI USE STATEMENT

In this work, we used generative AI tools for assisting with translation. We have not used generative AI tools for generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, or interpreting results. Additionally, we used generative AI tools for sourcing/searching for information. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

9 REPRODUCIBILITY STATEMENT

The main implementation of DisAb and the evaluation pipelines are available in an anonymous repository at `https://anonymous.4open.science/r/disab_anonymous-5F02/README.md`. The appendix provides detailed descriptions of the training procedures, data construction and filtering, evaluation protocols, and SMC inference settings used in our experiments. All underlying datasets used in this work are publicly available and are identified and cited in the appendix. To preserve double-blind anonymity during review, the final cleaned code release and model checkpoints will be publicly released upon publication.

REFERENCES

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 17981–17993. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/958c530554f78bcd8e97125b70e6973d-Paper.pdf.

Nathaniel R. Bennett, Joseph L. Watson, Robert J. Ragotte, Andrew J. Borst, DéJenaé L. See, Connor Weidle, Riti Biswas, Yutong Yu, Ellen L. Shrock, Russell Ault, Philip J. Y. Leung, Buwei Huang, Inna Goreshnik, John Tam, Kenneth D. Carr, Benedikt Singer, Cameron Criswell, Basile I. M. Wicky, Dionne Vafeados, Mariana Garcia Sanchez, Ho Min Kim, Susana Vázquez Torres, Sidney Chan, Shirley M. Sun, Timothy T. Spear, Yi Sun, Keelan O'Reilly, John M. Maris, Nikolaos G. Sgourakis, Roman A. Melnyk, Chang C. Liu, and David Baker. Atomically accurate de novo design of antibodies with RFdiffusion. *Nature*, 649(8095):183–193, January 2026. ISSN 1476-4687. doi: 10.1038/s41586-025-09721-5. URL https://www.nature.com/articles/s41586-025-09721-5.

Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 01 2000. ISSN 0305-1048. doi: 10.1093/nar/28.1.235. URL https://doi.org/10.1093/nar/28.1.235.

Paul J. Carter and Greg A. Lazar. Next generation antibody drugs: Pursuit of the 'high-hanging fruit'. *Nature Reviews Drug Discovery*, 17(3):197–223, 2018. doi: 10.1038/nrd.2017.227. URL https://doi.org/10.1038/nrd.2017.227.

The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, 01 2025. ISSN 1362-4962. doi: 10.1093/nar/gkae1010. URL https://doi.org/10.1093/nar/gkae1010.

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. doi: 10.1089/cmb.2024.0768. URL https://doi.org/10.1089/cmb.2024.0768.

J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022. doi: 10.1126/science.add2187. URL https://www.science.org/doi/abs/10.1126/science.add2187.

James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M. Deane. Sabdab: the structural antibody database. *Nucleic Acids Research*, 42(D1):D1140–D1146, 01 2014. ISSN 0305-1048. doi: 10.1093/nar/gkt1043. URL https://doi.org/10.1093/nar/gkt1043.

Zhangyang Gao, Cheng Tan, and Stan Z. Li. Foldtoken2: Learning compact, invariant and generative protein structure language. *bioRxiv*, 2024a. URL https://api.semanticscholar.org/CorpusID:270522256.

Zhangyang Gao, Cheng Tan, Jue Wang, Yufei Huang, Lirong Wu, and Stan Z. Li. Foldtoken: Learning protein language via vector quantization and beyond. In *AAAI Conference on Artificial Intelligence*, 2024b. URL https://api.semanticscholar.org/CorpusID:268510133.

Johnathan D. Guest, Thom Vreven, Jing Zhou, Iain Moal, Jeliazko R. Jeliazkov, Jeffrey J. Gray, Zhiping Weng, and Brian G. Pierce. An expanded benchmark for antibody-antigen docking and affinity prediction reveals insights into antibody recognition determinants. *Structure*, 29(6):606–621.e5, 2021. ISSN 0969-2126. doi: https://doi.org/10.1016/j.str.2021.01.005. URL https://www.sciencedirect.com/science/article/pii/S0969212621000058.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf A. Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025. doi: 10.1126/science.ads0018. URL https://www.science.org/doi/abs/10.1126/science.ads0018.

Filip Ekstrom Kelvinius and Fredrik Lindsten. Discriminator guidance for autoregressive diffusion models. In *International Conference on Artificial Intelligence and Statistics*, 2023. URL https://api.semanticscholar.org/CorpusID:264439058.

Xiangzhe Kong, Wenbing Huang, and Yang Liu. End-to-end full-atom antibody design. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17409–17429. PMLR, 23–29 Jul 2023a. URL https://proceedings.mlr.press/v202/kong23c.html.

Xiangzhe Kong, Wenbing Huang, and Yang Liu. Conditional antibody design as 3d equivariant graph translation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023b. URL https://openreview.net/forum?id=LFHFQbjxIiP.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

Zhihai Liu, Yan Li, Li Han, Jie Li, Jie Liu, Zhixiong Zhao, Wei Nie, Yuchen Liu, and Renxiao Wang. Pdb-wide collection of binding data: current status of the pdbbind database. *Bioinformatics*, 31 (3):405–412, 02 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu626. URL https://doi.org/10.1093/bioinformatics/btu626.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32819–32848. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/lou24a.html.

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 S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 9754–9767. Curran Associates, Inc., 2022. doi: 10.52202/068431-0709. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/3fa7d76a0dc1179f1e98d1bc62403756-Paper-Conference.pdf.

Jian Ma, Fandi Wu, Lin Yao, Jing Gao, Rubo Wang, Qifeng Li, Nianzu Yang, Songlin Jiang, Dawei Huang, Xiaoyong Pan, Yiheng Zhu, Tingjun Hou, Jianhua Yao, and Junchi Yan. Iggm2: An all-atom foundation model for adaptive immune receptor design. *bioRxiv*, 2026. doi: 10.64898/2026.07.09.737510. URL https://www.biorxiv.org/content/early/2026/07/09/2026.07.09.737510.

Karolis Martinkus, Jan Ludwiczak, WEI-CHING LIANG, Julien Lafrance-Vanasse, Isidro Hotzel, Arvind Rajpal, Yan Wu, Kyunghyun Cho, Richard Bonneau, Vladimir Gligorijevic, and Andreas Loukas. Abdiffuser: full-atom generation of in-vitro functioning antibodies. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 40729–40759. Curran Associates, Inc., 2023. doi: 10.52202/075280-1774. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/801ec05b0aae9fcd2ef35c168bd538e0-Paper-Conference.pdf.

Luis S. Mille-Fragoso, Claudia L. Driscoll, John N. Wang, Haoyu Dai, Talal Widatalla, Jim L. Zhang, Xiaowei Zhang, Bing Rao, Liang Feng, Brian L. Hie, and Xiaojing J. Gao. Efficient generation of epitope-targeted antibodies with Germinal. *Nature Biotechnology*, June 2026. ISSN 1546-1696. doi: 10.1038/s41587-026-03187-0. URL https://doi.org/10.1038/s41587-026-03187-0.

Zijing Ou, Chinmay Pani, and Yingzhen Li. Inference-time scaling of discrete diffusion models via importance weighting and optimal proposal design. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 36740–36775, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/3df1eca840e82b11bbc33f68c773c38e-Paper-Conference.pdf.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, David Kwabi-Addo, Dominique Beaini, Tommi Jaakkola, and Regina Barzilay. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*, 2025. doi: 10.1101/2025.06.14.659707. URL https://www.biorxiv.org/content/early/2025/06/18/2025.06.14.659707.

Mahdi Pourmirzaei, Alex Morehead, Farzaneh Esmaili, Jarett Ren, Mohammadreza Pourmirzaei, and Dong Xu. Gcp-vqvae: A geometry-complete language for protein 3d structure. *bioRxiv*, 2025. doi: 10.1101/2025.10.01.679833. URL https://www.biorxiv.org/content/10.1101/2025.10.01.679833v3.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 130136–130184. Curran Associates, Inc., 2024. doi: 10.52202/079017-4135. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/eb0b13cc515724ab8015bc978fdde0ad-Paper-Conference.pdf.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. Simplified and generalized masked diffusion for discrete data. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 103131–103167. Curran Associates, Inc., 2024. doi: 10.52202/079017-3277. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/bad233b9849f019aead5e5cc60cef70f-Paper-Conference.pdf.

Marta Skreta, Tara Akhound-Sadegh, Viktor Ohanesian, Roberto Bondesan, Alán Aspuru-Guzik, Arnaud Doucet, Rob Brekelmans, Alexander Tong, and Kirill Neklyudov. Feynman-kac correctors in diffusion: Annealing, guidance, and product of experts. *arXiv preprint arXiv:2503.02819*, 2025.

Pietro Sormanni, Francesco A. Aprile, and Michele Vendruscolo. Third generation antibody discovery methods: In silico rational design. *Chemical Society Reviews*, 47(24):9137–9157, 2018. doi: 10.1039/C8CS00523K. URL https://doi.org/10.1039/C8CS00523K.

Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, Augustin Žídek, Tim Green, Kathryn Tunyasuvunakool, Stig Petersen, John Jumper, Ellen Clancy, Richard Green, Ankur Vora, Mira Lutfi, Michael Figurnov, Andrew Cowie, Nicole Hobbs, Pushmeet Kohli, Gerard Kleywegt, Ewan Birney, Demis Hassabis, and Sameer Velankar. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, 01 2022. ISSN 0305-1048. doi: 10.1093/nar/gkab1061. URL https://doi.org/10.1093/nar/gkab1061.

Mihaly Varadi, Damian Bertoni, Paulyna Magana, Urmila Paramval, Ivanna Pidruchna, Malarvizhi Radhakrishnan, Maxim Tsenkov, Sreenath Nair, Milot Mirdita, Jingi Yeo, Oleg Kovalevskiy, Kathryn Tunyasuvunakool, Agata Laydon, Augustin Žídek, Hamish Tomlinson, Dhavanthi Hariharan, Josh Abrahamson, Tim Green, John Jumper, Ewan Birney, Martin Steinegger, Demis

Hassabis, and Sameer Velankar. Alphafold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Research*, 52(D1):D368–D375, 01 2024. ISSN 0305-1048. doi: 10.1093/nar/gkad1011. URL https://doi.org/10.1093/nar/gkad1011.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a.

Rubo Wang, Fandi Wu, Xingyu Gao, Jiaxiang Wu, Peilin Zhao, and Jianhua Yao. Iggm: A generative model for functional antibody and nanobody design. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 100801–100827, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/fa5b423e24b442180bcd4e13ae75a27f-Paper-Conference.pdf.

Wenda Wang, Yang Zhang, Zhewei Wei, and Wenbing Huang. Abflow: End-to-end paratope-centric antibody design by interaction enhanced flow matching. In Srinivasan Parthasarathy, David F. Gleich, Xiangliang Zhang, Wee Hyong Tok, Faisal Farooq, Qi He, Ambuj K. Singh, Haixun Wang, and Yan Liu (eds.), *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1, KDD 2026, Jeju Island, Korea, August 9-13, 2026*, pp. 1471–1482. ACM, 2026. doi: 10.1145/3770854.3780296. URL https://dl.acm.org/doi/10.1145/3770854.3780296.

Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Diffusion language models are versatile protein learners. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52309–52333. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/wang24ct.html.

Xinyou Wang, Zaixiang Zheng, Fei YE, Dongyu Xue, Shujian Huang, and Quanquan Gu. Dplm-2: A multimodal diffusion protein language model. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 35463–35490, 2025c. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/57c30b677add9aa78e1745f0643104d0-Paper-Conference.pdf.

Fandi Wu, Yu Zhao, Jiaxiang Wu, Biaobin Jiang, Bing He, Longkai Huang, Chenchen Qin, Fan Yang, Ningqiao Huang, Yang Xiao, et al. Fast and accurate modeling and design of antibody-antigen complex using tfold. *bioRxiv*, pp. 2024–02, 2024.

Jiaxiang Wu, Fandi Wu, Biaobin Jiang, Wei Liu, and Peilin Zhao. tfold-ab: fast and accurate antibody structure prediction without sequence homologs. *bioRxiv*, pp. 2022–11, 2022.

Luhuan Wu, Brian L Trippe, Christian Naesseth, David Blei, and John P Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. *Advances in Neural Information Processing Systems*, 2023.

Appendix Overview

The appendix provides additional qualitative results, implementation and evaluation details, SMC analyses, and ablation studies. Appendix A presents representative DisAb-VQ reconstructions and DisAb antibody designs. Appendix B details model training, dataset construction and filtering, evaluation cohorts, and metric definitions for both DisAb-VQ and DisAb. Appendix C describes the SMC implementation and provides equal-score-budget comparisons, alternative guidance analyses, and compute scaling experiments. Finally, Appendix D studies the effects of generic pretraining and the antigen-monomer structural condition.

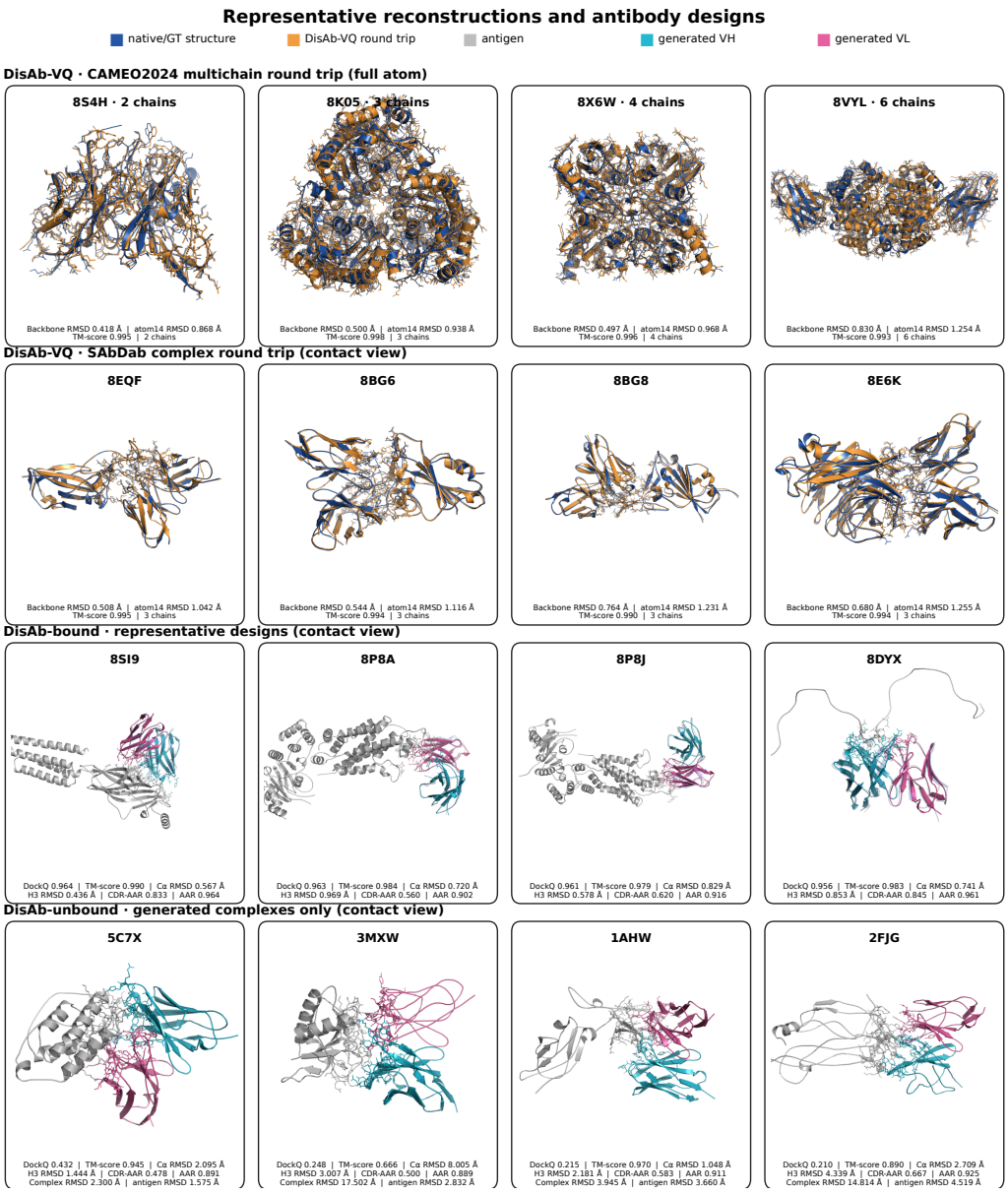


Figure 5: **Representative reconstructions and antibody designs.** Rows 1–2 show DisAb-VQ round-trip reconstruction on CAMEO2024 multi-chain complexes and SAbDab antibody–antigen complexes, respectively. Rows 3–4 show representative DisAb-bound and DisAb-unbound designs. Native structures are blue, DisAb-VQ reconstructions orange, antigen gray, generated VH cyan, and generated VL magenta.

A Qualitative Visualizations

Figure 5 shows representative DisAb-VQ reconstructions and DisAb designs. The first two rows illustrate representative multi-chain all-atom round-trip reconstructions on CAMEO2024 and SAbDab complexes. The last two rows show representative bound and unbound antibody designs. While DisAb-bound illustrates the geometry produced in the bound setting, the harder DisAb-unbound setting exhibits larger variation in the relative antibody–antigen pose, consistent with the quantitative results in Tables 2 and 3.

B Additional Training and Experiment Details

B.1 DisAb-VQ Training Details

Architecture and optimization. DisAb-VQ has 1,673,155,166 trainable parameters, an 8,192-entry cosine-VQ codebook, and a 256-dimensional code space. The Transformer hidden dimension is 1,184 and the model contains 42 encoder/decoder Transformer blocks in total, in addition to a five-layer invariant graph correction frontend. The maximum input size is 1,024 residues and eight chains.

Training uses BF16 mixed precision, while VQ distances and EMA statistics, Kabsch alignment, geometric calculations, gradient accumulation, and optimizer states are maintained in FP32. We use AdamW with weight decay 10^{-2} for decayed parameter groups and task-split gradient clipping at norm 1.0. Training uses one structure per GPU and gradient accumulation to obtain an effective global batch of 24–32, depending on the distributed configuration.

Training data. The DisAb-VQ training corpus combines predicted and experimental protein structures from AFDB (Varadi et al., 2024; 2022), the PDB (Berman et al., 2000), SAbDab (Dunbar et al., 2014), PDBbind (Liu et al., 2015), and UniProtKB/Swiss-Prot–AFDB (Consortium, 2025; Varadi et al., 2024; 2022). After filtering, the corpus includes 1,315,913 AFDB structures, 88,417 experimental training structures, 2,209 temporally selected SAbDab antibody–antigen assemblies, 511 PDBbind protein–protein complexes, together with curated SwissProt-AFDB, experimental-monomer, and non-antibody experimental-complex subsets.

Each training epoch samples 240,000 structures from this mixture: 88,609 AFDB structures, 72,000 experimental structures, 35,443 SwissProt-AFDB structures, 28,800 additional experimental monomers, 11,640 additional experimental complexes, 511 PDBbind complexes, 2,209 SAbDab assemblies, and 788 supplemental SAbDab assemblies. Small curated datasets are sampled at most once per epoch, with the remaining sampling mass assigned to the larger predicted-structure sources. For multi-chain examples, complex-to-single-chain dropout is applied with probability 0.5 to retain strong monomer reconstruction while preserving multi-chain supervision.

Training objective. The production DisAb-VQ is trained with a weighted combination of backbone, pairwise-distance, interface, local-geometry, torsional, latent-regularization, and vector-quantization objectives. Specifically, the backbone loss has weight 2.0, the all-pairs Cα distance and peptide-bond losses have weight 0.5, the interface, local-vector, side-chain local-geometry, and VQ objectives have unit weight, the χ- and ψ-torsion losses have weights 4.0 and 2.0, respectively, and the latent-spread regularizer has weight 0.25. The backbone term is evaluated after one whole-complex Kabsch alignment, while the interface term supervises native cross-chain residue pairs within 14 Å. Side-chain geometry is supervised in residue-local frames with symmetry-aware atom14 losses, together with explicit torsion supervision. Direct global all-atom and side-chain coordinate auxiliaries have zero production weight and are therefore excluded from the final objective. The codebook is initialized by K-means, updated with synchronized EMA statistics rather than direct gradient descent, and maintained with bounded dead-code recycling.

Structure filtering. All sources are converted to a common all-atom representation. We require valid backbone N/Cα/C atoms, supported residue types, at least 30 residues per retained chain, at most 1,024 residues and eight chains per sample, and sufficient heavy-atom completeness. Predicted AFDB and SwissProt-AFDB structures require mean pLDDT $\geq$ 70. Newly added experimental structures require a numerical resolution no worse than 3.0 Å. Multi-chain assemblies are additionally checked for abnormal adjacent-Cα gaps, inter-chain clashes, and contact connectivity. New experimental sources are deduplicated against existing training and validation pools using PDB identifiers and sequence/sample hashes before shard construction.

Downstream leakage audit. We explicitly checked the DisAb-VQ training corpus against the final downstream antibody evaluation sets. There is no exact-complex overlap with either the 232-complex leakage-filtered SAbDab bound test set or the 36-case bound/unbound docking benchmark. Thus, neither downstream benchmark contains a complex used to train the frozen structural tokenizer.

B.2 DISAB-VQ EVALUATION PROTOCOL

Single-chain cohorts. For CAMEO2024, all tokenizers are evaluated on the same 400 common-unseen target chains with complete residue correspondence and no unresolved residues. For SAbDab VH and VL, we use all 232 entries from the final temporally held-out, antigen-homology-filtered cohort.

The SAbDab antigen benchmark is further filtered to remove artifacts from incomplete or discontinuous experimental chains. The same 99 antigen chains are used for every tokenizer. We require $25 \leq L \leq 1024$, complete observed N/Cα/C coordinates, adjacent Cα distances no larger than 5 Å, and no unresolved residue-number intervals. Valid PDB insertion codes are retained. The minimum length of 25 residues is applied uniformly because it is required by the public GCP-VQ-VAE preprocessing pipeline.

Whole-complex cohorts. For CAMEO2024, target-chain identifiers are mapped to RCSB biological assembly 1 and filtered using the production complex-quality criteria, including valid protein chains, total context length within the model limit, sufficient heavy-atom completeness, absence of severe backbone discontinuities or inter-chain clashes, and a connected inter-chain contact graph. This yields 183 complexes.

The SAbDab whole-complex benchmark contains 180 antigen+VH+VL assemblies satisfying the corresponding structure-quality and context-length requirements. Whole-complex reconstruction is reported only for DisAb-VQ, because the other evaluated tokenizers do not expose an equivalent joint multi-chain all-atom decoding interface.

Tokenizer leakage audit. The frozen DisAb-VQ training corpus was checked against the downstream antibody evaluation sets. No exact complex from either the 232-case bound benchmark or the 36-case bound/unbound docking benchmark occurs in the tokenizer training corpus. This exact-complex audit is separate from the antigen-sequence homology filter used for DisAb-bound.

B.3 DISAB TRAINING DETAILS

Model and optimization. Both variants use a 16-block denoising Transformer with hidden dimension 1,024, SDPA attention, and untied two-layer output heads. The generator contains approximately 325.5M trainable parameters; the frozen 650M tfold sequence encoder is excluded from this count. Structural input embeddings are initialized from the 8,192-entry DisAb-VQ codebook, which remains frozen, while their projections into the shared hidden space remain trainable. The 1.67B-parameter DisAb-VQ itself is frozen throughout DisAb training and is therefore not included in the generator parameter count.

All production runs use BF16 mixed precision and AdamW with weight decay 0.01. The maximum joint sequence length is 1,024 residues. The amino-acid and structure diffusion losses use relative weights 0.5 and 1.5, respectively.

Diffusion objective. Both DisAb-bound and DisAb-unbound are trained with absorbing masked diffusion under the SUBS parameterization. At each training step, a diffusion time is sampled and generated-track tokens are independently replaced by MASK according to the log-linear masking schedule, while all conditioning positions remain clean. Cross-entropy is evaluated only on masked target positions, averaged within each generated track, and weighted by the corresponding diffusion-time factor. The amino-acid and structure-track objectives use relative weights of 0.5 and 1.5, respectively. DisAb-bound additionally uses weak geometric auxiliaries consisting of a 0.05-weighted sampled Cα distogram cross-entropy and a 0.02-weighted adjacent-residue Cα distance loss. These geometric auxiliaries are disabled for DisAb-unbound, whose production objective contains only the weighted masked-token diffusion losses.

DisAb-bound pretraining. The production bound model is initialized from generic joint sequence–structure pretraining and subsequently continued on a leakage-audited mixture of 70% AFDB monomers (Varadi et al., 2024; 2022), 10% historical PDB structures (Berman et al., 2000), and 20% explicit multi-chain structures. The largest source contains approximately 1.316M quality-

filtered AFDB structures. PDBbind and the supplemental SAbDab pool are excluded from this leakage-clean continuation.

Each logical pretraining epoch contains 320,000 sampled structures, and we train for five epochs. The four corruption/conditioning tasks—joint generation, sequence-to-structure, structure-to-sequence, and partial joint completion—are sampled with probabilities 0.35/0.25/0.20/0.20. Training is distributed over eight NVIDIA A100s with one sample per GPU and eight-step gradient accumulation, giving an effective batch size of 64. We use a peak learning rate of 3×10^{-5} with 5% warmup.

DisAb-bound antibody fine-tuning. We fine-tune on SAbDab using the temporal protocol of IgGM: training structures are deposited no later than 2022-12-31, validation spans 2023-01-01–2023-06-30, and the test interval is 2023-07-01–2023-12-30. We require paired VH/VL and antigen data, resolution no worse than 5 Å, total length at most 1,024, and a valid cached DisAb-VQ representation; at most three invalid cached structural tokens are tolerated.

Fine-tuning is configured for up to 30 epochs and selects the checkpoint with the lowest validation RMSD. It uses an effective batch size of 32 (one sample per GPU on eight GPUs with four-step gradient accumulation), a peak learning rate of 10^{-5}, 3% warmup, and gradient clipping at norm 1.0. The tfold encoder and VQ codebook embeddings remain frozen.

DisAb-unbound pretraining. DisAb-unbound is initialized from the bound model and introduces an input-only monomer-structure track. Generic unbound pretraining uses 11,640 quality-controlled experimental multi-chain complexes. For each complex, an independently encoded component is supplied as the structural condition and the model reconstructs the joint complex, yielding 26,344 chain-conditioned training examples per epoch. This stage is trained for five epochs with an effective batch size of 64, peak learning rate 3×10^{-5}, 5% warmup, and no auxiliary geometry losses.

DisAb-unbound antibody adaptation. Because most SAbDab complexes do not provide a paired unbound antigen structure, we construct monomer conditions by independently predicting the antigen with Boltz without MSA search and subsequently encoding the resulting antigen-only structure with DisAb-VQ. After quality filtering, this gives 2,992 training and 314 validation examples for the synthetic-unbound antibody task.

The first antibody adaptation stage is configured for up to 30 epochs with an effective batch size of 32, peak learning rate 10^{-5}, and 3% warmup. The target comprises the complete bound antigen+VH+VL structure together with VH/VL CDR amino acids, whereas antigen sequence, framework sequence, partition identities, hotspots, and the independently encoded monomer structure remain conditions.

Mixed training and recovery. We subsequently introduce explicit antigen-chain partition identities and perform a five-epoch mixed stage containing both generic complexes and antibody examples. Within the generic objective, single-component-to-complex conditioning and all-sequence-to-structure training are sampled with weights 0.60 and 0.40, respectively; antibody examples are interleaved every two generic steps. We found that this mixed stage improved generic complex reconstruction but partially degraded the antibody-specific objective. We therefore apply a final antibody-only recovery stage with peak learning rate 5×10^{-6}, effective batch size 32, and a maximum of 20 epochs.

Table 4: Training configuration of the production DisAb models. "Examples/epoch" denotes sampled presentations rather than unique structures.

Stage	Epochs	Effective batch	Peak LR	Warmup	Data / objective
Bound generic	5	64	3×10^{-5}	5%	70/10/20 AFDB/PDB/multi-chain; 320k/epoch
Bound SAbDab	≤ 30	32	1×10^{-5}	3%	temporal SAbDab; validation selection
Unbound generic	5	64	3×10^{-5}	5%	11,640 complexes; 26,344 conditioned examples/epoch
Unbound SAbDab	≤ 30	32	1×10^{-5}	3%	2,992 train / 314 validation
Unbound mixed	5	64	1×10^{-5}	3%	generic + antibody mixture
Unbound recovery	≤ 20	32	5×10^{-6}	3%	antibody only; validation selection

B.4 Antibody Evaluation Protocol

Bound benchmark. We evaluate DisAb-bound on the temporal SAbDab test interval 2023-07-01–2023-12-30. Entries must contain paired VH/VL and antigen chains, have experimental resolution no worse than 5 Å, fit within the 1,024-residue context, and admit a valid cached DisAb-VQ representation.

To reduce train–test homology leakage, we additionally remove a test complex if any antigen chain has at least 50% sequence identity to a training antigen chain over at least 80% bidirectional sequence coverage, as determined by MMseqs2. The resulting benchmark contains 232 complexes. This homology guarantee is defined with respect to the DisAb training set; baseline methods may have been trained on different corpora. We do not include IgGM2 (Ma et al., 2026) in the comparison because no public implementation or model checkpoint was available at the time of evaluation; we therefore evaluate the publicly available IgGM baseline instead.

DisAb-bound receives antigen sequence and its antibody-bound structure, together with the VH/VL framework sequence, but never native VH/VL framework coordinates. Because the task assumes an already known bound-antigen context, the absolute global rigid-body frame of the generated VH/VL structure is not treated as a docking variable. Accordingly, antibody-coordinate metrics use one joint VH+VL alignment, and Aligned DockQ applies the same joint VH+VL rigid transform before evaluating the interface against the fixed native antigen. No CDR or individual chain is independently realigned.

Unbound benchmark. We evaluate DisAb-unbound on 36 experimentally paired bound/unbound docking cases (Guest et al., 2021). Each case requires paired VH/VL, a compatible isolated-antigen structure, total context length at most 1,024 residues, and successful antigen sequence mapping between the bound and unbound structures. The preprocessing requires at least 0.85 antigen sequence identity and 0.70 coverage; alignment-based cropping is permitted when necessary.

Unlike the bound task, the unbound model must generate the relative antigen–VH–VL pose. Raw DockQ is therefore computed without first aligning the generated antibody onto the native antibody. Antibody-internal metrics use one joint VH+VL alignment, whereas complex RMSD metrics use one global antigen+VH+VL alignment. Per-chain alignment is never used for the primary complex metrics. RFantibody is reported separately because it receives native VH/VL framework coordinates, while Germinal is a de novo hallucination baseline and does not share exactly the same conditioning contract.

B.5 Evaluation Metrics and Alignment Protocols

This section specifies the computation of the evaluation metrics used in Tables 1–3. Coordinates are compared at fixed residue correspondence whenever possible; variable-length outputs are sequence-aligned before scoring common residues. Unless stated otherwise, RMSDs are reported in Å, with lower values being better, whereas higher values are better for TM-score, GDT-TS, DockQ, and amino-acid recovery.

Backbone and Cα RMSD. Backbone RMSD is computed over matched N/Cα/C coordinates,

$$\text{BB-RMSD} = \sqrt{\frac{1}{3L}\sum_{i=1}^{L}\sum_{a\in\{N,C_\alpha,C\}} \|\widetilde{\mathbf{x}}_{i,a} - \mathbf{y}_{i,a}\|_2^2}. \tag{12}$$

Cα RMSD is defined analogously using only Cα atoms. Single-chain tokenizer entries use one fixed-correspondence Kabsch fit per chain, whereas whole-complex tokenizer metrics use one global fit over all chains. For bound antibody evaluation, all antibody-coordinate metrics use one joint VH+VL fit. For unbound complex metrics, the fit is performed jointly over antigen+VH+VL.

TM-score and GDT-TS. TM-score is evaluated from matched Cα coordinates using the optimizer adopted by the GCP-VQ-VAE evaluation protocol. GDT-TS is computed after the stated Kabsch

alignment as

$$\text{GDT-TS} = \frac{1}{4} \sum_{r \in \{1,2,4,8\}\,\text{Å}} \frac{1}{L} \sum_{i=1}^{L} \mathbf{1}[d_i \le r], \tag{13}$$

where d_i is the aligned Cα distance. Whole-complex TM-score and GDT-TS use the same single global complex alignment rather than independent chain superposition.

All-atom RMSD. For tokenizer reconstruction, all-atom RMSD is evaluated in the atom14 representation after the same backbone alignment used for BB RMSD. Chemically equivalent side-chain atoms are symmetry corrected before the final distance calculation.

For antibody-design Tables 2 and 3, all-atom RMSD is computed over atoms present in both the generated and native structures after the corresponding backbone alignment. The bound metric uses the joint VH+VL transform, whereas the unbound "Complex all-atom" metric uses the single global antigen+VH+VL transform.

Code utilization. Tokenizer utilization is the fraction of codebook entries observed at least once in the corresponding evaluation cohort. Because the evaluated tokenizers use different codebook sizes, we report both the percentage and the number of occupied codes where available.

H3 RMSD. H3(C–Cα–N) is the N/Cα/C backbone RMSD restricted to H3 residues after the joint VH+VL alignment:

$$\text{H3-RMSD} = \sqrt{\frac{1}{3|\mathcal{H}_3|} \sum_{i \in \mathcal{H}_3} \sum_{a \in \{N, C_\alpha, C\}} \|\widetilde{\mathbf{x}}_{i,a} - \mathbf{y}_{i,a}\|_2^2}. \tag{14}$$

No additional H3-specific fit is applied.

DockQ and interface RMSD. DockQ is computed using the standard combination of native-contact recovery, interface RMSD, and ligand RMSD. DockQ$_{BB}$ applies the DockQ procedure to a backbone-only N/Cα/C representation.

For Table 2, Aligned DockQ is computed after one joint VH+VL Kabsch alignment while keeping the native antigen fixed. It therefore removes only the global antibody rigid-body frame and should not be interpreted as a raw docking-pose metric. The reported bound iRMS is evaluated under the same aligned interface protocol.

For Table 3, the primary DockQ score is computed directly on the generated antigen–antibody complex without antibody-to-native rigid-body alignment, because recovering the relative pose is part of the task. DockQ$_{BB}$ denotes the corresponding backbone-only calculation.

Boltz-2 backbone self-consistency. For a generated sequence–structure pair, Boltz-2 independently predicts a structure from the generated sequence. SC$_{BB}$ is the N/Cα/C RMSD between this prediction and the structure decoded from the generated structural tokens. Bound evaluation uses one joint VH+VL fit; unbound evaluation uses one global antigen+VH+VL fit. Boltz-2 is run without MSA search using 200 sampling steps, 10 recycling steps, and 5 diffusion samples.

Amino-acid recovery. Amino-acid recovery is the fraction of generated residues exactly matching the native sequence. CDR-AAR restricts this calculation to the union of H1–H3 and L1–L3,

$$\text{CDR-AAR} = \frac{\sum_i \mathbf{1}[\hat{s}_i = s_i]\mathbf{1}[i \in \text{CDR}]}{\sum_i \mathbf{1}[i \in \text{CDR}]}. \tag{15}$$

H3 diversity and novelty. H3 residues are defined using SAbDab Chothia numbering as H95–H102, including insertion positions such as H100A and H100B. Generated sequences are not renumbered; instead, H3 is extracted using the corresponding native antibody H3 span.

For two H3 sequences a and b, we first perform a global sequence alignment and define sequence identity as

$$\text{SeqId}(a, b) = \frac{\sum_k \mathbf{1}[a_k = b_k,\ a_k \ne -]}{\sum_k \mathbf{1}[a_k \ne - \vee b_k \ne -]}, \tag{16}$$

so insertions and deletions contribute to the alignment length and reduce identity.

For the K H3 candidates generated for condition c, sequence diversity is the mean pairwise dissimilarity

$$D_{\text{seq}}^{(c)} = 1 - \frac{2}{K(K-1)} \sum_{i<j} \text{SeqId}(s_{c,i}, s_{c,j}). \tag{17}$$

All cached SMC candidates are included without deduplication, so repeated sequences reduce diversity.

To measure proximity to the training distribution, each generated H3 sequence s is compared with a reference library of deduplicated training-set H3s:

$$\text{NNSeqId}(s) = \max_{t \in \mathcal{D}_{\text{train}}^{H3}} \text{SeqId}(s, t). \tag{18}$$

Lower NNSeqId therefore indicates greater novelty. The reference library is constructed from SAbDab entries deposited no later than 2022-12-31. Diversity uses the full SMC candidate pools, whereas the novelty analysis uses the single-generation caches from the bound and unbound evaluations and is stratified by H3 length into 4–8, 9–13, and ≥ 14 residues.

Aggregation. Metrics are first computed per example and then averaged over examples for which the corresponding quantity is valid. Failed predictions or non-finite values are excluded only from the affected metric, and valid counts are recorded by the evaluation manifests.

C SMC Implementation and Additional Analyses

C.1 Incremental weighting and resampling

We use the trained DisAb reverse transition as the SMC proposal, so the proposal path distribution is the original DisAb reverse process and no additional proposal-density correction is required. Let $\{x_k^{(n)}, w_k^{(n)}, h_k^{(n)}\}_{n=1}^N$ denote the particle population at diffusion level k. Each particle first takes one stochastic reverse transition,

$$\widetilde{x}_{k+1}^{(n)} \sim K_\theta(x_{k+1} \mid x_k^{(n)}, c). \tag{19}$$

Because Boltz-2 scoring requires a complete sequence–structure pair, all remaining masked target tokens in $\widetilde{x}_{k+1}^{(n)}$ are then filled greedily in one additional denoiser pass:

$$\hat{y}_{k+1}^{(n)} = G_\theta(\widetilde{x}_{k+1}^{(n)}, c). \tag{20}$$

The completion is used only for scoring and is not propagated as the next diffusion state.

With fixed inverse temperature β, the look-ahead potential is

$$h_{k+1}^{(n)} = \exp\left[-\beta\, \text{SC}_{BB}(\hat{y}_{k+1}^{(n)})\right]. \tag{21}$$

The unnormalized importance weight is updated through the incremental potential ratio

$$\widetilde{w}_{k+1}^{(n)} = w_k^{(n)} \frac{h_{k+1}^{(n)}}{h_k^{(n)}}. \tag{22}$$

Potentials and incremental weights are evaluated in log space for numerical stability. We initialize $h_0^{(n)} = 1$. Along an uninterrupted particle lineage,

$$\prod_{k=0}^{K-1} \frac{h_{k+1}}{h_k} = h_K, \tag{23}$$

so intermediate look-ahead potentials influence particle allocation without compounding absolute potentials across diffusion levels, while the cumulative incremental weight reduces to the desired

Table 5: **Equal-Boltz-2-score-budget comparison.** Both methods use 16 Boltz-2 evaluations per target on the same 30 held-out bound-design examples.

Method	DockQ ↑	TM ↑	Cα RMSD ↓	All-atom RMSD ↓	SC_{BB} ↓	CDR-AAR ↑
Independent best-of-16	0.659	0.955	1.532	2.035	**1.129**	0.480
SMC (4×4)	**0.667**	**0.959**	**1.489**	**1.956**	1.146	**0.538**

terminal self-consistency potential. At the final transition, the reverse process has produced a complete state and greedy completion therefore reduces to the identity.

Weights are normalized as

$$\bar{w}_{k+1}^{(n)} = \frac{\widetilde{w}_{k+1}^{(n)}}{\sum_m \widetilde{w}_{k+1}^{(m)}}, \tag{24}$$

with

$$\mathrm{ESS} = \left(\sum_{n=1}^{N} (\bar{w}_{k+1}^{(n)})^2 \right)^{-1}. \tag{25}$$

We apply systematic resampling whenever $\mathrm{ESS}/N < 0.70$. After resampling, selected particles inherit their associated partial states and look-ahead potentials and their weights are reset to $1/N$; otherwise, the normalized weights are carried to the next transition.

The production configuration uses $N = 8$ particles and $K = 16$ reverse transitions, with score temperature $T = 0.5$ and hence fixed inverse temperature $\beta = T^{-1} = 2$. Boltz-2 scoring uses no MSA search, 200 sampling steps, 10 recycling steps, and 5 diffusion samples.

C.2 Output semantics

If a terminal particle were sampled according to its normalized importance weight, the resulting particle system would provide the standard finite-particle SMC approximation to the tilted distribution in Equation 11. The greedy completions used at intermediate levels act only as bridging potentials and affect particle allocation efficiency; because their incremental ratios telescope, they do not alter the terminal potential. Our reported design, however, is selected by the best-ever-seen rule rather than terminal weighted sampling, and is therefore interpreted as SMC-guided design search rather than a sample from π_β.

C.3 Equal-score-budget comparison

To separate the benefit of sequential particle guidance from simply evaluating more candidates with Boltz-2, we compare SMC against independent best-of-16 sampling under the same scoring budget on 30 held-out SAbDab examples. Both methods use exactly 16 Boltz-2 evaluations per target: SMC uses four particles across four scored diffusion layers, whereas the baseline independently samples 16 complete trajectories and returns the candidate with the best Boltz-2 score.

Under the same external-scoring budget, SMC improves most reported metrics over independent best-of-16 selection (Table 5): DockQ increases from 0.659 to 0.667, TM-score from 0.955 to 0.959, Cα RMSD decreases from 1.532 to 1.489 Å, and all-atom RMSD from 2.035 to 1.956 Å. CDR-AAR also increases from 0.480 to 0.538. Backbone self-consistency is essentially unchanged (1.129 versus 1.146 Å). These results suggest that sequential allocation of the same Boltz-2 scoring budget can improve overall design quality beyond terminal best-of-N selection.

C.4 Uncertainty of SMC improvements

We additionally estimate uncertainty using paired target-level bootstrap resampling. Candidate selection is completed before resampling, and the same target indices are resampled jointly for the two compared methods. We use 100,000 bootstrap replicates and report percentile 95% confidence intervals. Improvement is defined such that positive values always favor SMC.

Table 6: **Paired-bootstrap uncertainty for SMC guidance.** Values are mean paired improvements with 95% bootstrap confidence intervals; positive values favor SMC.

Comparison	Metric	Improvement	95% CI
Unbound ($N = 36$)	DockQ	+0.0202	$[-0.0113,\ 0.0543]$
	Complex RMSD	+0.3376	$[-0.9136,\ 1.6812]$
	$SC_{complex,BB}$	+1.8985	$[-0.3110,\ 4.1146]$
	CDR-AAR	+0.0410	$[0.0134,\ 0.0678]$
Equal-budget ($N = 30$)	DockQ	+0.0080	$[-0.0482,\ 0.0647]$
	TM-score	+0.0041	$[0.0002,\ 0.0083]$
	SC_{BB}	-0.0167	$[-0.1101,\ 0.0715]$
	CDR-AAR	+0.0578	$[0.0320,\ 0.0830]$

The intervals show that the direction of the mean SMC effect is positive for most representative metrics, but the structural improvements are not precisely estimated on these small cohorts. Importantly, this uncertainty is largely a consequence of the limited sample sizes: on the substantially larger 232-case bound benchmark (Table 2), SMC yields clear and consistent improvements across the headline sequence–structure metrics. Taken together, the larger-scale bound evaluation and the smaller unbound/equal-budget analyses provide complementary evidence that SMC guidance improves the quality of DisAb generations, although the magnitude of the gain is less precisely estimated in the smaller cohorts.

C.5 Alternative guidance scores

ProteinMPNN compatibility score. We additionally evaluated ProteinMPNN (Dauparas et al., 2022) as an alternative inverse-folding compatibility signal. Given the VQ-decoded VH/VL backbone X, decoding order π, and generated sequence a, we compute

$$S_{\mathrm{MPNN}}(a, X; \pi) = -\frac{1}{L}\sum_{k=1}^{L} \log p_{\mathrm{MPNN}}(a_{\pi_k} \mid X, a_{\pi_{<k}}, \pi), \tag{26}$$

and average over four deterministic decoding orders using the official v48_020 checkpoint.

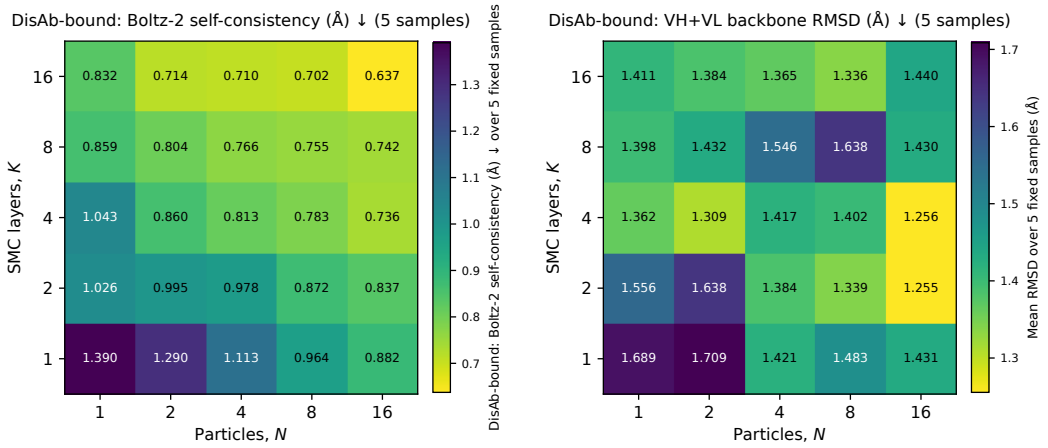
On the 50-case validation study, raw ProteinMPNN scoring and PDB-grouped isotonic fusion with Boltz-2 did not improve candidate selection over Boltz-2 alone: the best calibrated fusion obtained 1.7295 Å native VH+VL RMSD, compared with 1.7243 Å for raw Boltz-2 selection. We therefore use Boltz-2 alone as the production SMC potential.

Interface-energy guidance. For the unbound setting, an alternative direction is to guide SMC using an explicit antibody–antigen interface-energy objective, for example a PyRosetta-derived binding free-energy estimate after structural relaxation. Such a score may be closer to the ultimate design objective than structural self-consistency, since it directly favors energetically favorable interfaces rather than agreement between two structure predictions. We do not use this signal in the present SMC pipeline because reliable Rosetta interface energies generally require side-chain optimization and structural relaxation for each completed candidate. Repeating this procedure across particles and diffusion steps would make inference prohibitively expensive. Faster learned approximations to relaxed interface energy are therefore a promising direction for future guidance.

C.6 Effect of SMC compute

We study how the number of particles N and SMC layers K affect guided sampling in DisAb-bound. To isolate the effect of search compute, we use the same five fixed bound-design examples for every (N, K) configuration.

Figure 6(a) shows a clear decrease in Boltz-2 self-consistency as either the number of particles or the number of SMC layers increases. Averaged along either axis, the trend is monotonic, and the score drops from 1.390 Å at $N = K = 1$ to 0.637 Å at $N = K = 16$. This behavior is desirable independently of native-reference recovery: the purpose of joint sequence–structure generation is



(a) Boltz-2 self-consistency **(b)** Native VH+VL backbone RMSD

Figure 6: **Effect of SMC compute.** Mean performance over the same five fixed bound-design examples as the number of particles N and SMC layers K are varied. **(a)** Boltz-2 VH+VL backbone self-consistency. **(b)** Native VH+VL backbone RMSD. Lower is better in both panels.

Table 7: **Ablation of pretraining and unbound antigen-monomer conditioning.** Bound results use the 232-complex leakage-filtered SAbDab test set; unbound results use the 36-case docking benchmark.

Setting	Variant	Aligned DockQ ↑	TM ↑	Cα RMSD ↓	All-atom RMSD ↓	SC_{BB} ↓	CDR-AAR ↑
Bound	Full	**0.692**	**0.956**	**1.470**	**1.989**	**1.382**	**0.544**
	w/o bound pretraining	0.629	0.911	2.220	2.736	1.968	0.507
		DockQ ↑	VH+VL TM ↑	Complex RMSD ↓	Complex All-atom ↓	$SC_{complex,BB}$ ↓	CDR-AAR ↑
Unbound	Full	**0.129**	**0.861**	**10.712**	**10.910**	**14.362**	**0.517**
	w/o unbound pretraining	0.090	0.397	16.671	17.135	18.052	0.500
	w/o antigen-monomer track	0.115	0.842	13.125	13.521	14.497	0.502

to produce sequences and structures that are mutually compatible, and larger SMC search budgets increasingly favor such self-consistent candidates.

Native-reference RMSD behaves differently (Figure 6(b)). Increasing the search budget beyond a small 2×2 configuration does not yield a consistent additional reduction in RMSD, although several larger configurations remain competitive. This saturation is not necessarily inconsistent with the improvement in self-consistency. The two quantities measure different objectives: self-consistency compares the structure implied by the generated sequence with the jointly generated structure, whereas native RMSD compares the generated structure with one experimentally observed reference conformation.

Two factors may therefore limit the translation from lower self-consistency to lower native RMSD. First, a sequence–structure pair can become more internally consistent without becoming closer to the particular native structure. Second, Boltz-2 itself is an imperfect structure predictor, so its self-consistency estimate contains model error and cannot perfectly rank candidates by native RMSD. These results support using Boltz-2 as a reference-free guidance signal while also explaining why increasing SMC compute eventually yields diminishing returns in native-structure recovery.

D Ablation Studies

We ablate the generic pretraining stages and the isolated-antigen structural condition to identify their contributions to DisAb generation. For DisAb-bound, we remove generic bound pretraining and train only with the antibody-specific objective. For DisAb-unbound, we separately remove generic unbound pretraining or the antigen-monomer input track, while retaining the remaining training pipeline. Table 7 reports condition-matched structural and sequence metrics for the bound and unbound settings.

Effect of pretraining. Generic pretraining is important primarily for structural generation. Removing bound pretraining lowers TM-score from 0.956 to 0.911 and increases Cα/all-atom RMSD from 1.470/1.989 Å to 2.220/2.736 Å, while CDR-AAR decreases more modestly from 0.544 to 0.507. The effect is substantially larger in the unbound setting: removing unbound pretraining reduces VH+VL TM-score from 0.861 to 0.397 and increases whole-complex RMSD from 10.712 to 16.671 Å and all-atom RMSD from 10.910 to 17.135 Å. DockQ also decreases from 0.129 to 0.090, whereas CDR-AAR remains near 0.5. These results indicate that generic pretraining primarily establishes the structural prior required for joint geometry generation, with a smaller effect on local CDR sequence recovery.

Effect of the antigen-monomer track. Removing the isolated-antigen structural condition causes only a modest degradation of the internal VH+VL fold, with VH+VL TM-score decreasing from 0.861 to 0.842, but has a substantially larger effect on complex-level geometry. DockQ decreases from 0.129 to 0.115, while complex backbone and all-atom RMSD increase from 10.712/10.910 Å to 13.125/13.521 Å. Complex backbone self-consistency also worsens from 14.362 to 14.497 Å, whereas CDR-AAR changes only from 0.517 to 0.502. The antigen-monomer track therefore contributes primarily to modeling the target-specific antigen context and relative antigen–antibody arrangement, while also providing a smaller benefit to the internal VH+VL structure.