RiMES: Riemannian Flow Matching with Equivariance and Stereochemical Regularization for Antibody CDR Co-Design

Anonymous authors
Paper under double-blind review

Abstract

Antibody complementarity-determining region (CDR) co-design requires compatible sequence and structural predictions under a specified framework and antigen context. We present RiMES (Riemannian flow Matching with Equivariance and Stereochemical regularization), an iterative full-atom co-design model that couples geometric message passing to Riemannian flow matching on residue frames. A local-frame velocity head supplies translation corrections between refinement rounds, while a differentiable rotamer-compatibility term and a soft Cα overlap penalty connect sequence prediction to local geometry. The flow objective uses mixed prediction- and noise-based starts with detached targets and shared encoder representations. Under known-interface conditioning, we report PLM-free amino-acid recovery of 53.5% on RAbD CDR-H3, 77.1% on joint six-CDR design, and 47.9% on the cluster-date-ordered SAbDab evaluation. Component ablations support the mixed-start configuration and intermediate flow feedback, while exposing trade-offs between recovery and local structural accuracy. We separately evaluate parent-conditioned redesign using protein language model embeddings. Masking the designed CDR before PLM encoding and retraining removes the parent-conditioned recovery advantage, demonstrating the importance of distinguishing the two input regimes. Together, these results support flow-assisted conditional CDR co-design, with all-atom packing, antigen-relative placement, and experimental function remaining important limitations.

1 Introduction

Antibody function depends on the sequences and conformations of six complementarity-determining regions (CDRs). Designing these loops requires a joint sequence–structure model that uses framework and antigen context. We study conditional CDR co-design from antibody–antigen complexes, using the native framework, antigen, CDR lengths, and a reference-derived epitope as inputs. Our work focuses on local redesign under this conditioning. Generating complete antibodies or binding poses from an uncharacterized antigen remains outside the scope of these experiments.

Relationship to existing methods. Full-atom equivariant models such as dyMEAN (Kong et al., 2023) couple antibody geometry and sequence, while FlowDesign (Wu et al., 2025) applies flow matching with informed priors to CDR design. Residue-frame generative models (Yim et al., 2023; Bose et al., 2023) motivate learning on rotations and translations rather than unconstrained orientation coordinates. Our contribution is the coupling of iterative full-atom refinement, residue-frame supervision, and sequence-dependent stereochemical regularization within a single conditional co-design model. We use frame velocities to supervise shared representations and correct intermediate predictions; the iterative encoder produces the final full-atom structure.

Two information settings. In the **PLM-free setting**, we hide the target CDR sequence and supply no PLM embeddings. We use this setting for our primary architecture comparison. In the **parent-conditioned setting**, we encode an available parent sequence with ESM-2 and introduce its embeddings through a geometry-conditioned adapter. Although we zero target-position embeddings after encoding, framework embeddings retain native-CDR context. We therefore evaluate this variant,

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

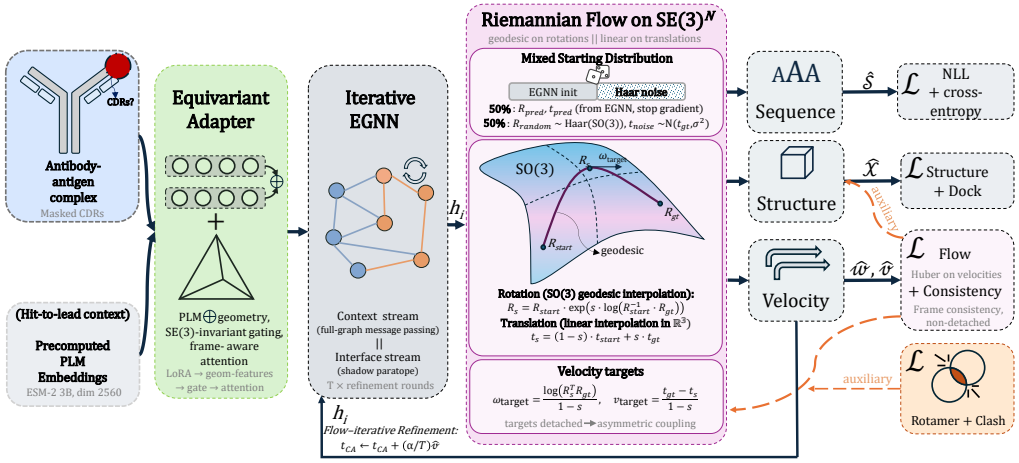


Figure 1: **RiMES workflow.** The adapter and parent PLM embeddings shown on the left apply only to RiMES*; PLM-free RiMES uses the EGNN residue features directly. Sequence and structure predictions read shared EGNN states. The velocity branch receives auxiliary frame-flow supervision and supplies translation corrections after the first two refinement rounds during training and inference. Mixed starts and ground-truth interpolation are used for training the velocity head. Frame consistency and rotamer/overlap penalties regularize predicted geometry. The reported generator uses iterative refinement without full ODE integration.

RiMES*, as parent-conditioned redesign. Recovery measures proximity to the parent in this setting; improved function requires separate validation.

Contributions. **(1) Coupled refinement and frame learning.** We share features between an EGNN and a residue-frame velocity predictor. Detached flow starts remove the target-side gradient path, while shared features and a non-detached consistency loss preserve encoder coupling. We use mixed starts and importance-weighted logit-normal time sampling for the auxiliary objective, and apply translation-only nudges between graph-refinement rounds during training and inference. **(2) Local physical regularization.** We couple sequence probabilities to predicted sidechain geometry through a differentiable χ_1 compatibility kernel and discourage coarse overlaps with Cα proximity penalties. These soft local constraints leave all-atom validity to be assessed separately. **(3) Evaluation under explicit input protocols.** We compare PLM-free and parent-conditioned design, include a control retrained with CDRs masked before ESM encoding, and report recovery, structure, naturalness, and interface-energy proxies separately.

On RAbD CDR-H3, PLM-free RiMES reaches 53.5% recovery versus 43.1% for dyMEAN, with 1.762 versus 1.573 Å Cα RMSD. Our main gain is in sequence recovery; structural rankings remain mixed. Our parent-conditioned model reaches 72.9% recovery, but masking the CDR before ESM extraction and retraining lowers it to 42.4% (App. E.1). This control motivates evaluating parent-conditioned redesign separately from PLM-free CDR design.

2 Method Overview and Workflow

RiMES refines sequences and full-atom coordinates over successive rounds (Fig. 1). A multi-channel EGNN encodes the supplied framework, antigen, and masked CDR context. Its shared residue features feed sequence, structure, and residue-frame velocity heads. The velocity head supplies translation corrections before the next refinement round during training and inference. We supervise this head with detached flow starts and targets while retaining encoder gradients through shared features and frame consistency. A Dunbrack-based rotamer term links sequence probabilities to predicted sidechain geometry. Parent-conditioned RiMES* additionally receives precomputed PLM embeddings through an optional adapter.

2.1 Problem Formulation and Preliminaries

We represent an antibody–antigen complex as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ whose nodes are residues. Each residue i carries an amino acid type $s_i \in \{1, \dots, 20\}$ and all-atom coordinates $\mathbf{X}_i \in \mathbb{R}^{C\times 3}$, where C covers backbone atoms (N, Cα, C, O) and up to ten sidechain slots. A subset $\mathcal{M} \subset \mathcal{V}$ marks the CDR residues to design, either CDR-H3 alone or all six loops (H1–H3, L1–L3). The epitope is selected as the 48 antigen residues closest to the reference CDR-H3. CDR lengths and framework context are supplied. Our goal is to jointly predict CDR sequences and structures under this known-interface conditioning:

$$p\big(\mathbf{s}_{\mathcal{M}},\, \mathbf{X}_{\mathcal{M}} \;\big|\; \mathbf{X}_{\mathcal{V}\setminus\mathcal{M}},\, \mathbf{s}_{\mathcal{V}\setminus\mathcal{M}}\big)\,. \tag{1}$$

Rigid-body geometry. Each residue i defines a local rigid frame $(R_i, \mathbf{t}_i) \in \mathrm{SE}(3) = \mathrm{SO}(3) \ltimes \mathbb{R}^3$, where R_i is built from the N–Cα–C triangle via Gram–Schmidt orthogonalization and $\mathbf{t}_i = \mathbf{C}\alpha_i$ (details in Appendix M.1) (Jumper et al., 2021). We use the exponential map $\exp : \mathbb{R}^3 \to \mathrm{SO}(3)$ (Rodrigues' formula) and its inverse, the logarithmic map, to move between axis-angle vectors and rotation matrices. These maps parameterize geodesics on $\mathrm{SO}(3)$ (Miller et al., 2024; Chen & Lipman, 2024; Cheng et al., 2026):

$$R(s) = R_0 \exp\big(s \cdot \log(R_0^\top R_1)\big), \quad s \in [0, 1], \tag{2}$$

which underpin the Riemannian flow matching formulation in §2.2.

2.2 Iterative Refinement with Residue-Frame Flow Supervision

We refine CDR predictions over T rounds. Each round embeds the current residue state, runs the multi-channel EGNN, and updates the predicted sequence and coordinates for the next round. The adapter participates only in the parent-conditioned variant (§2.4). Shared residue features also feed a velocity head on $\mathrm{SE}(3)^N$. Its flow loss trains these shared features, while frame consistency adds a direct gradient through predicted rotations (§2.3). The complete refinement loop is given in App. M.5.

Cross-round memory. To let later rounds build on earlier ones, we add a residual carry from the previous round's encoder output:

$$\mathbf{h}_i^{(0,r)} \leftarrow \mathbf{h}_i^{(0,r)} + \mathrm{FFN}\big(\mathbf{h}_i^{(L,r-1)}\big), \qquad r \geq 2. \tag{3}$$

Here $r = 1, \dots, T$ indexes refinement rounds, $\mathbf{h}_i^{(0,r)}$ is the post-adapter state, and $\mathbf{h}_i^{(L,r-1)}$ is the previous encoder output. The carry is a two-layer FFN; predicted sequences and coordinates also pass between rounds.

Sequence curriculum. During training, we reveal a fraction ρ_u of CDR sequence tokens as ground truth. We anneal this fraction from 0.9 to zero using $\rho_u = 0.5\big(\cos(\pi u/U_{\max}) + 1\big) \cdot 0.9$, where u indexes training steps and $U_{\max}$ is the training horizon. This index is distinct from refinement round r. Validation and inference use $\rho = 0$. The curriculum provides partly visible sequence context early in training and transitions to fully masked iterative design.

Multi-channel equivariant message passing. Each residue is modeled with C atom channels carrying learned importance weights w_i^c (Kong et al., 2023), so message passing operates on full-atom geometry rather than a single representative coordinate per residue. For every edge we compute cross-channel pairwise distances, weighted by atom importance and projected through learned atom attributes:

$$\mathbf{r}_{ij} = \frac{1}{\|\tilde{\mathbf{r}}_{ij}\| + 1} W_r\, \tilde{\mathbf{r}}_{ij}, \quad \tilde{\mathbf{r}}_{ij} = \mathbf{A}_i^\top \big[\|x_i^c - x_j^{c'}\| \cdot w_i^c w_j^{c'}\big]_{c,c'} \mathbf{A}_j, \tag{4}$$

Each layer then performs three updates. Edge messages aggregate pairwise information: $\mathbf{m}_{ij} = \mathrm{MLP}_{\mathrm{edge}}([\mathbf{h}_i \| \mathbf{h}_j \| \mathbf{r}_{ij}])$. Node features are updated with residual connections: $\mathbf{h}_i \leftarrow \mathbf{h}_i + \mathrm{MLP}_{\mathrm{node}}([\mathbf{h}_i \| \sum_j \mathbf{m}_{ij}])$. Coordinates are moved equivariantly (Tran et al., 2025):

$$\mathbf{x}_i^c \leftarrow \mathbf{x}_i^c + \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \phi_c(\mathbf{m}_{ij})\, (\mathbf{x}_i^c - \bar{\mathbf{x}}_j), \tag{5}$$

where $\phi_c(\mathbf{m}_{ij}) \in \mathbb{R}$ is a per-channel scalar gain produced by $\mathrm{MLP}_{\mathrm{coord}} : \mathbb{R}^d \to \mathbb{R}^C$ (one output per atom channel) and broadcast across the spatial axis, and $\bar{\mathbf{x}}_j$ is the channel-weighted centroid of neighbor j's atoms. The scalar-gain form preserves SE(3) equivariance because $Q\mathbf{x} - Q\bar{\mathbf{x}}$ co-rotates with the spatial vector while ϕ_c depends only on invariants. We run two parallel streams: a *context stream* over the full antibody–antigen graph and an *interface stream* restricted to the paratope–epitope subgraph. At each layer we synchronize paratope residues between the two, letting structural and binding-site information continuously flow into each other (Zeng et al., 2026).

Riemannian Flow Matching with Mixed Starting Distributions. Residue orientations lie on $\mathrm{SO}(3)$ (Brooks et al., 2019). Flow matching on $\mathrm{SE}(3)^N$ (Lipman et al., 2023; Chen & Lipman, 2024; Li et al., 2026; Park et al., 2022) supervises translation and rotation along their corresponding interpolation paths. We use a mixed starting distribution: perturbed EGNN predictions connect the auxiliary task to the refinement state, while noise-based starts provide additional training states (Yim et al., 2023). The component study compares this mixture with each single-start variant.

The starting frame is one of two modes, chosen per training graph by $m_g \sim \mathrm{Bernoulli}(0.5)$. The *prediction-mode* start uses the EGNN's current output with a stop-gradient,

$$R_i^{\mathrm{pred}} = \mathrm{sg}\big[\mathrm{FrameBuilder}(\hat{\mathbf{X}}_i)\big]_R, \quad \mathbf{t}_i^{\mathrm{pred}} = \mathrm{sg}\big[\hat{\mathbf{X}}_i^{\mathrm{C}\alpha}\big], \tag{6}$$

which makes the velocity targets $(\boldsymbol{\omega}^*, \mathbf{v}^*)$ constants w.r.t. the encoder; a small Gaussian perturbation ($\sigma_r = 0.1$ rad on $\mathrm{SO}(3)$, $\sigma_p = 0.5$ Å in $\mathbb{R}^3$) perturbs targets that would otherwise be zero at an exact prediction. The *noise-mode* start uses a Haar-uniform rotation on $\mathrm{SO}(3)$ (a rotation-invariant prior) and a Gaussian-perturbed ground-truth translation, providing a second source of training states. Given the starting frame, translations follow a linear path and rotations follow the geodesic on $\mathrm{SO}(3)$, with closed-form target velocities $(\mathbf{v}^*, \boldsymbol{\omega}^*)$; the full noiser, including the $(1-s)_\epsilon$ clip used near $s \to 1$, is in App. M.6.1 (Alg. 1).

Logit-normal time sampling. We sample interpolation times from a fixed logit-normal distribution (Esser et al., 2024),

$$s = \sigma(\mu + \sigma_z z), \qquad z \sim \mathcal{N}(0, 1), \tag{7}$$

and use self-normalized importance weights $w(s) \propto 1/p(s)$ to approximate a uniform-time objective. This finite-batch estimator is not exactly unbiased. The schedule changes sampling density rather than the geodesic path; its parameters are fixed hyperparameters, not a learned curriculum. We clip the denominator $(1-s)_\epsilon = \max(1-s, 10^{-3})$ when computing velocity targets near the endpoint. Along the ideal interpolation, target velocities remain finite; the clipping addresses numerical conditioning.

Equivariant velocity prediction. Our velocity head predicts body-frame angular velocities and ambient translation velocities from shared residue features and local frame geometry. Node inputs are $\mathbf{f}_i^{(0)} = \mathrm{MLP}_{\mathrm{in}}([\mathbf{h}_i \| \phi(s)])$. The message-passing layers use the invariant relative-frame features

$$\mathbf{e}_{ij} = \big(\|\mathbf{t}_j - \mathbf{t}_i\|, R_i^\top(\mathbf{t}_j - \mathbf{t}_i), \mathrm{vec}(R_i^\top R_j)\big). \tag{8}$$

Invariant message passing yields $\mathbf{f}_i^{(L_v)}$, with readout

$$\hat{\boldsymbol{\omega}}_i = W_\omega \mathbf{f}_i^{(L_v)}, \qquad \hat{\mathbf{v}}_i = R_i W_v \mathbf{f}_i^{(L_v)}. \tag{9}$$

With invariant encoder features, valid co-transforming frames, and a transform-independent neighborhood rule, the angular output is invariant and the translation output co-rotates under a global rigid transformation:

$$\hat{\boldsymbol{\omega}}_i(g \cdot x) = \hat{\boldsymbol{\omega}}_i(x), \qquad \hat{\mathbf{v}}_i(g \cdot x) = R_g \hat{\mathbf{v}}_i(x). \tag{10}$$

We prove this head-level property and compare the head with a coordinate-input predecessor in Appendix N. The proof applies to the velocity head; symmetry of the full generation pipeline requires separate analysis.

Flow-iterative refinement. We index refinement rounds by $r = 1, \ldots, T$ during training and inference. After each non-final round, we query the velocity head at $s = r/T$ and update Cα coordinates by

$$\mathbf{t}_i^{\mathrm{CA}} \leftarrow \mathbf{t}_i^{\mathrm{CA}} + \frac{\alpha}{T} \hat{\mathbf{v}}_i(s = r/T), \qquad r < T. \tag{11}$$

With $T = 3$ and $\alpha = 0.05$, the two queries occur at $s = 1/3$ and $s = 2/3$. Each nudge updates translations only; the final round produces the sequence and structure without another nudge. We do not use the optional post-EGNN midpoint integrator in our experiments.

Role of the flow objective. We use flow matching as an auxiliary representation-learning objective and to supply translation corrections during training and inference. Our EGNN decodes the final structure. The generator therefore does not sample a full ODE trajectory from the random training start distribution. We describe the optional midpoint integrator in App. M.10, but do not use it in our experiments. Our ablations examine the training objective and intermediate nudges within this coupled model.

2.3 Losses, Training, and Inference

Rotamer-Aware Joint Loss. We link predicted residue probabilities to sidechain geometry using a differentiable penalty based on the Dunbrack rotamer library (Shapovalov & Dunbrack, 2011). We compute the χ_1 sidechain dihedral directly from predicted coordinates:

$$\chi_1^{(i)} = \operatorname{atan2}\big((\mathbf{n}_1 \times \mathbf{n}_2) \cdot \hat{\mathbf{u}}_1,\ \mathbf{n}_1 \cdot \mathbf{n}_2\big), \tag{12}$$

where $\mathbf{n}_1, \mathbf{n}_2$ are normals to the N-Cα-Cβ and Cα-Cβ-X$_\gamma$ planes, and $\hat{\mathbf{u}}_1$ is the Cα-Cβ unit vector. We regularize χ_1, the first sidechain torsion, using predicted sidechain atoms. The backbone alone does not determine this angle. For an amino acid type a with rotamers $\{(\chi_a^{(r)}, p_a^{(r)})\}_{r=1}^{n_a}$, compatibility is computed via a probability-weighted von Mises kernel (bandwidth $\sigma_\chi = 30°$):

$$\operatorname{compat}(\chi, a) = \sum_{r=1}^{n_a} p_a^{(r)}\ \exp\left(\frac{\cos(\chi - \chi_a^{(r)}) - 1}{\sigma_\chi^2}\right). \tag{13}$$

The per-residue penalty weights these incompatibilities by the predicted sequence distribution $\mathbf{p}_i = \operatorname{softmax}(\hat{\mathbf{s}}_i)$:

$$\mathcal{L}_{\text{rotamer}} = \frac{\lambda_{\text{rot}}}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \sum_{a=1}^{20} p_i^{(a)} \big(1 - \operatorname{compat}(\chi_1^{(i)}, a)\big). \tag{14}$$

The loss sends gradients to both sequence probabilities and predicted sidechain coordinates. Glycine and alanine lack χ_1 and contribute zero penalty. Because penalty scales differ across residue types, its effect can include changes in composition as well as geometry (App. B).

Flow Matching Loss. We train the velocity network via a smooth L_1 (Huber) loss ℓ_δ (Huber, 1992) to reduce sensitivity to large prediction residuals:

$$\mathcal{L}_{\text{flow}} = \lambda_\omega \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} w_i\, \ell_\delta(\hat{\boldsymbol{\omega}}_i, \boldsymbol{\omega}_i^*) + \lambda_v \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} w_i\, \ell_\delta(\hat{\mathbf{v}}_i, \mathbf{v}_i^*). \tag{15}$$

We additionally regularize pairwise frame relationships using the *non-detached* predicted frames:

$$\mathcal{L}_{\text{consist}} = \frac{1}{|\mathcal{E}_\mathcal{M}|} \sum_{(i,j) \in \mathcal{E}_\mathcal{M}} \ell_\delta\left(\frac{\operatorname{tr}(\hat{R}_i^\top \hat{R}_j)}{3},\ \frac{\operatorname{tr}(R_i^{\text{gt}\top} R_j^{\text{gt}})}{3}\right). \tag{16}$$

Steric Clash Penalty. To discourage physically impossible overlaps, we apply a soft penalty on Cα-Cα distances:

$$\mathcal{L}_{\text{clash}} = \lambda_{\text{clash}}\Big[\lambda_{\text{inter}} \sum_{i \in \mathcal{M},\, j \in \mathcal{V}_{\text{ag}}} \operatorname{softplus}(d_{\text{thresh}} - d_{ij}) + \sum_{i,j \in \mathcal{M},\, i \neq j} \operatorname{softplus}(d_{\text{thresh}} - d_{ij})\Big], \tag{17}$$

using a coarse Cα proximity threshold $d_{\text{thresh}} = 3.0$ Å and an increased interface weight ($\lambda_{\text{inter}} = 2.0$).

Joint objective and training. We sum sequence, structure, and docking losses across the T refinement rounds and evaluate the auxiliary losses on the final outputs:

$$\mathcal{L} = \sum_{r=1}^{T} \left[\mathcal{L}_{\text{seq}}^{(r)} + \mathcal{L}_{\text{struct}}^{(r)} + \mathcal{L}_{\text{dock}}^{(r)}\right] + \lambda_f \mathcal{L}_{\text{flow}} + \lambda_c \mathcal{L}_{\text{consist}} + \mathcal{L}_{\text{rotamer}} + \mathcal{L}_{\text{clash}}. \tag{18}$$

The rotamer and clash terms include their stated weights. Stopping gradients through the flow starts removes the target-side gradient into the encoder. It does not detach the shared hidden features: the flow loss trains the velocity network and shared representations, while frame consistency additionally acts through the predicted rotations. We optimize our model with this joint objective and the cosine sequence-reveal curriculum. We precompute the optional ESM embeddings. The three refinement rounds and four adapter operations form the model computation; they are not separate pretraining stages.

2.4 Optional Parent-Conditioned Adapter

For RiMES*, we precompute ESM-2 embeddings of the available parent heavy and light chains (Lin et al., 2023; Hayes et al., 2025). The adapter has four blocks within a forward pass: a LoRA projection (Hu et al., 2022), an embedding of six invariant local-geometry scalars, a learned gate combining these features, and geometry-biased k-nearest-neighbor attention. The attention biases use distances and relative-frame traces. These inputs yield SE(3)-invariant features; equations and dimensions are in App. M.3.

At inference, we zero each designed CDR's embedding after PLM encoding. This removes its direct target-position vector, while framework embeddings can still carry parent-CDR context. We therefore evaluate RiMES* separately from the PLM-free model. Training uses the sequence-reveal curriculum described above. The four adapter blocks are jointly optimized operations, not separate training stages.

3 Experiments

3.1 Datasets

We use antibody-antigen complexes from SAbDab (Schneider et al., 2022; Dunbar et al., 2014) (yielding 3,779 filtered, unique full complexes) and the 60-complex RAbD benchmark (Adolf-Bryfogle et al., 2018). To reduce target-sequence overlap, we cluster CDR-H3 sequences via MMseqs2 at a 40% identity threshold and assign entire clusters to splits. This control does not certify antigen-level or structural independence.

We evaluate on three distinct tasks:

1. **Single CDR-H3 design (RAbD):** Training and validation use SAbDab, while testing uses the fixed RAbD set. Training clusters overlapping the test set are strictly removed, enabling direct comparison with standard baselines.
2. **Multi-CDR design (RAbD):** Uses identical cluster-based splits as above, but tasks the model with jointly designing all six CDR loops.
3. **Temporal-split single CDR-H3 (SAbDab):** A chronological split of SAbDab using date metadata, with each sequence cluster assigned wholly to one partition. The main evaluation includes all 754 test targets under the same structural conditioning.

3.2 Metrics

We evaluate sequence recovery, structural fidelity, and physical quality. For **sequence accuracy**, we report Amino Acid Recovery (AAR), Contact AAR (CAAR) for antigen-contacting residues, and BLOSUM62 for substitution plausibility. For **structural fidelity**, we measure global similarity via TM-score and local backbone deviation via $\text{RMSD}_{\text{H3/CDRs}}$ (Cα RMSD after independently fitting each designed CDR, in Å). For **local geometry**, we track steric conflicts ($\text{Clashes}_{\text{H3/CDRs}}$)

Table 1: Sequence and structural metrics across the three settings. RMSD is computed after independently fitting each designed CDR; multi-CDR RMSD averages the loop scores. Clashes denote intra-CDR Cα overlaps under the specified criterion. RiMES (unmarked) = PLM-free. **Bold** / underline = best / second-best per column. Markers: ‡ CDR-grafting, † additional inference-time conditioning, ⋆ ESM-2-3B (Appendix K.3). Every row, ours and baselines alike, is a single training run; subscripts are per-target standard deviations across the test targets, not seed-to-seed spread. Seed-to-seed variability for RiMES is reported separately in App. I.

Experiment	Method	AAR ↑	CAAR ↑	$\text{RMSD}_{\text{H3/CDRs}}$ ↓	$\text{Clashes}_{\text{H3/CDRs}}$ ↓
CDR-H3	dyMEAN	$\underline{0.431}_{\pm 0.138}$	$\underline{0.279}_{\pm 0.190}$	$\mathbf{1.573}_{\pm 0.622}$	$0.267_{\pm 0.667}$
	dyAb †	$0.335_{\pm 0.098}$	$0.204_{\pm 0.139}$	$1.727_{\pm 0.699}$	$0.635_{\pm 3.217}$
	AbEgDiffuser ‡⋆	$0.322_{\pm 0.151}$	$0.265_{\pm 0.178}$	$1.961_{\pm 0.917}$	$0.072_{\pm 0.393}$
	FlowDesign ‡†	$0.325_{\pm 0.171}$	$0.224_{\pm 0.181}$	$\underline{1.657}_{\pm 0.742}$	$\underline{0.036}_{\pm 0.202}$
	AbX ‡ † ⋆	$0.279_{\pm 0.095}$	$0.218_{\pm 0.150}$	$4.324_{\pm 3.920}$	$3.310_{\pm 3.455}$
	RiMES	$\mathbf{0.535}_{\pm 0.139}$	$\mathbf{0.334}_{\pm 0.209}$	$1.762_{\pm 0.636}$	$\mathbf{0.000}_{\pm 0.000}$
Multi-CDR	dyMEAN	$0.521_{\pm 0.186}$	$0.377_{\pm 0.273}$	$1.132_{\pm 0.459}$	$3.929_{\pm 3.192}$
	dyAb †	$0.271_{\pm 0.150}$	$0.176_{\pm 0.199}$	$1.204_{\pm 0.381}$	$0.879_{\pm 1.643}$
	AbEgDiffuser ‡⋆	$0.520_{\pm 0.179}$	$0.372_{\pm 0.259}$	$\underline{0.805}_{\pm 0.436}$	$\mathbf{0.131}_{\pm 0.490}$
	FlowDesign ‡†	$\underline{0.661}_{\pm 0.171}$	$0.506_{\pm 0.292}$	$0.946_{\pm 0.385}$	$\underline{0.144}_{\pm 0.219}$
	AbX ‡ † ⋆	$0.632_{\pm 0.083}$	$\underline{0.527}_{\pm 0.112}$	$\mathbf{0.729}_{\pm 0.112}$	$0.981_{\pm 2.342}$
	RiMES	$\mathbf{0.771}_{\pm 0.062}$	$\mathbf{0.687}_{\pm 0.094}$	$0.885_{\pm 0.343}$	$0.173_{\pm 0.385}$
Time Split	dyMEAN	$0.375_{\pm 0.116}$	$0.275_{\pm 0.206}$	$\underline{2.284}_{\pm 0.975}$	$1.016_{\pm 3.076}$
	dyAb †	$0.353_{\pm 0.121}$	$0.260_{\pm 0.194}$	$2.517_{\pm 1.235}$	$1.141_{\pm 9.867}$
	AbEgDiffuser ‡⋆	$0.289_{\pm 0.108}$	$0.196_{\pm 0.143}$	$2.584_{\pm 1.138}$	$0.312_{\pm 0.887}$
	FlowDesign ‡†	$\underline{0.451}_{\pm 0.131}$	$\underline{0.374}_{\pm 0.178}$	$\mathbf{1.108}_{\pm 0.524}$	$\underline{0.255}_{\pm 0.695}$
	AbX ‡ † ⋆	$0.423_{\pm 0.105}$	$0.321_{\pm 0.168}$	$2.824_{\pm 1.255}$	$0.887_{\pm 1.128}$
	RiMES	$\mathbf{0.479}_{\pm 0.125}$	$\mathbf{0.385}_{\pm 0.192}$	$2.407_{\pm 0.782}$	$\mathbf{0.130}_{\pm 1.122}$

and backbone stereochemistry via the coarse torsion-region fraction (Rama), a heuristic computed with the explicit angular rules in App. O, rather than a calibrated residue-specific Ramachandran validation score (Ramachandran, 1963). Finally, for **interface-energy plausibility**, we report the completion-corrected fraction of designs satisfying $\Delta\Delta G < 0$ under PyRosetta InterfaceAnalyzer after OpenMM relaxation (App. O.5). Multi-CDR RMSD averages the separately fitted loop scores; clash metrics cover all designed loops. We compute structural and physical metrics on raw outputs and use relaxed structures only for energy scoring. Full mathematical definitions are provided in Appendix O.

3.3 Results

Tables 1 and 2 report the PLM-free comparison. RiMES has higher reported sequence-recovery scores in these tables, with the PLM-free H3 comparison and component study forming our primary evidence. Structural and interface-energy rankings are mixed. These aggregate comparisons must be read together with the conditioning and output-paradigm differences in App. K.3.

Sequence recovery and seed variation. Our PLM-free model reaches H3 AAR 0.535 versus 0.431 for dyMEAN, with higher contact recovery and BLOSUM62 in the reported comparison. Across four training seeds using the same procedure within this variant, PLM-free H3 recovery ranges from 53.3% to 54.1% (mean 53.7%), indicating limited variation in aggregate recovery under this configuration (App. I). The single-run headline and across-seed mean are separate summaries. The multi-CDR aggregate scores and per-loop summary have an unresolved aggregation correspondence (App. E); comparisons in that setting remain descriptive. The separate paired FlowDesign analysis does not establish significance of its main-table margin (App. D).

PLM-free component evidence. The full PLM-free configuration has the highest AAR, contact recovery (CAAR), and whole-complex DockQ among the variants in Table 3. Removing inference-time feedback lowers AAR from 0.535 to 0.453 and DockQ from 0.366 to 0.331. Mixed starts

Table 2: Proxy metrics across the three settings (markers as in Table 1). $\Delta\Delta G < 0$ is the *completion-corrected* fraction of designs with lower predicted interface energy than the native complex under PyRosetta InterfaceAnalyzer (unscored designs count as $+\infty$ $\Delta\Delta G$); mean±std across targets. This statistic combines processing completion with the energy criterion and is not a standalone measure of molecular validity. TM-score is strongly affected by output paradigm because CDR-grafting methods reuse the native framework verbatim (Appendix K.4).

Experiment	Method	BLOSUM62 ↑	$\Delta\Delta G < 0$ ↑	TM-score ↑	Rama ↑
CDR-H3	dyMEAN	$\underline{0.552}_{\pm 0.124}$	$0.524_{\pm 0.307}$	$0.974_{\pm 0.009}$	$0.473_{\pm 0.157}$
	dyAb †	$0.465_{\pm 0.089}$	$0.511_{\pm 0.335}$	$0.953_{\pm 0.011}$	$0.331_{\pm 0.100}$
	AbEgDiffuser ‡⋆	$0.483_{\pm 0.122}$	$0.433_{\pm 0.409}$	$\underline{0.991}_{\pm 0.008}$	$0.251_{\pm 0.161}$
	FlowDesign ‡†	$0.478_{\pm 0.142}$	$\mathbf{0.680}_{\pm 0.271}$	$\mathbf{0.994}_{\pm 0.005}$	$0.295_{\pm 0.160}$
	AbX ‡ † ⋆	$0.248_{\pm 0.071}$	$0.610_{\pm 0.321}$	$0.967_{\pm 0.015}$	$\underline{0.489}_{\pm 0.150}$
	RiMES	$\mathbf{0.630}_{\pm 0.119}$	$\underline{0.658}_{\pm 0.122}$	$0.969_{\pm 0.009}$	$\mathbf{0.625}_{\pm 0.052}$
Multi-CDR	dyMEAN	$0.619_{\pm 0.050}$	$0.478_{\pm 0.275}$	$0.951_{\pm 0.012}$	$\underline{0.531}_{\pm 0.070}$
	dyAb †	$0.457_{\pm 0.047}$	$0.397_{\pm 0.299}$	$0.943_{\pm 0.011}$	$0.434_{\pm 0.049}$
	AbEgDiffuser ‡⋆	$0.668_{\pm 0.055}$	$0.598_{\pm 0.222}$	$\underline{0.984}_{\pm 0.010}$	$0.252_{\pm 0.055}$
	FlowDesign ‡†	$0.742_{\pm 0.061}$	$\mathbf{0.691}_{\pm 0.314}$	$\mathbf{0.989}_{\pm 0.007}$	$0.318_{\pm 0.057}$
	AbX ‡ † ⋆	$\underline{0.743}_{\pm 0.056}$	$0.599_{\pm 0.148}$	$0.977_{\pm 0.013}$	$\mathbf{0.581}_{\pm 0.080}$
	RiMES	$\mathbf{0.838}_{\pm 0.052}$	$\underline{0.642}_{\pm 0.324}$	$0.966_{\pm 0.012}$	$0.463_{\pm 0.037}$
Time Split	dyMEAN	$0.500_{\pm 0.108}$	$0.600_{\pm 0.327}$	$0.952_{\pm 0.027}$	$0.476_{\pm 0.118}$
	dyAb †	$0.484_{\pm 0.111}$	$0.593_{\pm 0.351}$	$0.937_{\pm 0.031}$	$0.433_{\pm 0.082}$
	AbEgDiffuser ‡⋆	$0.452_{\pm 0.089}$	$\underline{0.825}_{\pm 0.248}$	$\underline{0.985}_{\pm 0.012}$	$0.295_{\pm 0.119}$
	FlowDesign ‡†	$0.576_{\pm 0.106}$	$\mathbf{0.848}_{\pm 0.243}$	$\mathbf{0.993}_{\pm 0.008}$	$0.325_{\pm 0.129}$
	AbX ‡ † ⋆	$\underline{0.610}_{\pm 0.067}$	$0.653_{\pm 0.221}$	$0.951_{\pm 0.021}$	$\underline{0.599}_{\pm 0.077}$
	RiMES	$\mathbf{0.648}_{\pm 0.115}$	$0.650_{\pm 0.242}$	$0.952_{\pm 0.025}$	$\mathbf{0.640}_{\pm 0.126}$

Table 3: **PLM-free component ablations on RAbD CDR-H3** (60 targets, 40 samples each). CAAR measures contact-residue recovery. RMSD fits each CDR independently and measures local loop shape; DockQ uses full complexes (App. O.2). Overlaps count intra-CDR Cα pairs. The Euclidean variant retains flow training; the feedback ablation removes inference nudges. All values are from Table 14.

PLM-free configuration	AAR ↑	CAAR ↑	Self-fitted RMSD (Å) ↓	DockQ ↑	Cα overlaps ↓
RiMES	**0.535**	**0.334**	1.762	**0.366**	0.000
Uniform time sampling	0.529	0.314	1.760	0.356	0.019
Euclidean-flow variant	0.505	0.316	1.884	0.353	0.025
No inference-time feedback	0.453	0.307	1.844	0.331	0.053
Without rotamer loss	0.494	0.309	1.731	0.327	0.040
Prediction-only starts	0.442	0.220	1.810	0.345	0.019
Noise-only starts	0.413	0.309	1.853	0.349	0.019
Without frame consistency	0.501	0.316	1.825	0.347	0.005

outperform prediction-only and noise-only starts by 9.3 and 12.2 AAR points. The rotamer term exposes a different trade-off: removing it slightly improves self-fitted RMSD, but lowers CAAR from 0.334 to 0.309 and DockQ from 0.366 to 0.327. Thus the component evidence extends beyond overall sequence recovery. All seven shared ablation changes also reduce reported AAR in the parent-conditioned setting (App. F), although their effect sizes differ between regimes.

Recovery and candidate diversity. The PLM-free temperature sweep in App. G.4 shows that RiMES can produce a varied candidate pool while retaining much of its recovery. At $\tau = 0.3$, we obtain 31.5 unique sequences and 24.7 sequence clusters per target from 40 samples, compared with 1.8 sequences and 1.5 clusters under greedy decoding. AAR changes from 53.6% to 52.5% and CAAR from 33.4% to 31.6%. This sweep uses a separate training seed; temperature controls sequence diversity at generation time without retraining.

Structural and interface-quality trade-offs. On the chronological benchmark, RiMES has higher reported AAR than FlowDesign (47.9% versus 45.1%), alongside higher CDR-self-fitted RMSD (2.407 versus 1.108 Å). On RAbD H3, RiMES gives zero intra-CDR Cα overlaps under the specified criterion. These local scores coexist with substantial raw all-atom overlaps and antigen-relative placement errors (App. H). Table 2 reports completion-corrected interface-energy scores after relaxation, with unscored designs treated as $+\infty$ $\Delta\Delta G$; these computed scores do not measure binding affinity.

Local sidechain improvements. In the parent-conditioned loss ablation, including the rotamer term lowers type-matched χ_1 error from $76.4°$ to $63.9°$ and raw H3–antigen atom-pair clashes from 19.00 to 8.89 under the 2.0 Å criterion (Table 19). Native-angle recovery changes little (0.444 to 0.447). This supports local error reduction beyond the coarse Cα metric, while leaving exact native rotamer recovery and complete packing unresolved.

Parent-conditioned variant generation. With parent-sequence embeddings, H3 recovery is 72.9%, compared with 53.5% for PLM-free RiMES. Masking the CDR before PLM encoding and retraining gives 42.4%, removing this advantage (App. E.1). We therefore evaluate RiMES* separately as parent-conditioned variant generation. Some sequence models favor its variants, while AbLang2 favors the parent on average; the complete scorecard, including metric-specific best-of-40 selection, is in App. E.

Experimental measurements on a parent-conditioned panel. Ten CDR-H3 redesigns of trastuzumab were expressed and assayed against HER2 domain IV, with the native antibody produced and measured under the same protocol. All ten expressed (4.7–5.7 mg/ml; 90–94% SEC-HPLC purity) and all ten bound by bio-layer interferometry (Table 4). The strongest, a two-residue edit of the parent loop, binds at $K_D = 1.46$ nM against 0.98 nM for trastuzumab, with a similar association rate. Nine of ten show no self-association liability and all ten melt within 1.1 °C of the parent. No design exceeded native affinity. This is a single-target panel of ten candidates without a matched baseline arm, measured once per analyte, so it indicates that candidates from this regime expressed and bound under these conditions rather than establishing a general property of the method; the interface-energy proxy used for selection did not order the assayed candidates (Spearman $\rho = +0.05$). Full assay details, the developability panel and the proxy analysis are in App. C.6; Fig. 2 shows the modelled complexes.

Table 4: Measured affinity for the two strongest HER2 designs, with native trastuzumab under the same protocol. Bold marks the two CDR-H3 positions that differ from the parent. Single determination per analyte.

Design	CDR-H3	K_D (nM)	k_a (10^5/Ms)	k_d (10^{-4}/s)
s472	SRWGG**GG**YYAMDY	1.46	2.65	3.86
s343	SRWGG**YG**YYAMDY	2.93	3.11	9.11
Trastuzumab	SRWGGDGFYAMDY	0.98	2.55	2.50

4 Limitations and Conclusion

RiMES connects iterative full-atom prediction with residue-frame supervision and local sequence–geometry regularization. The reported benchmarks and component studies support this integrated approach to conditional CDR co-design, with the clearest gains in sequence recovery. Sampling provides additional control over candidate diversity. Structural results retain trade-offs in local accuracy, all-atom packing, and antigen-relative placement. Parent-conditioned variant generation is evaluated separately. Ten parent-conditioned CDR-H3 designs were expressed and assayed against HER2: all ten bound, the strongest at $K_D = 1.46$ nM against 0.98 nM for the parent, with developability retained (App. C.6). No design exceeded native affinity, and the interface-energy proxy used for selection did not rank the assayed candidates.

AI USE STATEMENT

Large language models assisted with wording, high-level ideation, code, and related-work discovery during preparation. They also assisted with a critical manuscript review, interpretation of reported results, mathematical consistency checks, and revision of the method and evaluation descriptions. This revision did not generate experimental data or rerun the scientific experiments. The authors remain responsible for validating the final claims, results, references, and AI-assisted changes.

ETHICS STATEMENT

This work is computational in nature and does not involve human participants, animal subjects, or the collection of personally identifiable information. All datasets used in this study were obtained from publicly available sources and were used in accordance with their respective terms and licenses. Our methods are intended to support scientific research in molecular modeling and drug discovery. We acknowledge that generative molecular design methods may potentially be applied beyond their intended research purposes; therefore, we encourage responsible use of the proposed methodology and appropriate experimental and safety validation before considering generated molecules for real-world applications. We have aimed to report the methodology, experimental settings, and evaluation procedures transparently to support reproducibility and research integrity.

REPRODUCIBILITY STATEMENT

Our methods and appendices describe the input protocols, sequence-cluster splits, model components, training objective, hyperparameters, and evaluation metrics. We distinguish PLM-free design from parent-conditioned redesign, report masked-input retraining and seed variability, and separate standard inference from optional flow integration. The velocity-head proof and initialization controls appear in App. N, with implementation details and evaluation definitions in Apps. M.6.2 and O. The source code will be shared publicly.

REFERENCES

Jared Adolf-Bryfogle, Oleks Kalyuzhniy, Michael Kubitz, Brian D Weitzner, Xiaozhen Hu, Yumiko Adachi, William R Schief, and Roland L Dunbrack Jr. Rosettaantibodydesign (rabd): A general framework for computational antibody design. *PLoS computational biology*, 14(4):e1006112, 2018.

Avishek Joey Bose, Tara Akhound-Sadegh, Guillaume Huguet, Kilian Fatras, Jarrid Rector-Brooks, Cheng-Hao Liu, Andrei Cristian Nica, Maksym Korablyov, Michael Bronstein, and Alexander Tong. Se (3)-stochastic flow matching for protein backbone generation. *arXiv preprint arXiv:2310.02391*, 2023.

Daniel Brooks, Olivier Schwander, Frédéric Barbaresco, Jean-Yves Schneider, and Matthieu Cord. Riemannian batch normalization for spd neural networks. *Advances in neural information processing systems*, 32, 2019.

Ricky TQ Chen and Yaron Lipman. Flow matching on general geometries. In *12th International Conference on Learning Representations, ICLR 2024*, 2024.

Chaoran Cheng, Yusong Wang, Yuxin Chen, Xiangxin Zhou, Nanning Zheng, and Ge Liu. Riemannian consistency model. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=SAlCQdk5lx.

James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M Deane. Sabdab: the structural antibody database. *Nucleic acids research*, 42 (D1):D1140–D1146, 2014.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=g97OHbQyk1.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pp. 492–518. Springer, 1992.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Foundations of Crystallography*, 32(5):922–923, 1976.

Xiangzhe Kong, Wenbing Huang, and Yang Liu. End-to-end full-atom antibody design. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 17409–17429, 2023.

Jack Kyte and Russell F Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, 157(1):105–132, 1982.

Marie-Paule Lefranc, Christelle Pommié, Manuel Ruiz, Véronique Giudicelli, Elodie Foulquier, Lisa Truong, Valérie Thouvenin-Contet, and Gérard Lefranc. Imgt unique numbering for immunoglobulin and t cell receptor variable domains and ig superfamily v-like domains. *Developmental & Comparative Immunology*, 27(1):55–77, 2003.

Zihao Li, Zhichen Zeng, Xiao Lin, Feihao Fang, Yanru Qu, Zhe Xu, Zhining Liu, Xuying Ning, Tianxin Wei, Ge Liu, et al. Flow matching meets biology and life science: a survey. *npj Artificial Intelligence*, 2(1):17, 2026.

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

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

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

Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. lddt: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, 2013.

Benjamin Kurt Miller, Ricky TQ Chen, Anuroop Sriram, and Brandon M Wood. Flowmm: Generating materials with riemannian flow matching. In *International Conference on Machine Learning*, pp. 35664–35686. PMLR, 2024.

Claudio Mirabello and Björn Wallner. Dockq v2: improved automatic quality measure for protein multimers, nucleic acids, and small molecules. *Bioinformatics*, 40(10):btae586, 2024.

Erik Nijkamp, Jeffrey A Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. Progen2: exploring the boundaries of protein language models. *Cell systems*, 14(11):968–978, 2023.

Tobias H Olsen, Fergus Boyles, and Charlotte M Deane. Observed antibody space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Science*, 31 (1):141–146, 2022.

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

Sung Woo Park, Hyomin Kim, Kyungjae Lee, and Junseok Kwon. Riemannian neural SDE: Learning stochastic representations on manifolds. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=1ryTomA0iKa.

Gopalasamudram Narayana Ramachandran. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, 7:95–99, 1963.

Constantin Schneider, Matthew IJ 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.

Maxim V Shapovalov and Roland L Dunbrack. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure*, 19(6):844–858, 2011.

Richard W Shuai, Jeffrey A Ruffolo, and Jeffrey J Gray. Iglm: Infilling language modeling for antibody sequence design. *Cell systems*, 14(11):979–989, 2023.

Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

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

Hoang V. Tran, Thieu Vo, An Nguyen The, Tho Tran Huu, Minh-Khoi Nguyen-Nhat, Thanh Tran, Duy-Tung Pham, and Tan Minh Nguyen. Equivariant neural functional networks for transformers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=uBai0ukstY.

Jun Wu, Xiangzhe Kong, Ningguan Sun, Jing Wei, Sisi Shan, Fuli Feng, Feng Wu, Jian Peng, Linqi Zhang, Yang Liu, et al. Flowdesign: Improved design of antibody cdrs through flow matching and better prior distributions. *Cell Systems*, 16(6), 2025.

Jason Yim, Andrew Campbell, Andrew YK Foong, Michael Gastegger, José Jiménez-Luna, Sarah Lewis, Victor Garcia Satorras, Bastiaan S Veeling, Regina Barzilay, Tommi Jaakkola, et al. Fast protein backbone generation with se (3) flow matching. *arXiv preprint arXiv:2310.05297*, 2023.

Cheng Zeng, Jirui Jin, Connor Ambrose, George Karypis, Mark Transtrum, Ellad B Tadmor, Richard G Hennig, Adrian Roitberg, Stefano Martiniani, and Mingjie Liu. Propmolflow: property-guided molecule generation with geometry-complete flow matching. *Nature Computational Science*, pp. 1–10, 2026.

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

Tian Zhu, Milong Ren, and Haicang Zhang. Antibody design using a score-based diffusion model guided by evolutionary, physical and geometric constraints. In *Forty-first International Conference on Machine Learning*, 2024.

Yibo Zhu, Xiumin Shi, Jingjuan Zhang, Weizhong Sun, and Lu Wang. Abegdiffuser: Antibody sequence-structure codesign with equivariant graph neural networks and diffusion models. *Journal of Chemical Theory and Computation*, 21(21):11307–11317, 2025.



(a) s472 in complex with HER2 (b) s472 loop, $K_D = 1.46$ nM (c) s343 loop, $K_D = 2.93$ nM

Figure 2: **Experimentally validated HER2 designs.** (a) The strongest design, s472, bound to HER2 domain IV: antibody F_v in cartoon, designed CDR-H3 in dark teal with side chains shown, antigen as a translucent surface with the epitope contacted by H3 in amber. (b,c) Designed CDR-H3 loops (teal) superposed on the native trastuzumab loop (grey) after Cα alignment of the heavy chain; sticks mark the two substituted positions, labelled by the substitution. The framework is essentially unchanged (whole-chain Cα RMSD 0.96 Å for s472, 1.41 Å for s343) while the designed loop deviates 2.10 Å and 3.49 Å respectively from the native loop in that frame, with a similar overall loop shape (1.66 and 2.20 Å self-fitted). Both substituted positions face the interface, with the F112Y ring occupying the same antigen pocket as the native phenylalanine in these models; the measured association rates are similar to the parent's (Table 4). These are modelled complexes, not experimentally determined structures of the designs.

A Notation

Table 5 summarises the symbols used throughout the paper. Evaluation-metric symbols are listed separately in Table 28 (Appendix O).

B Scope and Interpretation

Our primary result is improved sequence recovery with PLM-free RiMES under the supplied framework, loop-length, and native-epitope conditioning. Single-CDR-H3 design tests refinement of one loop, while multi-CDR design tests joint redesign of all six loops. Sequence clustering controls overlap in the target CDR-H3 sequence; it does not establish independence of the other loops, frameworks, antigen families, or structural folds. Recovery is lower on the chronological benchmark than on RAbD for our model, but the direction of this change differs across baselines, so the two test sets should not be treated as a uniform scale of difficulty.

Scope of the architectural evidence. The component ablations support the coupled configuration we evaluate. The Euclidean-flow alternative retains flow training, and the experiments do not isolate the effect of flow training itself or the stability effect of detaching targets. Detached velocity targets remove one gradient path, while shared features and frame consistency still couple the encoder and flow head. Our equivariance result concerns the velocity head under its stated assumptions, rather than the entire generation pipeline.

Structural and physical interpretation. Structural rankings remain mixed. Whole-antibody scores also reflect whether a method regenerates the framework or retains native coordinates; CDR-self-fitted RMSD instead measures local loop shape and removes rigid-body placement error. Local clash and torsion improvements do not establish correct antigen-relative placement or complete all-atom packing (App. H). The unnormalized rotamer penalty has an amino-acid-dependent floor, $1 - \max_\chi \mathrm{compat}(\chi, a)$, and zero contributions for Gly/Ala. Its effects on geometry and residue composition are therefore not separated by the present ablation.

Table 5: Summary of notation.

Symbol	Description
Graph and residue representation	
$\mathcal{G} = (\mathcal{V}, \mathcal{E})$	Antibody–antigen graph; k-NN edges on Cα ($k = 9$)
$\mathcal{M} \subset \mathcal{V}$, $\mathcal{V}_{\text{ag}}$, $\mathcal{E}_{\mathcal{M}}$	CDR design set, antigen residues, CDR–CDR edges
$s_i \in \{1, \dots, 20\}$	Amino acid type of residue i (plus special tokens BOA/BOH/BOL/mask/pad)
$\mathbf{X}_i \in \mathbb{R}^{C\times 3}$, $\mathbf{x}_i^c$, $\mathbf{w}_i$	All-atom coordinates ($C = 14$), per-channel position, atom-importance weights
SE(3) geometry	
$\mathrm{SO}(3)$, $\mathrm{SE}(3)$, $\mathfrak{so}(3)$	Rotation group, rigid-motion group, and Lie algebra of $\mathrm{SO}(3)$
$\widehat{\cdot}$, exp, log	Hat map and exponential/logarithmic maps on $\mathrm{SO}(3)$ (Rodrigues' formula)
$(R_i, \mathbf{t}_i)$	Local rigid frame: R_i via Gram–Schmidt on N–Cα–C, $\mathbf{t}_i = \mathrm{C}\alpha_i$
$R(s)$	Geodesic interpolation on $\mathrm{SO}(3)$, $s \in [0, 1]$
$g = (R_g, \mathbf{t}_g) \in \mathrm{SE}(3)$	Global transform acting as $(R_i, \mathbf{t}_i) \mapsto (R_g R_i,\ R_g \mathbf{t}_i + \mathbf{t}_g)$; the velocity head obeys the conditional transformation law (App. N)
Equivariant adapter	
$\mathbf{z}_i \in \mathbb{R}^{d_{\text{plm}}}$, d	PLM embedding ($d_{\text{plm}} = 2560$, ESM-2 3B); Hidden dimension $d = 64$
W_0, W_{down}, W_{up}	LoRA factors, rank $r = 16$; W_{up} zero-initialized
$\mathbf{g}_i \in \mathbb{R}^6$, $\mathbf{h}_i^{\text{plm}}$, $\mathbf{h}_i^{\text{geom}}$	SE(3)-invariant geometric features and the two adapter streams
$\boldsymbol{\sigma}_i$, $\mathbf{h}_i \in \mathbb{R}^d$	Per-residue gate; fused node feature shared across all modules
$\alpha_{ij}^{(h)}$, $\beta_{\text{dist}}^{(h)}$, $\beta_{\text{frame}}^{(h)}$	Geometry-biased attention logit; learnable distance/frame bias scalars per head ($n_h = 4$, $d_h = 16$)
Equivariant GNN and refinement	
L, L_v, T	EGNN layers ($L = 3$), velocity MP layers ($L_v = 2$), refinement rounds ($T = 3$)
$\mathbf{A}_i$, $\tilde{\mathbf{r}}_{ij}$, $\mathbf{r}_{ij}$	Learned atom attributes ($d_{\text{attr}} = 16$); raw and projected edge radial features
$\mathbf{m}_{ij}, \mathcal{N}(i)$	Edge message; k-NN neighborhood of node i
$\hat{\mathbf{s}}_i$, $\hat{\mathbf{X}}_i$	Predicted sequence logits and all-atom coordinates of residue i
Riemannian flow matching	
$s \in [0, 1]$, μ, σ_z, $w(s)$	Flow time; logit-normal schedule ($\mu = 0, \sigma_z = 1$); self-normalized importance weight
$m_g \sim \mathrm{Bern}(0.5)$, $\mathrm{sg}[\cdot]$	Per-graph mode indicator (prediction vs. noise); stop-gradient operator
$(R_i^0, \mathbf{t}_i^0)$, $(R_i^{\text{gt}}, \mathbf{t}_i^{\text{gt}})$	Starting frame (mixed distribution) and ground-truth frame for residue i
σ_r, σ_p, σ_t	Pred-mode rotation/translation perturbations (0.1 rad, 0.5 Å); noise-mode translation (2.0 Å)
$\boldsymbol{\omega}_i^*$, $\mathbf{v}_i^*$ / $\hat{\boldsymbol{\omega}}_i$, $\hat{\mathbf{v}}_i$, $\tilde{\mathbf{v}}_i$	Target / predicted angular and (ambient) translational velocities at time s; for the invariant velocity head, $\tilde{\mathbf{v}}_i$ is the local translation velocity, $\hat{\mathbf{v}}_i = R_i \tilde{\mathbf{v}}_i$ (App. N)
$\mathbf{e}_{ij}$	Invariant edge features of the velocity head $\big(\lVert\mathbf{t}_j - \mathbf{t}_i\rVert,\ R_i^\top(\mathbf{t}_j - \mathbf{t}_i),\ \mathrm{vec}(R_i^\top R_j)\big)$
α, K, Δs	Flow-iter step size ($\alpha = 0.05$); inference midpoint steps ($\Delta s = 1/K$)
Losses, weights, and rotamer geometry	
$\mathcal{L}_{\text{seq}}^{(r)}$, $\mathcal{L}_{\text{struct}}^{(r)}$, $\mathcal{L}_{\text{dock}}^{(r)}$	Per-round sequence cross-entropy, Kabsch-aligned coordinate, and docking losses
$\mathcal{L}_{\text{flow}}$, $\mathcal{L}_{\text{consist}}$	Velocity Huber loss (detached starts); pairwise frame-alignment consistency loss
$\mathcal{L}_{\text{rotamer}}$, $\mathcal{L}_{\text{clash}}$, ℓ_δ	Rotamer-aware joint loss; soft Cα–Cα clash penalty; smooth L_1 (Huber) loss
$\lambda_f, \lambda_c, \lambda_\omega, \lambda_v$, $\lambda_{\text{rot}}, \lambda_{\text{clash}}, \lambda_{\text{inter}}$	Loss weights: flow ($\lambda_f = 0.1$), consistency ($\lambda_c = 0.05$), rotational/translational velocity ($\lambda_\omega = \lambda_v = 1.0$), rotamer ($\lambda_{\text{rot}} = 0.1$), clash ($\lambda_{\text{clash}} = 0.5$, $\lambda_{\text{inter}} = 2.0$); $d_{\text{thresh}} = 3.0$ Å (Cα distance threshold)
χ_1, $\chi_a^{(r)}, p_a^{(r)}$, σ_χ	First sidechain dihedral; Dunbrack rotamer angle/probability; rotamer kernel bandwidth (30°)
$\mathrm{compat}(\chi, a)$, $p_i^{(a)}$	Von Mises rotamer-compatibility kernel; softmax probability of type a at position i
ρ_u	Sequence-curriculum reveal fraction (cosine schedule from 0.9 to 0); training only

Our evaluation is computational and single-target. Relaxed interface-energy scores, sequence-model likelihoods, and amino-acid frequency statistics do not measure affinity, specificity, or developability.

Completion-corrected energy fractions combine the score criterion with processing coverage and should be read alongside that coverage. The comparisons do not isolate template conditioning or output format as the cause of an energy gap.

C Supporting Measurements and Their Scope

This appendix collects measurements that are reported separately elsewhere in the paper, together with the scope of what each one supports. It is intended as a reading aid rather than an additional claim: the statements in the main text remain the operative ones, and where a measurement admits a narrower reading we give it.

The computational measurements (Apps. C.1–C.4, C.7–C.10) are separate from the experimental measurements in App. C.6 and can be read independently of them.

C.1 Per-loop breakdown of the multi-CDR comparison

The headline comparison is CDR-H3, the hardest and least templated loop. The per-loop breakdown in App. E shows the margin is not confined to it: in the multi-CDR setting PLM-free RiMES attains the best or second-best recovery on *all six* loops (H1 0.923, H2 0.877, H3 0.538, L1 0.902, L2 0.987, L3 0.723), and exceeds every baseline on every loop. The aggregate 0.771 versus 0.648 for the strongest baseline is accompanied by a positive per-loop difference at each position, rather than by a single loop dominating the average.

The short, near-germline loops are easier, so the all-CDR average should not be read as evidence about H3 specifically; the per-loop values are the appropriate comparison and are reported in full.

C.2 Dispersion of the physical metrics across targets

Because designs are produced one target at a time, the spread of a metric across targets is informative alongside its mean. On the CDR-H3 setting the physical criteria show a smaller across-target standard deviation for RiMES than for the baselines (Table 2):

Metric (CDR-H3)	RiMES (std across targets)	Baseline range
Ramachandran allowed	0.625 (± 0.052)	0.251–0.489 (± 0.100–0.161)
$\Delta\Delta G < 0$	0.658 (± 0.122)	0.433–0.680 (± 0.271–0.409)
Intra-CDR clashes	0.000 (± 0.000)	0.036–3.310 (± 0.202–3.455)

Figure 3 shows this directly. The torsion spread is $1.9\times$ tighter than the tightest baseline and the interface-energy spread $2.2\times$ tighter, while the clash count has zero variance because no design in the set clashes. Ramachandran and clash counts are measured on raw coordinates and $\Delta\Delta G$ on relaxed structures, as in the rest of the paper. Within this test set the physical quality therefore varies less from target to target than for the methods compared; the observation is limited to these targets and metrics. Sequence recovery retains a target-dependent spread (± 0.139) comparable to the baselines, so the observation concerns the physical criteria only.

C.3 Effect size of the torsion-supervision comparison

Extending torsion supervision from χ_1 to χ_1–χ_4 improves every CDR-H3 all-atom clash metric by 35–48%, with paired Wilcoxon p-values from 8×10^{-32} to 1×10^{-54} over 749 paired per-target means on the chronological split (Table 20). The comparison changes only the loss, holding model, data, initialization and seed fixed, on a held-out chronological set. The same table reports the metric that does not improve, whole-antibody antigen clashes ($p = 0.19$), which limits the observation to the designed loop rather than to global packing. A p-value here reflects consistency across targets rather than the practical size of the effect.

C.4 Seed variability relative to the reported margin

Across four training seeds under an identical procedure, PLM-free CDR-H3 recovery spans 0.533–0.541 (mean 0.537, coefficient of variation 0.59%), with a Friedman test giving $p = 0.49$: no

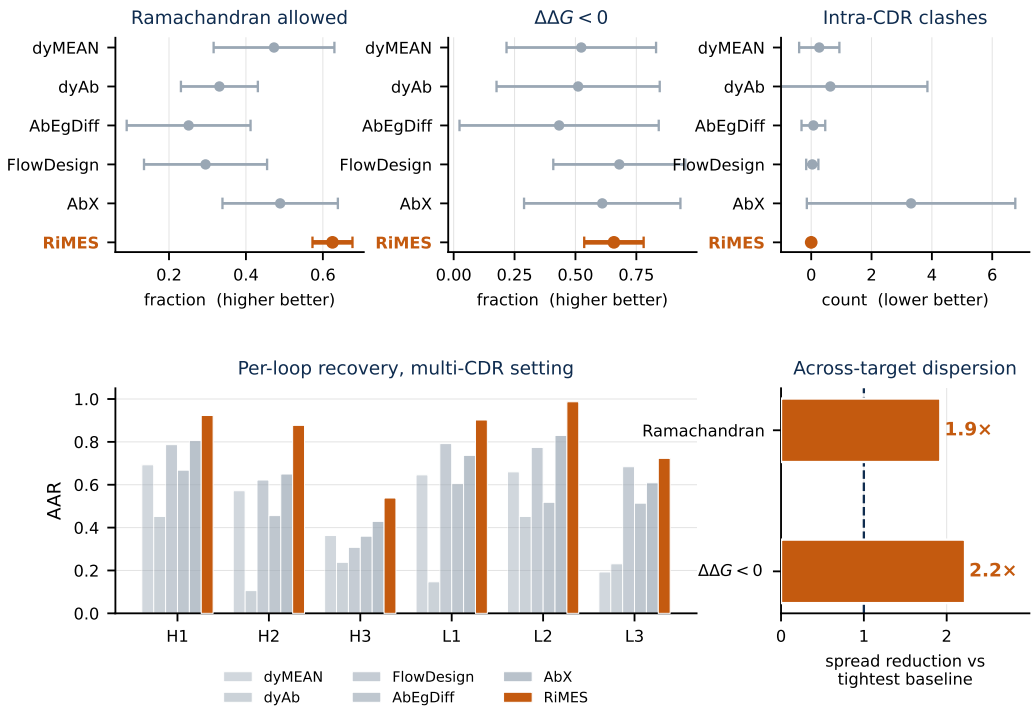


Figure 3: **Consistency and per-loop coverage.** Top: mean ± one standard deviation *across targets* for the three physical criteria on the RAbD CDR-H3 setting (Tables 1, 2). RiMES (orange) attains the best or second-best central value, with a shorter interval than the baselines shown; the clash interval reduces to a point because no design in this set clashes. Bottom left: per-loop recovery in the multi-CDR setting, where PLM-free RiMES is above the baselines at each of the six positions. Bottom right: ratio of the tightest baseline standard deviation to that of RiMES; values above the dashed line indicate a tighter spread. Dispersion is reported across targets, not across seeds; seed variability is separate (App. I). Ramachandran and clash counts use raw generated coordinates; $\Delta\Delta G$ uses relaxed structures (App. O).

detectable seed effect on the headline metric (App. I). Aligned RMSD_{H3} varies by 0.037 Å (CV 2.1%) and DockQ by 0.003 (CV 1.0%). For reference, the reported margin over the strongest PLM-free baseline is large relative to this seed-to-seed standard deviation. The four seeds share one training procedure, so this characterises run-to-run variation under that procedure rather than uncertainty over design choices.

C.5 Reading the component study as a group

The component study in Table 14, summarised in Fig. 4, reports each component separately. Grouped, six of the seven ablated components belong to the flow formulation, and these account for the larger effects in the table; we summarise what each comparison does and does not separate.

The Euclidean comparison varies the manifold with flow training held fixed. The Euclidean-flow variant *retains flow training in full* and changes only the space on which the velocity is defined. It is therefore a control on the $\mathrm{SE}(3)^N$ formulation itself rather than on the use of flow matching. Moving off the manifold costs 0.030 AAR in the PLM-free setting and 0.067 with the adapter, lowers the Ramachandran-allowed fraction by 0.124 and 0.082, and breaks clash-freedom in both ($0.000 \rightarrow 0.025$ and 0.027). It also produces the largest RMSD_{H3} degradation of any ablation we ran, $+0.122$ Å and $+0.106$ Å. Because this comparison alters the geometry of the flow and little else, it is the closest available separation of the manifold choice from the use of flow matching, and the differences have the same sign in both blocks. The two blocks share data and evaluation protocol, so this is agreement across training configurations rather than independent replication.

Both endpoints of the starting distribution are worse than the mixture. Prediction-only starts (p_{noise}=0) are 0.093 AAR below the full model and noise-only starts (p_{noise}=1) are 0.122 below, the two largest sequence differences in the PLM-free study. A possible reading is that prediction-only starts supply little trajectory for the velocity head to regress against, while noise-only starts train it on a distribution not encountered at inference; we have not tested this mechanism directly, and only the two endpoints and the chosen mixture were run.

Removing inference-time flow feedback has the largest effect on the physical metrics. The Ramachandran fraction falls by 0.174 and the intra-CDR clash count rises from 0.000 to 0.053, the largest clash value in the study. This suggests the integration step at inference contributes to the geometric metrics and not only to training, within this configuration.

Direction of the differences. Across the two blocks, all fifteen ablated configurations have lower AAR than their full model; fourteen of fifteen have a lower Ramachandran fraction and fourteen of fifteen have a non-zero clash count. Several differences are small relative to the across-target spread reported in Tables 1 and 2, and the study reports single runs per configuration, so the consistent sign is more informative than any individual magnitude.

What this does not establish. These controls compare configurations of a flow-based model; they do not include a no-flow baseline, so they bound the value of the manifold, the starting distribution, the inference-time feedback and the consistency term, but not the value of flow training as such. The detachment of velocity targets is likewise not isolated: shared features and the frame-consistency term still couple the encoder to the flow head after the gradient path is removed. The equivariance guarantee in App. N applies to the velocity head under its stated assumptions, not to the full generation pipeline.

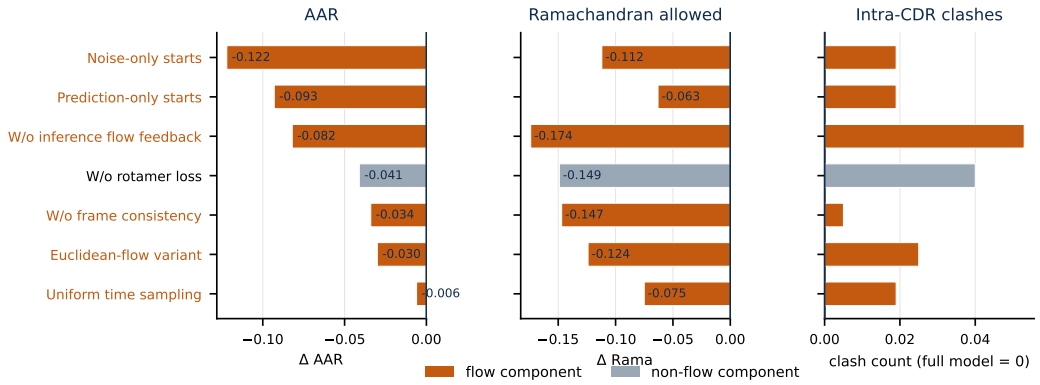


Figure 4: **Component effects in the PLM-free ablation** (Table 14), ordered by sequence-recovery loss. Orange marks components of the flow formulation, grey the one component that is not. Six of the seven ablated components belong to the flow design and account for the three largest recovery differences, the largest torsion difference and the largest clash value. The full model has zero intra-CDR clashes, so the right panel shows absolute counts. Each configuration is a single training run.

C.6 Experimental measurements on HER2

Ten CDR-H3 redesigns of trastuzumab, generated on the 1N8Z complex and selected by the interface proxy and developability filters, were expressed and assayed against HER2 domain IV. Native trastuzumab was produced and measured under the identical protocol as the positive control.

All ten expressed (4.7–5.7 mg/ml titre; 80–90% purity by SDS-PAGE, 90–94% by size-exclusion HPLC) and **all ten bound HER2 by BLI**. Table 6 reports the two strongest binders.

Three observations follow. First, the strongest design binds at $K_D = 1.46$ nM, within 1.5× of the parent, from a two-residue edit of the loop. Its association rate is close to the parent's (2.65 versus 2.55×10^5 $M^{-1}s^{-1}$) and the difference in affinity is carried mainly by a 1.5× faster off-rate, which

Table 6: **Measured affinity and developability of the two best HER2 designs**, with native trastuzumab under the same protocol. Both designs differ from the parent at two CDR-H3 positions only (bold). K_D, k_a, k_d by bio-layer interferometry; AC-SINS is a plasmon-resonance shift in nm, where larger values indicate self-association and > 10 nm is treated as high aggregation risk; T_{m1} is the first unfolding transition by differential scanning fluorimetry.

Design	CDR-H3	K_D (nM)	k_a (10^5/Ms)	k_d (10^{-4}/s)	AC-SINS	T_{m1} (°C)
s472	SRWGG**GG**YYAMDY	1.46	2.65	3.86	0.7	68.8
s343	SRWGG**YG**YYAMDY	2.93	3.11	9.11	4.0	68.6
Trastuzumab	SRWGGDGFYAMDY	0.98	2.55	2.50	−0.7	69.3

is consistent with a retained binding mode, though we have no structure of the design to confirm this. Second, nine of ten designs show no self-association liability and all ten melt within $1.1\,^{\circ}$C of the parent, so the substitutions did not measurably destabilise the domain in this panel. Third, no design exceeded native affinity. This is unsurprising given an affinity-matured parent and a model with no affinity objective, and the panel should be read as showing that variants in a narrow sequence neighbourhood expressed and bound, rather than as evidence of affinity maturation.

The interface-energy proxy filters but does not rank. Across the ten assayed variants the PyRosetta $\Delta\Delta G$ used for selection showed no rank correlation with measured affinity (Spearman $\rho = +0.05$, $n = 10$). For context, the best binder scored $\Delta\Delta G = +28.8$ REU and would have been excluded by a $\Delta\Delta G < 0$ filter, while the most favourable computed score (-106.9 REU) measured 14.68 nM. With ten points the correlation estimate is imprecise, so we draw only the narrow conclusion that this proxy did not order these candidates, and we use it as a coarse filter rather than a ranking signal. Because every assayed design bound, the observation concerns the scoring function rather than the generated candidates.

C.7 Metric structures are unrelaxed

The sequence, structural, clash and torsion metrics in this paper are computed on *raw generated coordinates*: no OpenMM relaxation or Rosetta side-chain repacking is applied on that path (App. O), so the reported values describe what the model emits rather than what a force field can rescue. The interface-energy analysis is the documented exception and uses separately relaxed structures, with the same distinction applied to the baseline that ships its own relaxed samples. Under that convention RiMES attains zero intra-CDR Cα clashes and the highest Ramachandran-allowed fraction among the methods compared. These values describe the generated coordinates directly; they do not indicate how any method would compare after a relaxation or repacking step.

C.8 Joint behaviour of recovery and geometry

Higher sequence recovery frequently trades against structural plausibility. In the per-target diagnostics (App. G.5) the AAR $\times$ RMSD$_{H3}$ distribution forms a single negatively-sloped cluster, without a high-AAR/high-RMSD population in these targets. At the aggregate level the method lies on the reported recovery versus interface-energy frontier. Both are descriptions of the present evaluation and do not exclude such a trade-off on other targets.

C.9 Where parent-conditioned variation is concentrated

In the parent-conditioned setting the adapter does not merely reproduce the parent, nor does it rewrite the loop. Position-wise, it holds the parent's flanking scaffold fixed and concentrates its variation on the apex positions that contact the antigen. This pattern is compatible with a hit-to-lead setting, where the validated binding mode is retained and the contact residues varied. The PLM-free model under the same conditions also varies the scaffold, as expected when no parent is assumed. The two regimes therefore differ in what they hold fixed as well as in their recovery scores.

C.10 TRANSLATION BEHAVIOUR OF THE VELOCITY HEAD

The equivariant velocity head consumes only relative frame displacements, so a rigid translation of the input leaves its output unchanged exactly rather than approximately. The check in App. N reports invariance to machine precision across shift magnitudes spanning three orders of magnitude, independent of the learned weights. The property follows from the parameterisation; it applies to the velocity head under the stated assumptions and not to the generation pipeline as a whole.

D PAIRED TARGET-LEVEL COMPARISONS

We report a separate analysis using records that can be paired by target. Positive intervals describe an advantage on those matched records; they neither identify its mechanism nor extend to unscored outputs. The RiMES* comparisons additionally include parent-sequence information, as described below.

Table 7: **Separate paired-record analysis on RAbD H3.** Target-level AAR differences use $10{,}000$ paired bootstrap resamples over shared targets. Baseline sequences are recovered from design structures. FlowDesign contributes $57/60$ targets and $2{,}280/2{,}400$ designs; AbEgDiffuser is not included. These intervals describe the matched records used here. Their exact correspondence to the aggregate evaluation in Table 1 remains unresolved, so they are not significance intervals for the main-table margins.

Baseline	RiMES* ΔAAR [95% CI]	PLM-free RiMES ΔAAR [95% CI]
dyMEAN	$+0.295$ $[+0.263, +0.328]$	$+0.100$ $[+0.074, +0.126]$
dyAb †	$+0.391$ $[+0.357, +0.425]$	$+0.195$ $[+0.163, +0.228]$
AbX ‡ † ⋆	$+0.462$ $[+0.424, +0.499]$	$+0.266$ $[+0.230, +0.301]$
FlowDesign ‡†	$+0.287$ $[+0.252, +0.322]$	$+0.096$ $[+0.065, +0.126]$

The evaluation mapping between this analysis and the main aggregate scores is unresolved. FlowDesign's main AAR of 0.325 gives a PLM-free margin of $0.535 - 0.325 = 0.210$, whereas the paired margin is 0.096. The completed-target summary reports AAR approximately 0.44. Under common scoring and equal target weights, zero-filling the three missing targets would give $(57/60) \times 0.44 = 0.418$, rather than 0.325. Completion alone therefore does not explain the difference. The correspondence of generated, readable, and scored designs, residue mapping, and aggregation remains unspecified. We retain these as separate reported summaries and do not use the paired interval to establish significance of the main-table margin.

E PARENT-CONDITIONED REDESIGN AND CONTROLS

RiMES* is a secondary, parent-conditioned setting. It uses the sequence of the native reference antibody in a supplied complex, so recovery measures proximity to that parent. Although designed-position embeddings are zeroed before the adapter, ESM-2's bidirectional framework embeddings retain native-CDR context. We distinguish this setting from our primary PLM-free comparison throughout.

E.1 INPUT-MASKED CONTROL

We retrain with designed CDR tokens hidden before ESM-2 encoding to test dependence on native input context. H3 AAR falls from 0.729 to 0.424, below the PLM-free value of 0.535; multi-CDR AAR falls from 0.830 to 0.601, below 0.771. On the chronological split, masked-input AAR is 0.410, compared with PLM-free 0.479 (Table 8). In the multi-CDR control, the available loop-level values also fall below the PLM-free results: H3 0.37, L3 0.55, H1 0.78, L1 0.73, and L2 0.79. These findings identify native input context as central to the parent-conditioned recovery gain. They do not establish membership in, or independence from, the PLM pretraining corpus.

Separate paired contrast between our variants. With $20{,}000$ target-level bootstrap replicates, the paired parent-conditioned minus PLM-free differences are AAR $+0.196$ (95% CI $[+0.164, +0.228]$),

Table 8: **Input-masked retraining and parent conditioning.** RiMES is PLM-free; RiMES* uses parent-sequence ESM-2 context. The masked-input control hides designed CDR tokens before ESM-2 encoding. Within each setting, the two PLM runs differ in the sequence supplied to ESM-2; the PLM-free run removes the adapter. Each row is a single run. Multi-CDR AAR is pooled over redesigned positions, and RMSD averages independent loop fits; the other settings use CDR-H3. Rama is the same coarse torsion-region score used in the main evaluation. Masked-input AAR is below PLM-free AAR in all three settings.

Experiment	Method	AAR ↑	CAAR ↑	RMSD ↓	BLOSUM62 ↑	Rama ↑
CDR-H3	Masked-input	0.424	0.262	1.807	0.545	0.486
CDR-H3	RiMES* (PLM)	**0.729**	**0.591**	1.773	**0.802**	0.567
CDR-H3	RiMES (PLM-free)	0.535	0.334	**1.762**	0.630	**0.625**
Multi-CDR	Masked-input	0.601	0.508	0.915	0.708	0.396
Multi-CDR	RiMES* (PLM)	**0.830**	**0.757**	**0.879**	**0.880**	**0.551**
Multi-CDR	RiMES (PLM-free)	0.771	0.687	0.885	0.838	0.463
Time Split	Masked-input	0.410	0.333	**2.367**	0.546	0.396
Time Split	RiMES* (PLM)	**0.624**	**0.480**	2.403	**0.721**	0.486
Time Split	RiMES (PLM-free)	0.479	0.385	2.407	0.648	**0.640**

CAAR $+0.254$ ($[+0.204, +0.305]$), and BLOSUM62 $+0.169$ ($[+0.144, +0.194]$). This analysis changes both information access and adapter use. Its AAR difference is not the aggregate difference $+0.194$ in Table 8; the precise correspondence of the evaluation records and aggregation remains unresolved. We retain the interval as a separate descriptive comparison, without attributing the difference to an isolated adapter mechanism.

E.2 Per-CDR breakdown

Table 9 retains the loop-level results and their supplied "All" summary. Within the RiMES columns, parent conditioning improves most loop scores but slightly reduces L2 recovery from 0.987 to 0.981.

Table 9: **Per-CDR recovery in the multi-CDR setting** on RAbD ($N = 60$ targets, 40 samples each). Bold and underline mark the highest and second-highest entries within each row of this analysis. We retain the "All" row as supplied. Several baseline values in that row differ from the main multi-CDR results, and the aggregation linking these tables is unresolved; see the accompanying text. This table is not an independent validation of the main aggregate comparisons.

CDR	dyMEAN	dyAb †	FlowDesign ‡†	AbEgDiff ‡⋆	AbX ‡ † ⋆	**RiMES**	**RiMES ⋆**
H1	0.693	0.452	0.787	0.668	0.806	<u>0.923</u>	**0.944**
H2	0.572	0.106	0.622	0.456	0.650	<u>0.877</u>	**0.903**
H3	0.363	0.238	0.308	0.360	0.429	<u>0.538</u>	**0.651**
L1	0.646	0.147	0.792	0.605	0.737	<u>0.902</u>	**0.938**
L2	0.660	0.452	0.774	0.518	0.830	**0.987**	<u>0.981</u>
L3	0.193	0.231	0.684	0.514	0.609	<u>0.723</u>	**0.813**
All	0.478	0.251	0.648	0.516	0.632	<u>0.771</u>	**0.830**

Aggregation correspondence. The baseline "All" entries for dyMEAN, dyAb, FlowDesign, and AbEgDiffuser differ from their main-table multi-CDR AAR values: respectively, 0.478 versus 0.521, 0.251 versus 0.271, 0.648 versus 0.661, and 0.516 versus 0.520. For these four baselines, the main values coincide, after rounding, with the unweighted means of the six displayed loop scores. By contrast, the RiMES columns and AbX agree with their "All" summaries rather than those unweighted means. This arithmetic identifies an unresolved aggregation correspondence; it does not establish which aggregation was applied to the underlying records, validate either set of main scores, or provide a basis for replacing them.

E.3 SEQUENCE-QUALITY AND LOCAL-GEOMETRY DIAGNOSTICS

We evaluate parent-conditioned candidates with sequence-model scores, composition statistics, parent proximity, and local geometry. Table 10 summarizes the sequence-quality comparisons, including the less favorable AbLang2 result. Table 11 gives complementary composition, geometry, and diversity measurements. These are computational diagnostics of the candidate pool, without a matched parent-conditioned redesign baseline. Experimental results for ten designs from this regime are reported in App. C.6. Best-of-40 summaries select candidates independently for each metric; they do not identify a single candidate that improves all criteria.

Table 10: **Sequence-quality proxies for RiMES*** on RAbD CDR-H3 (60 targets, 40 designs each). Upper rows are reference-free; "% designs > parent" compares individual designs with the parent and "best-of-40" reports targets with at least one improvement at $\tau = 1.0$. Pseudo-PPL and autoregressive PPL are compared only within rows. H3 identity measures parent proximity, not de-novo recovery; the PLM-free reference is 0.535 AAR (Table 1). Best-of-40 is evaluated separately per metric and need not select the same candidate. Details in App. E.4.

Metric	Parent	RiMES*	% designs > parent	best-of-40
ESM-2 pseudo-PPL ↓ (general)	21.88	**19.42**	74.2%	95.0%
ProGen2 PPL ↓ (general)	37.44	**35.45**	65.4%	95.0%
IgLM PPL ↓ (antibody)	1.525	**1.519**	56.4%	98.3%
AbLang2 pseudo-PPL ↓ (antibody)	**38.66**	41.84	41.8%	88.3%
Human-repertoire AA frequency proxy ↑	0.0626	**0.0699**	83.8%	98.3%
Parent proximity (reference-based)				
H3 identity to parent (AAR) ↑	1	0.729	—	—

Table 11: **Complementary diagnostics for parent-conditioned RiMES*** on RAbD CDR-H3 (60 targets, 40 designs per target). Parent denotes the native reference antibody. Charge and hydrophobicity follow App. O.6; BLOSUM62 measures parent-sequence similarity, and RMSD uses an independent CDR fit. Cα clashes and Rama are local geometry proxies. Unique-design counts come from the parent-conditioned temperature sweep, separately from the PLM-free sweep in Table 18. Naturalness, human-frequency, identity, and per-metric best-of-40 results are collected in Table 10.

Axis	Metric	Parent	RiMES*	Interpretation
Composition	CDR-H3 net charge (pH 7)	−0.45	−0.49	reference scale $\sigma = 1.11$
	CDR-H3 hydrophobicity	−0.73	−0.77	reference scale $\sigma = 0.53$
Parent proximity	BLOSUM62 ↑	—	0.802	substitution similarity
	H3 backbone RMSD to parent (Å) ↓	—	1.77	CDR-self-fitted score
Local geometry	Intra-CDR Cα clashes ↓	—	**0.000**	Cα criterion
	Coarse Rama ↑	—	0.567	coarse torsion proxy
Sampling diversity	Unique designs / 40 ($\tau = 0.5$ / 1.0) ↑	—	24.3 / 33.4	temperature-dependent pool

E.4 SEQUENCE SCORES AND SAMPLING TEMPERATURE

We evaluate parent-conditioned variants from 60 RAbD targets with 40 designs per target under the default configuration and temperatures $T \in \{0.1, 0.3, 0.5, 1.0\}$. We score native and generated sequences with the same routines: ESM-2-650M Lin et al. (2023) and AbLang2 pseudo-perplexity, ProGen2-small and IgLM autoregressive perplexity, a human-repertoire amino-acid frequency proxy, net charge at pH 7, and mean Kyte–Doolittle hydrophobicity. Perplexities are compared within each scoring model. The frequency proxy uses marginal amino-acid frequencies and does not establish clinical humanness.

Mean scores. Table 12 reports the full temperature sweep. ESM-2 and ProGen2 assign lower mean perplexity to generated variants, whereas AbLang2 assigns lower mean perplexity to the parent (38.66 versus 41.84 under the default configuration). IgLM means are close (1.525 for the parent versus

Table 12: Reference-free designability of RiMES* versus the natural antibody on RAbD (n=60 targets, 40 samples / target). ↓ / ↑ indicates the better direction. Bold marks the best column per row. Pseudo-PPL (masked PLMs) and true PPL (autoregressive PLMs) are not directly comparable across rows; each row compares GT against RiMES* in isolation.

Metric	GT	Default	T=0.1	T=0.3	T=0.5	T=1.0
Sequence naturalness (perplexity, ↓)						
ESM-2 (general, masked)	21.88	**19.42**	19.53	20.07	20.56	21.49
ProGen2 (general, autoregressive)	37.44	**35.45**	35.50	35.91	36.25	37.07
AbLang2 (antibody, masked)	**38.66**	41.84	41.68	41.72	42.01	43.52
IgLM (antibody, autoregressive)	1.525	1.519	**1.518**	1.519	1.520	1.522
Other designability						
Human-repertoire AA frequency (↑)	0.0626	**0.0699**	0.0697	0.0688	0.0676	0.0640
CDR-H3 charge	−0.45	−0.49	−0.48	−0.48	−0.50	−0.52
CDR-H3 hydrophobicity	−0.73	−0.77	−0.78	−0.80	−0.80	−0.78

Table 13: Fraction of designs that beat the natural antibody on the target's own GT score under four PLMs and the human-repertoire amino-acid frequency proxy. Top: averaged over all 40 designs per target. Bottom: fraction of targets where the *best* of the 40 designs beats GT. Default leads the ESM-2, ProGen2, and frequency-proxy per-design comparisons; the best per-design setting varies by scorer. At T=1.0, best-of-40 success is highest for each of the four PLMs, with selection performed separately per metric.

Metric	Default	T=0.1	T=0.3	T=0.5	T=1.0
Per-design success: fraction of designs better than GT					
ESM-2 PPL (↓)	**0.742**	0.719	0.675	0.643	0.535
ProGen2 PPL (↓)	**0.654**	0.642	0.608	0.579	0.509
AbLang2 PPL (↓)	0.418	0.418	**0.425**	0.410	0.382
IgLM PPL (↓)	0.564	**0.574**	0.554	0.523	0.485
Human-repertoire AA frequency (↑)	**0.838**	0.832	0.797	0.742	0.576
Best-of-40 success: fraction of targets where any design beats GT					
ESM-2 PPL (↓)	0.750	0.767	0.800	0.867	**0.950**
ProGen2 PPL (↓)	0.700	0.717	0.767	0.817	**0.950**
AbLang2 PPL (↓)	0.517	0.617	0.700	0.717	**0.883**
IgLM PPL (↓)	0.633	0.750	0.850	0.917	**0.983**
Human-repertoire AA frequency (↑)	0.867	0.883	0.933	0.950	**0.983**

1.518–1.522 for variants). The amino-acid frequency proxy increases from 0.0626 to 0.0699 in the default setting. Mean charge and hydrophobicity shifts remain within one native standard deviation (1.11 and 0.53, respectively). These are computational score comparisons, not measurements of biological improvement.

Per-design and per-metric selection. Table 13 reports the fraction of individual designs beating their target's parent and, separately, the fraction of targets with at least one better-scoring candidate. Under the default configuration, these per-design fractions are 74.2% for ESM-2, 65.4% for ProGen2, 56.4% for IgLM, 41.8% for AbLang2, and 83.8% for the frequency proxy. At $T = 1.0$, best-of-40 success is 95.0%, 95.0%, 98.3%, 88.3%, and 98.3%, respectively. Selection is performed independently for each metric: these fractions do not imply that one candidate improves all scores or retains an acceptable interface.

F Ablation Study

To assess the components of the integrated model, we conduct a systematic ablation study on the RAbD benchmark for CDR-H3 design under both training regimes (PLM-free and PLM-augmented; each ablation is run in the regimes where the component is applicable). We evaluate leave-one-out variants of (1) **geometric features** within the adapter; (2) the **rotamer loss**; (3) **Riemannian flow**

Table 14: **Ablation study on RAbD CDR-H3 design** (60 targets, 40 samples each). The upper block is the PLM-free regime and the lower block the PLM-augmented regime (RiMES*); each block is headed by its own reference model. **Bold** marks that reference model, not the column-wise best: a few single columns improve when a component is removed (BLOSUM62 without the rotamer loss, RMSD_{H3} for several variants, Rama under the pred-only start), and these exceptions are discussed in the text. RMSD_{H3}: Cα RMSD of CDR-H3 after alignment (Å); $\text{Clashes}_{\text{H3}}$: intra-CDR C$\alpha$ steric clashes; Rama: coarse torsion-region fraction.

Model	AAR↑	CAAR↑	BLOSUM62↑	RMSD_{H3} ↓	DockQ↑	$\text{Clashes}_{\text{H3}}$ ↓	Rama↑
RiMES	**0.535**	**0.334**	**0.630**	**1.762**	**0.366**	**0.000**	**0.625**
Uniform time sampling	0.529	0.314	0.625	1.760	0.356	0.019	0.550
Euclidean-flow variant	0.505	0.316	0.607	1.884	0.353	0.025	0.501
Without inference-time flow feedback	0.453	0.307	0.557	1.844	0.331	0.053	0.451
Without rotamer loss	0.494	0.309	0.657	1.731	0.327	0.040	0.476
Prediction-only starts ($p_{\text{noise}} = 0$)	0.442	0.220	0.603	1.810	0.345	0.019	0.562
Noise-only starts ($p_{\text{noise}} = 1$)	0.413	0.309	0.598	1.853	0.349	0.019	0.513
Without frame consistency	0.501	0.316	0.587	1.825	0.347	0.005	0.478
RiMES* (Full model)	**0.729**	**0.591**	**0.802**	**1.773**	**0.350**	**0.000**	**0.567**
w/o Geometric feat.	0.680	0.530	0.760	1.762	0.351	0.001	0.502
Uniform time sampling	0.661	0.504	0.743	1.804	0.348	0.012	0.392
Euclidean-flow variant	0.662	0.494	0.754	1.879	0.342	0.027	0.485
Without inference-time flow feedback	0.682	0.511	0.760	1.781	0.331	0.022	0.464
Without rotamer loss	0.682	0.510	0.772	1.774	0.351	0.005	0.463
Prediction-only starts ($p_{\text{noise}} = 0$)	0.654	0.474	0.747	1.768	0.331	0.037	0.570
Noise-only starts ($p_{\text{noise}} = 1$)	0.650	0.473	0.737	1.695	0.341	0.001	0.484
Without frame consistency	0.702	0.542	0.779	1.793	0.346	0.000	0.520

matching (reverting to Euclidean); (4) **flow-iterative refinement** at inference; and (5) **logit-normal time sampling** (reverting to uniform), and additionally compare (6) **pred-only flow** (p_{noise}=0, removing the Haar/Gaussian noise-mode start); (7) **random-only flow** (p_{noise}=1, removing the detached EGNN-prediction start); and (8) removal of the **frame consistency loss** $\mathcal{L}_{\text{consist}}$ (an additional direct frame-gradient pathway). The **equivariant adapter** is not a separate row: ablating it removes the PLM pathway entirely, which is precisely the PLM-free model, so its contribution is read as the contrast between the two blocks of Table 14 and, with PLM access held fixed, from the encoder ablation of Table 17. The stated training budget is up to 200 epochs with matched hyperparameters (Table 27) and evaluated on 60 test targets (40 samples each, 2,400 designs per ablation). Backbone-level sequence/structure/stereochemistry metrics are reported in Table 14; sidechain-level metrics (χ_1 MAE, rotamer accuracy, all-atom interface clashes) are reported separately in Table 19.

Interpretation of component ablations. The PLM-free block provides our primary evidence for the coupled architecture. RiMES leads its seven variants on AAR, CAAR, and whole-complex DockQ. This agreement across scores is useful because recovery at all CDR positions, recovery at contact positions, and complex-level structural agreement measure different aspects of co-design. The full model's DockQ is 0.366, versus 0.327–0.356 across these variants. This is a comparison within the component study, not evidence of uniformly accurate binding poses.

Mixed starts and intermediate feedback. The mixed-start configuration reaches AAR 0.535, compared with 0.442 for prediction-only and 0.413 for noise-only starts. The corresponding CAAR values are 0.334, 0.220, and 0.309. Removing inference feedback lowers AAR by 8.2 points and DockQ by 0.035. Together, these comparisons support both the evaluated start mixture and use of predicted velocities during refinement. They do not isolate target detachment as the cause of a training-stability improvement. Replacing the Riemannian objective with the Euclidean-flow variant also lowers AAR (0.535 to 0.505) and increases self-fitted RMSD (1.762 to 1.884 Å), with flow training retained in both conditions.

Consistent recovery directions across input regimes. All seven shared ablation changes lower reported AAR in both the PLM-free and parent-conditioned blocks. The decreases span 0.6–12.2 percentage points without PLM input and 2.7–7.9 points with parent context. Their relative sizes differ: removing inference feedback costs 8.2 versus 4.7 points, whereas uniform time sampling costs 0.6 versus 6.8 points. We therefore observe a consistent direction of recovery change across the

two regimes, without assuming a universal ranking of component importance. These are single-run component comparisons; their effects are not additive, and the seed study does not provide uncertainty estimates for individual ablations.

Local regularization and competing objectives. Removing the rotamer loss improves PLM-free self-fitted RMSD from 1.762 to 1.731 Å, but lowers AAR from 0.535 to 0.494, CAAR from 0.334 to 0.309, and DockQ from 0.366 to 0.327. Its benefit is therefore clearest in joint recovery and the reported local/complex diagnostics, rather than in optimizing every structural score. Frame consistency has a smaller recovery effect (0.535 to 0.501), and uniform time sampling changes AAR from 0.535 to 0.529. In the parent-conditioned sidechain diagnostic, retaining the rotamer term reduces χ_1 MAE by 12.5° and the mean 2.0 Å atom-pair clash count by approximately 53% (Table 19); native-angle recovery changes only from 0.444 to 0.447. These improvements remain local, and the loss's composition and geometric effects are not separated.

The between-regime contrast (0.729 versus 0.535) changes both parent information and adapter use. We use the within-regime ablations above to assess components and the masked-input control to assess dependence on parent context. The conditional head theorem, seed analysis, and component studies answer separate questions; none substitutes for a matched detachment or no-flow-training control.

G The Importance of Flow

We compare the reported sequence, structure, physical-geometry, and developability-proxy metrics of RiMES* with the configuration labeled RiMES* (no-flow). The reference values below belong to this comparison and are kept separate from other reported runs.

G.1 Interpretation

The largest difference is in sequence recovery. RiMES* achieves H3 AAR of 72.850%, compared with 58.980% for no-flow, a difference of 13.670 percentage points. Contact recovery falls from 59.070% to 39.760%, approximately 19.300 points, and BLOSUM62 decreases from 0.802 to 0.677. The recovery difference therefore extends to antigen-contacting positions. These scores measure agreement with the reference sequence; they do not measure binding affinity or specificity.

Aligned structural scores show much smaller differences. Independently aligned H3 RMSD changes from 1.773 to 1.778 Å, approximately 0.005 Å. Antibody-aligned Cα RMSD and TM-score also differ little at the reported precision. Thus, the sequence-recovery gap coexists with similar aggregate scores for fitted loop shape and antibody structure. In contrast, H3 RMSD without an additional fit favors no-flow (8.817 versus 9.110 Å). This distinction matters: independently fitting H3 removes its rigid-body placement error, so fitted loop similarity alone does not establish correct placement relative to the framework or antigen. Missing no-flow DockQ and lDDT prevent comparisons on those two metrics. Similar point estimates are not a statistical equivalence test.

Physical geometry shows trade-offs. RiMES* increases the coarse torsion-region fraction from 0.497 to 0.567, with fewer torsion outliers (4.635 versus 5.280) and fewer intra-CDR Cα clashes at the displayed precision (0.000 versus 0.003). No-flow has fewer interface clashes under both the Cα and heavy-atom criteria, higher bond-length validity, and fewer bond-length violations. In particular, the low bond-validity fractions in both runs (0.155 and 0.182) leave substantial local geometry errors. These results support a selective torsion and intra-CDR benefit for RiMES*, while retaining trade-offs in bond geometry and interface packing.

Sequence-quality proxies also expose trade-offs. No-flow has a higher human-repertoire amino-acid frequency proxy (0.080 versus 0.070), whereas RiMES* has fewer total PTM-liability counts and methionine oxidation-liability counts. The no-flow sequences are less negatively charged and have higher mean Kyte–Doolittle hydrophobicity. Neither shift has a universal favorable direction. The frequency score and motif counts are computational proxies; they do not establish clinical humanness, experimental stability, or developability.

Table 15: Reported comparison of RiMES* and RiMES (no-flow). Differences use Δ = no-flow − RiMES*. Arrows indicate the preferred direction of each score; charge and hydrophobicity are descriptive quantities without a universal preferred direction. Missing scores are shown as an em dash.

Metric	Direction	RiMES*	No-flow	Δ
Sequence				
AAR H3	↑	0.729	0.590	−0.137
CAAR H3	↑	0.591	0.398	−0.193
BLOSUM62	↑	0.802	0.677	−0.125
Structure				
H3 RMSD, independently aligned (Å)	↓	1.773	1.778	+0.005
H3 RMSD, without an additional fit (Å)	↓	9.110	8.817	−0.293
Cα RMSD, antibody-aligned (Å)	↓	1.083	1.085	+0.002
TM-score	↑	0.972	0.972	+0.000
Physical geometry				
Intra-CDR Cα clashes	↓	0.000	0.003	+0.003
Interface Cα clashes	↓	0.570	0.502	−0.068
Interface heavy-atom clashes	↓	8.891	8.730	−0.162
Coarse torsion-region fraction	↑	0.567	0.497	−0.070
Coarse torsion outlier count	↓	4.635	5.280	+0.645
Backbone bond-length validity	↑	0.155	0.182	+0.028
Backbone bond-length violations	↓	31.947	31.024	−0.923
Developability proxies and sequence descriptors				
Human-repertoire AA frequency proxy	↑	0.070	0.080	+0.010
Total PTM-liability count	↓	0.397	0.415	+0.018
Methionine oxidation-liability count	↓	0.183	0.251	+0.068
CDR net charge	—	−0.487	−0.219	+0.267
CDR hydrophobicity	—	−0.773	−0.664	+0.109

Numerical reporting. Values and supplied differences are rounded independently to three decimal places; differences are not recomputed from the displayed values. The original summary also contains last-digit discrepancies between some supplied differences and subtraction of its reported means. An em dash for DockQ or lDDT denotes an unavailable no-flow score (reported as NaN), not zero performance.

Scope of the comparison. The clearest advantage of RiMES* in these results is sequence and contact recovery, accompanied by improved coarse torsion statistics. Structural and physical-quality outcomes remain mixed. These aggregate values do not include uncertainty estimates. Attributing the differences specifically to flow requires the two runs to share parent-sequence access, training and evaluation protocols, and a documented definition of the disabled flow components. The no-flow label here identifies the reported comparison configuration without assuming which losses or refinement updates were removed.

G.2 PLM Embedding Ablation

To assess the contribution of protein language model (PLM) representation quality, we substitute the PLM backbone while keeping all other architectural components fixed: SE(3) equivariant adapter, Riemannian flow matching with logit-normal time sampling, rotamer-aware loss, and steric clash penalty. All models use 200 training epochs on the SAbDab training split and are evaluated on RAbD with 40 samples per target for CDR-H3 design.

PLM substitution protocol. For all general-protein PLMs (ESM-2 and ESMc variants), per-residue embeddings are precomputed independently for the antigen, heavy, and light chains and then concatenated in the order consumed by the adapter. AbLang2 Olsen et al. (2024) is antibody-specific (trained on paired OAS sequences) and does not embed non-antibody chains. We therefore use the *paired* AbLang2 model with concatenated input `<HC>|<LC>`, recover per-residue embeddings for

each chain by slicing out the encoder states between the corresponding boundary tokens, and replace the antigen embedding with a zero tensor of matching dimension (480). This compares a different antibody-side PLM and removes antigen embeddings while retaining the downstream tensor layout (and therefore the adapter, GNN, and flow modules) identical to all other rows of Table 16; only the embedding dimensionality differs across rows.

Table 16: **PLM embedding ablation** on RAbD CDR-H3 design. All architectural components are held constant; only the PLM embedding source varies. AAR / CAAR: (contact-weighted) amino acid recovery. BLOSUM62: normalised BLOSUM62 substitution score on CDR-H3. RMSD_{H3}: Cα RMSD of CDR-H3 after alignment (Å). $\text{Clashes}_{\text{H3}}$: intra-CDR C$\alpha$ steric clashes. Rama: coarse torsion-region fraction of backbone (ϕ, ψ) dihedral angles.

	Sequence Design			Structure (H3)			Stereochemistry
Model	AAR↑	CAAR↑	BLOSUM62↑	RMSD_{H3}↓	DockQ↑	$\text{Clashes}_{\text{H3}}$↓	Rama↑
ESM-2-3B (2560-d) (**RiMES***)	**0.729**	**0.591**	**0.802**	**1.773**	**0.350**	**0.00**	0.567
ESM-2-650M (1280-d)	0.634	0.474	0.729	1.864	0.328	0.07	0.572
ESMc-600M (1152-d)	0.623	0.461	0.716	1.791	0.338	0.00	0.588
ESMc-300M (960-d)	0.589	0.393	0.686	1.790	0.333	0.00	0.622
AbLang2-paired (480-d)	0.543	0.369	0.640	1.791	0.345	0.02	**0.729**

Table 16 compares the parent-context encoders we evaluate. Recovery decreases from 0.729 with ESM-2-3B to 0.589 with ESMc-300M, while structural metrics vary less. We interpret this as a comparison of encoder choices, rather than evidence for a particular co-evolutionary mechanism, because the comparison does not isolate family, capacity, representation, or native sequence context. AbLang2 also changes antigen embedding availability and input format, preventing a capacity-only or antigen-removal interpretation. Its higher coarse Rama score and lower recovery show a trade-off in our setup; they do not establish that the flow prior dominates or that one PLM architecture better matches the adapter.

G.3 Encoder Ablation: PLM Access Is Not Sufficient

Table 16 varies the PLM with the encoder fixed. We now run the converse: hold the PLM fixed at ESM-2-3B and vary the encoder. We construct a dyMEAN + ESM-2-3B variant that projects and concatenates the same ESM-2-3B residue embeddings used by RiMES into the AM-EGNN input features, leaving the rest of the dyMEAN training recipe unchanged, and compare it to RiMES under identical PLM access in Table 17.

Table 17: **Encoder ablation** on RAbD CDR-H3 design (PLM fixed at ESM-2-3B). The dyMEAN + ESM-2-3B row is *our own* encoder-ablation variant, not a competitor baseline; it exhibits unusually low bond-length validity ($\sim$0.11) for a full-structure generator and ESMFold scoring fails on its outputs, so its raw stereochemistry numbers should be read with caution.

	Sequence Design			Structure (H3)			Stereochemistry
Encoder (PLM = ESM-2-3B)	AAR↑	CAAR↑	BLOSUM62↑	RMSD_{H3}↓	DockQ↑	$\text{Clashes}_{\text{H3}}$↓	Rama↑
dyMEAN encoder (AM-EGNN)	0.405	0.288	0.577	**1.732**	0.325	0.580	**0.584**
RiMES *	**0.729**	**0.591**	**0.802**	1.773	**0.350**	**0.000**	0.567

Our dyMEAN variant with the same cached ESM features reaches AAR 0.405, compared with 0.431 for the PLM-free dyMEAN reference and 0.729 for parent-conditioned RiMES. Its low bond-length validity and failed structure-prediction scoring limit its usefulness as a competitive control. We cannot separate architecture, feature integration, optimization, and implementation as explanations for the shortfall. This experiment therefore leaves open whether our adapter is necessary and whether the recovery gain extends beyond native-parent context.

Table 18: Temperature ablation on RAbD test set (CDR-H3, 40 samples per target), for **PLM-free RiMES** (an independently retrained seed, so $\tau = 0$ here is that run's greedy decoding and differs slightly from the run reported in Table 1). Temperature scaling is applied to the sequence logits at generation time. The corresponding diversity figures for RiMES* are in Table 11.

Metric	$\tau = 0$	$\tau = 0.1$	$\tau = 0.3$	$\tau = 0.5$	$\tau = 1.0$
Sequence Quality					
AAR H3	$0.536_{\pm 0.140}$	$0.537_{\pm 0.139}$	$0.525_{\pm 0.139}$	$0.504_{\pm 0.138}$	$0.460_{\pm 0.129}$
CAAR H3	$0.334_{\pm 0.210}$	$0.336_{\pm 0.209}$	$0.316_{\pm 0.211}$	$0.287_{\pm 0.206}$	$0.235_{\pm 0.189}$
Diversity					
Avg. clusters	1.5	5.2	24.7	34.7	38.8
Avg. unique seqs	1.8	10.4	31.5	37.1	38.9
Cluster ratio	0.038	0.130	0.617	0.867	0.970
Targets w/ 1 cluster	$36/60$	$7/60$	$2/60$	$1/60$	$0/60$

G.4 Temperature ablation

Temperature Ablation. We vary the sampling temperature τ to characterize the recovery–diversity trade-off (Table 18). Temperature scales the sequence logits: $\tau = 0$ uses greedy decoding, while larger values introduce more stochasticity.

At $\tau = 0.1$, mean AAR and CAAR are similar to greedy decoding ($0.536 \rightarrow 0.537$ and $0.334 \rightarrow 0.336$), while unique sequences increase from 1.8 to 10.4 per target and single-cluster targets decrease from $36/60$ to $7/60$. These are descriptive comparisons of means; variation within generated samples does not establish statistical equivalence between temperatures.

Increasing τ to 1.0 yields 38.9 unique sequences per target, with AAR 0.460 and CAAR 0.235. Relative to greedy decoding, this is a recovery cost of 7.6 AAR points and approximately 10 CAAR points. The additional diversity becomes smaller at higher temperatures: the displayed means increase from 37.1 unique sequences at $\tau = 0.5$ to 38.9 at $\tau = 1.0$, while AAR decreases by another 4.4 points.

For our model on this benchmark, $\tau = 0.3$ offers a useful balance: 31.5 unique sequences and 24.7 clusters per target, with reductions of 1.1 AAR points and 1.8 CAAR points from greedy decoding. Clusters use the 90% identity/80% coverage rule of App. O.7; they complement exact sequence counts with diversity under that similarity criterion. Only $2/60$ targets remain in a single cluster at this temperature, compared with $36/60$ under greedy decoding. Thus the mean increase accompanies a reduction in the number of targets with a single sequence cluster.

At $\tau = 0.5$, broader exploration costs 3.2 AAR points. Moving further to $\tau = 1.0$ adds only 1.8 unique sequences per target on average while losing another 4.4 AAR points. These results give a practical choice between recovery and candidate diversity for one trained PLM-free model. They do not establish probability calibration, structural diversity, or retention of physical quality across temperatures, for which this table reports no measurements.

G.5 Per-target diagnostic analysis

The RAbD panels summarize per-target metrics for 60 complexes with 40 designs each. The temporal correlation panel covers all 754 targets in the chronological benchmark of Table 1 and describes target-level associations among the reported metrics. Torsion-specific eligibility filtering applies only to the separate 749-target analysis in App. H.

We use these plots to examine associations among recovery, CDR structure, and other scores. CDR-self-fitted RMSD describes local loop shape; it cannot validate framework- or antigen-relative placement, or folding of the designed sequence. Similarly, an absence of high-AAR/high-RMSD points cannot establish physical validity or isolate the rotamer loss's effect. Per-loop summaries are descriptive, and their correspondence to the main aggregate scores remains unresolved (App. E). Masked-input retraining tests the native-context path.

H3 recovery ranks correlate at $\rho = 0.69$ between H3-only and multi-CDR tasks, and H3-only versus pooled multi-CDR recovery has $\rho = 0.62$. These descriptive associations indicate partially shared target difficulty; they do not imply that the tasks are interchangeable.

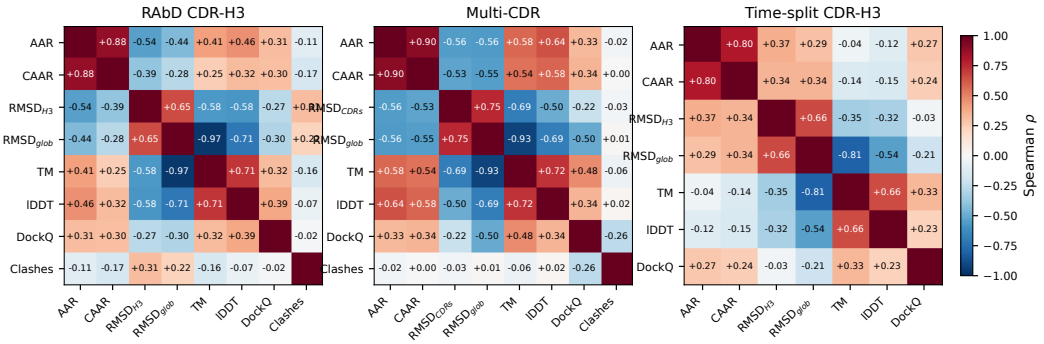


Figure 5: **Per-target Spearman correlations among sequence- and structure-quality metrics.** Three panels: RAbD CDR-H3 (60 targets), multi-CDR (60 targets), and chronological time-split (754 targets, as in the main benchmark). Sequence (AAR, CAAR) and structure (TM, LDDT, RMSD, DockQ) form weakly correlated axes; clashes drop to near-zero variance and decouple from every other metric.

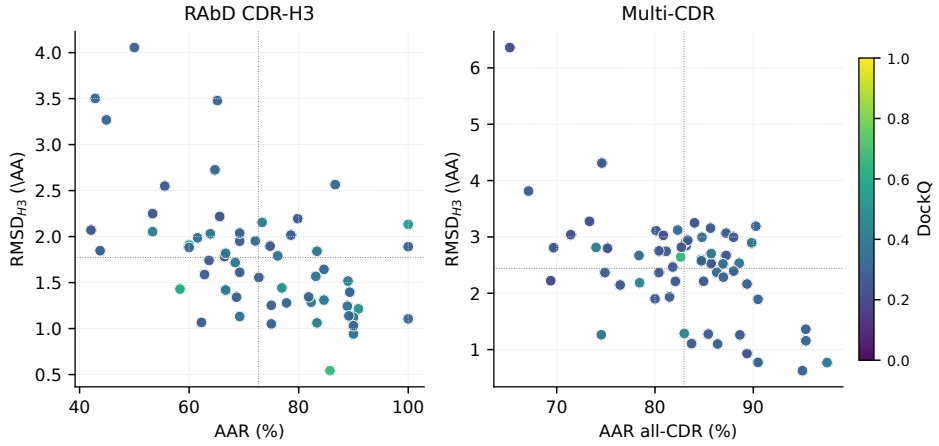


Figure 6: **AAR vs** RMSD_{H3} **per RAbD target.** Color: per-target DockQ. Dotted lines: panel-wise means. Higher recovery is associated with lower CDR RMSD in both settings. This target-level association is descriptive and does not isolate the effect of an individual model component.

H Effectiveness and Scope of the Physical Grounding

Here we examine the physical-grounding term of §2.3 within its intended scope (Cα-clash and χ_1), complementing the sidechain ablation in Table 19. We consider both the local improvements and the outcomes that remain unfavorable to our model.

H.1 Isolating the grounding term

We compare torsion supervision in the no-adapter (PLM-free) setting on 749 eligible targets from the 754-target chronological test split. Five targets are excluded because their PDB structures do not meet the requirements for torsion evaluation; the main benchmark retains all 754 targets. Each condition contains 40 designs per eligible target, totaling 29,960 designs. All metrics in this analysis use raw generated coordinates, without relaxation or sidechain repacking.

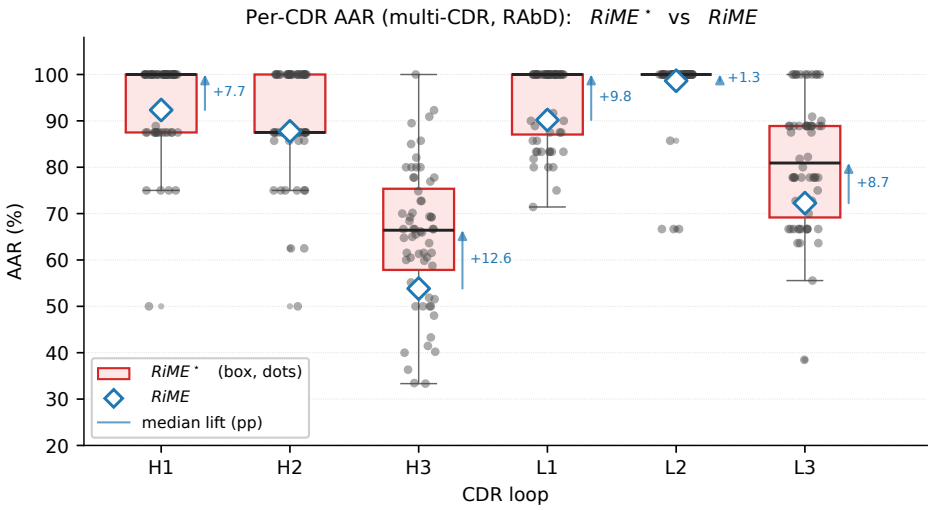


Figure 7: **Per-CDR AAR breakdown on multi-CDR design.** Boxes and grey dots: RiMES* per-target AAR over the 60 RAbD targets. Diamond markers: PLM-free RiMES per-CDR mean (per-target spread is unavailable for that run, hence the single-value marker). Blue arrows annotate the per-target *median* lift in percentage points; across-target *means* are in Table 9 (mean and median differ slightly on L2, where a small subset of PLM-free-favouring targets pulls the mean below zero while the median remains positive). The PLM contribution concentrates on the hardest loops (H3, L3) while H1, L1, and L2 are near-saturated regardless of the PLM.

Table 19: **Sidechain quality metrics** on CDR-H3 residues, for the PLM-augmented model (RiMES*) and its variants. χ_1 MAE: angular error of the first sidechain dihedral vs. the native. Rotamer accuracy: fraction within the native-angle tolerance defined in App. O. Interface clashes (all-atom) counts non-hydrogen atom pairs below 2.0 Å across the CDR–antigen interface, including sidechain atoms.

Model	χ_1 MAE (°) ↓	Rot. Acc. ↑	Interface clashes$_{aa}$ ↓
RiMES*	**63.9**	**0.447**	**8.89**
Without rotamer loss	76.4	0.444	19.00
Other ablations (for reference):			
Euclidean-flow variant	77.8	0.451	16.37
Without flow feedback (inference)	74.3	0.452	14.12
Uniform time sampling	77.8	0.443	9.72

For each metric, we first average the 40 designs within each target. We then apply a paired Wilcoxon signed-rank test to the matched target means for the χ_1 and χ_1–χ_4 conditions, with $n = 749$ target pairs (Table 20). Individual designs from the same target are not treated as independent statistical pairs.

These results show local all-atom improvements: CDR-H3 clash counts improve by 35–48% and the worst overlap by 14%, whereas the whole-complex clash count does not improve significantly.

Remaining antigen-relative placement error. In the antigen coordinate frame, framework Cα RMSD is 28 Å for RiMES, compared with 24 Å for dyMEAN and 25 Å for dyAb. Our model therefore retains a substantial placement error that CDR-aligned metrics can obscure. These measurements do not separate rigid-body placement from internal geometry as sources of all-atom clashes. The local torsion prior alone does not guarantee a correct binding pose.

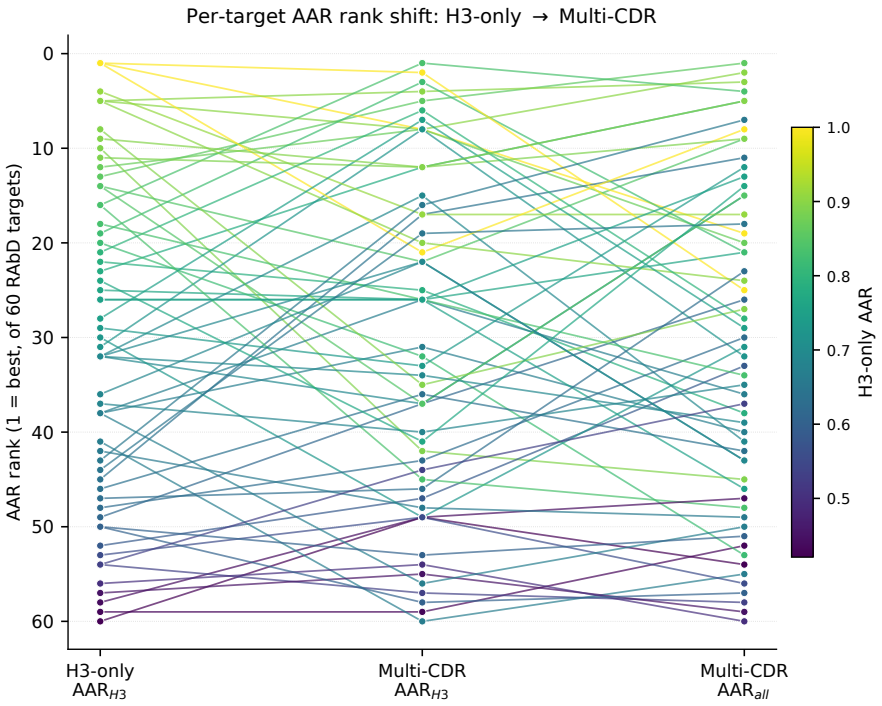


Figure 8: **Per-target AAR rank shift, H3-only → multi-CDR.** Each line is one of the 60 RAbD targets, colored by H3-only AAR. Rank Spearman across the two H3-AAR columns is $+0.69$; comparing H3-only AAR to all-CDR AAR yields $+0.62$ with mean absolute rank shift 11.7.

Table 20: Isolating the torsion supervision (no-adapter setting, chronological split). The reported extension of torsion supervision from χ_1 to χ_1–χ_4 improves the three CDR-H3 clash counts by 35–48% and the worst overlap by 14%, while the whole-complex clash count does not improve significantly. The complete result is reported, including the unfavourable last row; "Native" is the crystal-structure reference on the same metric. The structures are unrelaxed. Overlap uses the van der Waals definition in App. O.3; the worst overlap is a maximum within each structure, averaged across designs and targets for this table. Tests use 749 paired target means.

All-atom metric	χ_1 only	χ_1–χ_4	p-value	Native
H3 antigen clashes	74.8	39.1 (−48%)	4×10^{-34}	0.50
H3 internal clashes	50.4	27.1 (−46%)	1×10^{-54}	0.81
H3 framework clashes	19.2	12.5 (−35%)	8×10^{-32}	1.13
Worst H3 antigen overlap (Å)	1.64	1.42 (−14%)	4×10^{-12}	0.06
Whole-antibody antigen clashes	307.6	342.5 (+11%)	0.19 (n.s.)	1.63

Rotamer canonicality. Consistently, RiMES produces the most canonical χ_1 conformations among the compared methods: 99.6% of eligible designed positions lie within 40° of a Dunbrack rotamer basin, with a mean deviation of 5.7°, versus 10.0° for crystal sidechains and canonicality rates of 90.8–93.7% for the baselines. The *unsupervised* χ_2 occupancy is 69.6%, versus 89.7% for native structures, consistent with the intended scope of the loss. We interpret this as rotamer *canonicality*, which is distinct from native rotamer recovery. On the latter criterion, our model is comparable to or below a shuffled control using residues of the same type. Basin occupancy includes every designed residue with a defined torsion and a rotamer-library entry, irrespective of native residue identity. Native-angle recovery instead requires a matching amino-acid type and defined angles in both structures. The shuffled control pairs a designed torsion with a native torsion from a different position of the same amino-acid type, preserving the marginal angle distributions while removing the position-specific pairing.

Scope of the torsion-extension evidence. Table 20 supports a local effect of extended torsion supervision, while whole-antibody–antigen clashes do not improve significantly. The local metrics remain far from native references, and placement errors persist.

An additional reported summary of extended χ_2–χ_4 supervision gives 98.5% χ_2 basin occupancy, a 9% reduction in interface clashes that is not statistically significant, and an AAR decrease from 0.473 to 0.423. Its configuration, target membership, and interface-clash definition have not been linked to the 749-target comparison in Table 20. We retain this unfavorable summary separately; the 9% and 48% reductions cannot be interpreted as the same comparison without that mapping. These results do not establish that extended torsion supervision consistently improves recovery or interface quality.

I Seed Stability

To assess sensitivity to random initialization, we retrain each of two model variants four independent times with seeds $\{0, 1, 2, 3\}$. The two variants are the PLM-free baseline RiMES (without the equivariant adapter) and the full model RiMES*, the parent-conditioned model trained with the PLM adapter for hit-to-lead redesign. Within each variant, all four seeds use the same training procedure as the corresponding main model. The headline RiMES* H3 AAR is 0.729; the across-seed mean is reported separately. Each of the eight resulting models generates 40 samples per target on the RAbD-60 test set. Table 21 reports per-target metrics for all eight runs, together with the Friedman test p-value computed from the four sets of paired per-target outcomes within each variant.

Table 21: Per-target metrics on RAbD-60 across four independent seeds for each of two model variants (RiMES, upper; RiMES*, lower). CV is the coefficient of variation across runs; Friedman $p > 0.05$ indicates no statistically significant difference between seeds at $\alpha = 0.05$.

Metric	Seed 0	Seed 1	Seed 2	Seed 3	Mean	Std	CV (%)	Friedman p
AAR H3	0.541	0.539	0.536	0.533	0.537	0.003	0.592	0.490
CAAR H3	0.335	0.320	0.310	0.325	0.323	0.010	3.189	0.208
RMSD_{H3} aligned (Å)	1.807	1.824	1.848	1.760	1.810	0.037	2.055	0.011
TM-score	0.969	0.970	0.969	0.970	0.969	0.001	0.085	< 0.001
LDDT	0.816	0.828	0.824	0.833	0.825	0.007	0.873	< 0.001
DockQ	0.339	0.335	0.339	0.343	0.339	0.003	1.012	0.868
AAR H3	0.729	0.697	0.706	0.695	0.706	0.014	2.052	0.102
CAAR H3	0.591	0.535	0.556	0.547	0.557	0.024	4.321	0.290
RMSD_{H3} aligned (Å)	1.773	1.754	1.785	1.793	1.776	0.017	0.955	0.910
TM-score	0.972	0.972	0.972	0.972	0.972	0.000	0.018	0.365
LDDT	0.843	0.845	0.844	0.844	0.844	0.001	0.132	0.002
DockQ	0.350	0.334	0.336	0.347	0.342	0.008	2.375	0.313

Coefficients of variation are below 5% for all metrics in this seed analysis, in both variants. PLM-free H3 AAR ranges from 0.533 to 0.541 (mean 0.537); parent-conditioned H3 AAR ranges from 0.695 to 0.729 (mean 0.706). We find limited variation in aggregate recovery under these settings, alongside significant seed differences in some structural metrics: H3 RMSD, TM-score, and lDDT for PLM-free RiMES, and lDDT for RiMES*. Aggregate stability can thus coexist with systematic target-level differences. We do not interpret a nonsignificant Friedman test as equivalence, and four runs cannot cover all possible initialization variability.

J Dataset Construction and Evaluation Settings

J.1 Source Data and Preprocessing

All experiments are based on antibody–antigen complex structures derived from SAbDab, which remains one of the most established large-scale structural resources for antibody modeling and design. Starting from a local SAbDab summary snapshot, we preprocess each complex into the format required by our co-design pipeline. Although recent public releases of SAbDab contain more than 10,000 antibody-related structures, the effective dataset used in this work is smaller for several

task-specific reasons. Our local snapshot contains 13,278 summary rows corresponding to 6,741 unique PDB identifiers, since multiple rows may describe different chain annotations of the same deposited structure. We then collapse entries to one complex per PDB and discard structures that cannot be reliably parsed, renumbered, or mapped to valid heavy/light chain definitions, leaving 5,371 cleaned antibody structures. Finally, because our task focuses on antibody–antigen co-design, we retain only full complexes with a valid heavy chain, light chain, and antigen, resulting in 3,779 usable examples.

J.2 Cluster-Based Leakage Control

To reduce information leakage across data splits, we perform sequence-based clustering before assigning examples to train, validation, and test sets. In the single-CDR design setting considered in this work, the clustering key is the target CDR-H3 amino acid sequence. For each complex, the CDR-H3 sequence is written as a FASTA entry and clustered with MMseqs2 using a minimum sequence identity threshold of 0.4. We use the resulting MMseqs2 clusters as groups; the threshold alone does not establish an exhaustive all-pairs cross-split identity bound.

Clusters, rather than individual complexes, are treated as the atomic unit of split assignment. This prevents recorded cluster overlap between partitions. The split key is CDR-H3; it does not establish held-out sequence identity for the other five CDRs, frameworks, or antigen families. In the external benchmark setting, SAbDab and benchmark examples are clustered jointly, and any cluster containing a benchmark member is removed from the training pool. In the chronological setting, clustering is performed before temporal partitioning, and each entire cluster is assigned to exactly one split. This provides a stricter estimate of generalization than example-level random splitting.

J.3 Evaluation Settings

We consider two complementary evaluation settings, summarized in Table 22. The first is an *external benchmark split*, where training and validation examples are drawn from SAbDab and the test set is given by the fixed RAbD benchmark. This setting is useful because it follows the standard evaluation protocol adopted in prior antibody CDR design studies and therefore enables direct comparison with previously reported results. However, because the test set is externally specified, it does not measure prospective generalization to future structures.

The second is a *chronological time split*, where all train, validation, and test examples are drawn from SAbDab and ordered using SAbDab date metadata. We assign whole sequence clusters to training, validation, and test partitions in temporal order, targeting an approximate $70/10/20$ split by sample count. This setting is intended to better approximate deployment, where a model trained on historical structures is applied to newly released antibody complexes.

Table 22: Comparison of the two evaluation settings considered in this work.

Aspect	**External Benchmark Split**	**Chronological Time Split**
Training set	SAbDab	SAbDab
Validation set	SAbDab	SAbDab
Test set	RAbD	SAbDab
Test construction	Fixed external benchmark	Chronologically ordered hold-out set
Data separation	Benchmark-based separation	Temporal separation
Leakage prevention	Cluster overlap removed from training	Whole clusters assigned to a single split
Split criterion	Benchmark membership	SAbDab date metadata
Nominal proportion	Benchmark-dependent	Approximately 70/10/20
Evaluation focus	Comparison with prior benchmark results	Forward generalization to newer structures

J.4 External Benchmark Split Statistics

The external benchmark split consists of SAbDab-derived training and validation sets together with the fixed RAbD test benchmark. Its detailed statistics are reported in Table 23. Because the test set is defined by benchmark membership rather than a target ratio, the resulting split is heavily skewed toward training data. Nevertheless, cluster-based filtering ensures that no benchmark-overlapping

Table 23: Summary statistics of the external benchmark split (SAbDab train / validation and RAbD test). CDR lengths, chain lengths, and antigen lengths are reported as mean ± std. No PDB or cluster overlap is present between any pair of splits.

Statistic	Train	Valid	Test
# Complexes	3,224	371	60
# Unique PDBs	3,224	371	60
# Unique clusters	1,632	181	59
Proportion (%)	88.2	10.2	1.6
CDR-H1 length	8.2 ± 0.7	8.1 ± 0.7	8.1 ± 0.4
CDR-H2 length	7.8 ± 0.8	7.9 ± 0.5	7.8 ± 0.4
CDR-H3 length	14.3 ± 4.4	15.1 ± 4.5	13.0 ± 3.2
CDR-L1 length	7.6 ± 2.2	7.3 ± 2.3	6.9 ± 1.7
CDR-L2 length	3.1 ± 0.5	3.1 ± 0.6	3.1 ± 0.5
CDR-L3 length	9.3 ± 1.3	9.1 ± 1.6	9.2 ± 0.9
Heavy chain length	121.5 ± 5.2	122.2 ± 4.9	120.6 ± 3.4
Light chain length	109.6 ± 3.1	109.3 ± 3.7	109.5 ± 2.2
Total antigen length	335.3 ± 337.7	339.5 ± 296.9	228.5 ± 150.6
PDB overlap with other splits	0	0	0
Cluster overlap with other splits	0	0	0

Table 24: Summary statistics of the chronological SAbDab train / validation / test split. CDR lengths, chain lengths, and antigen lengths are reported as mean ± std. No PDB or cluster overlap is present between any pair of splits.

Statistic	Train	Valid	Test
# Complexes	2,647	378	754
# Unique PDBs	2,647	378	754
# Unique clusters	1,541	179	152
Proportion (%)	70.0	10.0	20.0
CDR-H1 length	8.2 ± 0.7	8.1 ± 0.5	8.0 ± 0.4
CDR-H2 length	7.8 ± 0.9	7.9 ± 0.5	7.9 ± 0.5
CDR-H3 length	14.1 ± 4.5	14.9 ± 3.6	14.7 ± 4.2
CDR-L1 length	7.3 ± 2.1	7.6 ± 2.3	8.0 ± 2.2
CDR-L2 length	3.1 ± 0.6	3.0 ± 0.4	3.0 ± 0.4
CDR-L3 length	9.3 ± 1.4	9.3 ± 1.4	9.3 ± 1.0
Heavy chain length	121.5 ± 5.3	121.7 ± 4.2	121.6 ± 4.4
Light chain length	109.7 ± 3.2	109.6 ± 3.4	109.2 ± 2.7
Total antigen length	289.2 ± 299.3	452.7 ± 404.0	432.2 ± 348.5
PDB overlap with other splits	0	0	0
Cluster overlap with other splits	0	0	0

cluster appears in the training pool. The total $3{,}224 + 371 + 60 = 3{,}655$ examples are smaller than the 3,779 usable SAbDab complexes (§J.1) because 124 examples whose CDR-H3 cluster overlaps a RAbD test member are removed from both training and validation.

J.5 CHRONOLOGICAL SABDAB TIME SPLIT STATISTICS

For the chronological setting, we sort cluster-safe SAbDab examples using the SAbDab date field and assign whole clusters to train, validation, and test in temporal order. We target an approximate 70/10/20 split by sample count. Since clusters are kept intact, the final proportions are approximate rather than exact. The resulting split statistics are shown in Table 24.

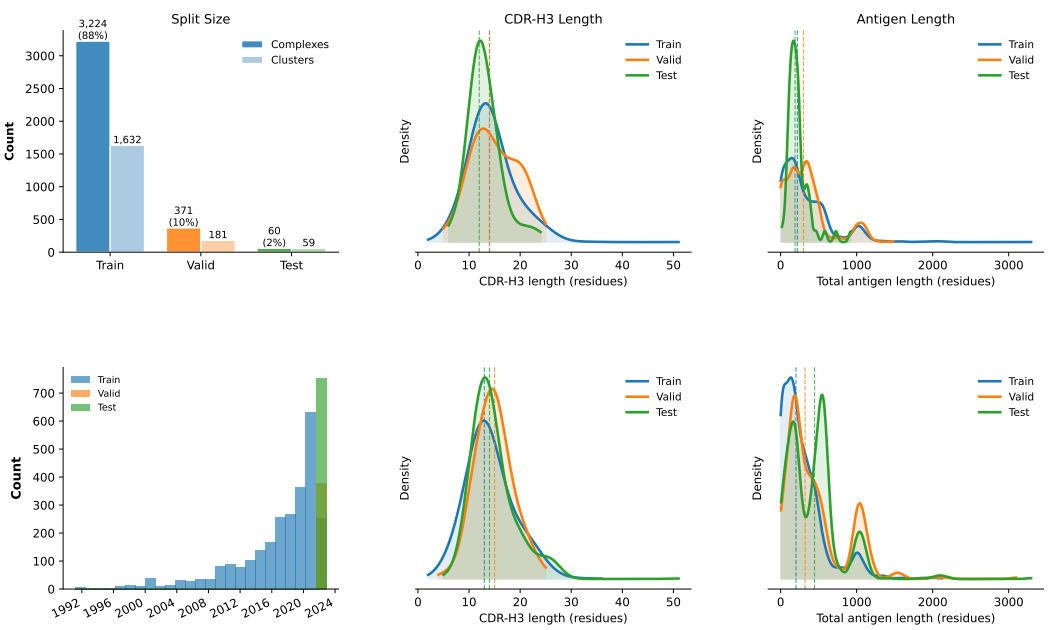


Figure 9: Dataset analysis for the two evaluation settings. **Top row (Setting I):** the external benchmark split, where training and validation data come from SAbDab and the test set is the fixed RAbD benchmark. **Bottom row (Setting II):** the chronological SAbDab time split, where sequence clusters are assigned to partitions using SAbDab date metadata. Dashed vertical lines in the KDE panels indicate the median per split.

J.6 Discussion

The two settings probe different aspects of model behavior. The external benchmark split measures comparability with prior work under the standard RAbD protocol, whereas the chronological split evaluates forward generalization to clusters with later maximum member dates. Because older structures remain with their cluster, this is not a strict individual-complex time cutoff. It also introduces a distribution shift beyond the grouping date. For example, the validation and test sets in the chronological split contain longer antigens on average than the training set. We therefore consider the two settings complementary: one for benchmark comparability, and one for realistic temporal evaluation; Figure 9 summarises both.

K Baselines

We compare against five published antibody-design baselines spanning the dominant methodological families currently used for antibody CDR design: bilevel equivariant graph diffusion (Zhu et al., 2025), rectified-flow inpainting with template conditioning (FlowDesign (Wu et al., 2025)), score-based diffusion with IPA heads and language-model features (Zhu et al., 2024), iterative equivariant full-atom refinement (Kong et al., 2023), and flow-matching with AlphaFold2-augmented antigens (Tan et al., 2025). In the RAbD (Adolf-Bryfogle et al., 2018) / SAbDab (Schneider et al., 2022) comparisons, all methods receive native antibody framework and antigen information as conditioning for CDR design. Their structural outputs differ: AbEgDiffuser, FlowDesign, and AbX graft generated CDRs onto the native framework, whereas dyMEAN, dyAb, and RiMES regenerate antibody structures (App. K.2). These are framework- and interface-conditioned tasks. The baseline comparisons use $N = 40$ samples per target; for AbX (which natively generates 100 samples per target) we subsample uniformly at random to $N = 40$ with a fixed seed to equalize the sample budget.

AbEgDiffuser (Zhu et al., 2025). Bilevel equivariant graph diffusion operating at both residue- and atom-level granularity, conditioned on ESM-2-T36 (Lin et al., 2023) embeddings. Standard sequence and structural metrics use the authors' raw generated PDBs, matching the raw-output

protocol for RiMES. The supplied OpenMM-relaxed PDBs are stored separately and used only for interface-energy scoring (App. O.5); those pre-relaxed structures enter the scoring stage directly without another relaxation step.

FlowDesign (Wu et al., 2025). A rectified-flow generator built on the DiffAb (Luo et al., 2022) codebase. In addition to the native framework, FlowDesign consumes a *per-target template antibody structure* loaded from `./template/dict.pkl` at inference time. The template is fed to the encoder as 3D conditioning through `construct_3d_basis` on the template's heavy-atom coordinates (`rectflow.py` L82–91), providing the generator with a pre-retrieved, already-folded antibody backbone not available to every method in our comparison. We mark this as additional inference-time conditioning and discuss its likely effect on the reported metrics in Appendix K.4. Generated with the official `design_testset.py` driver and the `codesign_single_H3` configuration.

AbX (Zhu et al., 2024). Score-based diffusion with a seqformer (Jumper et al., 2021) backbone and an Invariant Point Attention head for SE(3) coordinate prediction. Conditions on ESM-2-T36 (3B) residue embeddings and a supervised contact-regression head pretrained on ESM-2 features. We mark the contact head as additional supervised conditioning; ESM-2-3B itself is used by multiple methods (including ours) and is not marked separately.

dyMEAN (Kong et al., 2023). End-to-end full-atom antibody design with an iterative equivariant message-passing encoder (AM-EGNN) applied for three refinement rounds. Inference is conditioned on both the framework/antigen and an explicit epitope mask identifying the antigen residues known to contact the antibody in the native complex. Generated with the official `generate.py` with `-n_samples 40`.

dyAb (Tan et al., 2025). Flow-matching generator with the dyAbEGNN equivariant backbone. At inference, the native antigen is augmented with AlphaFold2-predicted structures aligned to the crystal via Kabsch superposition, giving the generator access to multi-view geometric context for the antigen. We mark the AF2-predicted antigen as additional model-derived structural conditioning. Generated with `generate.py -num_samples 40 -use_af2ag True`.

Comparison to published headline numbers. We evaluate the baselines under the generation, conditioning, sample-budget, and scoring protocols described here. These choices may explain differences from published results, but we have not isolated their contributions through matched ablations. We therefore leave attribution to template access, oracle selection, or sample count open.

K.1 Generation Paradigm: An Empirical Audit of Baseline Output PDBs

To interpret full-structure comparisons, we examine which coordinates each method writes to its output PDB. Antibody-design methods differ both in the regions they generate and in the native coordinates they retain. We document these output choices below and explain how they affect downstream evaluation.

K.2 Two Output Paradigms

Across the six methods considered, we observe a clean bimodal split.

CDR-grafting paradigm. AbEgDiffuser, FlowDesign, and AbX all produce output PDBs in which the light chain and the antigen chain(s) have *exactly* 0.000 Å CA RMSD to the native, to three decimal places, for every one of the 20 sampled PDBs we examined. This is consistent with these methods generating only the designed CDR region (sequence and structure) and then programmatically splicing it into an otherwise bit-identical copy of the native framework and antigen.

Full-structure paradigm. dyMEAN, dyAb, and our method regenerate the entire antibody backbone (both heavy and light chains) in one forward pass. The light chain, which is *not* the design target in single-CDR evaluations, nonetheless exhibits a non-zero CA RMSD (0.65 ± 0.08 Å for

dyMEAN, 0.90 ± 0.15 Å for dyAb, and 0.70 ± 0.11 Å for our method), reflecting small per-residue perturbations that accumulate along the generated backbone. The antigen chain remains at 0.000 Å for all three methods; none of them regenerate the antigen.

K.3 Inputs and Additional Conditioning

The two paradigms are mostly orthogonal to the *inputs* each method consumes at inference time. Table 25 summarizes both dimensions. The framework, antigen, and epitope mask (where used) are standard inputs in the RAbD benchmark and are available to all methods; the remaining columns identify inputs that only a subset of methods consume. We mark such inputs with † when they provide additional structural or supervised conditioning beyond the common framework/antigen setup.

Table 25: Inference-time inputs and output-PDB paradigm for each baseline. Paradigm and framework-in-output columns are empirically verified by the per-chain CA RMSD measurements reported in Appendix K.2. The annotations from this table are surfaced at the result tables (Tables 1 and 2) as ‡ for CDR-grafting paradigm (deduced from the Framework-in-output column), † for additional conditioning, and ⋆ for use of the ESM-2-3B PLM. PLM = pretrained protein language model residue embeddings. CR = supervised contact-regression head pretrained on ESM-2. † = additional conditioning not available to every method. **Note on PLM identity.** ESM-2-T36 (`esm2_t36_3B_UR50D`) *is* the 3B-parameter ESM-2 model; AbEgDiffuser, AbX, and RiMES⋆ therefore consume the same PLM weights, although AbX additionally uses a contact-regression head (CR) supervised on top of those embeddings, which is why its PLM column is marked with †.

Method	Paradigm	Framework in output	Antigen	Epitope	PLM	Template	AF2 ag
AbEgDiffuser	CDR-grafting	exact native	✓	no	ESM-2-3B (T36)	no	no
FlowDesign	CDR-grafting	exact native	✓	no	no	†	no
AbX	CDR-grafting	exact native	✓	no	† ESM-2-3B + CR	no	no
dyMEAN	Full-structure	regenerated (∼0.7 Å)	✓	✓	no	no	no
dyAb	Full-structure	regenerated (∼0.9 Å)	✓	✓	no	no	†
RiMES	Full-structure	regenerated	✓	✓	no	no	no
RiMES⋆	Full-structure	regenerated	✓	✓	ESM-2-3B	no	no

K.4 Interpreting output-paradigm differences

Whole-antibody RMSD, TM-score, and lDDT reflect both the generated region and the output protocol. CDR-grafting methods retain the native framework, while full-structure methods predict its coordinates. Framework reuse therefore contributes to whole-antibody comparisons (App. K.2).

CDR-local metrics reduce the contribution of unchanged framework coordinates. Their interpretation still requires explicit conditioning, residue selection, alignment, and aggregation. Our CDR-self-fitted RMSD measures local shape after independent loop alignment; it does not assess placement relative to the framework or antigen. The aggregation correspondence for the multi-CDR sequence summaries remains unresolved as stated in App. E.

Interface-energy comparisons also depend on generated coordinates, conditioning, relaxation, scoring, and completion. We retain failed attempts in the completion-corrected denominator and do not infer that scores become independent of the output paradigm after filtering. The available comparisons do not isolate architecture or output scope as the cause of failure rates or energy differences.

L Antibody Representation and Data

L.1 Amino Acid and Atom Vocabulary

Each residue is represented with up to $C = 14$ atom channels: 4 backbone atoms (N, Cα, C, O) and up to 10 sidechain atoms in a fixed per-amino-acid ordering. The 20 standard amino acids are indexed in the order G, A, V, L, I, F, W, Y, D, H, N, E, K, Q, M, R, S, T, C, P. Five special tokens are appended: `PAD`, `MASK`, `BOA` (beginning of antigen), `BOH` (beginning of heavy chain), and `BOL` (beginning of light chain).

Each atom is typed into one of 7 categories (pad, mask, C, N, O, S, other) and assigned one of 9 positional labels (pad, mask, backbone, β, γ, δ, ϵ, ζ, η) based on the Greek-letter naming convention.

Missing atom coordinates are filled with the Cα position of the same residue. An `xloss_mask` binary tensor indicates which atom slots contain real coordinates and should contribute to the structure loss.

L.2 Graph Construction

The antibody–antigen complex is represented as a graph with three types of global nodes (BOA, BOH, BOL) whose coordinates are set to the centroid of their respective chains. Two sets of edges are constructed:

- **Context edges** connect residues within the same segment (antigen–antigen, heavy–heavy, or light–light) via k-nearest neighbors ($k = 9$) computed from minimum cross-channel atom–atom distances.
- **Interface edges** connect residues across segments (paratope–epitope) similarly via k-NN, optionally using predicted inter-chain edge distances when the edge-distance prediction module is enabled.

L.3 CDR Identification and Masking

CDR regions are identified using IMGT numbering (Lefranc et al., 2003) applied to the renumbered antibody sequences. Table 26 lists the IMGT residue number ranges for each region.

Table 26: IMGT numbering ranges for antibody variable domain regions.

Region	Heavy chain	Light chain
FR1	1–26	1–26
CDR1	27–38	27–38
FR2	39–55	39–55
CDR2	56–65	56–65
FR3	66–104	66–104
CDR3	105–117	105–117
FR4	118–129	118–129

Three binary masks govern the design process:

- **Coordinate mask** (`cmask`): 1 for all antibody Fv residues (excluding global nodes), 0 for antigen and global nodes. The model predicts coordinates at masked positions.
- **Sequence mask** (`smask`): 1 only for CDR residues to be designed. In single-CDR mode, only CDR-H3 is masked; in multi-CDR mode, all six CDRs (H1, H2, H3, L1, L2, L3) are masked. Framework sequences are always provided as ground truth.
- **Paratope mask**: 1 for CDR residues involved in binding interface supervision (default: H3; multi-CDR: L3 and H3). This mask controls the docking loss and the shadow paratope stream in the dual-stream encoder.

L.4 Epitope Definition

The antigen epitope is defined as the 48 antigen residues closest to CDR-H3, measured by minimum heavy-atom distance between each antigen residue and any CDR-H3 residue. This interface-centric representation focuses the model's attention on the binding-relevant portion of the antigen surface.

L.5 Framework Templates

CDR coordinates are initialized from a conserved framework template rather than random noise. The template is built offline by averaging backbone atom positions at highly conserved IMGT positions across the training set, then linearly interpolating coordinates between known anchor positions along

each chain. Sidechain slots are initialized at the Cα position. At training time, the template is *not* aligned to the ground truth (no Kabsch), so the model must learn to place CDRs in the correct global context.

M Complete Model Specification

We describe RiMES component by component, following the flow of data during training and inference and giving the dimensions, hyperparameters, and numerical details needed to reproduce our model.

M.1 SE(3) Background and Notation

Rotation group. $\mathrm{SO}(3) = \{R \in \mathbb{R}^{3\times 3} : R^\top R = I,\ \det(R) = 1\}$ is the special orthogonal group. Its Lie algebra $\mathfrak{so}(3)$ is the space of skew-symmetric matrices, isomorphic to $\mathbb{R}^3$ via the *hat* map $\widehat{\cdot}: \mathbb{R}^3 \to \mathfrak{so}(3)$:

$$\widehat{\boldsymbol{\omega}} = \begin{pmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{pmatrix}. \tag{19}$$

Exponential and logarithmic maps. Rodrigues' formula gives the exponential:

$$\exp(\widehat{\boldsymbol{\omega}}) = I + \frac{\sin\theta}{\theta}\widehat{\boldsymbol{\omega}} + \frac{1-\cos\theta}{\theta^2}\widehat{\boldsymbol{\omega}}^2, \quad \theta = \|\boldsymbol{\omega}\|, \tag{20}$$

with Taylor expansion for $\theta < 10^{-4}$. The inverse is

$$\log(R) = \frac{\theta}{2\sin\theta}(R - R^\top), \quad \theta = \arccos\left(\tfrac{\mathrm{tr}(R)-1}{2}\right), \tag{21}$$

with arccos input clamped to $[-1{+}10^{-7},\ 1{-}10^{-7}]$ and an eigendecomposition fallback for $\theta > \pi - 0.2$.

Rigid frames. Each residue has a local frame $(R_i, \mathbf{t}_i) \in \mathrm{SE}(3)$:

$$\begin{aligned} \mathbf{e}_1 &= (\mathrm{C}_i - \mathrm{C}\alpha_i)/\|\cdot\|_\epsilon, \quad \mathbf{v}_2 = \mathrm{N}_i - \mathrm{C}\alpha_i, \\ \mathbf{e}_2 &= (\mathbf{v}_2 - (\mathbf{v}_2\cdot\mathbf{e}_1)\mathbf{e}_1)/\|\cdot\|_\epsilon, \quad \mathbf{e}_3 = \mathbf{e}_1 \times \mathbf{e}_2, \\ R_i &= [\mathbf{e}_1\ \mathbf{e}_2\ \mathbf{e}_3], \quad \mathbf{t}_i = \mathrm{C}\alpha_i, \end{aligned} \tag{22}$$

with $\|\cdot\|_\epsilon = \max(\|\cdot\|, 10^{-6})$.

Geodesic. The shortest path on $\mathrm{SO}(3)$ between R_0 and R_1 is $R(s) = R_0 \exp(s \log(R_0^\top R_1))$ for $s \in [0, 1]$.

M.2 Input Representation

We represent each antibody–antigen complex as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ in which each node i carries:

- **Amino acid type** $s_i \in \{0, \ldots, 24\}$ (20 standard + special tokens for BOA/BOH/BOL/mask/pad).
- **Atom coordinates** $\mathbf{X}_i \in \mathbb{R}^{C\times 3}$ with $C = 14$ channels: 4 backbone (N, Cα, C, O) + up to 10 sidechain slots, padded where absent.
- **Atom importance weights** $\mathbf{w}_i \in [0, 1]^C$ with zeros for pad slots.
- **Residue position** $r_i \in \mathbb{Z}$ (sequence index within its chain).
- **Segment ID** $g_i \in \{\text{heavy}, \text{light}, \text{antigen}\}$.
- **PLM embedding** $\mathbf{z}_i \in \mathbb{R}^{d_{\text{plm}}}$ with $d_{\text{plm}} = 2560$ (ESM-2 3B).

The CDR mask $\mathcal{M} \subset \mathcal{V}$ selects residues to design (H3 only, or H1–H3 + L1–L3 for six-CDR design). We initialize CDR coordinates from framework templates via conserved-position interpolation and replace CDR sequence tokens with the `[MASK]` token.

M.3 Equivariant Adapter: Full Specification

Our adapter transforms PLM embeddings $\mathbf{z}_i$ into SE(3)-invariant structure-aware features of width $d = 64$. We optimize its four forward-pass blocks through the joint objective.

Block 1: LoRA projection.

$$\mathbf{h}_i^{\text{plm}} = W_0 \mathbf{z}_i + W_{\text{up}} W_{\text{down}} \mathbf{z}_i, \quad W_{\text{down}} \in \mathbb{R}^{r \times d_{\text{plm}}},\ W_{\text{up}} \in \mathbb{R}^{d \times r}, \tag{23}$$

with rank $r = 16$. W_{up} is zero-initialized.

Block 2: Geometric features. Six SE(3)-invariant scalars per residue:

$$\mathbf{g}_i = [d_{NC\alpha},\ d_{C\alpha C},\ d_{CO},\ d_{NC},\ \cos\angle_{NC\alpha C},\ \cos\angle_{C\alpha CO}] \in \mathbb{R}^6, \tag{24}$$

projected by a two-layer MLP with SiLU to $\mathbf{h}_i^{\text{geom}} \in \mathbb{R}^d$.

Block 3: Gating.

$$\boldsymbol{\sigma}_i = \sigma\big(\text{MLP}([\mathbf{h}_i^{\text{plm}} \| \mathbf{h}_i^{\text{geom}}])\big), \quad \mathbf{h}_i = \boldsymbol{\sigma}_i \odot \mathbf{h}_i^{\text{plm}} + (1 - \boldsymbol{\sigma}_i) \odot \mathbf{h}_i^{\text{geom}}, \tag{25}$$

where $\sigma(\cdot)$ is the elementwise logistic sigmoid (matching the notation used in the main paper; the gate output $\boldsymbol{\sigma}_i$ is a separate vector and not the function).

Block 4: Geometry-biased attention. Over a k-NN graph ($k = 9$ on Cα distances), $n_h = 4$ heads with $d_h = 16$:

$$\alpha_{ij}^{(h)} = \frac{\mathbf{q}_i^{(h)\top} \mathbf{k}_j^{(h)}}{\sqrt{d_h}} + \beta_{\text{dist}}^{(h)}(-\|\mathbf{t}_i - \mathbf{t}_j\|) + \beta_{\text{frame}}^{(h)} \frac{\text{tr}(R_i^\top R_j)}{3}, \tag{26}$$

with $\beta_{\text{dist}}^{(h)}, \beta_{\text{frame}}^{(h)}$ learnable per head. Softmax over j, then standard value aggregation. A residual FFN (two linear layers, SiLU, dropout $p = 0.1$) follows.

SE(3)-invariance. Under $\mathbf{X}_i \mapsto Q\mathbf{X}_i + \mathbf{u}$ with $Q \in \text{SO}(3)$: PLM embeddings (block 1) depend only on sequence, so they are invariant. Geometric features (block 2) use bond lengths and angles, both invariant. Gating (block 3) is a pointwise function of invariants. In attention (block 4), the distance $\|\mathbf{t}_i - \mathbf{t}_j\|$ is invariant, and $\text{tr}(R_i^\top R_j) = \text{tr}((QR_i)^\top QR_j)$ because $Q^\top Q = I$. The k-NN graph uses distances only. Hence $\mathbf{h}_i$ is SE(3)-invariant.

M.4 Multi-Channel Equivariant Message Passing (EGNN)

The EGNN operates on the full atom representation $\mathbf{X}_i \in \mathbb{R}^{C \times 3}$ with importance weights $\mathbf{w}_i$. We use $L = 3$ message-passing layers with two parallel streams.

Edge radial features. For edge (i, j), cross-channel distances weighted by atom importance:

$$\tilde{\mathbf{r}}_{ij} = \mathbf{A}_i^\top \left[\|\mathbf{x}_i^c - \mathbf{x}_j^{c'}\|_\epsilon \cdot w_i^c w_j^{c'}\right]_{c,c'} \mathbf{A}_j, \quad \mathbf{r}_{ij} = \frac{W_r \tilde{\mathbf{r}}_{ij}}{\|\tilde{\mathbf{r}}_{ij}\| + 1}, \tag{27}$$

where $\mathbf{A}_i \in \mathbb{R}^{C \times d_{\text{attr}}}$ are learned atom attributes, $d_{\text{attr}} = 16$. The distance is computed as $\sqrt{\|\mathbf{x}_i^c - \mathbf{x}_j^{c'}\|^2 + 10^{-8}}$ for gradient stability at coincident atoms.

Edge message. $\mathbf{m}_{ij} = \text{MLP}_{\text{edge}}([\mathbf{h}_i \| \mathbf{h}_j \| \mathbf{r}_{ij}])$ with attention $\alpha_{ij} = \text{sigmoid}(W_\alpha \mathbf{m}_{ij})$, then $\mathbf{m}_{ij} \leftarrow \alpha_{ij} \mathbf{m}_{ij}$.

Node update. $\mathbf{h}_i \leftarrow \mathbf{h}_i + \text{MLP}_{\text{node}}([\mathbf{h}_i \| \sum_{j \in \mathcal{N}(i)} \mathbf{m}_{ij}])$.

Coordinate update (equivariant). For each channel c of node i:

$$\mathbf{x}_i^c \leftarrow \mathbf{x}_i^c + \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \phi_c(\mathbf{m}_{ij}) \, (\mathbf{x}_i^c - \bar{\mathbf{x}}_j), \tag{28}$$

where $\phi_c(\mathbf{m}_{ij}) \in \mathbb{R}$ is a per-channel scalar gain produced by $\mathrm{MLP}_{\mathrm{coord}} : \mathbb{R}^d \to \mathbb{R}^C$ (one output per atom channel) and broadcast across the spatial axis, and $\bar{\mathbf{x}}_j = \sum_{c'} w_j^{c'} \mathbf{x}_j^{c'} / \max(\sum_{c'} w_j^{c'}, 1)$ is the importance-weighted centroid. The scalar-gain form preserves SE(3) equivariance: under $\mathbf{x} \mapsto Q\mathbf{x}$, the spatial vector $(\mathbf{x}_i^c - \bar{\mathbf{x}}_j)$ co-rotates while ϕ_c depends only on invariants.

Dual streams. A *context stream* operates over the full k-NN graph on all residues (antibody + antigen). An *interface stream* operates on the paratope + epitope subgraph only. At each layer, paratope residues synchronize hidden states between streams. For empty edge lists (rare with small per-GPU batches), we bypass the edge operations and run only the node update with zero aggregation. Figure 10 shows our complete adapter stack.

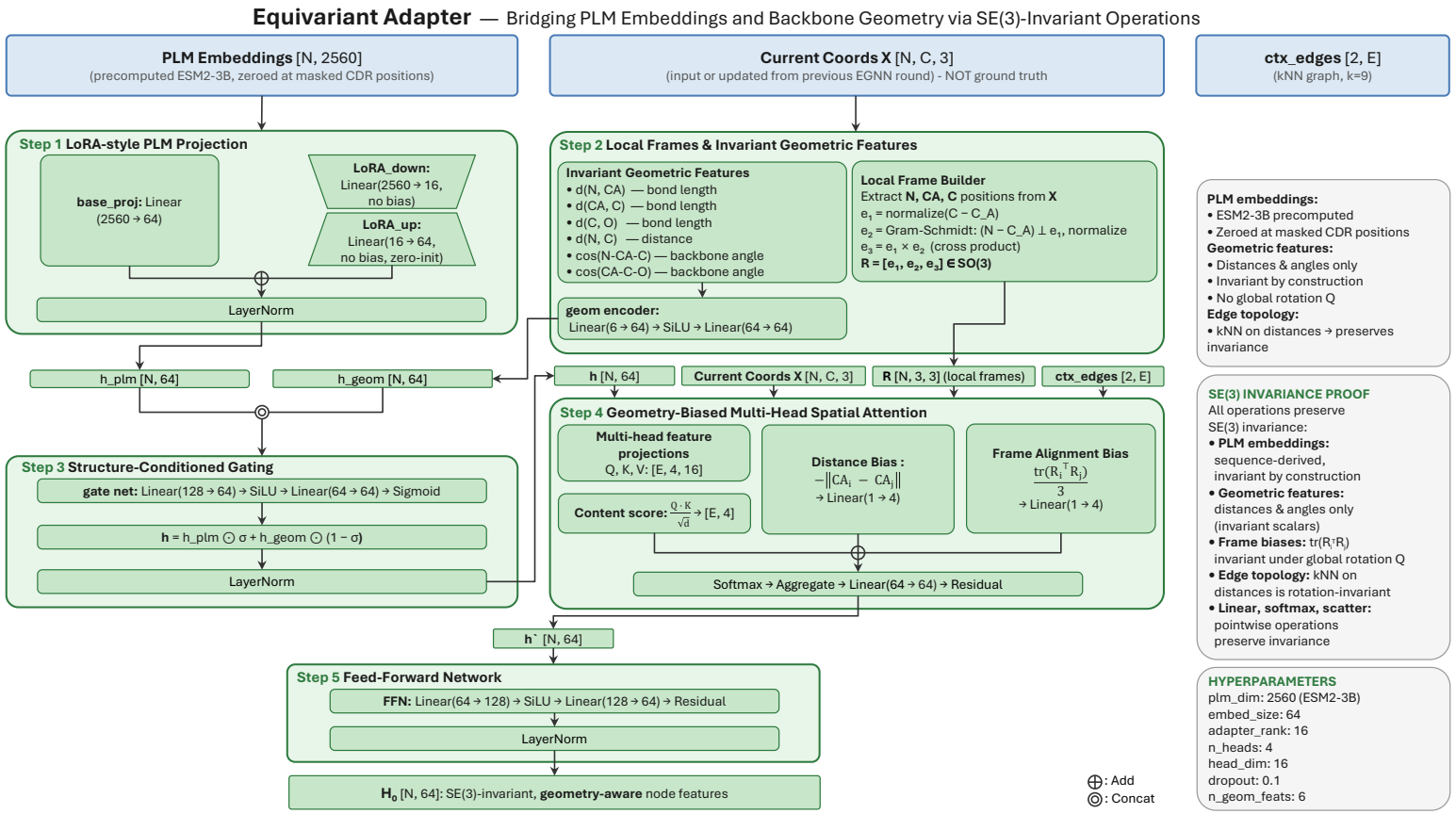


Figure 10: **Equivariant adapter, detailed implementation (App. M.3).** ESM-2-3B embeddings $\mathbf{z}_i$ are first compressed by a rank-16 LoRA projection with zero-init residual (**Stage 1**; Eq. 23). Six SE(3)-invariant local-geometry scalars (four bond lengths and two bond angles) are MLP-projected to a parallel feature stream (**Stage 2**). A residue-wise sigmoid gate mixes the two streams (**Stage 3**; Eq. 25), and a k-NN attention block with two learned per-head geometric biases (negative distance, frame trace) produces the final SE(3)-invariant feature $\mathbf{h}_i$ (**Stage 4**; Eq. 26). The stage labels denote operations within a forward pass, not separate training phases. d=64 throughout; W_{up} zero-initialised; k=9, n_h=4 heads, d_h=16.

M.5 Iterative Refinement Loop

We run $T = 3$ rounds indexed $r = 1, 2, 3$. At each round, the optional adapter combines parent embeddings and local geometry. At inference, raw embeddings at all designed CDR positions are zeroed. During training, the sequence-reveal curriculum supplies partial ground-truth sequence context (App. M.9). From round 2 onward, the memory FFN carries the preceding encoder state. The context and interface streams each apply $L = 3$ message-passing layers and decode sequence logits and coordinates. We propagate intermediate sequence predictions as soft distributions and decode the final sequence greedily or with the stated temperature rule. During training and inference, non-final rounds receive the translation nudge of Eq. 11, at $s = 1/3$ and $s = 2/3$. Base losses supervise each round, while the final outputs supply the auxiliary frame, rotamer, and clash losses.

Flow component	RiMES invariant velocity head
Start distribution	Per-graph mixture of perturbed detached predictions and Haar rotations/noisy reference translations
Interpolation	Per-residue logit-normal times and self-normalized inverse-density weights
Velocity input	Shared state and time embedding; $96 = 64 + 32$ node inputs
Edge features	Distance, relative displacement in the local frame, and relative rotation (13 entries)
Velocity output	Body-frame angular velocity and local translation mapped to ambient coordinates by R_i
Training gradients	Shared hidden features and direct frame consistency; starts are detached
Intermediate feedback	Translation-only nudges after non-final rounds during training and inference

Figure 11: RiMES's residue-frame flow module. The shared-feature and frame-consistency paths couple auxiliary frame learning to the encoder. The coordinate-input predecessor is used only in the diagnostic comparison of App. N.1.

M.6 Riemannian Flow Module

We organize our flow module into three parts: a noiser that produces training interpolants, a velocity network that predicts velocities, and a loss that matches them.

M.6.1 Frame Noiser with Mixed Starting Distribution

Given ground-truth frames $(R_{\mathrm{gt}}, \mathbf{t}_{\mathrm{gt}})$ and EGNN predictions $(\hat{R}_{\mathrm{pred}}, \hat{\mathbf{t}}_{\mathrm{pred}})$:

Algorithm 1 Mixed-mode frame noiser

Require: R_{gt}, $\mathbf{t}_{\mathrm{gt}}$, $\hat{R}_{\mathrm{pred}}$, $\hat{\mathbf{t}}_{\mathrm{pred}}$, batch_id (per-residue graph assignment)
$z \sim \mathcal{N}(\mathbf{0}, I_N)$; $s \leftarrow \mathrm{clamp}(\sigma(\mu + \sigma_z z), 10^{-4}, 1 - 10^{-4})$ ▷ independent per-residue times
$w \leftarrow \exp(-\log p(s))$; $w \leftarrow w / \max(\overline{w}, 10^{-8})$ ▷ self-normalized importance weights
$R_{\mathrm{start}} \leftarrow \mathrm{sg}[\hat{R}_{\mathrm{pred}}] \cdot \exp(\widehat{0.1 \cdot \boldsymbol{\xi}})$, $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, I_3)$
$\mathbf{t}_{\mathrm{start}} \leftarrow \mathrm{sg}[\hat{\mathbf{t}}_{\mathrm{pred}}] + \mathcal{N}(\mathbf{0}, 0.5^2 I_3)$
$m_g \sim \mathrm{Bernoulli}(0.5)^B$; $m_{\mathrm{res}} \leftarrow m_g[\mathrm{batch_id}]$
if any m_{res} **then**
    $R_{\mathrm{noise}} \sim \mathrm{Haar}(\mathrm{SO}(3))^N$ via quaternions
    $\mathbf{t}_{\mathrm{noise}} \leftarrow \mathbf{t}_{\mathrm{gt}} + \mathcal{N}(\mathbf{0}, 2.0^2 I_3)$
    $R_{\mathrm{start}}[m_{\mathrm{res}}] \leftarrow R_{\mathrm{noise}}[m_{\mathrm{res}}]$
    $\mathbf{t}_{\mathrm{start}}[m_{\mathrm{res}}] \leftarrow \mathbf{t}_{\mathrm{noise}}[m_{\mathrm{res}}]$
end if
$R_s \leftarrow R_{\mathrm{start}} \exp(s \log(R_{\mathrm{start}}^{\top} R_{\mathrm{gt}}))$
$\mathbf{t}_s \leftarrow (1 - s)\mathbf{t}_{\mathrm{start}} + s\,\mathbf{t}_{\mathrm{gt}}$
return $(R_s, \mathbf{t}_s, R_{\mathrm{start}}, \mathbf{t}_{\mathrm{start}}, s, w)$ ▷ w is the normalized weight above

$\mathrm{sg}[\cdot]$ denotes `detach`. Haar sampling uses $\mathbf{q} \sim \mathcal{N}(\mathbf{0}, I_4)$, $\mathbf{q} \leftarrow \mathbf{q}/\|\mathbf{q}\|$, then quaternion-to-matrix conversion.

M.6.2 Velocity Network

Invariant node input. $\mathbf{f}_i^{(0)} = \mathrm{MLP}_{\mathrm{in}}([\mathbf{h}_i \| \phi(s)])$, where $\phi(s)$ is a 32-dimensional time embedding (Linear(1,32), SiLU, Linear(32,32)). With $d = 64$, the 96-dimensional concatenation is projected to width d with LayerNorm and SiLU. Global rotation entries and translations are not concatenated to node features.

Relative-frame message passing. We use $L_v = 2$ message-passing layers on a k-NN graph ($k = 9$) built from current translations. Edges use the 13-dimensional invariant features of Eq. 8. Each layer computes

$$\mathbf{m}_{ij}^{(\ell)} = \mathrm{MLP}_{\mathrm{edge}}^{(\ell)}([\mathbf{f}_i^{(\ell)} \| \mathbf{f}_j^{(\ell)} \| \mathbf{e}_{ij}]),$$
$$\mathbf{f}_i^{(\ell+1)} = \mathbf{f}_i^{(\ell)} + \mathrm{MLP}_{\mathrm{node}}^{(\ell)}([\mathbf{f}_i^{(\ell)} \| \mathrm{mean}_{j \in \mathcal{N}(i)} \mathbf{m}_{ij}^{(\ell)}]).$$

Local-frame readout. The angular output is a body-frame tangent vector, $\hat{\boldsymbol{\omega}}_i = W_\omega \mathbf{f}_i^{(L_v)}$. The translation output is predicted locally and mapped to ambient coordinates: $\tilde{\mathbf{v}}_i = W_v \mathbf{f}_i^{(L_v)}$ and $\hat{\mathbf{v}}_i = R_i \tilde{\mathbf{v}}_i$. Both readouts are zero-initialized, including biases. Appendix N states the assumptions and proves the resulting transformation law.

M.6.3 Flow Matching Loss

Velocity targets at time s:

$$\boldsymbol{\omega}_i^* = \log(R_i(s)^\top R_i^{\mathrm{gt}})/(1-s)_\epsilon, \tag{29}$$
$$\mathbf{v}_i^* = (\mathbf{t}_i^{\mathrm{gt}} - \mathbf{t}_i(s))/(1-s)_\epsilon, \tag{30}$$

with $(1-s)_\epsilon = \max(1-s, 10^{-3})$.

$$\mathcal{L}_{\mathrm{flow}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} w_i \Big[\lambda_\omega \, \ell_\delta(\hat{\boldsymbol{\omega}}_i, \boldsymbol{\omega}_i^*) + \lambda_v \, \ell_\delta(\hat{\mathbf{v}}_i, \mathbf{v}_i^*) \Big], \tag{31}$$

with Huber loss ℓ_δ and per-residue importance weights w_i. Because starting frames are detached, this loss carries no gradient back to the encoder via its velocity targets $(\boldsymbol{\omega}^*, \mathbf{v}^*)$; a residual gradient still reaches the encoder through the shared representation $\mathbf{h}_i$ that the velocity network consumes (§2.3).

M.6.4 Frame Consistency Loss

Using *non-detached* predicted frames $\hat{R}_i$ from $\hat{\mathbf{X}}_i$:

$$\mathcal{L}_{\mathrm{consist}} = \frac{1}{|\mathcal{E}_{\mathcal{M}}|} \sum_{(i,j) \in \mathcal{E}_{\mathcal{M}}} \ell_\delta \left(\frac{\mathrm{tr}(\hat{R}_i^\top \hat{R}_j)}{3}, \, \frac{\mathrm{tr}(R_i^{\mathrm{gt}\top} R_j^{\mathrm{gt}})}{3} \right), \tag{32}$$

over CDR–CDR context edges. This provides a direct frame-coordinate signal to the encoder ($\hat{R}_i \to \hat{\mathbf{X}}_i$); the residual $\mathbf{h}_i$ pathway from $\mathcal{L}_{\mathrm{flow}}$ is not gated by detach (App. M.6.3).

M.7 Rotamer-Aware Joint Loss

χ_1 computation. Using backbone atoms N, Cα, Cβ and first sidechain atom X$_\gamma$:

$$\mathbf{n}_1 = (\mathrm{N} - \mathrm{C}\alpha) \times (\mathrm{C}\beta - \mathrm{C}\alpha),$$
$$\mathbf{n}_2 = (\mathrm{C}\alpha - \mathrm{C}\beta) \times (\mathrm{X}_\gamma - \mathrm{C}\beta),$$
$$\hat{\mathbf{u}}_1 = (\mathrm{C}\alpha - \mathrm{C}\beta)/\|\cdot\|_\epsilon,$$
$$\chi_1 = \mathrm{atan2}((\mathbf{n}_1 \times \mathbf{n}_2) \cdot \hat{\mathbf{u}}_1, \; \mathbf{n}_1 \cdot \mathbf{n}_2).$$

Dunbrack rotamer library. For each of 18 types with χ_1 (all except Gly, Ala), we use the backbone-independent table of (Shapovalov & Dunbrack, 2011), which gives up to 3 canonical rotamers $\{(\chi_a^{(r)}, p_a^{(r)})\}$ with probabilities. Compatibility kernel:

$$\mathrm{compat}(\chi, a) = \sum_{r=1}^{n_a} p_a^{(r)} \exp\left(\frac{\cos(\chi - \chi_a^{(r)}) - 1}{\sigma_\chi^2}\right), \tag{33}$$

with kernel bandwidth $\sigma_\chi = 30^\circ$.

Loss.

$$\mathcal{L}_{\mathrm{rotamer}} = \frac{\lambda_{\mathrm{rot}}}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \sum_{a=1}^{20} p_i^{(a)} \big(1 - \mathrm{compat}(\chi_1^{(i)}, a)\big), \tag{34}$$

with $p_i^{(a)} = \mathrm{softmax}(\hat{\mathbf{s}}_i)_a$. Glycine and alanine contribute zero. Note that $\mathrm{compat}(\chi, a) \in [0, 1]$ for normalized mixture weights, and its exact penalty floor is $1 - \max_\chi \mathrm{compat}(\chi, a)$; the loss is monotone in rotamer compatibility but its absolute scale depends on the rotamer distribution of the amino acid (see also the Limitations note in §B).

M.8 Steric Clash Penalty

$$\mathcal{L}_{\mathrm{clash}} = \lambda_{\mathrm{clash}} \Big[\lambda_{\mathrm{inter}} \sum_{\substack{i \in \mathcal{M} \\ j \in \mathcal{V}_{\mathrm{ag}}}} \mathrm{softplus}(d_{\mathrm{thresh}} - d_{ij}) + \sum_{\substack{i,j \in \mathcal{M} \\ i \neq j}} \mathrm{softplus}(d_{\mathrm{thresh}} - d_{ij})\Big], \tag{35}$$

with d_{ij} the Cα–Cα distance, $d_{\mathrm{thresh}} = 3.0\,\text{Å}$, $\lambda_{\mathrm{inter}} = 2.0$, $\lambda_{\mathrm{clash}} = 0.5$.

M.9 Base Losses and Total Objective

Sequence cross-entropy (per round). Let $\mathcal{M}$ be the full design set and $\mathcal{S}_{u,r} \subseteq \mathcal{M}$ the sequence-supervised positions at optimization step u and refinement round r. The reveal curriculum supplies partial ground-truth context with reveal probability $\rho_u = 0.5(\cos(u\pi/U_{\max}) + 1) \cdot 0.9$. The per-round cross-entropy is

$$\mathcal{L}_{\mathrm{seq}}^{(r)} = -\frac{1}{\max(1, |\mathcal{S}_{u,r}|)} \sum_{i \in \mathcal{S}_{u,r}} \log[\mathrm{softmax}(\hat{\mathbf{s}}_i^{(r)})]_{s_i^{\mathrm{gt}}}. \tag{36}$$

This notation distinguishes the sequence-supervision set from the full design set $\mathcal{M}$ used in the geometric losses; it does not equate the sequence and PLM masks.

Structure loss (per round). $\mathcal{L}_{\mathrm{struct}}^{(r)} = \ell_\delta(\hat{\mathbf{X}}_{\mathcal{M}}^{(r)}, \mathrm{Kabsch}(\mathbf{X}_{\mathcal{M}}^{\mathrm{gt}} \to \hat{\mathbf{X}}_{\mathcal{M}}^{(r)}))$ plus a bond-length regularizer. Kabsch alignment handles degenerate cases (NaN covariance or fewer than 2 atoms) by returning identity rotation and zero translation.

Docking loss (per round). $\mathcal{L}_{\mathrm{dock}}^{(r)} = \ell_\delta(\hat{\mathbf{X}}_{\mathrm{paratope}}^{(r)}, \mathbf{X}_{\mathrm{paratope}}^{\mathrm{gt}}) + \ell_\delta(\hat{d}_{\mathrm{inter}}, d_{\mathrm{inter}}^{\mathrm{gt}})$ penalizing paratope atom positions and predicted inter-chain distances.

Total objective.

$$\begin{aligned}\mathcal{L} = \sum_{r=1}^{T} \big[\mathcal{L}_{\mathrm{seq}}^{(r)} + \mathcal{L}_{\mathrm{struct}}^{(r)} + \mathcal{L}_{\mathrm{dock}}^{(r)}\big] \\ + \lambda_f \mathcal{L}_{\mathrm{flow}} + \lambda_c \mathcal{L}_{\mathrm{consist}} + \mathcal{L}_{\mathrm{rotamer}} + \mathcal{L}_{\mathrm{clash}}.\end{aligned} \tag{37}$$

The auxiliary terms use the final outputs. Starting frames and interpolation targets are detached. Hidden features supplied to the velocity network are not detached, so flow loss trains that network and shared representations. Consistency additionally acts through predicted rotations. The adapter blocks and refinement rounds belong to the jointly optimized model; they are not separate pretraining stages. ESM embeddings are precomputed. At inference, all designed CDR-position vectors are zeroed before adaptation, while framework vectors retain parent context.

Training terminology. An adapter block is an operation within one forward pass; a refinement round is a repeated model computation within that pass; and the reveal curriculum changes the sequence context across optimizer updates. None denotes a separately pretrained model. The from-scratch and fine-tuning experiments in App. N.1 compare alternative initialization routes for the revised head, rather than consecutive stages required by the joint objective.

M.10 Inference

1. Initialize CDR coordinates from framework templates (conserved-position interpolation).
2. Run $T = 3$ EGNN refinement rounds, applying translation nudges after rounds 1 and 2 at flow times $1/3$ and $2/3$, and decode $\hat{\mathbf{X}}, \hat{\mathbf{s}}$.
3. Decode the final sequence: greedy ($\arg\max$) or temperature-τ sampling.

Optional integration switch. The implementation also exposes post-EGNN midpoint flow integration through `flow_refine_steps`. This option is disabled in every reported experiment. The evaluated model uses only the intermediate translation feedback described above.

M.11 Training Details and Hyperparameters

Table 27: Complete hyperparameter list.

Group	Symbol	Value
Architecture	Hidden dim d	64
	EGNN layers L	3
	Velocity MP layers L_v	2
	Refinement rounds T	3
	k-NN neighbors	9
Adapter	LoRA rank r	16
	Attention heads n_h	4
	Dropout	0.1
	Atom attribute dim d_{attr}	16
Flow	Time schedule μ, σ_z	0.0, 1.0
	Time clamp $\sigma_{\min}$	10^{-4}
	$(1-s)$ clamp	10^{-3}
	Pred-mode perturbation σ_r, σ_p	0.1 rad, 0.5 Å
	Noise-mode translation σ_t	2.0 Å
	Mode mixing p_{pred}	0.5
	Flow-iter strength α	0.05
Loss weights	Flow λ_f	0.1
	Consistency λ_c	0.05
	Flow rot λ_ω, trans λ_v	1.0, 1.0
	Rotamer λ_{rot}	0.1
	Rotamer kernel σ_χ	30°
	Clash $\lambda_{\text{clash}}, \lambda_{\text{inter}}, d_{\text{thresh}}$	0.5, 2.0, 3.0 Å
	Sequence curriculum ρ_u	$0.5(\cos(\pi u/U_{\max})+1)\cdot 0.9$
Training	Optimizer	Adam
	LR schedule	$10^{-3} \to 10^{-4}$ exponential decay
	Grad clip	1.0
	Batch size (global)	32
	Epochs	200 (RAbD), 300 (six-CDR)

Hardware. Training uses 8 NVIDIA A100 80GB GPUs with PyTorch DDP. RAbD single-CDR training takes ∼6 hours; six-CDR training ∼9 hours. Inference is ∼1.2 s per target per sample (40 samples per target, single GPU).

Numerical stability. For stable training on small per-GPU batches, we rely on three protections: (i) we bypass edge operations in the EGNN and velocity network for empty edge lists; (ii) we compute distance norms as $\sqrt{\|\cdot\|^2 + 10^{-8}}$ for gradient stability at coincident atoms; (iii) we fall back to identity in Kabsch alignment when the covariance matrix is singular or has NaN entries.

N VELOCITY-HEAD COMPARISON AND CONDITIONAL EQUIVARIANCE

N.1 ARCHITECTURE COMPARISON AND INITIALIZATION CONTROLS

The RiMES architecture specified here uses the invariant relative-frame head of App. M.6.2. A coordinate-input predecessor serves as a diagnostic comparison: it concatenates shared features, the time embedding, global rotation entries, and translations, giving $d + 32 + 9 + 3 = 108$ inputs for $d = 64$. Its messages use pairwise distances and its translation output is read directly in ambient coordinates. This unconstrained coordinate dependence does not enforce Eq. 10. The invariant and predecessor heads have $224{,}230$ and $222{,}694$ parameters, respectively ($+0.69\%$).

With the revised head, we obtain CDR-H3 AAR 0.678 in an independent from-scratch run, within the predecessor's reported 0.65–0.73 seed range. Fine-tuning this head from the released model gives 0.729. We use these experiments to compare initialization routes for the head revision; equal AAR values can arise from different designs and training histories.

The following proof establishes the canonical head's transformation law under the stated assumptions. It does not establish symmetry of the complete generation pipeline.

N.2 REQUIRED TRANSFORMATION LAW

Let $g = (R_g, \mathbf{t}_g) \in \mathrm{SE}(3)$ with $R_g \in \mathrm{SO}(3)$, $\mathbf{t}_g \in \mathbb{R}^3$ act on the per-residue frames $\{(R_i, \mathbf{t}_i)\}_{i=1}^N$ by $g \cdot (R_i, \mathbf{t}_i) = (R_g R_i,\, R_g \mathbf{t}_i + \mathbf{t}_g)$. The rotation velocity $\hat{\boldsymbol{\omega}}_i$ is a *body-frame* tangent vector and the translation velocity $\hat{\mathbf{v}}_i$ is an ambient vector, so equivariance of the velocity field requires Eq. equation 10: $\hat{\boldsymbol{\omega}}_i$ invariant (it lives in the body frame) and $\hat{\mathbf{v}}_i$ co-rotating with R_g (and invariant to $\mathbf{t}_g$).

N.3 INVARIANT CONSTRUCTION

Let $\mathbf{H} = (\mathbf{h}_i) \in \mathbb{R}^{N \times d}$ be the backbone node features.

Assumption 1. *The backbone encoder is* $\mathrm{SE}(3)$*-invariant on node features:* $\mathbf{h}_i(g \cdot x) = \mathbf{h}_i(x)$ *for all* $g \in \mathrm{SE}(3)$. *This is an assumption of the proof; the fixed-feature head diagnostics below do not measure encoder-feature invariance.*

The head uses only the $\mathrm{SE}(3)$-invariant edge features $\mathbf{e}_{ij} = (\|\mathbf{t}_j - \mathbf{t}_i\|,\, R_i^\top(\mathbf{t}_j - \mathbf{t}_i),\, \mathrm{vec}(R_i^\top R_j))$ of Eq. equation 8, the invariant node features $\mathbf{h}_i$, and a transform-independent time embedding; message passing and readout are given by Eqs. equation 8–equation 9.

N.4 PROOF OF EQUIVARIANCE

Lemma 1 (Invariant edge features)**.** *For all* $g \in \mathrm{SE}(3)$, $\|\mathbf{t}_j - \mathbf{t}_i\|$, $R_i^\top(\mathbf{t}_j - \mathbf{t}_i)$ *and* $R_i^\top R_j$ *are invariant, and the* k*-NN neighborhood* $\mathcal{N}(i)$ *is invariant when distance ties use a transform-independent rule.*

Proof. Under g, $\mathbf{t}_j - \mathbf{t}_i \mapsto R_g(\mathbf{t}_j - \mathbf{t}_i)$, so $\|\mathbf{t}_j - \mathbf{t}_i\|$ is preserved (R_g orthogonal); $R_i^\top(\mathbf{t}_j - \mathbf{t}_i) \mapsto (R_g R_i)^\top R_g(\mathbf{t}_j - \mathbf{t}_i) = R_i^\top R_g^\top R_g(\mathbf{t}_j - \mathbf{t}_i) = R_i^\top(\mathbf{t}_j - \mathbf{t}_i)$; and $R_i^\top R_j \mapsto (R_g R_i)^\top(R_g R_j) = R_i^\top R_j$. Pairwise distances are preserved, so the set of k nearest neighbors of each i is unchanged. □

Lemma 2 (Invariant hidden states)**.** *Under Assumption 1,* $\mathbf{f}_i^{(\ell)}(g \cdot x) = \mathbf{f}_i^{(\ell)}(x)$ *for all layers* ℓ *and all* $g \in \mathrm{SE}(3)$.

Proof. By induction. Base: $\mathbf{f}_i^{(0)} = \mathrm{MLP}_{\mathrm{in}}([\mathbf{h}_i \| \phi(s)])$ depends only on the invariant $\mathbf{h}_i$ (Assumption 1) and the transform-independent time s, hence invariant. Step: if $\mathbf{f}_i^{(\ell)}$ is invariant, then by

Lemma 1 the arguments of $\mathrm{MLP}_{\mathrm{edge}}$ and the neighborhood $\mathcal{N}(i)$ are invariant, so each message $\mathbf{m}_{ij}^{(\ell)}$ and their mean $\bar{\mathbf{m}}_i^{(\ell)}$ are invariant, and hence $\mathbf{f}_i^{(\ell+1)}$ is invariant. □

Theorem 1. *The velocity head of Eqs. equation 8–equation 9 satisfies the transformation law equation 10 for all $g \in \mathrm{SE}(3)$ and all parameter values, provided the frames are valid and co-transform, encoder features are invariant, and graph construction is transform-independent.*

Proof. By Lemma 2, $\mathbf{f}_i^{(L_v)}$ is invariant, so $\hat{\boldsymbol{\omega}}_i(g \cdot x) = W_\omega \mathbf{f}_i^{(L_v)} = \hat{\boldsymbol{\omega}}_i(x)$ and $\tilde{\mathbf{v}}_i(g \cdot x) = \tilde{\mathbf{v}}_i(x)$. Therefore

$$\hat{\mathbf{v}}_i(g \cdot x) = R_i(g \cdot x)\, \tilde{\mathbf{v}}_i(g \cdot x) = (R_g R_i)\, \tilde{\mathbf{v}}_i(x) = R_g\big(R_i\, \tilde{\mathbf{v}}_i(x)\big) = R_g\, \hat{\mathbf{v}}_i(x).$$

No assumption is made on the weights, so equivariance holds throughout training and at convergence. □

N.5 Head-level translation invariance and perturbation sensitivity

Let $F_\theta(H, R, \mathbf{t}, s, \mathbf{b}) = (\hat{\boldsymbol{\omega}}, \hat{\mathbf{v}})$ denote the velocity head, with batch assignments $\mathbf{b}$. In this diagnostic, we hold H, R, s, and $\mathbf{b}$ fixed and change only the residue origins $\mathbf{t}$. We compare the angular and translational output tensors using

$$\mathrm{rel}(a, b) = \frac{\|a - b\|}{\|b\| + 10^{-30}}. \tag{38}$$

The output readouts receive nonzero random weights for this test. Their default zero initialization would otherwise yield a trivial comparison between identically zero outputs.

Test A: common translation. We add the same vector $\mathbf{c}$ to every residue origin. Pairwise displacements, distances, and relative rotations are unchanged, so the invariant head satisfies

$$F_\theta(H, R, \mathbf{t} + \mathbf{c}, s, \mathbf{b}) = F_\theta(H, R, \mathbf{t}, s, \mathbf{b})$$

for every parameter setting, under the graph-construction and frame assumptions stated above. This architectural translation-invariance property also applies to trained weights. A finite-precision output comparison is a numerical check of the property.

Test B: independent residue perturbations. We perturb each origin by an independent identically distributed displacement $\boldsymbol{\epsilon}_i$. These perturbations change relative geometry and can change the outputs. The relative differences illustrate sensitivity for the chosen weights and perturbations; they are weight-dependent and are not required to vanish by equivariance.

In both tests, we evaluate head outputs without alignment or structure decoding. We do not test global rotation equivariance, encoder-feature invariance, or end-to-end generated structures. Our rotation-equivariance statement is the conditional architectural result proved above. Kabsch placement in the ordinary sampler (App. P) remains separate from these diagnostics.

O Evaluation Metrics

We evaluate generated antibody CDR designs along seven axes: *sequence recovery* (§O.1), *structural accuracy* (§O.2), *physical plausibility* (§O.3), *sidechain quality* (§O.4), *interface-energy proxy* (§O.5), *reference-free designability* (§O.6), and *sequence diversity* (§O.7). We compute all metrics per sample unless noted otherwise and report the mean ± standard deviation across the test set. For sequence, structural, clash, and torsion metrics, we use raw generated structures without OpenMM relaxation or Rosetta sidechain repacking. For the interface-energy analysis in App. O.5, we use separate relaxed structures. We apply the same distinction to AbEgDiffuser, using its raw samples for the standard metrics and its separately supplied relaxed samples for the energy analysis.

O.1 Sequence Recovery Metrics

Amino Acid Recovery (AAR ↑). AAR measures the fraction of CDR positions whose predicted amino acid identity exactly matches the native sequence. For a predicted CDR sequence $\hat{\mathbf{s}} = (\hat{s}_1, \ldots, \hat{s}_L)$ and native sequence $\mathbf{s}^* = (s_1^*, \ldots, s_L^*)$,

$$\mathrm{AAR} = \frac{1}{L} \sum_{i=1}^{L} \mathbf{1}[\hat{s}_i = s_i^*], \tag{39}$$

where L is the CDR length. When multiple CDR loops are designed simultaneously, we concatenate all designed positions and compute a single AAR over the union. Per-CDR AAR is also reported separately (e.g., AAR H3).

Missing-output convention. For the main tables, we use the fixed sampling budget and assign zero recovery to missing generations. In the paired target-level analysis in Table 7, we use matched recovered records and report the corresponding evaluation counts. These summaries answer different statistical questions; a completion fraction alone does not support converting the paired intervals into main-table margins.

Contact Amino Acid Recovery (CAAR ↑). CAAR restricts the AAR calculation to *paratope* residues, i.e., CDR positions that are in contact with the antigen epitope. A CDR residue i is defined as contacting the antigen if $\min_{j \in \text{epitope}} d(i,j) < d_{\text{contact}}$, where $d(i,j)$ is the minimum heavy-atom distance between residues i and j, and $d_{\text{contact}} = 6.6\,\text{Å}$. Formally,

$$\mathrm{CAAR} = \frac{\sum_{i=1}^{L} \mathbf{1}[\hat{s}_i = s_i^*] \cdot c_i}{\sum_{i=1}^{L} c_i + \epsilon}, \tag{40}$$

where $c_i \in \{0, 1\}$ indicates whether position i contacts the antigen and $\epsilon = 10^{-10}$ prevents division by zero. CAAR captures sequence recovery at the binding interface, which is more functionally relevant than overall AAR.

BLOSUM62 Score (↑). Binary AAR treats all mismatches equally. The BLOSUM62 score captures *biochemical similarity*: conservative substitutions (e.g., Leu → Ile) receive higher scores than disruptive ones (e.g., Leu → Asp). We normalize the raw BLOSUM62 matrix entry to $[0, 1]$ per position:

$$\mathrm{BLOSUM62} = \frac{1}{L} \sum_{i=1}^{L} \frac{B(\hat{s}_i, s_i^*) - \min_a B(s_i^*, a)}{\max\big(B(s_i^*, s_i^*),\, B(\hat{s}_i, \hat{s}_i)\big) - \min_a B(s_i^*, a)}, \tag{41}$$

where $B(a, b)$ is the standard BLOSUM62 substitution score between amino acids a and b, and the minimum is taken over all 20 standard amino acids. A score of 1.0 indicates perfect identity; conservative mutations score close to 1; disruptive mutations score close to 0.

O.2 Structural Accuracy Metrics

RMSD (CDR) (↓). For each designed CDR, we independently fit its predicted and native Cα coordinates using the Kabsch algorithm (Kabsch, 1976). Let $\mathcal{M}_c$ denote the residues of loop c. Its alignment and RMSD are

$$(R_c^*, \mathbf{t}_c^*) = \arg \min_{R \in \mathrm{SO}(3),\, \mathbf{t}} \sum_{i \in \mathcal{M}_c} \|R\hat{\mathbf{x}}_i + \mathbf{t} - \mathbf{x}_i^*\|^2,$$

$$\mathrm{RMSD}_c = \sqrt{\frac{1}{|\mathcal{M}_c|} \sum_{i \in \mathcal{M}_c} \|R_c^*\hat{\mathbf{x}}_i + \mathbf{t}_c^* - \mathbf{x}_i^*\|^2}. \tag{42}$$

For the aligned-CDR metric in `cal_metrics_all.py`, we use this independent fit and keep the whole-antibody fit separate; we do not reuse it for the CDR score. For multi-CDR summaries, we average the separately fitted loop RMSDs. This measures local loop shape rather than placement relative to the framework or antigen. We report antigen-frame placement errors separately in App. H.

DockQ (↑). DockQ (Mirabello & Wallner, 2024) compares modeled and native complex interfaces. The supplied `cal_metrics_all.py` calls the evaluator with `cdrh3_only=False`, passing complete complexes rather than an isolated CDR. Its wrapper invokes DockQ with `-allowed_mismatches 50 -short` and reads the total score over native interfaces, with automatic chain mapping. This aggregate can include antibody heavy–light interfaces as well as antibody–antigen interfaces; it is not an H3-specific placement score. The wrapper falls back to the first parsed interface score if the total line is absent, and returns NaN on execution or parsing failure. We use the reported DockQ values as a secondary complex-level diagnostic alongside the CDR-specific metrics, without inferring per-interface accuracy from the aggregate.

TM-score (↑). TM-score (Zhang & Skolnick, 2004) provides a length-normalized measure of structural similarity between the model and reference antibody (both heavy and light chains merged into a single chain). Unlike RMSD, TM-score is length-independent and ranges in $(0, 1]$:

$$\text{TM-score} = \frac{1}{L_{\text{ref}}} \sum_{i=1}^{L_{\text{aligned}}} \frac{1}{1 + \big(d_i / d_0(L_{\text{ref}})\big)^2}, \tag{43}$$

where d_i is the distance between i-th aligned residue pair after optimal superposition, and $d_0(L) = 1.24\sqrt[3]{L - 15} - 1.8$ is a length-dependent normalization. We use the Zhang Lab TM-score executable `https://aideepmed.com/TM-score/` applied to the full antibody Fv region.

lDDT (↑). The local Distance Difference Test (lDDT) (Mariani et al., 2013) evaluates structural accuracy without requiring global superposition, making it robust to domain-level rigid body shifts. For each residue i, lDDT examines all inter-residue distances d_{ij} within an inclusion radius $R_0 = 15$ Å in the reference structure and checks how many of these distances are preserved (within tolerance thresholds $\{0.5, 1, 2, 4\}$ Å) in the model structure:

$$\text{lDDT} = \frac{1}{4|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \sum_{\tau\in\{0.5,1,2,4\}} \mathbf{1}\big[|d_{ij}^{\text{model}} - d_{ij}^{\text{ref}}| < \tau\big], \tag{44}$$

where $\mathcal{P} = \{(i, j) : d_{ij}^{\text{ref}} < R_0\}$. We use the `lddt` binary with the `-x` flag (stereochemical check) applied to the full antibody. The global lDDT score is the average over all residues.

O.3 Physical Plausibility Metrics

CDR Clashes (↓). Steric clashes indicate physically implausible atom overlaps. We compute **intra-CDR Cα clashes**: the number of Cα–Cα pairs within the designed CDR loop(s) with distance below 3.0 Å (i.e., $2 \times r_{\text{vdW}}^{\text{C}} - \delta$, where $r_{\text{vdW}}^{\text{C}} = 1.7$ Å and $\delta = 0.4$ Å tolerance), excluding sequential neighbors with residue separation < 2:

$$N_{\text{CDR-clash}} = \big|\{(i, j) : i, j \in \text{CDR},\ i < j,\ |i - j| \geq 2,\ \|\mathbf{x}_i^{C\alpha} - \mathbf{x}_j^{C\alpha}\| < 3.0\,\text{Å}\}\big|. \tag{45}$$

Interface Clashes[aa] (↓). We count all-atom (heavy atom) clashes between the designed CDR residues and antigen residues, using a stricter threshold of 2.0 Å:

$$N_{\text{interface}}^{\text{aa}} = \big|\{(a, b) : a \in \text{CDR}_{\text{heavy}},\ b \in \text{Ag}_{\text{heavy}},\ \|\mathbf{x}_a - \mathbf{x}_b\| < 2.0\,\text{Å}\}\big|, \tag{46}$$

We include only non-hydrogen atoms in these sets and count atom pairs, not residues, using a fixed distance cutoff rather than atom-type-dependent van der Waals radii. We also report a separate Cα-only interface clash count (threshold 3.0 Å).

Van der Waals overlap in the grounding diagnostic. The all-atom grounding analysis in Table 20 uses atom-type-dependent overlap. For heavy atoms a, b at distance d_{ab}, the positive overlap is

$$o_{ab} = \max\{0, r_a + r_b - 0.4\,\text{Å} - d_{ab}\}. \tag{47}$$

The radii are $r_{\text{C}} = 1.70$, $r_{\text{N}} = 1.55$, $r_{\text{O}} = 1.52$, and $r_{\text{S}} = 1.80$ Å, with a 3.2 Å neighbor-search cutoff. Positive overlaps count as clashes. For one generated structure, the worst H3–antigen overlap is the maximum o_{ab} over pairs with one atom in CDR-H3 and the other in the antigen, or zero if no pair overlaps. It measures the deepest single contact; the table averages these per-structure maxima within targets and then across targets. This diagnostic differs from the fixed 2.0 Å pair count in Eq. 46.

Coarse torsion-region fraction (Rama ↑). We compute the fraction of CDR residues whose backbone dihedral angles (ϕ, ψ) pass the angular rules specified below. The dihedral angles are calculated from four consecutive backbone atoms using the standard atan2-based formula:

$$\phi_i = \mathrm{atan2}\big((\mathbf{n}_1 \times \mathbf{n}_2) \cdot \hat{\mathbf{u}}_1,\ \mathbf{n}_1 \cdot \mathbf{n}_2\big), \tag{48}$$

where $\mathbf{n}_1, \mathbf{n}_2$ are normals to consecutive peptide planes and $\hat{\mathbf{u}}_1$ is the normalized bond vector. ϕ_i is defined by atoms C_{i-1}–N_i–$\mathrm{C}\alpha_i$–C_i and ψ_i by N_i–$\mathrm{C}\alpha_i$–C_i–N_{i+1}.

Outlier classification follows simplified Ramachandran rules:

- **Glycine**: no restriction (almost all (ϕ, ψ) allowed).
- **Proline**: outlier if $\phi > 0^\circ$ or $\phi < -100^\circ$.
- **General**: $\phi > 0^\circ$ is outlier unless $30^\circ \leq \phi \leq 90^\circ$ and $10^\circ \leq \psi \leq 70^\circ$ (left-handed α-helical region).

The coarse torsion-region fraction is reported as $\mathrm{Rama} = 1 - N_{\mathrm{outlier}}/N_{\mathrm{total}}$.

O.4 Sidechain Quality Metrics

χ_1 MAE (°) (↓). The χ_1 dihedral angle is the first sidechain torsion (N–Cα–Cβ–X$_\gamma$). We compute the mean absolute error (in degrees) between predicted and native χ_1 angles over eligible, type-matched CDR positions:

$$\chi_1\text{-MAE} = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \min\big(|\hat{\chi}_{1,i} - \chi^*_{1,i}|,\ 360^\circ - |\hat{\chi}_{1,i} - \chi^*_{1,i}|\big), \tag{49}$$

where $\mathcal{S}$ contains positions with the same predicted and native amino-acid identity and a defined χ_1 in both structures. We require all four dihedral atoms to be present. We exclude positions with a type mismatch, an undefined angle, or missing required atoms from both numerator and denominator, rather than counting them as failed recoveries. Gly and Ala have no defined χ_1. The minimum accounts for angular periodicity. The dihedral is computed from four atom positions using the atan2-based formula (Eq. 48), applied to atoms N, Cα, Cβ, X$_\gamma$.

Rotamer Accuracy (Rot. Acc. ↑). Rotamer accuracy measures the fraction of eligible CDR residues whose predicted χ_1 has circular angular error below 40° from the native value. This is a native-angle recovery criterion with a strict 40° threshold. It is distinct from assigning the predicted and native angles to the same Dunbrack rotamer basin, and from the reference-free canonicality statistic in App. H:

$$\mathrm{Rot.\,Acc.} = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \mathbf{1}\Big[\min\big(|\hat{\chi}_{1,i} - \chi^*_{1,i}|,\ 360^\circ - |\hat{\chi}_{1,i} - \chi^*_{1,i}|\big) < 40^\circ\Big]. \tag{50}$$

O.5 Interface-Energy Proxy

We assess binding-energy plausibility through the change in interface energy between the designed and the native antibody–antigen complex under a relax-then-score in-silico pipeline.

Pipeline. The energy protocol consists of two operations. (i) *Energy minimisation* in implicit solvent using OpenMM with the Amber14 forcefield, 1,000 L-BFGS steps at constant temperature, with the antibody–antigen heavy atoms restrained at 5 kcal mol^{-1} Å^{-2} to avoid gross conformational drift. (ii) *Interface scoring* using the PyRosetta `InterfaceAnalyzerMover` with the standard `ref2015` energy function, taking the `dG_separated` field as the interface free-energy estimate ΔG (computed as the energy difference between the bound complex and the same chains held apart while sidechains are repacked). For AbEgDiffuser, we use the supplied OpenMM-relaxed structures directly in operation (ii), without further minimization. We keep them separate from the raw structures used for its sequence, structural, clash, and torsion metrics (App. K). For other generated outputs, we relax the structures before scoring.

Per-design $\Delta\Delta G$. For each design i we compute

$$\Delta\Delta G_i \ = \ \Delta G_i^{\mathrm{design}} \ - \ \Delta G^{\mathrm{native}}, \tag{51}$$

where ΔG^{native} is obtained by passing the native complex of the same target through the stated energy pipeline. Designs with $\Delta\Delta G_i < 0$ have a lower computed interface-energy score than the native reference; this is not a measured binding comparison.

Completion correction. We aggregate energy results as the fraction of the configured attempt budget for which a design is successfully processed and has a lower interface score than its reference:

$$\%\,\text{improved}^{\text{cc}} = \frac{|\{i : \text{scored}_i \text{ and } \Delta\Delta G_i < 0\}|}{N_{\text{attempted}}}. \tag{52}$$

We use a RAbD attempt budget of $60 \times 40 = 2400$. Missing scores contribute zero to the numerator, equivalent to assigning them $+\infty$ for this indicator. Our fixed-denominator measure combines scoring completion with the energy condition; we do not interpret it as a direct measurement of molecular validity or binding affinity.

O.6 Reference-Free Designability Metrics

The recovery metrics in §O.1 are bounded above by the ground-truth sequence and therefore cannot detect designs that score *better* than the natural antibody. We complement them with seven *reference-free* scores that depend only on the designed CDR-H3 sequence: four protein language model perplexities (ESM-2, ProGen2, AbLang2, IgLM), a human-repertoire amino-acid frequency proxy, CDR-H3 net charge, and mean Kyte–Doolittle hydrophobicity.

ESM-2 Pseudo-Perplexity (↓). ESM-2 (Lin et al., 2023) is a transformer protein language model trained on UniRef50 with a masked-residue objective. We use `esm2_t33_650M_UR50D` (650M parameters). For a CDR-H3 sequence $\hat{\mathbf{s}} = (\hat{s}_1, \ldots, \hat{s}_L)$ we compute the pseudo-perplexity, masking one position at a time:

$$\text{PPL}_{\text{ESM-2}}(\hat{\mathbf{s}}) = \exp\left(-\frac{1}{L} \sum_{i=1}^{L} \log P_{\text{ESM-2}}\big(\hat{s}_i \mid \hat{\mathbf{s}}_{\setminus i}\big) \right), \tag{53}$$

where $\hat{\mathbf{s}}_{\setminus i}$ denotes the sequence with position i replaced by the `[MASK]` token. Lower values indicate sequences that the model finds more typical of natural proteins. A general (non-antibody) protein PLM scores "protein-like" naturalness rather than "antibody-like".

ProGen2 Perplexity (↓). ProGen2 (Nijkamp et al., 2023) is an autoregressive protein language model trained on UniRef90. We use `progen2-small` (151M parameters) and compute the standard left-to-right perplexity:

$$\text{PPL}_{\text{ProGen2}}(\hat{\mathbf{s}}) = \exp\left(-\frac{1}{L} \sum_{i=1}^{L} \log P_{\text{ProGen2}}\big(\hat{s}_i \mid \hat{s}_{<i}\big) \right). \tag{54}$$

Like ESM-2, ProGen2 was not trained specifically on antibodies and therefore measures general protein naturalness.

AbLang2 Pseudo-Perplexity (↓). AbLang2 (Olsen et al., 2024) is a masked language model trained on paired heavy- and light-chain antibody sequences from the OAS database (Olsen et al., 2022). We compute its pseudo-perplexity analogously to Eq. 53, replacing $P_{\text{ESM-2}}$ with P_{AbLang2}. These scores measure model-specific antibody-sequence likelihood. They do not establish that an evaluated sequence was present in the training corpus or that a higher-likelihood variant has improved function.

IgLM Perplexity (↓). IgLM (Shuai et al., 2023) is an autoregressive antibody language model conditioned on a chain token (`[HEAVY]`/`[LIGHT]`) and species token (`[HUMAN]`/etc.). We score with `[HEAVY][HUMAN]` conditioning and compute the analogue of Eq. 54 under P_{IgLM}.

For all four PPL metrics we evaluate on the CDR-H3 segment only, so the two general PLMs and the two antibody PLMs all see the same input and compute perplexity over the same residues; pseudo-PPL and true PPL are on different scales and we never compare across rows of Table 12.

Human-repertoire amino-acid frequency proxy (↑). We compute a human-repertoire amino-acid frequency proxy as the geometric mean of per-position human CDR-H3 amino-acid frequencies:

$$\text{Humanness}(\hat{\mathbf{s}}) = \exp\left(\frac{1}{L}\sum_{i=1}^{L}\log f_{\text{human}}(\hat{s}_i)\right), \tag{55}$$

where $f_{\text{human}}(a)$ is the marginal frequency of amino acid a at any CDR-H3 position in the OAS-derived human heavy-chain distribution; rare amino acids receive a Laplace floor of $\epsilon = 5{\times}10^{-3}$ to avoid undefined logarithms. This is a deliberately simple proxy: for clinical-grade humanness one would use Hu-mAb or OASis on the full Fv.

CDR-H3 Net Charge. Net charge at pH 7, computed by summing per-residue charges over the CDR-H3:

$$\text{Charge}(\hat{\mathbf{s}}) = \sum_{i=1}^{L} q_{\text{pH7}}(\hat{s}_i), \tag{56}$$

with $q_{\text{pH7}}(\text{K}) = q_{\text{pH7}}(\text{R}) = +1$, $q_{\text{pH7}}(\text{D}) = q_{\text{pH7}}(\text{E}) = -1$, $q_{\text{pH7}}(\text{H}) = +0.5$, all others 0. We report this as a constraint rather than an optimisation target: developability typically requires the design's charge to lie within the natural range, not to deviate strongly in either direction.

Mean Kyte–Doolittle Hydrophobicity. We average the Kyte–Doolittle (Kyte & Doolittle, 1982) hydrophobicity scale over the CDR-H3:

$$\text{Hydro}(\hat{\mathbf{s}}) = \frac{1}{L}\sum_{i=1}^{L} h_{\text{KD}}(\hat{s}_i), \tag{57}$$

with values $h_{\text{KD}} \in [-4.5, 4.5]$ ranging from arginine (-4.5) to isoleucine ($+4.5$). As with charge we treat this as a distributional constraint, not a target.

Aggregation across designs. For a single setting (e.g. T=0.1) we report the mean over all $N_{\text{targets}} \times N_{\text{samples}}$ designs. We additionally report two GT-comparison metrics that exploit the reference-free property: (i) the fraction of designs whose score beats the GT for the same target, averaged over targets, and (ii) the fraction of targets for which at least one of the N_{samples} designs beats the GT (best-of-N).

O.7 Sequence Diversity Metrics

When generating multiple candidates per target, we measure both exact sequence diversity and clustering under the selected decoding rule.

Sequence Diversity (MMseqs2 Cluster Ratio ↑). For each target antigen, we cluster all generated CDR sequences using MMseqs2 (Steinegger & Söding, 2017) with a sequence identity threshold of 90% and coverage of 80% (greedy set cover mode). The **cluster ratio** is defined as:

$$\text{Diversity}_{\text{cluster}} = \frac{N_{\text{clusters}}}{N_{\text{samples}}}, \tag{58}$$

where N_{clusters} is the number of sequence clusters and N_{samples} is the total number of generated sequences for that target. A ratio of 1.0 means all sequences are distinct at 90% identity; a ratio close to 0 indicates a concentrated candidate pool under the chosen decoding rule. Concentration under greedy decoding alone does not establish collapse during training. We also report the number of unique sequences ($N_{\text{unique}} = |\{\hat{\mathbf{s}}\}|$).

Vendi Diversity Score (↑). The Vendi score (Friedman & Dieng, 2023) provides a kernel-based measure of effective sample diversity. Given N generated sequences, we construct an $N \times N$ pairwise similarity matrix K where K_{ij} is the normalized Needleman–Wunsch alignment score between sequences i and j:

$$K_{ij} = \frac{\text{Align}(\hat{\mathbf{s}}_i, \hat{\mathbf{s}}_j)}{\max(|\hat{\mathbf{s}}_i|, |\hat{\mathbf{s}}_j|)}. \tag{59}$$

The Vendi score is then the exponential of the von Neumann entropy of the normalized kernel matrix:

$$\mathrm{Vendi}(K) = \exp\Big(-\sum_k \lambda_k \log \lambda_k\Big), \tag{60}$$

where λ_k are the eigenvalues of K/N. The Vendi score ranges from 1 (all samples identical) to N (all samples maximally dissimilar), and can be interpreted as the effective number of distinct samples.

AA Distribution KL Divergence (↓). To assess whether generated sequences exhibit natural amino acid usage, we compute the Kullback–Leibler divergence between the amino acid frequency distribution of the generated CDR sequences and the natural human CDR-H3 distribution aggregated from the Observed Antibody Space (OAS) database (Olsen et al., 2022):

$$D_{\mathrm{KL}}(P_{\mathrm{gen}} \| P_{\mathrm{nat}}) = \sum_{a \in \mathcal{A}} P_{\mathrm{gen}}(a) \log \frac{P_{\mathrm{gen}}(a)}{P_{\mathrm{nat}}(a)}, \tag{61}$$

where $\mathcal{A}$ is the set of 20 standard amino acids. Laplace smoothing ($\epsilon = 10^{-8}$) is applied to P_{gen} to avoid undefined logarithms. We additionally report the symmetric Jensen–Shannon divergence $D_{\mathrm{JS}} = \frac{1}{2} D_{\mathrm{KL}}(P_{\mathrm{gen}} \| M) + \frac{1}{2} D_{\mathrm{KL}}(P_{\mathrm{nat}} \| M)$ where $M = \frac{1}{2}(P_{\mathrm{gen}} + P_{\mathrm{nat}})$.

O.8 Summary of Metrics and Notation

Table 28 summarizes all evaluation metrics, their desired direction, and the section in which each is defined.

Table 28: Summary of evaluation metrics.

Category	Metric	Direction	Description
Sequence	AAR	↑	Exact amino acid match rate
	CAAR	↑	Match rate at antigen-contact positions
	BLOSUM62	↑	Biochemically-aware substitution score
Structure	RMSD (CDR)	↓	Cα RMSD after per-CDR Kabsch fitting
	TM-score	↑	Length-normalized structural similarity
	lDDT	↑	Local distance difference test
Physics	CDR clashes	↓	Intra-CDR Cα steric clashes
	Interface clashes[aa]	↓	CDR–antigen all-atom clashes
	Rama	↑	Coarse torsion-region fraction
Sidechain	χ_1 MAE (°)	↓	Mean absolute χ_1 dihedral error
	Rot. Acc.	↑	Fraction with native χ_1 error $< 40°$
Binding	$\Delta\Delta G < 0$ (%, cc)	↑	PyRosetta InterfaceAnalyzer; completion-corrected
PLM PPL	ESM-2 PPL	↓	650M masked PLM, pseudo-PPL on CDR-H3
	ProGen2 PPL	↓	151M autoregressive PLM, true PPL
	AbLang2 PPL	↓	Antibody masked PLM, pseudo-PPL
	IgLM PPL	↓	Antibody autoregressive PLM, true PPL
Designability	Human-repertoire AA frequency	↑	Geometric-mean human CDR-H3 AA frequency
	Charge	constraint	CDR-H3 net charge at pH 7
	Hydrophobicity	constraint	Mean Kyte–Doolittle
Diversity	Cluster ratio	↑	MMseqs2 clusters / samples (90% id)
	Vendi score	↑	Effective number of distinct sequences
	D_{KL}(AA)	↓	AA distribution vs. natural CDR-H3

P Inference Pipeline

Sampling. The model's `sample` method runs $T = 3$ iterative refinement rounds. CDR coordinates are initialized from the framework template. In each round:

1. Feature extraction with the equivariant adapter (in PLM-augmented mode, PLM embeddings are zeroed at CDR positions; PLM-free mode uses no PLM input).
2. Message passing through the dual-stream EGNN.
3. Coordinate update at CDR positions.
4. Sequence update: $\arg\max$ on the final round; softmax distribution on intermediate rounds (for label smoothing in subsequent rounds).
5. After each non-final round, apply the translation-only Cα nudge in Eq. 11. No nudge follows the final round.

Kabsch alignment. After all refinement rounds, we align the predicted CDR backbone (paratope) to the shadow paratope coordinates via Kabsch superposition and apply the resulting rigid transformation to the entire antibody. This fits the antibody to our model's predicted interface coordinates; it does not by itself establish a correct binding pose or the symmetry of the complete pipeline (App. N).

Selection criterion. We compute RiMES's aggregate metrics without selecting a single design per target. Unless stated otherwise, we use a budget of 40 designs per target; in the torsion-supervision study in Table 20, we use the same budget on 749 torsion-eligible targets, for $29{,}960$ designs per condition. We identify best-of-40 summaries explicitly. Our implementation also exposes a lowest-pRMSD selection across multiple trials for backwards compatibility with the dyMEAN codebase, but we do not use it in any reported number.

Temperature sampling. When $\tau > 0$, we decode sequences via temperature-scaled sampling: $s_i \sim \text{Categorical}(\text{softmax}(\hat{\mathbf{s}}_i/\tau))$. $\tau = 0$ (default) corresponds to greedy $\arg\max$ decoding.

Optional integration. The implementation exposes post-refinement flow integration and a noise-start option (App. M.10). Neither is used in the reported evaluations; standard inference follows the iterative procedure above.

Q Reproducibility

The accompanying source snapshot contains the available model modules, metric routines, training and generation entry points, command examples, and data-preprocessing scripts, with a SHA-256 file manifest. It is a documented source snapshot; its identity with the historical implementation of every reported run has not been established. The manuscript specifies the selected model conventions, including active training nudges, the joint objective, and the evaluation definitions.

The unresolved aggregate/paired, multi-CDR aggregation, and torsion-extension correspondences are stated alongside those analyses. Unreconciled perplexity-stratified records are excluded from evidence for the PLM effect; the masked-input retraining control is retained. These reporting limits do not justify substituting inferred scores for the supplied measurements.

SAbDab data access: `https://opig.stats.ox.ac.uk/webapps/sabdab-sabpred/sabdab`. The separate interface-energy evaluation uses PyRosetta; installation instructions are at `https://www.pyrosetta.org/`. No public or anonymous code URL is specified in this manuscript version.