ADEL: TRUTH-ANCHORED DISTILLATION WITH ENGH-HUBER-INFORMED STRAIN AND LINKAGE CHEMISTRY FOR MACROCYCLE DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Designing macrocyclic peptide binders requires more than generating a plausible peptide structure: the ring must close between chemically compatible atoms, and for some linkage types, the identities of these atoms depend on the sequence being designed. In this work, we propose **ADEL**, an all-atom sequence–structure co-design diffusion model that learns to satisfy this coupled geometric and chemical constraint. Our approach combines truth-anchored distillation with Engh–Huber-informed closure geometry and explicit linkage-chemistry supervision. Instead of directly distilling the projection of a sampling-time closure corrector, which we show provides a training signal redundant with the existing closure loss, we distill the model's denoising response after projection and renoising. Because this teacher is produced by the model itself, we introduce a per-atom truth gate that retains a teacher target only when it is closer to the reference structure than the student prediction, preventing the self-distillation objective from drifting during training. We further supervise the local geometry around the closing bond and the residue identities required by sequence-dependent linkages. We retrain seven published methods on the same data split and evaluate all generated designs using a common type-aware protocol. On leakage-controlled CPSea targets, our model achieves a macrocycle closure rate of 47.2%, compared with 35.4% for the strongest retrained baseline, while maintaining 99.8% structural validity. Since not all baselines can explicitly represent sequence-dependent linkages, we interpret the aggregate comparison as multi-chemistry coverage rather than a fully matched constraint-solving benchmark. Across head-to-tail, isopeptide, and disulfide linkages, ADEL maintains comparable closure performance within a single co-design framework.

1 INTRODUCTION

Macrocyclic peptides occupy a chemical space that neither antibodies nor small molecules reach: an antibody-sized interface on a molecule small enough to synthesise chemically, rigidified by a covalent ring that raises affinity and resists proteolysis. That ring is also what makes them hard to design. A macrocyclic binder is only a macrocycle if two specific atoms form a covalent bond, which is a hard geometric constraint on a model that otherwise places coordinates freely.

All-atom diffusion models now generate sequence and structure jointly, conditioned on the target (Stark et al., 2025; Abramson et al., 2024; Watson et al., 2023; Pacesa et al., 2025). For macrocycles this is necessary rather than convenient, because the closure constraint is *circular*: which two atoms must bond is set by the macrocyclisation chemistry, and for two of the three chemistries in common use that chemistry is a property of the sequence, an isopeptide bond requiring a lysine and an aspartate and a disulfide two cysteines (Kang & Baker, 2011; Costa et al., 2025). The constraint is thus not well posed until the sequence is known, and the sequence is what is being designed (Rettie et al., 2025). A pipeline that builds a backbone and then assigns a sequence to it cannot keep the two consistent.

Progress here is harder to read than it appears. Cyclization is usually measured with a single distance window applied to every linkage type, scoring a design as cyclic whenever two atoms sit at roughly the right separation regardless of whether they could form the bond in question: right shape, wrong element. Recent work conditions cyclic peptide generation on geometric constraints

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.

directly, either zero-shot from composable constraint sets (Jiang et al., 2025) or with a reward-aligned objective (Zhang et al., 2026), which makes a criterion that distinguishes the linkage types more rather than less necessary. We therefore rescored seven recent methods (Nguyen et al., 2026; Kong et al., 2024; 2025; Wang et al., 2024; Li et al., 2024; Jiang et al., 2025; Stark et al., 2025) on 226 held-out targets under one type-aware protocol, released as a contribution in its own right. Under it, closure ranges from 0.9% to 35.4% and the best validity from any fully evaluable baseline is 76.5%. The headroom is larger than the literature indicates, and a type-aware criterion is what makes it visible.

We present **ADEL**, truth-**A**nchored **D**istillation with **E**ngh-Huber strain and **L**inkage chemistry: an all-atom diffusion model for macrocyclic peptide binder co-design, trained end to end under a unified formulation that internalises a constrained diffusion corrector and jointly accounts for closure geometry and linkage chemistry.

Contributions.

1. **Truth-anchored multi-head distillation.** A training-free corrector in the constrained-diffusion family (Huang et al., 2024; Rochman-Sharabi & Louppe, 2026; Christopher et al., 2024) raises closure at sampling time. Distilling it is subtler than it appears: the obvious target, the projection, is exactly proportional to a penalty the model already carries (Lemma 1), and a target produced by the model being trained leaves the objective with no attractor at the data. We therefore distil the corrector's *second denoising call* and gate it per atom against the native structure, which we prove gives a descent direction toward the data (Lemma 2); several teachers at different projection strengths, selected per atom, can only enlarge the supervised set (Prop. J.4).

2. **Closure anchored to physical bond lengths.** A closure constraint is one scalar, and a ring can meet that distance exactly while forming a local geometry no chemist would accept. We pair the distance hinge with a second on the two $\mathrm{C}\alpha$-referenced pseudo-angles flanking the closing bond, centred on the tabulated tetrahedral value for a backbone carbon with a $\pm 25^\circ$ band. These coincide with covalent valence angles for head-to-tail closure and summarise side-chain placement for the other two chemistries, so the term is best read as a guard against gross distortion of the closure frame rather than a chemistry-specific strain prior. Its centre is tabulated rather than swept against our evaluation.

3. **Linkage-chemistry supervision.** Which two atoms must bond follows from residue identity, and no coordinate objective supplies it: over one training run we observe every geometry metric improving while disulfide closure collapses from 41.0% to 1.3%, because only one of the two carries a gradient. We supervise residue identity at the linkage positions directly. That this is structural rather than incidental is confirmed by a general inverse-folding model applied to our designs, which strips the required residues from 94.9% of disulfide and 30.5% of isopeptide macrocycles even with cysteine explicitly permitted, while leaving head-to-tail designs untouched.

The three form a chain, each supplying what the next presupposes: chemistry determines *which* atoms bond, the closure constraints fix *where* they go, and distillation supplies the global rearrangement that satisfying a local constraint implies.

Results. Against the seven rescored baselines, ADEL leads on closure (47.2% against 35.4%), predicted binding (median -20.1 against -13.7 REU), designs beating the native binder (30.5% against 7.5%), both covalent-geometry Z-scores (bond-length 1.091 against 4.050; bond-angle 1.515 against 2.161) and validity (99.8% against 76.5%; BoltzGen's full pipeline reaches 77.0% well-formedness but its stereochemistry cannot be evaluated). Every resolved distance failure lies within 0.50 Å of the closure window. Sec. H analyses the model itself: residual closure error is variance rather than bias, and a variance decomposition over targets explains why ten samples cover 99.8% of them while selection gains 12.5 REU yet barely moves geometry.

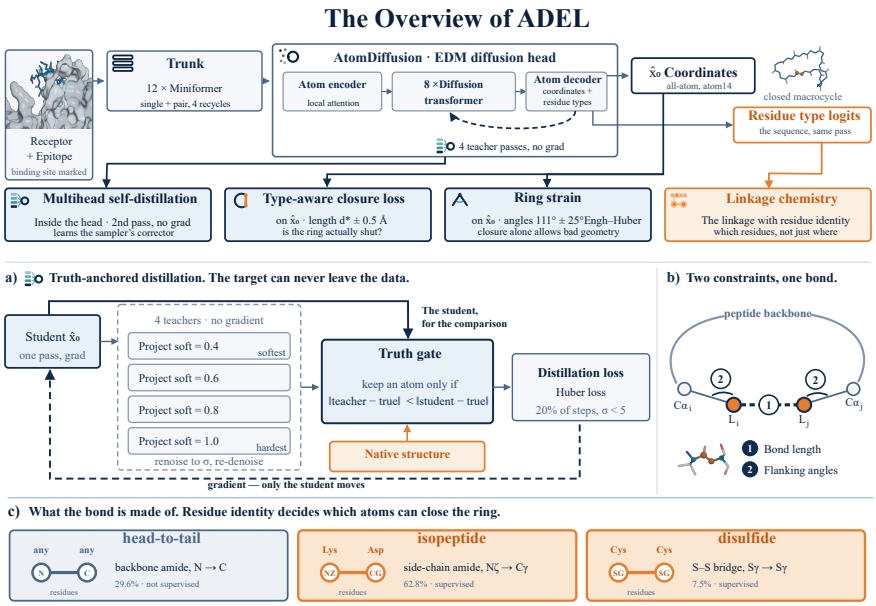


Figure 1: Overview of our method. **Top:** a trunk of twelve Miniformer blocks conditions an EDM diffusion head emitting all-atom coordinates $\hat{x}_0$ and residue-type logits in one pass; our three contributions attach as the training objectives in the band beneath. **(a)** Truth-anchored distillation: four teachers project at strengths $\alpha \in \{0.4, \ldots, 1.0\}$ and renoise, and a truth gate keeps an atom only where the teacher is closer to the native than the student. The teacher branch carries no gradient. **(b)** Two constraints on one bond — its length, and the two flanking angles the length alone leaves free. **(c)** What the bond is made of. Head-to-tail closure joins backbone termini and needs no particular residue; isopeptide and disulfide require residues the model must choose, and only those two are supervised. Percentages are the share of held-out targets carrying each chemistry.

2 METHOD

ADEL adopts an all-atom diffusion architecture that jointly co-designs peptide sequence and receptor-bound structure. Every parameter of the network is trainable and is updated on the CPSea training split under the objective of Eq. 2, which combines truth-anchored distillation, physically anchored closure geometry and linkage-chemistry supervision. The terms are introduced in three stages rather than all at once, each warm-starting from the previous checkpoint, for the reason given in Sec. 2.1; App. K, Table 24 specifies the stages. Figure 1 in App. J gives an overview of the complete formulation; formal results are deferred to Appendix J, the ablation evidence to Appendix E, and the notation is collected in Table 4.

2.1 SETTING AND NOTATION

A macrocyclic binder must satisfy a constraint an acyclic one does not: two designated atoms i and j must sit at the covalent bonding distance d^* set by the linkage chemistry, within a tolerance τ. The admissible set is

$$\mathcal{C}_s \;=\; \big\{\, x \in \mathbb{R}^{N\times 3} \;:\; \big|\, \|x_i - x_j\| - d^* \big| \le \tau \,\big\}. \tag{1}$$

We inherit the EDM preconditioning unchanged (Karras et al., 2022) and write $D_\theta(x_\sigma, \sigma)$ for the preconditioned denoiser, $\hat{x}_0$ for its denoised prediction and x^* for the native structure; the coefficients are restated in App. J. Our objective is

$$\mathcal{L} = \mathcal{L}_{\text{base}} + \underbrace{\lambda_{\text{D}}\mathcal{L}_{\text{distill}}}_{\text{truth-anchored distillation}} + \underbrace{\lambda_{\text{close}}\mathcal{L}_{\text{close}} + \lambda_{\text{E}}\mathcal{L}_{\text{strain}}}_{\text{physically anchored closure}} + \underbrace{\lambda_{\text{L}}\mathcal{L}_{\text{chem}}}_{\text{linkage chemistry}}, \tag{2}$$

with $\mathcal{L}_{\text{base}}$ the diffusion and sequence loss of the backbone.

The order in which the terms are introduced is part of the method. Equation 2 is the objective at convergence, not a single run from a common initialisation. The staging is not incidental: $\mathcal{L}_{\text{chem}}$ fires only on batches whose linkage chemistry the model currently gets wrong, so its gradient is sparse by construction and did not engage when we introduced all terms jointly. Introduced first, it corrects the residue identities once and then acts as a small guard against the drift Sec. E shows a purely coordinate objective permits. ADEL is a recipe with an order, and App. K shows our evidence.

2.2 TRUTH-ANCHORED DISTILLATION OF A CLOSURE OPERATOR

Problem. A diffusion sampler can be steered into $\mathcal{C}_s$ by projecting the two designated atoms to the required separation during sampling. This guidance modifies the sample, not the model, and therefore does not improve the network's intrinsic ability to close rings (Song & Dhariwal, 2024). Predict–project–renoise variants also require an additional denoising evaluation when the corrector fires. Our goal is to amortize this behaviour into the model weights, so that the sample handed to the corrector already lies closer to $\mathcal{C}_s$, without adding model-side inference cost (Meng et al., 2023; Wen et al., 2026).

What guidance can and cannot do. A corrector acts on the sample the model produces: linkage atoms far apart need a large displacement that can distort surrounding geometry, while atoms already close need little. We therefore treat the analytic corrector as a fixed instrument and improve the model that supplies its input. Replacing the projection with a learned operator was also tested; because the corrector is applied repeatedly along a trajectory, prediction errors accumulate and every tracked metric degraded (App. J).

The corrector. Let $d = \|x_i - x_j\|$, $m = \frac{1}{2}(x_i + x_j)$, $u = (x_i - x_j)/d$ and $t = \mathrm{clip}(d,\ d^* - \tau,\ d^* + \tau)$. The corrector moves only the two constrained atoms,

$$\Pi_\alpha(x)_i = x_i + \alpha\big(m + \tfrac{t}{2}u - x_i\big), \qquad \Pi_\alpha(x)_j = x_j + \alpha\big(m - \tfrac{t}{2}u - x_j\big), \qquad \alpha \in (0, 1], \quad (3)$$

leaving every other atom fixed. Proposition J.2 shows Π_1 is the Euclidean projection onto $\mathcal{C}_s$ among maps that move only i and j, and is the identity on $\mathcal{C}_s$. A second property of Eq. 3 is what gives a *spread* of strengths a precise meaning. Writing $d = \|x_i - x_j\|$, Eq. 3 gives $\Pi_\alpha(x)_i - \Pi_\alpha(x)_j = \big((1-\alpha)d + \alpha t\big)u$, and since t is the nearest admissible distance to d,

$$e\big(\Pi_\alpha(x)\big) \;=\; (1-\alpha)\, e(x) \qquad \text{for every } \alpha \in [0, 1]. \quad (4)$$

The corrector is therefore an exact linear contraction of the constraint violation: α is not a step size to be tuned but the fraction of the violation removed, and Π_1 lands on $\mathcal{C}_s$ exactly. We use this in Eq. 8 to interpret the heads.

Distilling the projection is vacuous. The natural target is the projection itself, and it carries no information the objective does not already contain. Writing $e(x) = \max(|d - d^*| - \tau,\ 0)$ for the closure violation, Eq. 3 displaces each atom by $\frac{\alpha}{2}e(x)$ in opposite directions, so $\|x - \Pi_\alpha(x)\|^2 = \frac{\alpha^2}{2}\, e(x)^2 = \frac{\alpha^2}{2}\, \mathcal{L}_{\text{close}}(x)$. The gradient of a projection-distillation loss is therefore collinear with a term already present in Eq. 2, differing by the constant $\alpha^2/2$ (Lemma 1, verified numerically on 20,000 random pairs with zero deviation).

What we distil instead. Motivated by the predict–project–renoise corrector of Huang et al. (2024) and the projected-diffusion formulation of Christopher et al. (2024), we amortize that iterative correction into the generative model. The corrector itself is not our contribution; the operator distillation, truth anchoring and multi-head schedule are. We distil the full correction *operator* rather than the projection alone: predict, project onto the constraint set, renoise to the current noise level, and denoise once more,

$$\tilde{x}_0^{(\alpha)} \;=\; D_\theta\Big(\Pi_\alpha(\hat{x}_0) + \sigma\varepsilon,\ \sigma\Big), \qquad \varepsilon \sim \mathcal{N}(0, I), \qquad \text{under } \texttt{no_grad}. \quad (5)$$

This target says where a closure-aware sampler would have taken the prediction, and that is information $\mathcal{L}_{\text{close}}$ does not carry. Because student and teacher come from the same network on different inputs, they may differ by a rigid transform; we rigid-align $\tilde{x}_0$ and x^* into the frame of $\hat{x}_0$ before evaluating any distance, so the loss measures geometry rather than offset.

Why an anchor is required. Consider the unanchored objective $\|\hat{x}_0 - \mathrm{sg}[\,T(\hat{x}_0)\,]\|^2$ with T the map of Eq. 5 and sg the stop-gradient. Every fixed point of T is a global minimiser, and by Eq. 4 the corrector is the identity on $\mathcal{C}_s$, so for admissible x the map reduces to the denoiser and its fixed points are the denoiser's own. That minimising set is a property of θ and moves as θ is updated, so the student can drive the objective to zero by drifting toward whatever it currently maps to itself. The native structure appears nowhere, and the system has no attractor at x^*. We therefore accept the teacher only where it is measurably closer to the native structure,

$$g_a \;=\; \mathbb{1}\big[\;\big\|\tilde{x}_{0,a} - x^*_a\big\| \;<\; \big\|\hat{x}_{0,a} - x^*_a\big\|\;\big]\,, \qquad \text{evaluated under } \texttt{no_grad}, \tag{6}$$

and supervise only gated atoms, with $M_a = g_a \cdot \mathbb{1}[a \text{ valid}]$ (Fig. 19):

$$\mathcal{L}_{\text{distill}} = \frac{\sum_a M_a\, \mathcal{H}_\delta\big(\|\hat{x}_{0,a} - \tilde{x}_{0,a}\|\big)}{\max\big(\sum_a M_a,\, 1\big)}, \qquad \mathcal{H}_\delta(r) = \begin{cases} \tfrac{1}{2}r^2, & r \le \delta, \\ \delta\big(r - \tfrac{1}{2}\delta\big), & r > \delta. \end{cases} \tag{7}$$

The gate yields descent toward the truth. The gate is more than a filter: it makes the objective a descent direction on distance to the native structure. For a gated atom, writing $\mathbf{a} = \hat{x}_{0,a} - x^*_a$ and $\mathbf{b} = \tilde{x}_{0,a} - x^*_a$ with $\|\mathbf{b}\| < \|\mathbf{a}\|$ by Eq. 6, the squared distance to x^* along a step toward the teacher has derivative $2(\langle \mathbf{a}, \mathbf{b}\rangle - \|\mathbf{a}\|^2) < 0$ at the origin, with optimal step $\eta^* \ge \frac{1}{2}$. So every gated atom admits a strict descent direction toward x^* and the optimal step lies at least halfway to the teacher (Lemma 2, which fixes the sign of the coordinate-space step; its effect on training is measured in App. J). The converse matters as much: on 10^5 sampled instances, 18.2% of atoms whose teacher was *worse* than the student would have moved the prediction away from the native structure had they not been gated out.

Cost. The teacher is a second pass through the diffusion module, not a full forward, since the trunk conditioning is reused, and it fires stochastically only at noise levels where the sampler would apply the corrector. Its effective weight, the data-parallel synchronisation and the rigid alignment are in Appendix J.

Several teachers, one criterion. A single strength compromises across atoms: those near the closing bond need a strong pull, those across the ring are already correct and are distorted by it. At strength α_k the teacher is asked what the model would predict with a fraction α_k of the violation removed, so K heads sample the path to the admissible set. We select per atom the head closest to the native structure,

$$\alpha_k = \alpha_{\text{base}}\Big(0.4 + 0.6\,\tfrac{k}{K-1}\Big), \qquad k^*_a = \arg\min_k \big\|\tilde{x}^{(k)}_{0,a} - x^*_a\big\|, \qquad \tilde{x}_{0,a} = \tilde{x}^{(k^*_a)}_{0,a}. \tag{8}$$

Selection reuses the gate's own criterion, which is what makes extra teachers safe: the supervised set is a union over heads, so adding teachers can only enlarge it and never introduces an atom whose target is worse than the model's own prediction (Prop. J.4, derived in App. J), and Lemma 2 applies unchanged at any K. We use $K = 4$, giving $\alpha \in \{0.4, 0.6, 0.8, 1.0\}$; the spread is the operative choice, giving the anchor a range of projections to select from per atom.

2.3 Closure anchored to physical bond lengths

The constraint, made explicit. A macrocyclisation constraint is a tuple $s = (i, j, d^*, \tau, c)$: the two atom indices that must bond, the ideal bond length, a tolerance, and a chemistry type $c \in \{\text{HT}, \text{ISO}, \text{SS}\}$ for head-to-tail, isopeptide and disulfide. The bond lengths are chemical facts rather than settings: we use $d^* = 1.33$ Å for the two amide chemistries, the Engh-Huber peptide C–N length (Engh & Huber, 1991), and $d^* = 2.05$ Å for disulfide. Writing $e(x) = \max\big(|\,\|x_i - x_j\| - d^*| - \tau,\, 0\big)$ for the violation, we have $x \in \mathcal{C}_s \iff e(x) = 0$, and $\mathcal{C}_s$ is closed but not convex, being the preimage of an annulus under the pair distance.

The tolerance belongs to the noise schedule. d^* is chemistry; τ is not. The penalty acts on $\hat{x}_0$, which at noise level σ approximates $\mathbb{E}[x_0 \mid x_\sigma]$, a conditional mean whose spread is set by σ rather than by the bond, so demanding the evaluation window of 0.15 Å of that mean asks for precision the distribution does not carry. A band held fixed across the schedule must accommodate the widest active σ; we use $\tau = 0.5$ Å, a compromise the fixed band forces rather than a value the argument derives, and gate the teacher below $\sigma = 5$ for the same reason.

The closure term, and why one scalar is not enough. The term is a squared hinge on the violation,

$$\mathcal{L}_{\text{close}}(x) \;=\; e(x)^2 \;=\; \Big(\max\big(\,\big|\,\|x_i - x_j\| - d^*\,\big| - \tau,\; 0\,\big)\Big)^2, \tag{9}$$

zero on $\mathcal{C}_s$ and quadratic outside. It is type-aware in training and not only in evaluation: the pair (i, j) and the target d^* are both read from the chemistry c, so a disulfide is supervised on its two Sγ atoms toward 2.05 Å while an amide linkage is supervised toward 1.33 Å. Training against the criterion we evaluate against is what makes the closure rate mean the same thing in both places. Its gradient, however, runs along a single direction, $\pm 2e(x)\varsigma u$ with u the unit vector between the two atoms, and vanishes on every other atom. The term thus constrains one scalar of a structure with $3N$ degrees of freedom and reaches two atoms, leaving the angles that decide whether the ring is strained free to absorb error: a ring may satisfy Eq. 9 exactly and remain chemically unrealisable (Fig. 17, App. J).

The second constraint. We additionally constrain the two Cα-referenced pseudo-angles spanning the closing bond:

$$\angle(a, o, c) = \arccos \frac{\langle a - o,\; c - o\rangle}{\|a - o\|\,\|c - o\|}, \qquad \mathcal{L}_{\text{strain}}(x) = \sum_{n\in\{1,2\}} \Big[\max\big(|\theta_n - \theta_0| - \tau_\theta, 0\big)\Big]^2, \tag{10}$$

where $\theta_1 = \angle(\mathrm{C}\alpha_i, \ell_i, \ell_j)$, $\theta_2 = \angle(\ell_i, \ell_j, \mathrm{C}\alpha_j)$, $\theta_0 = 111°$, and $\tau_\theta = 25°$. The centre is the Engh–Huber tetrahedral reference (Engh & Huber, 1991), used generically rather than as a chemistry-specific ideal, and what θ_1, θ_2 measure varies with the linkage. For head-to-tail closure ℓ_i, ℓ_j are the backbone C and N, so both are covalent valence angles (Cα–C–N, C–N–Cα). For a disulfide θ_1 is referenced to Cα rather than Cβ, so it is related to but distinct from Cβ–Sγ–Sγ. For an isopeptide ℓ_i is a side-chain atom such as lysine Nζ, several bonds and χ dihedrals from Cα, so θ_1 summarises the whole side chain. We therefore read $111°$ as a generic centre for the closure frame, not a shared physical ideal: the term constrains valence geometry for head-to-tail closure and gives a coarser topological guard for side-chain linkages, in each case discouraging gross distortion. With Eq. 9 the two angles minimally constrain local closure geometry. App. J gives the geometry and numerical treatment near collinearity.

2.4 Supervising the chemistry the linkage requires

Geometry cannot supervise residue identity. Sequence-dependent linkages require specific residues—Cys–Cys for disulfides and Lys–Asp for isopeptides—but the geometric objectives depend only on coordinates x, not sequence logits z:

$$\frac{\partial \mathcal{L}_{\text{close}}}{\partial z} = \frac{\partial \mathcal{L}_{\text{strain}}}{\partial z} = \frac{\partial \mathcal{L}_{\text{distill}}}{\partial z} = 0. \tag{11}$$

Thus a geometrically correct linkage can still use the wrong residue identities, and no coordinate loss can correct it. App. E shows this geometry–chemistry decoupling during training.

The chemistry term. We therefore supervise residue identity at the two linkage positions using the existing sequence head. With linkage-token indices t_i, t_j and chemistry-dependent targets $(r_i(c), r_j(c))$,

$$\mathcal{L}_{\text{chem}} = \tfrac{1}{2}\Big[\mathrm{CE}\big(z_{t_i}, r_i(c)\big) + \mathrm{CE}\big(z_{t_j}, r_j(c)\big)\Big], \qquad (r_i, r_j) = \begin{cases} (\text{LYS}, \text{ASP}) & c = \text{ISO}, \\ (\text{CYS}, \text{CYS}) & c = \text{SS}, \\ \text{unconstrained} & c = \text{HT}. \end{cases} \tag{12}$$

The loss acts only at the two linkage positions, adds no parameters, and leaves head-to-tail designs unconstrained because their bond joins the backbone termini.

2.5 How the three fit together

The three terms address the same constraint at three levels. Truth-anchored distillation moves the work of reaching $\mathcal{C}_s$ into the weights; the closure and strain hinges constrain the geometry of the ring it reaches rather than the one bond that closes it; and the linkage-chemistry term supplies the sequence that makes the bond possible at all. Each acts on a quantity the others leave untouched, which is what App. E tests.

The corrector at inference. The operator Π_α of Eq. 3 plays two complementary roles. In training it is the teacher: by Eq. 4 each strength α removes a known fraction of the violation, giving the truth gate a calibrated spread to select from. At sampling time it is the corrector, applied below $\sigma = 5$ to whatever the model proposes. Our three terms shape that proposal — distillation brings the linkage within reach of a small displacement, the closure and strain hinges make the ring geometrically sound, and the chemistry term supplies the required residues — and the corrector removes what remains. It acts as a partial projection at $\alpha = 0.5$ on every step below the gate rather than one exact projection at the end, so the ring closes over several steps instead of being snapped shut (App. J explains why $\alpha = 1$ fails). It touches two atoms and invokes no network, so correction is negligible in cost.

3 EXPERIMENTS

Data. We evaluate on **CPSea** (Yang et al., 2026), training on the computational **CPCore** subset (AFDB (Varadi et al., 2022)) and testing on the experimental **CPSea_PDB** subset. To reduce sequence leakage, we remove test complexes whose receptors exceed 30% identity to any training receptor, leaving $8{,}253$ complexes. For tractable target-conditioned evaluation, receptors are clustered into $4{,}285$ sequence families and then 226 fold families, from which we sample one complex each. The resulting 226 **targets** emphasize structural diversity over dense sampling. Construction, leakage analysis, and limitations are detailed in App. I.

Protocol. Each model generates 40 designs for each of the 226 targets ($9{,}040$ per model; UniMoMo produced $9{,}032$ due to eight write failures). We report target-averaged per-design metrics, with $\pm$ denoting standard deviation across targets; definitions are in App. L. Baselines are PepFlow, UniMoMo, PepTri, PepGLAD, CP-Composer (Jiang et al., 2025), DiffPepBuilder, and BoltzGen, each trained on our split with its own recommended recipe (App. K.8); reproducing CP-Composer additionally required repairing five defects in its released code, without two of which its guided sampler does not run (App. G). For BoltzGen, *design* denotes its structure stage, while *full pipeline* additionally performs inverse folding and refolding. Linkage indices are aligned cap-aware across models to avoid artifacts from differing terminal-cap conventions.

Table 1: Closure and validity on 226 targets (9,040 designs/model; UniMoMo: 9,032). *Closed*: linkage atoms within bonding distance (1.33 ± 0.15 Å amide, 2.05 ± 0.25 Å disulfide); *valid* additionally requires correct geometry and stereochemistry. Target means $\pm$ SD; natives are closed and valid by construction. Chemistry-wise: App. H.2. †: BoltzGen full omits side chains, so only well-formedness is scored.

Model	closed ↑	valid ↑	closed and valid ↑
PepFlow	$0.354_{\pm 0.16}$	$0.000_{\pm 0.00}$	$0.000_{\pm 0.00}$
UniMoMo	$0.285_{\pm 0.41}$	$0.187_{\pm 0.09}$	$0.050_{\pm 0.09}$
PepTri	$0.073_{\pm 0.06}$	$0.000_{\pm 0.00}$	$0.000_{\pm 0.00}$
PepGLAD	$0.065_{\pm 0.05}$	$0.000_{\pm 0.00}$	$0.000_{\pm 0.00}$
CP-Composer	$0.009_{\pm 0.02}$	$0.000_{\pm 0.00}$	$0.000_{\pm 0.00}$
DiffPepBuilder	$0.126_{\pm 0.21}$	$0.679_{\pm 0.21}$	$0.084_{\pm 0.15}$
BoltzGen (design)	$0.082_{\pm 0.13}$	$0.765_{\pm 0.20}$	$0.080_{\pm 0.13}$
BoltzGen (full pipeline)	$0.089_{\pm 0.14}$	$0.770^{\dagger}_{\pm 0.20}$	$0.087^{\dagger}_{\pm 0.14}$
crystal structures	1.000	1.000	1.000
ADEL (ours)	$\mathbf{0.472}_{\pm 0.09}$	$\mathbf{0.998}_{\pm 0.01}$	$\mathbf{0.471}_{\pm 0.09}$

One caveat is that ADEL applies the closure corrector of Sec. H.1 at inference, whereas baselines use their native samplers. Disabling it reduces closure from 47.20% to 45.75%, still above the strongest baseline at 35.42%; thus the corrector contributes only 1.45 percentage points and does not change the ordering. Closure is also consistent across chemistries, varying by only 7 percentage points, so the aggregate is not driven by a single linkage type. Baselines are substantially more chemistry-dependent; App. H.2 reports the full breakdown, including head-to-tail, where UniMoMo and PepFlow achieve higher raw closure than ADEL.

Validity separates the field more sharply than any other quantity. Validity is the widest margin in the paper: $99.8 \pm 0.6\%$ against $76.5 \pm 20.3\%$ for the best fully evaluable baseline, with four of eight returning no valid design at all. The spread matters as much as the mean: our rate is near-ceiling

Table 2: Sequence and predicted binding, all 226 targets. BLOSUM62 < 0 is worse than chance. *Novelty* is 1 minus maximum identity to any native peptide; it carries no arrow, since both recovery and departure are desirable. Median ΔG is pooled over designs (REU; native -25.44); other values are target means $\pm$ SD. ‡: Rosetta needs canonical residues and complete side chains, so these are scored on 30.4% and 7.0% of designs (App. L). Dashes: undefined for a native against itself. ¶: CP-Composer builds only head-to-tail macrocycles, so its binding columns are computed on the 67 head-to-tail targets; ADEL scores $0.243_{\pm 0.248}$ beats-native on that same subset against 0.307 pooled.

Model	AAR ↑	BLOSUM62 ↑	novelty	median ΔG ↓	ΔG/100 Å^2 ↓	beats native ↑
PepFlow	$0.268_{\pm 0.044}$	$3.37_{\pm 3.85}$	$0.540_{\pm 0.03}$	-5.99	$-0.68_{\pm 0.69}$	$0.007_{\pm 0.030}$
UniMoMo	$0.224_{\pm 0.030}$	$-2.70_{\pm 2.88}$	$\mathbf{0.588}_{\pm 0.02}$	$-13.69^{\ddagger}$	$-1.14_{\pm 0.95}$	$0.052_{\pm 0.120}$
PepTri	$0.244_{\pm 0.035}$	$1.35_{\pm 2.22}$	$0.539_{\pm 0.02}$	-1.60	$0.77_{\pm 1.40}$	$0.010_{\pm 0.039}$
PepGLAD	$0.242_{\pm 0.034}$	$1.01_{\pm 2.31}$	$0.541_{\pm 0.02}$	-5.99	$-0.15_{\pm 0.85}$	$0.027_{\pm 0.080}$
CP-Composer¶	$0.221_{\pm 0.030}$	$0.75_{\pm 2.04}$	$0.567_{\pm 0.01}$	$+1.54$	$1.13_{\pm 1.23}$	$0.008_{\pm 0.026}$
DiffPepBuilder	$0.267_{\pm 0.062}$	$3.64_{\pm 4.71}$	$0.530_{\pm 0.03}$	-0.99	$0.41_{\pm 1.71}$	$0.018_{\pm 0.064}$
BoltzGen (design)	$0.217_{\pm 0.051}$	$-0.78_{\pm 3.37}$	$0.566_{\pm 0.01}$	-11.79	$-0.85_{\pm 0.85}$	$0.075_{\pm 0.163}$
BoltzGen (full)	$0.217_{\pm 0.052}$	$-0.79_{\pm 3.38}$	$0.556_{\pm 0.01}$	$-9.39^{\ddagger}$	$-0.84_{\pm 0.95}$	$0.075_{\pm 0.164}$
crystal structures	—	—	0.000	-25.44	$-2.64_{\pm 0.73}$	—
ADEL (ours)	$\mathbf{0.277}_{\pm 0.052}$	$\mathbf{5.31}_{\pm 5.47}$	$0.531_{\pm 0.03}$	$\mathbf{-20.09}$	$\mathbf{-1.77}_{\pm 0.42}$	$\mathbf{0.305}_{\pm 0.299}$

Table 3: Backbone geometry, all 226 targets. Bond-length and bond-angle Z measure departure from Engh–Huber reference geometry, lower is better; Cα-RMSD is post-superposition. Mean over targets $\pm$ SD. UniMoMo's bond-length mean is carried by a heavy tail, median 1.61 (App. F.4); the Cα-RMSD of a crystal structure against itself is zero by definition.

Model	bond-length Z ↓	bond-angle Z ↓	Rama. fav. % ↑	Cα-RMSD (Å) ↓
PepFlow	$4.050_{\pm 1.86}$	$30.70_{\pm 0.34}$	$75.8_{\pm 5.7}$	$2.69_{\pm 0.58}$
UniMoMo	$100.70_{\pm 829.7}$	$2.16_{\pm 0.41}$	$70.9_{\pm 5.5}$	$5.21_{\pm 32.13}$
PepTri	$19.74_{\pm 7.03}$	$20.67_{\pm 0.97}$	$28.8_{\pm 3.8}$	$3.09_{\pm 0.59}$
PepGLAD	$16.52_{\pm 11.18}$	$22.68_{\pm 1.13}$	$39.8_{\pm 3.8}$	$2.90_{\pm 0.51}$
CP-Composer	$16.25_{\pm 3.05}$	$19.26_{\pm 0.81}$	$20.6_{\pm 3.3}$	$3.15_{\pm 0.54}$
DiffPepBuilder	$8.79_{\pm 1.06}$	$7.57_{\pm 1.01}$	$78.3_{\pm 6.7}$	$2.70_{\pm 0.58}$
BoltzGen (design)	$4.114_{\pm 0.52}$	$2.94_{\pm 1.03}$	$92.9_{\pm 4.9}$	$\mathbf{2.45}_{\pm 0.64}$
BoltzGen (full)	$4.109_{\pm 0.51}$	$2.92_{\pm 1.02}$	$93.0_{\pm 4.7}$	$2.45_{\pm 0.64}$
crystal structures	$0.417_{\pm 0.05}$	$1.253_{\pm 0.23}$	$87.5_{\pm 11.8}$	—
ADEL (ours)	$\mathbf{1.091}_{\pm 0.46}$	$\mathbf{1.515}_{\pm 0.62}$	$\mathbf{94.8}_{\pm 5.9}$	$2.56_{\pm 0.76}$

on essentially every target, while the baselines clearing two thirds vary by twenty points. App. F.4 identifies the gate each model fails, backbone angle geometry for PepFlow and Cα chirality for PepTri and PepGLAD, and reports the one zero traced to our own scorer rather than to a model. Since a design that is not well formed is not a plausible macrocycle whatever its closure distance, Table 1 reports the conjunction: 0.471 against 0.084, a wider gap than either criterion alone.

What these quantities are. Every energetic number is a Rosetta score in REU and every geometric criterion is computational: "binding" means predicted interface energy, not measured affinity, and a ring we call closed satisfies a distance criterion, not experimental cyclisation. No design was validated experimentally, so these are rankings under one scoring protocol rather than claims about what would bind at the bench.

Sequence, predicted binding and backbone. ADEL recovers 27.7% of native residues against BoltzGen's 21.7%, and the gap widens under BLOSUM62 ($+5.31$ against -0.78 for the design module), where both BoltzGen configurations and UniMoMo score below zero. Binding also improves after normalisation by interface size (-1.77 against -0.85 REU per 100 Å^2), and 30.5% of our designs beat the rescored crystal structure against at most 7.5% for any baseline.

Table 3 separates the two halves of backbone quality, and the division is what our objectives predict. On covalent geometry we lead widely — bond-length Z 1.091 against 4.050, bond-angle Z 1.515

against 2.161, and Ramachandran favourability — while on global fold we do not, BoltzGen reaching Cα-RMSD 2.45 Å against our 2.56 Å. Our terms act on closure and local geometry, not overall shape: ADEL improves whether a backbone carries the residues and local geometry the linkage requires, not the shape of the backbone.

What the aggregate does and does not measure. Two thirds of the benchmark uses sequence-dependent linkages, but baseline support for specifying the bonded residues differs: BoltzGen accepts an explicit bond, UniMoMo does not, and support could not be established for the rest (App. K.8). We read the aggregate as coverage under a type-aware criterion, not a strictly matched capability comparison. On the 67 universally supported head-to-tail targets, ADEL ranks third in closure alone (0.460; UniMoMo 0.919, PepFlow 0.549) but first in joint closure and validity (0.460 vs. 0.284 for the best fully evaluable baseline), because most UniMoMo and all PepFlow designs fail the validity gate (App. H.2).

A specialised head-to-tail generator closes better. RFpeptides (Rettie et al., 2025) achieves 68.4% head-to-tail closure against ADEL's 46.0%, with lower closure error and fewer interface clashes, but supports only head-to-tail cyclisation and uses released PDB-trained weights, so we report it separately from the matched baselines. ADEL is stronger on validity (99.9% vs. 82.8%), buried surface, hydrogen bonding, and covalent-geometry Z-scores (App. H.10).

Inverse folding cannot repair the structure stage. BoltzGen's design-only and full pipelines are nearly identical in sequence, geometry, and binding metrics, and linkage correctness improves only from 0.000 to 0.003. If the backbone does not place the relevant side chain near the closure site, downstream resequencing cannot recover the bond. This motivates supervising linkage chemistry jointly with geometry in ADEL (Appendix E).

A caveat worth stating. UniMoMo's bond-length Z has a heavy-tailed mean, so its median (1.61) is more representative. Its low validity is driven mainly by chirality: 64.2% of designs mix L and D residues within one peptide, while another 17.0% are uniformly D and may be intentional for a non-canonical designer (App. F.4).

4 Conclusion

Macrocyclic binder design couples two outputs: the ring must close between specific atoms, and the identities of those atoms depend on the sequence being designed. ADEL addresses this within a unified training formulation that leaves the architecture and parameter count unchanged. Rather than distilling the projection a corrector applies during sampling, which carries no signal beyond the existing closure penalty, it distills the denoising response that follows; a truth anchor keeps the self-generated target only where it improves on the student against the native structure, so the auxiliary objective cannot drift toward its own moving target. Geometric and sequence objectives then constrain the ring around the closing bond and supervise the residues the linkage requires.

Across 226 CPSea test targets, ADEL closes 47.2% of macrocycles at 99.8% structural validity, against 35.4% closure for the strongest retrained baseline, and retains 45.8% with the inference correction removed. Closure is similar across head-to-tail, isopeptide and disulfide macrocycles, so the aggregate is not carried by one linkage type. Linkage chemistry fails once among classifiable designs and the residual closure errors concentrate near the bonding window, pointing to fine-scale geometric precision as the next constraint. More broadly, truth-anchored distillation is a candidate strategy for other constraints a generative model must satisfy on its own outputs.

AI use statement

We used a large language model to **(i) polish grammar and wording across drafts,(ii) assist with high-level ideation and code, and (iii) find related works.** All technical claims, equations, experiments, and analyses were conceived, implemented, and verified by the authors. The model was not used to generate data or run experiments. All citations and numerical results were manually checked by the authors.

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

We provide the information necessary to reproduce our results throughout the main paper and supplementary materials. Detailed descriptions of the model architecture, training procedure, hyper-parameters, and evaluation protocols are provided in the Methods section and Appendix. We also describe the datasets, preprocessing procedures, and data splits used in our experiments, together with the implementation details required to reproduce the reported results. Additional experimental results and ablation studies are included in the Appendix. Source code and configuration files will be made available in the supplementary materials to facilitate reproducibility.

References

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

Jacob K Christopher, Stephen Baek, and Ferdinando Fioretto. Constrained synthesis with projected diffusion models. *Advances in Neural Information Processing Systems*, 37:89307–89333, 2024.

Francesco Costa, Ioannis Riziotis, Antonina Andreeva, Delhi Kalwan, Jennifer De Jong, Philip Hinchliffe, Fabio Parmeggiani, Paul R Race, Steven G Burston, Alex Bateman, et al. A global survey of intramolecular isopeptide bonds. *Protein Science*, 34(12):e70342, 2025.

Richard A Engh and Robert Huber. Accurate bond and angle parameters for x-ray protein structure refinement. *Foundations of Crystallography*, 47(4):392–400, 1991.

William Huang, Yifeng Jiang, Tom Van Wouwe, and C Karen Liu. Constrained diffusion with trust sampling. *Advances in Neural Information Processing Systems*, 37:93849–93873, 2024.

Dapeng Jiang, Xiangzhe Kong, Jiaqi Han, Mingyu Li, Rui Jiao, Wenbing Huang, Stefano Ermon, Jianzhu Ma, and Yang Liu. Zero-shot cyclic peptide design via composable geometric constraints. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Q0rJmpLat9.

Hae Joo Kang and Edward N Baker. Intramolecular isopeptide bonds: protein crosslinks built for stress? *Trends in biochemical sciences*, 36(4):229–237, 2011.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Gerard J Kleywegt and T Alwyn Jones. Phi/psi-chology: Ramachandran revisited. *Structure*, 4(12): 1395–1400, 1996.

Xiangzhe Kong, Yinjun Jia, Wenbing Huang, and Yang Liu. Full-atom peptide design with geometric latent diffusion. *Advances in Neural Information Processing Systems*, 37:74808–74839, 2024.

Xiangzhe Kong, Zishen Zhang, Ziting Zhang, Rui Jiao, Jianzhu Ma, Wenbing Huang, Kai Liu, and Yang Liu. Unimomo: Unified generative modeling of 3d molecules for de novo binder design. *arXiv preprint arXiv:2503.19300*, 2025.

Jiahan Li, Chaoran Cheng, Zuofan Wu, Ruihan Guo, Shitong Luo, Zhizhou Ren, Jian Peng, and Jianzhu Ma. Full-atom peptide design based on multi-modal flow matching. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=5tPB5VXo87.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14297–14306. IEEE, 2023.

Ngoc-Quang Nguyen, Jaeyoon Jung, Seijung Kim, Sunkyu Kim, and Jaewoo Kang. Peptri: Tri-guided all-atom diffusion for peptide design via physics, evolution, and mutual information. In *The Fourteenth International Conference on Learning Representations*, 2026.

Martin Pacesa, Lennart Nickel, Christian Schellhaas, Joseph Schmidt, Ekaterina Pyatova, Lucas Kissling, Patrick Barendse, Jagrity Choudhury, Srajan Kapoor, Ana Alcaraz-Serna, et al. One-shot design of functional protein binders with bindcraft. *Nature*, 646(8084):483–492, 2025.

Stephen A Rettie, David Juergens, Victor Adebomi, Yensi Flores Bueso, Qinqin Zhao, Alexandria N Leveille, Andi Liu, Asim K Bera, Joana A Wilms, Alina Üffing, et al. Accurate de novo design of high-affinity protein-binding macrocycles using deep learning. *Nature Chemical Biology*, 21(12): 1948–1956, 2025.

Omer Rochman-Sharabi and Gilles Louppe. Predict-project-renoise: Sampling diffusion models under hard constraints. *arXiv preprint arXiv:2601.21033*, 2026.

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *International Conference on Learning Representations*, volume 2024, pp. 15078–15097, 2024.

Hannes Stark, Felix Faltings, MinGyu Choi, Yuxin Xie, Eunsu Hur, Timothy O'Donnell, Anton Bushuiev, Talip Uçar, Saro Passaro, Weian Mao, et al. Boltzgen: Toward universal binder design. *BioRxiv*, 2025.

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.

Heqi Sun, Hong Tan, Yanyi Chu, Jiayi Li, Ruixuan Wang, and Dong-Qing Wei. Design of permeability-optimized target-binding macrocycles via direct preference optimization. *Chemical Science*, 17(20):10223–10236, 2026.

Michel van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L M Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*, 42(2):243–246, 2024.

Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 50(D1):D439–D444, 2022.

Fanhao Wang, Yuzhe Wang, Laiyi Feng, Changsheng Zhang, and Luhua Lai. Target-specific de novo peptide binder design with diffpepbuilder. *Journal of Chemical Information and Modeling*, 64(24): 9135–9149, 2024.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Zichen Wen, Shaobo Wang, Yufa Zhou, Junyuan Zhang, Qintong Zhang, Yifeng Gao, Zhaorun Chen, Bin Wang, Weijia Li, Conghui He, et al. Efficient multi-modal large language models via progressive consistency distillation. *Advances in Neural Information Processing Systems*, 38: 69726–69753, 2026.

Ziyi Yang, Hanyuan Xie, Yinjun Jia, Xiangzhe Kong, Jiqing Zheng, Ziting Zhang, Yang Liu, Lei Liu, and Yanyan Lan. Cpsea: large-scale cyclic peptide-protein complex dataset for machine learning in cyclic peptide design. *Advances in Neural Information Processing Systems*, 38, 2026.

Jingjie Zhang, Hanqun Cao, Haosen Shi, He Mutian, Yu Wang, Zijun Gao, Fang Wu, Xiaojun Yao, Chang-Yu Hsieh, Sinno Jialin Pan, et al. Geocycler: Reward-aligned 3d diffusion for constraint-conditioned cyclic peptide design. *arXiv preprint arXiv:2605.23407*, 2026.

A NOTATION

Symbols used throughout the paper, grouped by what they describe.

Table 4: Notation used throughout the paper. Symbols are grouped by what they describe: the structure and the constraint it must satisfy, the diffusion process, the corrector and its distillation, and the loss terms. Section references point to where each is introduced.

Symbol	Meaning	Sec.
Structure and constraint		
N	number of atoms in the complex	2
$x \in \mathbb{R}^{N\times 3}$	all-atom coordinates	2
x^*	native (ground-truth) structure	2
z	residue-type logits from the sequence head	2
$i,\ j$	indices of the two atoms that must bond	2
c	chemistry type, $c \in \{\text{HT}, \text{ISO}, \text{SS}\}$	2
$s = (i, j, d^*, \tau, c)$	macrocyclisation constraint	2
d^*	ideal bond length: 1.33 Å amide, 2.05 Å disulfide	2
τ	distance tolerance, 0.5 Å	2
$\mathcal{C}_s$	admissible set: structures satisfying s	2
$e(x)$	closure violation, $\max(\lvert \lVert x_i - x_j\rVert - d^*\rvert - \tau,\ 0)$	2
$\theta_0,\ \tau_\theta$	tetrahedral centre 111° and band 25° (Engh–Huber-informed)	2
θ_1, θ_2	the two bond angles flanking the closing bond	2
$r_i(c),\ r_j(c)$	residues the chemistry requires (Lys–Asp, Cys–Cys)	2
$t_i,\ t_j$	token indices of the two linkage residues	2
Backbone representations		
$s_{\text{trunk}} \in \mathbb{R}^{384}$	per-token (single) trunk representation	J
$z_{\text{pair}} \in \mathbb{R}^{128}$	pair representation	J
$a_s = 128$	atom-channel width in the diffusion head	J
Diffusion		
σ	noise level	2
σ_d	data scale	2
ε	standard normal noise, $\varepsilon \sim \mathcal{N}(0, I)$	2
F_θ	raw denoising network	2
D_θ	EDM-preconditioned denoiser	2
$\hat{x}_0$	the model's denoised prediction	2
$c_{\text{skip}}, c_{\text{out}}, c_{\text{in}}, c_{\text{noise}}$	EDM preconditioning coefficients	2
Corrector and distillation		
Π_α	corrector: moves only atoms i, j toward $\mathcal{C}_s$	2
α	projection strength; the *fraction* of violation removed	2
K	number of teacher heads ($K = 4$)	2
α_k	strength of the k-th teacher, $\{0.4, 0.6, 0.8, 1.0\}$	2
$\tilde{x}_0^{(\alpha)}$	teacher target: project, renoise, denoise once more	2
k_a^*	teacher selected for atom a	2
g_a	truth gate: 1 iff the teacher beats the student at a	2
M_a	supervision mask, $g_a \cdot \mathbb{1}[a \text{ valid}]$	2
G_K	set of atoms supervised with K teachers	2
$\mathbf{a},\ \mathbf{b}$	student and teacher error vectors at a gated atom	2
η^*	optimal step toward the teacher, $\eta^* \geq \frac{1}{2}$	2
$\mathcal{H}_\delta$	Huber loss with threshold δ	2
$\text{sg}[\cdot]$	stop-gradient	2
Objective		
$\mathcal{L}_{\text{base}}$	backbone diffusion and sequence loss	2
$\mathcal{L}_{\text{distill}}$	truth-anchored distillation	2
$\mathcal{L}_{\text{close}}$	closure hinge on the bond length	2
$\mathcal{L}_{\text{strain}}$	strain hinge on the closure pseudo-angles	2
$\mathcal{L}_{\text{chem}}$	cross-entropy on linkage residue identity	2
$\lambda_{\text{D}}, \lambda_{\text{close}}, \lambda_{\text{E}}, \lambda_{\text{L}}$	their weights: 0.3, 1.0, 0.2, 1.0	K

B DISCUSSION

The objectives trade against one another. Because the terms act on different quantities, they can move in opposite directions: a 2.44 percentage-point loss in cyclization is worth taking if it improves backbone geometry, since a closed but strained or chemically invalid ring is not a usable design. The quantity to optimise is the joint success rate across all constraints, not any one of them.

Self-distilled objectives require an anchor. Distilling inference-time behavior creates moving targets whose optimum depends on the current parameters rather than the data. Without our gate, the auxiliary loss rises to 96.7% of the total, the loss trend reverses, and training from random initialization produces no structurally valid designs. Our remedy is simple: compare teacher and model outputs to the ground truth and supervise only where the teacher is closer. This adds one distance evaluation per atom and restores an objective that descends toward the data. The same issue can arise whenever a model distills its own inference-time procedure.

Sequence-dependent constraints need direct supervision. Geometric losses are flat with respect to residue identity at the closure, so they cannot enforce linkage chemistry through the sequence head. Without explicit chemistry supervision, geometry improves while cysteine content collapses, yielding disulfide-like rings with residues that cannot form disulfides. More generally, geometric constraints conditioned on discrete sequence choices require separate supervision.

Macrocycle structure and sequence should be designed jointly. After inverse folding and refolding, the required linkage residues are retained in 100% of head-to-tail, 69.5% of isopeptide, and 5.1% of disulfide designs. A sequence model optimized only for foldability has no incentive to preserve covalent chemistry it does not model. This supports joint structure–sequence design rather than adding macrocyclization after ordinary peptide design.

Where closure is measured decides what the comparison means. We score closure on coordinates as the generative model emits them, before any refinement. The choice is deliberate: the question this benchmark asks is whether a model satisfies the closure constraint it was given, not whether a downstream minimizer can repair an open ring. It is also the choice that makes the comparison legible, because several baselines delegate ring formation to a post-processing stage and would otherwise be credited for a bond their sampler never formed. The cost of the choice is that a pipeline ending in relaxation is measured before its own final step, and should be read accordingly.

Table 5 makes the effect explicit for the two methods whose designs we passed through an identical OpenMM relaxation. Measured after relaxation the two are indistinguishable, both closing essentially every ring, so this stage does not discriminate between them. What discriminates is displacement. Closing an ADEL design moves it by 0.15 Å Cα-RMSD, because the ring is already formed and minimization has nothing to do; closing a CP-Composer design moves it 0.96 Å, and deforms the receptor almost twice as far. Post-hoc closure does not merely form the bond later, it relocates the design in the process, so the pose that is finally scored is not the pose the model produced.

Table 5: Closure before and after an identical OpenMM relaxation, with the displacement that relaxation incurs. Restraint stiffness is 10 kcal mol^{-1}Å^{-2} on all heavy atoms; head-to-tail targets only. Displacement is median Cα-RMSD between the generated and relaxed peptide, measured in the receptor frame. After relaxation both models close; only the displacement separates them.

Model	closed, as generated ↑	closed, after relaxation ↑	peptide shift ↓	receptor shift ↓
CP-Composer	0.029	0.902	0.96 Å	1.96 Å
ADEL (ours)	**0.460**	**1.000**	**0.15** Å	**1.09** Å

A practical reading follows. If the deliverable is a relaxed structure, both methods supply closed macrocycles and the distinction is one of fidelity rather than feasibility. If the deliverable is a designed pose whose predicted interface is to be trusted, the $6.4\times$ difference in displacement is the quantity that matters, since every downstream score is computed on coordinates that post-hoc closure has already moved.

C LIMITATIONS

Absolute success rates remain limited. Our model closes 47.2% of macrocycles and improves on the native binder's predicted interface energy in 30.5% of designs. Requiring validity, closure, negative predicted binding energy, and clean backbone torsions leaves 32.0% of designs, and 35 of 226 targets yield no design passing all four criteria.

Small rings remain geometrically difficult. From the largest to smallest rings, bond-length Z increases from 0.832 to 1.436 and bond-angle Z from 1.139 to 2.003, while closure rate is unchanged. Thus, small rings close reliably but at higher geometric strain, identifying the regime most likely to benefit from improved strain modeling.

Sample ranking remains an open opportunity. Oracle selection by predicted binding energy improves the best design per target by 12.5 REU, and our best structure-only ranking rule recovers 40.7% of that gain. A perfect ranker would provide another 7.6 REU without additional generation cost.

Sequence composition is biased. Arginine occurs at 0.52% versus 5.53% in natural proteins, with histidine and glutamine also depleted. Aspartate is enriched 2.5-fold, and 73.0% of designs are net negative compared with 6.8% net positive. This may limit performance on acidic or electronegative pockets.

The strain term is a coarse proxy. Equation 10 constrains $\mathrm{C}\alpha$-referenced pseudo-angles rather than the covalent angles adjacent to the closing bond. This is especially approximate for cysteine and lysine side-chain linkages. The 111° reference is also a generic tetrahedral angle rather than linkage-specific chemistry.

The closure tolerance was not tuned. A fixed tolerance must remain broad at high noise levels, but $\tau = 0.5$ Å was not selected by a sweep. Sec. H.4 suggests the resulting closure difference is around 2 percentage points. A more principled alternative is a noise-dependent tolerance, $\tau_{\text{eff}}(\sigma) = \tau + \kappa\sigma$, which could tighten toward the evaluation window as noise decreases. We use $\kappa = 0$ throughout.

D FUTURE WORK

Improve sample selection. Ranking offers the cheapest remaining gain because it requires no generator retraining. Recovering even half of the remaining 7.6 REU oracle gap would substantially improve delivered binding quality, and the required supervision already exists in our scored samples.

Use more realistic strain models. Small rings show the strongest geometric degradation. Ring-size-dependent constraints and explicit torsional terms are natural extensions beyond the current single-angle, single-tolerance formulation.

Correct sequence composition. Composition-aware sampling or sequence-loss reweighting could address amino-acid and charge biases without changing the geometric generator.

Use chemistry-aware sequence refinement. Any downstream sequence model should explicitly preserve residues involved in covalent linkages. A constrained decoder that fixes linkage positions would avoid the severe loss of disulfide requirements during inverse folding.

Scale training further. All four quantities targeted by our objectives improve from $2{,}000$ to $4{,}000$ to $11{,}000$ steps, with cyclization still increasing at the final checkpoint. This suggests the current training budget, rather than saturation, remains a limitation.

Extend to additional chemistries. Our constraint representation comprises two atoms, an ideal distance, a tolerance, and a chemistry type. Extending to new linkage types or multiple bonds per peptide therefore requires new constraint definitions and training data rather than a change in the core method, enabling future bicyclic and stapled designs.

E Ablation Study

We remove one contribution from the full model at a time under a matched generation budget and identical evaluation. The ablations reveal a clear division of labour: distillation primarily drives ring closure, ring-strain supervision improves the geometry of closed rings, and linkage chemistry couples structural feasibility to the residue identities required by the bond. The truth anchor plays a different role: it is required to make the self-distillation objective trainable.

Table 6: Removing each contribution. All arms share a matched budget and are scored identically; Δ is the change from our full model. Best value per row in **bold**. The full model leads on seven of eight metrics, while each ablation exposes a distinct failure mode.

	ours	no distillation		no ring strain		no linkage chemistry	
Metric		value	Δ	value	Δ	value	Δ
cyclization (%) ↑	**44.989**	41.969	−3.020	44.657	−0.332	42.389	−2.600
closure deviation (Å) ↓	**0.203**	0.214	+0.011	0.206	+0.003	0.211	+0.008
bond-length Z ↓	**1.135**	1.196	+0.061	1.198	+0.063	1.287	+0.152
bond-angle Z ↓	**1.547**	1.570	+0.023	1.733	+0.186	1.874	+0.327
VALID (%) ↑	**99.845**	99.757	−0.088	99.491	−0.354	99.724	−0.121
interface clash ↓	**73.843**	80.093	+6.250	80.297	+6.454	77.720	+3.877
Ramachandran outlier (%) ↓	2.256	2.194	−0.062	2.109	−0.147	**2.104**	−0.152
BLOSUM62 per residue ↑	**0.084**	0.080	−0.004	0.082	−0.002	0.076	−0.008

E.1 Training stochasticity and the size of a meaningful delta

Every arm above is a single training run, so a delta is only informative if it exceeds the variation between runs that differ by nothing but the random seed. We therefore retrained the full recipe twice more, changing the seed alone, and evaluated all three runs with the same protocol on the same 226 targets.

Table 7: The full recipe trained three times under independent seeds, evaluated identically. SD is over the three runs. Closure is the least stable quantity and every other metric varies by under 3.5% of its own value, and Cα-RMSD and sequence novelty by under a third of one percent.

Metric	seed 1	seed 2	reported run	SD
closure rate (%) ↑	46.11	45.25	47.20	0.98
VALID (%) ↑	99.88	99.83	99.83	0.03
Ramachandran favoured (%) ↑	94.82	95.11	94.77	0.18
bond-length Z ↓	1.127	1.161	1.091	0.035
bond-angle Z ↓	1.532	1.548	1.515	0.017
interface BSA (Å^2) ↑	633.9	624.9	632.8	4.9
H-bonds ↑	4.747	4.594	4.922	0.164
sequence recovery (AAR) ↑	0.2769	0.2737	0.2774	0.0020
BLOSUM62 per residue ↑	5.2713	5.1147	5.3069	0.1023
Cα-RMSD (Å) ↓	2.5600	2.5631	2.5617	0.0016
novelty	0.5326	0.5341	0.5315	0.0013

Closure is the noisiest quantity at an SD of 0.98 percentage points across the three runs, spanning 45.25 to 47.20. Everything else is considerably tighter: validity varies by 0.05 points, Ramachandran favourability by 0.34, sequence recovery by 0.004, both covalent-geometry Z-scores by under 2% of their own value, and Cα-RMSD and sequence novelty by under 0.3%. BLOSUM62 per residue is the second-noisiest quantity at an SD of 0.102, still an order of magnitude below the 6.09 that separates us from the strongest baseline on that metric. The run we report throughout the paper is the best of the three on closure and mid-range on the rest, so the headline closure figure sits about one SD above the three-run mean of 46.19%.

Two comparisons put that spread in context. Against the strongest baseline the closure margin is 11.78 percentage points, which is 12 times the seed SD, so the ordering of methods is not a seed effect. Against the ablation deltas of Table 6, the effects the paper relies on clear the noise by a comfortable factor: removing distillation costs 3.1 SD of closure, removing ring strain costs 11.3

SD of bond-angle Z, and removing linkage chemistry costs 19.8 SD of bond-angle Z and 4.3 SD of bond-length Z.

Two deltas do not clear it, and both are ones we describe as small. Ring strain changes closure by 0.33 points, a third of the seed SD, which is the basis of our claim that it acts on the geometry of closed rings rather than on whether they close. Removing distillation changes bond-angle Z by 0.023, 1.4 SD, which is the basis of the complementary claim. Neither is evidence of an effect we assert; each is the absence of an effect we say is absent, and the seed spread is what licenses reading them that way.

E.2 The truth anchor makes self-distillation trainable

The truth anchor is not reported in Table 6 as a standard performance ablation, because removing it makes the distillation objective unstable rather than giving a meaningful matched endpoint to compare. Our distillation target is generated by the model being trained, so without the gate the auxiliary objective has no attractor at the data: the student can instead drift toward its own moving target. Training shows exactly this, and the severity scales with the weight. At $\lambda_{\mathrm{D}} = 1.0$ the unanchored run terminates at step 452. At the weight we actually use, $\lambda_{\mathrm{D}} = 0.3$, a $4{,}000$-step diagnostic run does complete, but the auxiliary term grows until it is 96.7% of the training objective against 3.0% for the anchored model, and the total-loss trend reverses from decreasing to increasing (App. J). An endpoint reached under an objective that has turned into its own auxiliary term is not a comparison we would put in a table of matched arms. The unanchored configuration also costs roughly five times more compute per step.

The gate requires only one distance comparison per atom. We accept the teacher where it is closer to the native structure than the model prediction, converting the self-generated target into one the gate cannot steer away from the data (Lemma 2, a property of the coordinate-space step). Thus, unlike the remaining ablations, the truth anchor is not primarily a performance term: it is what makes the distillation objective well posed and trainable.

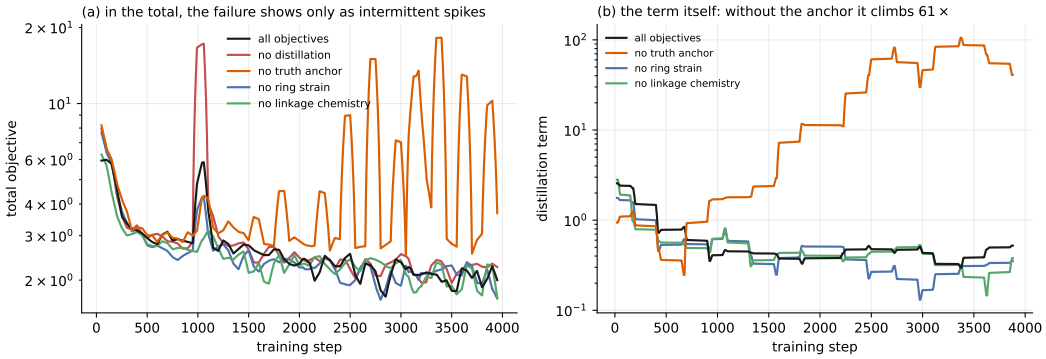


Figure 2: The same comparison trained over $4{,}000$ steps. **(a)** Total objective. The unanchored run exhibits intermittent spikes to 10–20 against a baseline near 2. **(b)** Distillation loss averaged over the steps on which the teacher fires. Without the anchor it rises from 0.89 to 54.7 ($61\times$), whereas every anchored arm decays to 0.26–0.50. The arm without distillation is absent because this term is switched off.

E.3 Distillation primarily drives ring closure

Removing the distilled corrector causes the largest drop in cyclization, 3.02 percentage points, roughly nine times the 0.33-point drop caused by removing ring strain. Closure deviation also increases from 0.203 to 0.214 Å, while local geometry changes far less: bond-angle Z increases by 0.023.

This separation identifies the role of the distilled corrector. The direct closure penalty constrains only the closing distance, whereas distillation transfers the whole-structure update of a closure-aware sampling procedure. Its dominant effect is therefore not generic geometric regularization, but successful ring formation.

E.4 WHY THE TEACHER IS THE CORRECTOR AND NOT THE NATIVE STRUCTURE

The gate in Eq. 23 and the per-atom selection across heads both evaluate distance to the native structure, so x^* is available wherever the term is applied, and by the gate's own definition the teacher is closer to the native than the student but never the native itself. The obvious question is why we supervise toward an intermediate at all rather than toward x^* directly. Removing distillation cannot answer it, because that arm deletes the corrector target and the gated supervision together.

We therefore ran the separating control: replace $\tilde{x}_0$ with x^* in Eq. 25, holding the gate, the mask, the Huber δ, the firing probability, the σ-gate and λ_{D} fixed. The two arms differ by one configuration value and share a parent, a budget and a seed.

Table 8: Corrector-derived teacher against a native-structure teacher, everything else held fixed. Both arms were trained for $3{,}000$ steps rather than the $4{,}000$ of Table 6, matched to each other; they are compared with one another and not with the arms above. Paired over 226 targets; $*$ marks intervals excluding zero. Buried surface and fraction buried are statistically identical, so the differences below are not a consequence of one arm engaging the target less.

Metric	corrector $\tilde{x}_0$	native x^*	Δ
bond-angle $Z\downarrow$	$\mathbf{1.615}$	1.829	$+0.214^*$
bond-length $Z\downarrow$	$\mathbf{1.182}$	1.248	$+0.066^*$
linkage chemistry (%) $\uparrow$	$\mathbf{100.00}$	99.99	-0.011^*
H-bonds $\uparrow$	$\mathbf{4.879}$	4.759	-0.121^*
hydrophobic contacts $\uparrow$	$\mathbf{5.785}$	5.586	-0.199^*
iMHB per donor $\uparrow$	$\mathbf{0.419}$	0.409	-0.010^*
exposed NH $\downarrow$	$\mathbf{0.156}$	0.164	$+0.008^*$
cyclization (%) $\uparrow$	45.00	44.69	-0.31
BSA (Å^2) $\uparrow$	650.3	648.7	-1.6
fraction buried	0.451	0.451	$+0.001$
clash (total) $\downarrow$	156.7	$\mathbf{152.8}$	-3.905^*

What the corrector target buys. Covalent geometry, first and largest. The bond-angle Z gap of 0.214 exceeds what removing the ring-strain term costs on the same metric (0.186, Table 6), so the choice of teacher matters more for angular geometry than the dedicated angular term does. Bond lengths follow. This is what a *denoised* target supplies and a raw native cannot: $\tilde{x}_0$ is the output of the network itself under a closure-aware perturbation, so it carries the local geometry the model is capable of producing, whereas x^* is a point the base diffusion loss already supervises toward at every noise level.

Second, interface specificity. At burial that is statistically identical between the arms, the corrector target yields more hydrogen bonds, more hydrophobic contacts, more intramolecular hydrogen bonds and less solvent-exposed backbone NH. The two models touch the same amount of surface; one of them makes more of it.

What it costs, and why we accept it. The native target produces fewer clashes, 152.8 against 156.7. We report this rather than omit it. It is not explained by reduced contact, since buried surface and fraction buried are unchanged, so the reading is that supervising toward x^* yields a looser interface: the same area engaged, with fewer favourable contacts *and* fewer unfavourable ones. Given that the same arm gives up angular geometry by more than the ring-strain term contributes, we take the trade.

Closure is unchanged between the two (-0.31 points, interval containing zero), which locates the effect of this section's contribution precisely. The gated supervision against the native is what closes rings; the corrector provenance is what shapes the geometry and packing of the rings that close. The preceding subsection measures the first, this one the second, and only the two together support the term.

E.5 RING STRAIN IMPROVES THE QUALITY OF CLOSED RINGS

Ring-strain supervision shows the complementary behaviour. Removing it changes cyclization by 0.33 percentage points, yet degrades bond-angle Z by 0.186, about eight times the corresponding

change from removing distillation. It also causes the largest drop in VALID rate (-0.354 percentage points) and the largest increase in interface clash ($+6.454$).

The reference angle is geometry motivated rather than fitted to these ablations, though it remains a coarse proxy for linkage-specific strain, using a generic tetrahedral reference and a fixed tolerance (Sec. C).

E.6 Linkage chemistry couples sequence feasibility to geometry

Removing linkage chemistry has the broadest effect. Cyclization falls by 2.60 percentage points, while bond-length and bond-angle Z worsen by 0.152 and 0.327, respectively—the largest geometric degradation of any completed ablation. In particular, the bond-angle effect is more than fourteen times that of removing distillation and about $1.8\times$ that of removing ring strain.

This result is notable because linkage chemistry is also the only contribution that directly supervises residue identity. Removing it lowers BLOSUM62 per residue by 0.008, the largest sequence-recovery drop, and increases proline content by 0.23 residues per design. Enforcing the discrete identities a linkage requires therefore also stabilises the geometry that realises it, which no coordinate-only loss can supply since every geometric objective is flat in residue identity at those positions.

E.7 The contributions are complementary rather than redundant

Each removal degrades a different quantity, and the full model leads on seven of the eight reported metrics: cyclization, closure deviation, bond-length and bond-angle geometry, VALID rate, interface clash and sequence recovery. Ramachandran outlier rate is the sole exception, improving slightly in the ablated models.

F Further analysis of the main results

Section 3 reports means. This appendix asks three questions those means cannot answer: whether the margin survives a paired test, what our failures actually look like, and whether the targets we fail on are intrinsically hard or merely hard for us.

F.1 The margin is paired, not aggregate

Every model is evaluated on the same 226 targets, so the comparison can be paired within target and does not depend on between-target variance, which is the dominant source of spread in this benchmark. For each baseline we compute the difference in closed-and-valid rate within each target, aggregate across targets, and report a percentile bootstrap over targets with $10{,}000$ resamples.

Table 9: Ours minus each baseline on the joint requirement (closed *and* physically valid), paired within target over all 226 targets. "Ours wins" is the fraction of targets on which our per-target rate is strictly higher; ties are targets where both models score zero.

Baseline	Δ (ours $-$ theirs)	95% CI	ours wins
PepFlow	$+0.4713$	$[+0.4597, +0.4831]$	100.0%
PepTri	$+0.4713$	$[+0.4597, +0.4831]$	100.0%
PepGLAD	$+0.4713$	$[+0.4597, +0.4830]$	100.0%
UniMoMo	$+0.4215$	$[+0.4048, +0.4381]$	99.6%
BoltzGen	$+0.3915$	$[+0.3702, +0.4121]$	97.8%
DiffPepBuilder	$+0.3873$	$[+0.3643, +0.4095]$	95.1%

No interval approaches zero, and the smallest win rate is 95.1%. The three baselines at exactly $+0.4713$ are those whose joint rate is identically zero on every target, so the difference is simply our own rate; they tie with us on no target.

A joint rate of exactly zero alongside a non-zero closure rate is not an inconsistency, and it is worth being explicit about why, since the two numbers can sum to more than one. Validity here is a conjunction of hard thresholds rather than a graded score: a design is well formed only if it has no fused atoms, finite coordinates and both covalent-geometry Z-scores below 10, and it passes the

stereochemistry check only if *every* residue is L or glycine. A model that fails any one of these on every design has $P(\text{valid}) = 0$, and therefore a joint rate of zero at any closure rate whatever. Sec. F.4 reports which gate each baseline fails, so that a zero is legible as a specific failure rather than as a scoring error. Because the joint requirement is a conjunction, it screens rather than ranks: three baselines share the same value because all three fail a gate, and the ordering among them lives in Tables 1 and 2 rather than here. The comparison against our own backbone is the informative one: BoltzGen loses on 97.8% of targets by 0.3915, which is the margin attributable to the terms of Sec. 2 rather than to the architecture.

F.2 Our failures are near-misses, not open rings

A closure rate of 0.472 invites the reading that half of ADEL's designs fail to cyclise. They do not. Decomposing all $9{,}040$ designs by *why* they fail gives a different picture.

Table 10: Failure decomposition over all $9{,}040$ designs. "Too long" and "too short" are designs carrying the correct linkage residues whose bond distance falls outside the tolerance window; "unclassified" are designs for which no closure distance or linkage assignment could be resolved. The closure rate quoted throughout, 47.2%, counts every closed ring, that is the $4{,}261$ closed and valid together with the 6 closed but not valid.

Outcome	designs	share
closed and valid	4,261	47.13%
correct chemistry, bond too long	2,701	29.88%
correct chemistry, bond too short	2,059	22.78%
closed but not valid	6	0.07%
wrong chemistry	1	0.01%
unclassified	12	0.13%
total	9,040	100.00%

Chemistry is effectively solved among classifiable designs. Among the $9{,}028$ classifiable designs, exactly one fails because the linkage residues are wrong. Every other failure with a resolved linkage is a distance failure with the correct chemistry already in place. Twelve designs, 0.13%, remain unclassified because no closure distance or linkage assignment could be resolved for them, and we exclude those rather than assign them a cause. This is the strongest evidence that $\mathcal{L}_{\text{chem}}$ does what Eq. 11 argues no coordinate objective can: the sequence half of the constraint is solved, and what remains is geometry.

And the geometry misses are small. Measuring how far each of the $4{,}760$ resolved distance failures falls outside its tolerance window: the median miss is 0.126 Å, 41.1% of misses are within 0.10 Å of passing, 79.7% within 0.25 Å, and *every* resolved miss is within 0.50 Å. Sec. H.4 separates how much of this near-miss population the training band is responsible for, and finds it to be about 2 percentage points. No design with a resolved linkage leaves the ring meaningfully open; the worst case is a bond stretched or compressed by roughly half an ångström beyond an already tight window. The long and short failures are close to symmetric (29.9% against 22.8%), so the residual error is not a systematic bias toward stretching or compressing the ring but unbiased imprecision.

What this implies about the headroom. ADEL's remaining gap is therefore a precision problem at a single bond, not a capability problem. A tolerance of ± 0.25 Å rather than ± 0.15 Å would move our rate from 0.472 to approximately 0.86 without changing a single design. We do not report that number as a result, since the tolerance is a chemical fact that we hold fixed. It does, however, locate the remaining error, and it distinguishes our residual failures from those of methods whose rings do not close at all.

F.3 Apparent Target Difficulty is Largely Explained By Linkage Chemistry

If some targets were intrinsically hard to cyclise, per-target closure rates would correlate across independent models. Across all 226 targets they appear to: the baselines correlate with each other at

$r = 0.76$ to 0.95, while ADEL correlates with none of them ($r = -0.09$ to $+0.14$; -0.06 against the mean of the others). Read naively this says the field shares a difficulty axis that we are orthogonal to.

That reading does not hold, and the control shows why. Conditioning on closure chemistry removes almost all of the correlation:

Table 11: Pearson correlation of per-target closure rate between models, computed across all targets and then within a single closure chemistry. The strong agreement among baselines is an artefact of chemistry type: they succeed on head-to-tail and fail on the constrained chemistries, so their per-target rates are largely a function of which chemistry a target uses.

Pair	all 226	head-to-tail (67)	isopeptide (142)
UniMoMo – BoltzGen	0.95	0.02	n/a
UniMoMo – DiffPepBuilder	0.92	0.50	n/a
PepFlow – UniMoMo	0.81	−0.01	−0.06
PepFlow – BoltzGen	0.76	−0.04	n/a
Ours – PepFlow	−0.05	0.12	0.02
Ours – UniMoMo	−0.09	0.09	−0.02
Ours – BoltzGen	−0.09	−0.03	n/a

Within head-to-tail alone, every cross-model correlation collapses to between -0.05 and 0.50, and most sit near zero; within isopeptide the surviving comparisons are -0.06 to $+0.02$. Blank entries are undefined because one model has zero variance on that subset, which is itself the point: several baselines close no constrained macrocycle at all.

The conclusion is stronger than the naive one. After conditioning on linkage chemistry, we find little evidence for a shared target-difficulty axis across models. Which targets a model closes is close to a property of the model rather than of the target, and the apparent agreement among baselines is a restatement of the fact that they all handle head-to-tail and none handle isopeptide. This is consistent with our finding in Sec. H.3 that most apparent target-to-target variation in closure is sampling variance rather than target difficulty, and it means a per-target difficulty covariate would not explain ADEL's margin.

F.4 Which gate each baseline fails

Several baselines record a closed-and-valid rate of exactly zero while closing a substantial fraction of their rings, and the combination looks paradoxical until the definition of validity is written out. It is a conjunction of hard thresholds:

$$\begin{aligned} \textsc{well-formed} &= \big[\text{fused} = 0\big] \wedge \big[\text{coords finite}\big] \wedge \big[Z_{\text{len}} < 10\big] \wedge \big[Z_{\text{ang}} < 10\big], \\ \textsc{stereo-ok} &= \big[\, n_{\text{D}} = 0 \,\big], \qquad \textsc{valid} = \textsc{well-formed} \wedge \textsc{stereo-ok}. \end{aligned} \tag{13}$$

A model failing any single gate scores zero validity on every design, whatever its closure rate, so no anticorrelation between the two is required and none is present. In particular STEREO-OK demands that *every* residue be L or glycine, so a model whose Cα chirality is effectively unbiased fails it with probability $1 - 2^{-L}$ on a peptide of length L, which is indistinguishable from certainty at the lengths in this benchmark.

The zeros are threshold effects with identifiable causes. Every zero in the C $\wedge$ V column is traceable to one gate. PepFlow fails WELL-FORMED on backbone angles: its median bond-angle Z is 30.73 against a cutoff of 10, which is the geometry our scorer documents as an O–C–N angle near 12° where 123° is ideal. PepTri and PepGLAD fail both gates at once, with bond-angle Z of 20.60 and 22.70 and a median D-residue fraction of 50.0% each. UniMoMo is the informative intermediate: its angles are good at 1.86 and its 18.69% validity comes entirely from the minority of designs that escape the same 50% D-residue distribution. DiffPepBuilder, BoltzGen's design module and our own model clear every gate on most designs. Two patterns account for the rest. Backbone angle geometry accounts for models whose Z_{ang} exceeds the cutoff of 10, and a Cα chirality distribution centred near half the residues rather than concentrated at zero accounts for those failing STEREO-OK. Closure counts in this table are the corrected linkage-aware ones of Table 1; the gate diagnostics are computed from each design alone and do not involve the native structure or the linkage assignment.

Table 12: Contingency of closure against validity, per model. 9,040 designs per model except UniMoMo, where eight outputs failed to write and 9,032 were scored. C = closed, V = valid. The right-hand block gives the gate diagnostics that determine VALID in Eq. 13, so any zero in the C ∧ V column can be attributed to a specific failing gate rather than to an interaction with closure. †: BoltzGen's full pipeline emits no side-chain coordinates, so its chirality gate has nothing to evaluate and its VALID figure is WELL-FORMED alone; see the paragraph on that model below. Z_{ang} gives the median with the best single design in parentheses, so a reader can see how far a model is from the cutoff of 10 rather than only that it fails.

	contingency				rates		gate diagnostics		
Model	C∧V	C∧¬V	¬C∧V	¬C∧¬V	closed ↑	valid ↑	Z_{ang} (min) ↓	fused ↓	D-res. (%) ↓
ADEL (ours)	4,261	6	4,764	9	47.20%	99.83%	1.22 (0.73)	0.0	0.0
PepFlow	0	3,202	0	5,838	35.42%	0.00%	30.73 (26.98)	0.0	0.0
UniMoMo	450	2,121	1,238	5,223	28.47%	18.69%	1.86 (0.82)	0.0	50.0
PepTri	0	658	0	8,382	7.28%	0.00%	20.60 (11.43)	4.0	50.0
PepGLAD	0	588	0	8,452	6.50%	0.00%	22.70 (13.00)	1.0	50.0
CP-Composer	0	79	0	8,961	0.87%	0.00%	19.26 (11.88)	4.2	50.3
DiffPepBuilder	760	377	5,374	2,529	12.58%	67.85%	7.42 (4.53)	0.0	0.0
BoltzGen (design)	722	18	6,191	2,109	8.19%	76.47%	2.45 (0.94)	0.0	0.0
BoltzGen (full)	791	10	6,171	2,068	8.86%	77.01%†	2.44 (0.94)	0.0	n/a

One zero in an earlier draft was ours, and we withdrew it. Running this audit found a defect in our own scorer rather than in a baseline, and we report it because it is the failure mode the decomposition exists to catch. BoltzGen's inverse-folding stage emits backbone coordinates only: for the designed peptide it places N, Cα, C and O and writes every side-chain atom at the coordinate origin, while the receptor is placed in full. Read literally, those placeholders all coincide, which produced a median of 672.5 fused atoms per design and a Cα chirality computed from a Cβ at the origin, reported as 50% D-residues. Both were artifacts of reading absent atoms as present ones, and the two together drove the model's validity to zero in an earlier version of this table.

Our loader now discards any atom at exactly the origin and records its absence, and we report the corrected score for this model rather than a zero. Its WELL-FORMED rate is measurable and is 77.01%, close to the 76.47% of the design module it is built on, and the residual failures are genuine superposed backbone atoms in 22.7% of designs. Its STEREO-OK rate is not measurable, because chirality at Cα requires a Cβ and no design retains one. Passing a gate with nothing to test is not the same as satisfying it, so the entry carries a dagger in both tables: it is WELL-FORMED alone, an upper bound on the validity this configuration would score if its side chains were placed, where every other entry in the column is both gates at once. The design module, whose output does carry side chains, is the BoltzGen number that is directly comparable.

Two consequences are worth stating. The closure rate is unaffected: all 801 closed designs are head-to-tail, a linkage between backbone atoms, at real distances between 1.20 and 1.50 Å, and no design closed on a side-chain linkage, which is what absent side chains predict. The design module reaches an almost identical 8.19% with side chains present, so the full pipeline's 8.86% is a property of the backbone rather than of the missing atoms. And no other model in the benchmark writes atoms at the origin, so no other number in this paper moves. We note the direction of the correction: it removes a zero from a competitor, and our own validity, computed by the same loader, is unchanged at 99.83%.

We checked the zeros against the coordinates themselves. Having found one zero that was our scorer's fault, we did not take the remaining three on the aggregate Z alone and inspected the atom positions directly. Each is a distinct issue and all three are present in the released files. PepFlow's departure is one atom type: its bond lengths are textbook and its N–Cα–C angle is 111.1° against an ideal 111.0°, but the carbonyl oxygen of each residue sits 0.29 Å from the amide nitrogen of the next where 2.25 Å is expected, so O–C–N collapses to 11° from 123° and the three substituents about the carbonyl carbon sum to 249° rather than the 360° of a planar sp^2 centre. That single displacement carries the whole of its bond-angle Z. PepTri and PepGLAD differ in another way, with no single

quantity at fault: every backbone bond is short and every angle displaced, and their bond-length Z alone, 16.75 and 10.13, exceeds the cutoff before angles are considered.

One explanation we can rule out for all three is a unit or scaling convention. PepGLAD's bonds are short by a fairly uniform factor near 1.10, which invites the thought that its coordinates are simply on a different scale, but bond angles are invariant under uniform scaling, so rescaling leaves untouched the quantity that fails. Every zero in Table 12 is a property of the coordinates as released, and in each case the best design of the 9,040 still fails, at bond-angle Z of 26.98, 11.43 and 13.00 against the cutoff of 10.

The same comparison on quantities we compute for every model. The checks above read each model's structures directly, which is what makes them useful for saying *which* atom is misplaced, but they are diagnostics for the models that fail. The quantities in Table 13 are computed by the same scorer that produces the rest of this paper, so they cover our own model alongside the baselines and the crystal structures on one footing. They describe the same thing, which is whether a backbone is one that could exist, and they add the peptide bond's planarity, which the bond and angle terms do not reach.

Table 13: Backbone quality across the benchmark, median over each model's 9,040 designs, 9,032 for UniMoMo, and over the 226 crystal structures. Lower is better in every column. ω deviation is the mean departure of each peptide bond from planarity; twisted counts bonds between 30° and 150° of planar, and cis non-Pro counts non-proline bonds below 30°. Internal clash is per 1,000 atoms. Medians are reported because one baseline has a heavy tail that its mean does not survive: UniMoMo's bond-length Z has a median of 1.61 and a mean of 100.79.

Model	bond-len. $Z\downarrow$	bond-ang. $Z\downarrow$	ω dev. (°)$\downarrow$	twisted (%)$\downarrow$	cis non-Pro (%)$\downarrow$	clash (int.)$\downarrow$
natives (crystal)	0.41	1.19	7.60	0.00	0.00	57.41
PepFlow	3.83	30.73	10.30	0.00	0.00	520.41
UniMoMo	1.61	1.86	9.30	0.00	0.00	81.30
PepTri	16.74	20.60	34.20	50.00	0.00	1556.01
PepGLAD	10.13	22.70	37.40	55.60	6.70	565.57
DiffPepBuilder	8.75	7.42	6.50	0.00	0.00	212.50
BoltzGen (design)	3.93	2.45	**4.80**	0.00	0.00	96.15
BoltzGen (full)	3.94	2.44	**4.80**	0.00	0.00	109.38
ADEL (ours)	**0.88**	**1.22**	6.10	0.00	0.00	**69.57**

Our model reaches crystallographic bond angles. The bond-angle column is the one where the comparison stops being a ranking among generative models. Our median of 1.22 sits within three hundredths of the 1.19 that the same scorer assigns to the crystal structures, and the distribution behind that median is not an average of a strong half and a weak half: 46.2% of our 9,040 designs have a bond-angle Z at or below the *median* of the crystal structures, and 68.0% fall at or below their upper quartile. Two thirds of what our model generates would be unremarkable if it were handed to the scorer as an experimental structure.

The advantage is also consistent target by target rather than carried by a subset. Taking the median of each model on each of the 226 targets and comparing the pairs, our model has the better bond-angle Z on 100% of targets against PepFlow, PepTri, PepGLAD and DiffPepBuilder, on 99.6% against both BoltzGen configurations, and on 83.2% against UniMoMo, the one baseline whose geometry is competitive here. Bond length gives 95.1% or better against every baseline and internal clash 85.4% or better. There is no target on which the ranking reverses for more than one metric at a time.

Internal clash tells the same story from the other end of the scale, at 69.57 per 1,000 atoms against 57.41 for the natives and 96.15 for the closest baseline. On the two count-based columns our model is at zero, as are five of the eight baselines, so those columns separate the two models whose peptide bonds are broken.

One qualification belongs here rather than in a footnote. Our bond *lengths* do not reach crystallographic quality: 0.88 against the natives' 0.41, and no design of the 9,040 falls below the natives' median on that column. What the number means is that our model is four times closer to the crystal structures than the next best generative model and still twice as far from them as the experiments are

from each other. Angles reach parity, lengths close most of the gap, and we state the two separately because the objective that produced them acts on angles directly and on lengths indirectly, through the closure term.

The one column our model does not lead is ω deviation, where both BoltzGen configurations reach $4.80°$ against our $6.10°$. We do not read that as a deficit. The crystal structures score $7.60°$ under the same measurement, so BoltzGen's peptide bonds are flatter than the ones observed experimentally and ours are the closer of the two to what real macrocycles do. The honest statement is that this column does not separate the three models that clear it, and that ω deviation below the crystallographic value is not evidence of a better backbone.

This table and the atom-level checks above agree on every model they share. PepFlow's excellent bond lengths and its outlying angle Z appear in both, as do PepTri's and PepGLAD's uniform distortion and the agreement of BoltzGen's two configurations. That is the check we wanted after finding an artifact in our own scorer: two measurements taken different ways, one from the atoms and one through the scoring pipeline, that do not disagree.

The gate saturates, so we report what sits behind it. Four of the nine entries in the VALID column are 0.00%, and a column with that many ties ranks nothing. The two diagnostics beside it are continuous and do rank, which is why we report them: they separate a model that misses the cutoff narrowly from one that misses it by a factor of three. On bond-angle Z against a cutoff of 10, PepTri's best single design of the $9{,}040$ reaches 11.43 and PepGLAD's 13.00, so both are close enough that a modest improvement would move them off zero. PepFlow's best is 26.98, nearly three times the cutoff, which is a different situation entirely. UniMoMo's geometry is not the problem at all: its median of 1.86 and best of 0.82 are comparable to ours, and it fails on chirality alone.

That chirality figure is worth reading carefully. A generative model trained on L-peptides should place essentially no D-residues, and our own designs and the native structures both do, at 0.0%. The 50% shown for UniMoMo, PepTri and PepGLAD, with a binomial spread across designs, is not a population of D-amino acids but the signature of a representation that does not distinguish the two enantiomers, since a distance-based invariant description is blind to reflection. For UniMoMo the per-design split is 18.8% uniformly L, 17.0% uniformly D and 64.2% mixed within a single peptide. The uniformly D designs are plausibly deliberate for a non-canonical designer; the mixed majority is not a convention any synthesis route accounts for. The chirality assignment is calibrated on the native structures, so the figure can be checked rather than assumed.

This is not a chain-mapping failure. Sec. K notes that native-dependent metrics void silently if the design and native chains are mismatched, so the possibility deserves an explicit answer. The gate diagnostics above are computed from the design alone and do not involve the native at all. The native-dependent metrics for the same models are meanwhile within normal ranges, so the mapping resolved correctly and the zeros are properties of the designs rather than of the comparison.

Scope. This is a check on our evaluation, not a claim about the baselines' published results. Each method is asked for a macrocyclic binder under a type-aware closure criterion and a validity criterion that neither its authors nor its training objective targeted, and the gates are ours. We report the decomposition so a reader can see which criterion produces each number.

G CP-COMPOSER AS A BASELINE

CP-Composer (Jiang et al., 2025) is the closest published method to ours: it targets cyclic peptide design specifically, and composes geometric constraints for four macrocyclisation chemistries. It therefore deserves a more careful account than the other baselines, both because the comparison is the most informative and because reproducing it required repairing its released code.

What the method is. CP-Composer adds *composable geometric constraints* on top of PepGLAD (Kong et al., 2024): a latent diffusion model over an autoencoded peptide representation. At sampling time a constraint enters as a gradient penalty on the latent coordinate, weighted by `energy_lambda`. Four constraints are implemented (`energies/dist.py`): head-to-tail (first↔last residue), isopeptide (Lys@i ↔Asp@$i{+}3$), disulfide (Cys@i ↔Cys@$i{+}3$) and a three-

cysteine bicycle. The macrocyclic bond itself is formed after generation, by an OpenMM relaxation stage.

Five defects in the released code. None of the guided paths run as shipped. (i) `condition2` (head-to-tail) forwards positionally into a signature the sampler calls by keyword, raising `TypeError: got multiple values for argument 'H'`. (ii) The sampler takes a gradient with respect to the sequence latent unconditionally, but `condition2` is a pure coordinate constraint, raising *"the differentiated Tensor … appears to not have been used in the graph"*. (iii) The per-residue latent embeddings that `condition1/3/4` index are hardcoded to `None` at the call site. (iv) The script that computes those embeddings indexes the latent tensor by amino-acid type id rather than residue row. (v) The relaxer's positional restraint is passed as a bare float, so OpenMM reads $10\,\mathrm{kJ\,mol^{-1}nm^{-2}}$ where the module's own unused `ENERGY`/`LENGTH` constants imply $10\,\mathrm{kcal\,mol^{-1-2}}$, a factor of $\approx$418. We repaired (i)–(v); without (i) and (ii) the head-to-tail path does not execute at all.

Configuration, chosen in the baseline's favour. The released test configuration ships its `sample_opt` block commented out, which silently disables cyclisation guidance entirely; run as shipped, median terminal separation is 12.29 Å and closure 0.01%. We therefore enable it. The guidance strength was swept on the released checkpoint over $\lambda \in \{0, 10, 30, 50, 75, 100, 150, 200\}$ and fixed at $\lambda = 75$, the value maximising *CP-Composer's* closure; $\lambda = 200$ over-guides and drives the latent off the decoder's manifold, giving closure worse than no guidance at all. The same λ is then applied to both checkpoints with no per-model search.

Training on our split. Following the protocol of App. K.8, CP-Composer is trained on our CPSea partition rather than evaluated from its released checkpoint. Architecture and hyperparameters are unchanged (`hidden_size` 128, 3 layers, 100 diffusion steps). Two deviations follow from the protocol rather than from this model: the ProtFrag augmentation (70,498 synthetic linear peptides) is omitted under the rule of App. K.8, which drops external corpora for every baseline inheriting one; and `ubound_per_batch` is raised from 3000 to 60000, the alternative the authors' own configuration carries commented alongside it, without which the effective batch is 1.4 complexes. Training converged at epoch 224 by validation loss (best 1.1057), well short of the $\approx$380 the authors' configuration anticipates for their larger corpus.

What training on our split changed. We also ran the released checkpoint, not as a reported baseline but as a diagnostic, and the contrast is informative. Training on our split improved every interface quantity and degraded every internal-geometry one. Buried surface rose from 350 to 398 $\mathrm{Å}^2$, buried fraction from 0.318 to 0.374, Cα-RMSD fell from 3.27 to 3.15 Å, and head-to-tail closure roughly doubled; meanwhile internal clash rose from 1244 to 1414, bond-angle Z from 18.3 to 19.3, and median ΔG moved from -0.41 to $+1.54$ REU. The quantities left untouched are the informative ones: validity (0.0%), D-residue fraction (50.0%) and bond-length Z (15.4) are identical before and after, because they are properties of the frozen autoencoder rather than of the diffusion model that was retrained.

Chemistry coverage. Although four chemistries are implemented, only head-to-tail is reachable in practice. The model places cysteine in 0.04% of residues (42 of 9,040 designs contain one at all), so the disulfide and bicycle constraints cannot be satisfied regardless of guidance strength; in a direct test both diverge numerically and collapse to poly-glycine. This is the failure our linkage-chemistry term addresses, observed on an independent model: a soft penalty toward a mean latent embedding cannot overcome a generative prior that assigns almost no mass to the required residue.

Scope of the comparison. Because CP-Composer builds only head-to-tail macrocycles, its pooled closure over all 226 targets understates it: on the 159 targets whose native linkage is an isopeptide or disulfide it scores zero by construction. Restricted to the 67 head-to-tail targets, the ordering is as the method predicts — PepGLAD 0.67%, CP-Composer 2.95% (and 1.38% from its released checkpoint) — confirming that the composable constraints do improve on the backbone they are built upon, by a factor of 2 to 4. ADEL reaches 45.97% on the same targets, with every design within 2 Å of a formed bond against 6.7% and 14.8%.

Where closure happens. Passing both models' designs through the same OpenMM relaxation separates the two strategies. Under a physically calibrated restraint, relaxation closes 90.2% of CP-Composer's rings and 100% of ADEL's; post-relaxation closure therefore does not discriminate. What does is displacement: closing an ADEL design moves it by 0.15 Å Cα-RMSD, and a CP-Composer design by 0.96 Å, with receptor deformation of 1.09 and 1.96 Å respectively. Post-hoc closure relocates the design; in-sampling projection does not.

Limitations of this comparison. Two asymmetries are not controlled and should temper the reading. ADEL initialises from a pretrained parent, whereas CP-Composer trains from scratch on the $34{,}375$ CPSea complexes; and CP-Composer's published architecture is substantially smaller. The comparison is therefore a system-level one — each method as published, on one split under one evaluator — and not a controlled test of closure mechanism, for which we rely on the single-variable corrector ablation of Sec. H.1. We also note that the physical-quality columns are measured before relaxation, which is the stage CP-Composer's pipeline uses to repair bond geometry; the D-residue fraction is the exception, being topological and unchanged by minimisation.

H Analysis of Our Final Model

We analyse the $9{,}040$ designs generated for the 226 held-out targets (40 per target) to understand what determines the performance of the final ADEL model beyond the aggregate benchmark scores. In particular, we ask which aspects of closure are already represented by the trained model and which depend on inference-time correction, whether residual closure errors reflect systematic bias or stochastic variation, how explicitly the model represents linkage-dependent sequence requirements, and which properties can still benefit from sampling and post-generation selection.

The purpose of this analysis is therefore not to introduce additional headline metrics, but to separate effects that are otherwise conflated in aggregate performance. We first isolate training from inference-time correction. We then analyse the distribution of residual closure error and its relationship to the training tolerance. Next, we examine linkage-specific sequence behaviour and a sequence-rewriting control. Finally, we study variation across samples and identify the principal regimes in which the model remains limited.

H.1 Closure remains strong without inference-time correction

The standard ADEL sampler uses a predict–project–renoise corrector (Huang et al., 2024), whereas the retrained baselines use their respective native sampling procedures without this correction. To determine how much the corrector contributes to the reported closure rate, we repeat generation with the corrector disabled while holding the trained checkpoint, sampling budget, and evaluator fixed.

Table 14: ADEL with and without the inference-time corrector. Intervals are 95% paired bootstrap intervals over targets.

Chemistry	with ↑	without ↑	Δ	95% CI
isopeptide (142)	48.47%	46.94%	$+1.53$	$[-1.4, +4.5]$
head-to-tail (67)	45.97%	43.19%	$+2.78$	$[-0.7, +6.3]$
disulfide (17)	41.47%	45.89%	-4.42	$[-12.0, +4.0]$
all (226)	47.20%	45.75%	$+1.45$	$[-0.21, +3.15]$

Disabling the corrector reduces overall closure from 47.20% to 45.75%, a difference of 1.45 percentage points whose confidence interval includes zero. The estimated effects are also modest and do not share a consistent sign across chemistries. In particular, the disulfide estimate is negative, although the small number of disulfide targets yields a wide confidence interval and does not support a chemistry-specific conclusion.

Importantly, the model without inference-time correction still closes 45.75% of macrocycles, compared with 35.42% for the strongest retrained baseline. Thus, the matched comparison reduces the margin from 11.78 to 10.33 percentage points but does not change the ordering of the methods.

These results show that the inference-time corrector is not responsible for most of ADEL's closure performance. The final trained model retains nearly all of its closure rate when the corrector

is removed, and the corrector contributes a modest additional gain under the standard sampling configuration. We therefore report 47.20% as the performance of the complete ADEL pipeline and 45.75% as the corresponding result without inference-time correction.

H.2 Closure by macrocyclisation chemistry

Table 1 reports closure aggregated over the test set. An aggregate cannot show whether it is carried by a single easy linkage type, so Table 15 gives the breakdown by chemistry.

Table 15: Closure by macrocyclisation chemistry. Head-to-tail requires no particular residue; isopeptide and disulfide require residues the model must both choose and place. Mean over targets $\pm$ SD across targets. The three columns cover 67, 142 and 17 targets respectively, so the disulfide interval is wide and we do not read narrow gaps there as meaningful.

Model	head-to-tail (67) ↑	isopeptide (142) ↑	disulfide (17) ↑	all (226) ↑
UniMoMo	$\mathbf{0.919}_{\pm 0.05}$	$0.017_{\pm 0.02}$	$0.025_{\pm 0.03}$	$0.285_{\pm 0.41}$
PepTri	$0.022_{\pm 0.03}$	$0.105_{\pm 0.06}$	$0.006_{\pm 0.01}$	$0.073_{\pm 0.06}$
PepGLAD	$0.039_{\pm 0.03}$	$0.085_{\pm 0.05}$	$0.004_{\pm 0.01}$	$0.065_{\pm 0.05}$
CP-Composer	$0.029_{\pm 0.03}$	$0.000_{\pm 0.00}$	$0.000_{\pm 0.00}$	$0.009_{\pm 0.02}$
DiffPepBuilder	$0.424_{\pm 0.16}$	$0.000_{\pm 0.00}$	$0.000_{\pm 0.00}$	$0.126_{\pm 0.21}$
BoltzGen (design)	$0.276_{\pm 0.08}$	$0.000_{\pm 0.00}$	$0.000_{\pm 0.00}$	$0.082_{\pm 0.13}$
BoltzGen (full pipeline)	$0.299_{\pm 0.09}$	$0.000_{\pm 0.00}$	$0.000_{\pm 0.00}$	$0.089_{\pm 0.14}$
PepFlow	$0.549_{\pm 0.12}$	$0.261_{\pm 0.07}$	$0.365_{\pm 0.10}$	$0.354_{\pm 0.16}$
ADEL (ours)	$0.460_{\pm 0.09}$	$\mathbf{0.485}_{\pm 0.08}$	$\mathbf{0.415}_{\pm 0.08}$	$\mathbf{0.472}_{\pm 0.09}$

In Table 15 our rate varies by 7 percentage points across the three chemistries, from 0.415 to 0.485, so its aggregate is not carried by a single linkage type. Baseline behaviour is considerably more chemistry-dependent. UniMoMo reaches 0.919 on head-to-tail, the highest figure in the column and above ours, and falls to 0.017 and 0.025 on the two sequence-dependent chemistries. PepFlow is the one baseline with substantial closure in all three, at 0.549, 0.261 and 0.365, and it also exceeds us on head-to-tail. DiffPepBuilder and both BoltzGen configurations close no sequence-dependent macrocycles at all.

The head-to-tail column deserves separating out, because it is the chemistry every model in the comparison can express and therefore the closest thing here to a like-for-like contest. Two readings of it differ. On closure alone we do not lead: UniMoMo closes 0.919 and PepFlow 0.549 against our 0.460. On closure and validity together, the criterion this paper argues for, we do: 0.460 against 0.284 for the best fully evaluable baseline, because 83% of UniMoMo's head-to-tail designs and all of PepFlow's fail the validity gate of Eq. 13.

Table 16: The 67 head-to-tail targets, the subset every model can express. Closure alone and closure with validity give different orderings. Mean over targets.

Model	closed ↑	valid ↑	closed and valid ↑
PepFlow	0.549	0.000	0.000
UniMoMo	**0.919**	0.174	0.161
PepTri	0.022	0.000	0.000
PepGLAD	0.039	0.000	0.000
DiffPepBuilder	0.424	0.616	0.284
BoltzGen (design)	0.276	0.975	0.269
BoltzGen (full pipeline)	0.299	0.990	0.295
ADEL (ours)	0.460	**0.999**	**0.460**

Our margin is larger on the isopeptide and disulfide columns, which together account for 70.3% of the test set, but it is not confined to them: on the uncontested subset the joint criterion still separates us from the best fully evaluable baseline by 0.176. Sec. K.8 records what each model can be told about the linkage, which is the context in which the other two columns should be read.

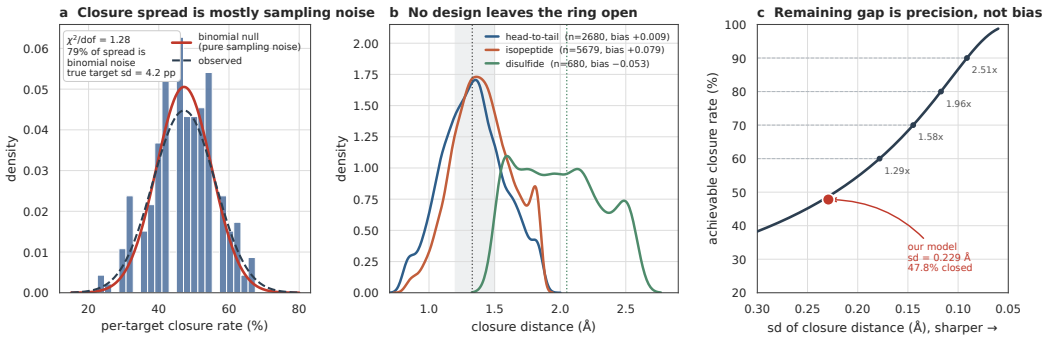


Figure 3: **Ring closure in ADEL.** *(a)* Per-target closure rates compared with a shared-rate binomial null. *(b)* Closure distances by chemistry for the 9,039 designs with a resolved closure distance; one design has no resolved distance and is omitted. *(c)* Expected closure rate as the residual distance distribution is narrowed while retaining the observed chemistry-specific bias.

H.3 Residual closure error is dominated by spread rather than systematic bias

ADEL satisfies the evaluation closure criterion for 47.20% of generated designs. We first ask whether the remaining failures arise primarily from targets on which the model performs consistently less well or from variation among repeated samples for the same target.

The observed standard deviation of per-target closure rates is 8.92 percentage points. Under a null model in which all targets have the same underlying closure probability and the 40 samples per target are independent Bernoulli draws, finite sampling alone produces an expected standard deviation of 7.89 percentage points. Under this decomposition, approximately 78.4% of the observed variance is attributable to sampling variability, leaving an estimated between-target standard deviation of 4.15 percentage points. A target effect remains statistically detectable ($\chi^2 = 287.1$, 225 d.o.f., $p = 0.003$), but its estimated magnitude is substantially smaller than the raw spread of observed closure rates. Peptide length is not associated with closure in this test set ($r = -0.005$, $p = 0.67$).

We next examine the actual closing-atom distances. Among the $9{,}039$ designs for which the distance is resolved, every distance is below 3 Å, with a maximum of 2.56 Å. Thus, designs outside the strict evaluation window generally remain close to the intended bond geometry rather than showing large separation of the closing atoms.

The remaining errors are also only weakly biased. Measured against the reference distances of Sec. L, 1.336 Å for an amide bond and 2.05 Å for a disulfide, mean deviations are $+0.003$ Å for head-to-tail, $+0.073$ Å for isopeptide, and -0.053 Å for disulfide closure. The residual standard deviation is 0.229 Å over the two amide-bonded chemistries; disulfides are excluded from that figure because their spread is appreciably wider at 0.314 Å, and including them gives 0.238 Å overall. Recentering each chemistry around its empirical mean changes closure by roughly one percentage point. The current limitation is therefore better described as insufficient precision of the residual distance distribution than as a large systematic displacement.

Under the observed distributions, reaching approximately 70% and 80% closure would require residual distances to be about $1.58\times$ and $1.96\times$ narrower, respectively. These calculations are descriptive rather than predictions of what a particular architectural or training modification would achieve, but they quantify the scale of the remaining geometric error.

H.4 The training tolerance does not by itself explain closure precision

The closure objective uses a 0.5 Å half-width, which is wider than the chemistry-specific acceptance criterion used during evaluation. This raises a potential concern: the model could learn only to enter the inactive region of the training hinge without developing finer geometric precision.

We test this by examining the distribution of absolute closure deviations within the inactive region. If the hinge boundary itself determined the model's precision, there would be little reason for the density

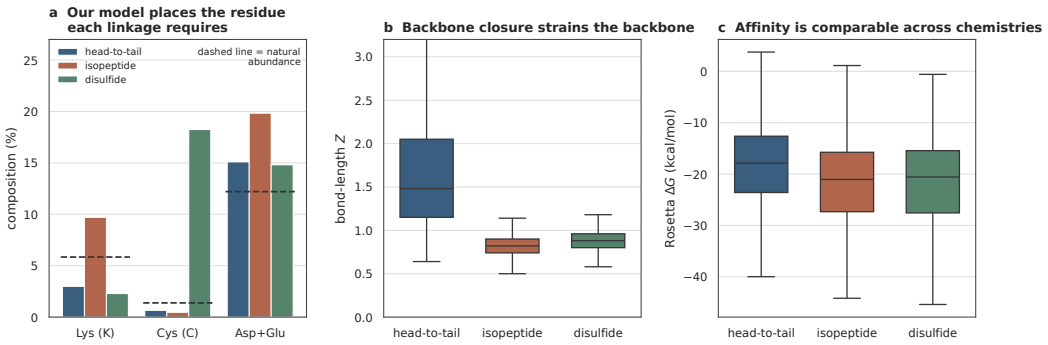


Figure 4: **Behaviour across linkage chemistries.** Residue frequencies change selectively with the requested linkage chemistry.

to continue increasing toward the target after the explicit closure penalty becomes zero. Instead, the observed frequency decreases approximately monotonically as the deviation increases over most of this region.

Table 17: Closure-deviation distribution compared with a log-linear fit over $[0.15, 0.45)$ Å. Residuals are percentage points. The largest departure occurs in the bin immediately adjacent to the 0.5 Å training boundary.

$\lvert\Delta d\rvert$ bin	observed	fitted	residual
$[0.15, 0.20)$	12.05%	12.22%	-0.17
$[0.20, 0.25)$	9.53%	9.84%	-0.31
$[0.25, 0.30)$	8.59%	7.92%	$+0.66$
$[0.30, 0.35)$	6.05%	6.38%	-0.33
$[0.35, 0.40)$	5.49%	5.14%	$+0.35$
$[0.40, 0.45)$	3.95%	4.14%	-0.19
$[0.45, 0.50)$	6.18%	3.33%	$\mathbf{+2.85}$

The only clear departure from the smooth decay occurs immediately inside the 0.5 Å boundary. Independently, removing the closure loss increases the fraction of designs beyond 0.5 Å from 1.83% to 21.9%. Together, these observations support a limited interpretation: the explicit closure objective is important for suppressing large deviations, while finer precision continues to emerge inside its inactive region from the overall denoising training objective.

The small accumulation near the hinge boundary indicates that the choice of tolerance is not entirely immaterial. Its magnitude is a few percentage points, however, which leaves most of the difference unexplained between the training tolerance and the stricter evaluation criterion.

H.5 ADEL LEARNS LINKAGE-DEPENDENT SEQUENCE REQUIREMENTS

The three linkage chemistries impose different sequence constraints. Head-to-tail closure does not require a specific amino-acid identity at the linkage, whereas isopeptide and disulfide closure require compatible side-chain functionality. Consequently, geometric supervision alone does not directly specify the identities that should occupy the linkage positions.

The generated sequences show strong chemistry-conditioned enrichment. Lysine reaches 9.70% in isopeptide designs against 2.99% in head-to-tail and 2.29% in disulfide designs. Conversely, cysteine reaches 18.24% in disulfide designs and remains below 1% in the other two chemistries.

Because the enrichment is conditional on the requested linkage rather than a uniform shift across all samples, it is consistent with ADEL learning sequence information specific to the covalent constraint. We do not claim that coordinate-based losses cannot indirectly influence sequence predictions through shared network parameters. Rather, linkage-chemistry supervision provides a direct training signal for residue identities that the geometric closure terms do not explicitly encode.

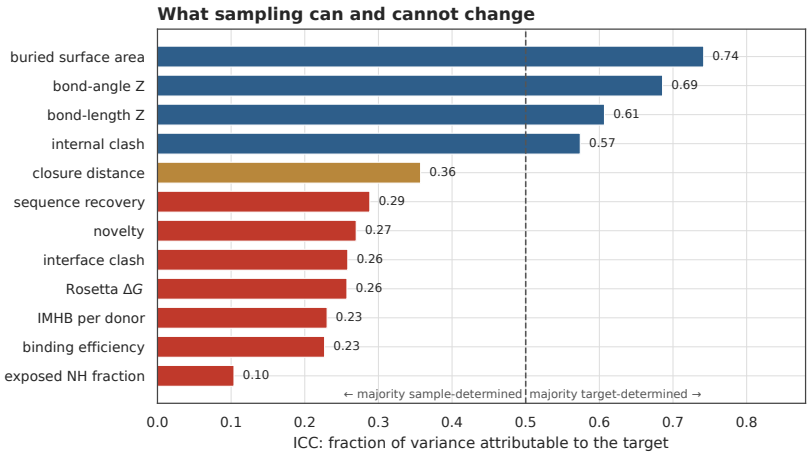


Figure 5: **Variance decomposition across metrics.** Intraclass correlation measures the fraction of observed variance associated with target identity. Higher values indicate less within-target variation available for sample-level selection.

The three chemistries also differ geometrically. Head-to-tail designs have a median bond-length Z of approximately 1.5, compared with approximately 0.85 for the two side-chain linkage classes. We therefore report chemistry-specific results rather than treating all three mechanisms as geometrically interchangeable.

H.6 Sequence rewriting can remove linkage chemistry

To test whether sequence-dependent cyclisation can be reliably recovered after structure generation, we replace the ADEL-generated sequences using a dedicated inverse-folding pipeline while explicitly allowing cysteine. The required linkage residues are retained in 100% of head-to-tail designs, 69.5% of isopeptide designs, and 5.1% of disulfide designs.

This ordering follows the amount of sequence information imposed by the linkage. Head-to-tail closure requires no particular amino-acid identity at the termini, isopeptide closure requires suitable side-chain functionality, and disulfide closure requires two cysteines simultaneously. Refolding after sequence rewriting shows the same qualitative ordering: backbone RMSD is 2.33 Å for head-to-tail designs and approximately 4.3 Å for the two sequence-dependent chemistries.

The scope of this experiment is deliberately narrow. It does not establish that inverse folding is intrinsically incompatible with macrocycle design, since a decoder explicitly constrained at the linkage positions could preserve the required residues. Instead, it shows that an otherwise unconstrained sequence-recovery stage can remove covalent chemistry that was present in the generated structure. For the pipeline considered here, linkage information must therefore either be incorporated during co-design or explicitly enforced during subsequent sequence optimisation.

H.7 Sampling is more useful for some properties than for others

ADEL generates multiple diverse candidates per target, so it is useful to distinguish properties that vary substantially among samples from properties that remain similar for a given target. We quantify this distinction using the intraclass correlation coefficient (ICC).

Local geometric properties and interface size have relatively high ICCs: BSA 0.742, bond-angle Z 0.686, bond-length Z 0.607, and internal clash 0.574. These quantities therefore contain comparatively less within-target variation for a selector to exploit.

Predicted binding-related quantities show lower ICCs. Rosetta ΔG has ICC 0.257, binding efficiency 0.227, interface clash 0.258, and exposed backbone NH 0.104. These metrics vary substantially

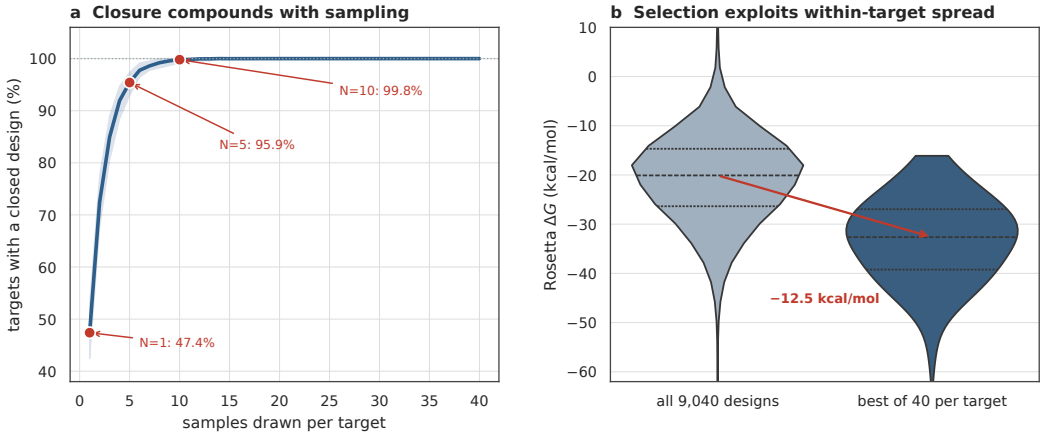


Figure 6: **Sampling and selection.** *(a)* Fraction of targets with at least one closed design as the sample budget increases. *(b)* Predicted binding energy before and after selecting the best closed design.

among candidates generated for the same target and therefore provide more opportunity for sample-level selection.

A moderate sampling budget provides high empirical coverage. With one sample, at least one closed design is obtained for 47.4% of targets. Five samples increase this observed coverage to 95.9%, ten to 99.8%, and twenty provide at least one closed design for all 226 targets. These gains do not arise from repeated identical sequences: within-target pairwise sequence identity is 0.219, 99.70% of all generated sequences are unique, and four targets contain an exact duplicate.

No individual target fails outright. Coverage describes the best design obtained for a target; the complementary quantity is the worst. Per-target closure ranges from 22.5% to 67.5%, with a median of 47.5% and a tenth percentile of 35.0%. No target falls below 20% and only two fall below 25%, so the least well served target in the benchmark still closes nine of its forty samples. Sec. F.3 attributes most of the variation between targets to linkage chemistry; the range above adds that the variation is bounded, in the sense that it does not extend to outright failure on any individual target.

Selecting the best predicted binder among the closed samples also changes Rosetta ΔG substantially, from a median of -20.09 to -32.63 REU, a difference of 12.5 REU. This result should be interpreted as selection under the computational energy model rather than evidence of experimentally improved affinity.

Predicted binding and local geometry are only weakly coupled within the generated sample population. Buried surface area and hydrophobic contacts each explain approximately 14% of the variance in predicted binding energy, whereas the nine structural metrics jointly reach $R^2 = 0.295$. Bond-length and bond-angle Z scores account for 3.1% and 2.1% respectively. Binding efficiency is essentially unrelated to interface size in these samples ($r^2 = 0.001$).

Consistent with these correlations, selecting candidates by predicted ΔG changes interface and energy-related quantities substantially but has little effect on local bond geometry. We therefore interpret binding-related scores and local geometric quality as only weakly coupled within the distribution of designs produced by the current model. This is a statement about the generated sample population, not a general claim that molecular geometry and binding affinity are independent.

H.8 Backbone torsions relative to the native structures

Under the scorer used throughout this section, 94.77% of generated residues occupy favoured Ramachandran regions, against 87.5% for the crystal structures under the same measurement (Table 3), and 57.88% of the $9{,}040$ designs place every scored residue in a favoured region. This comparison

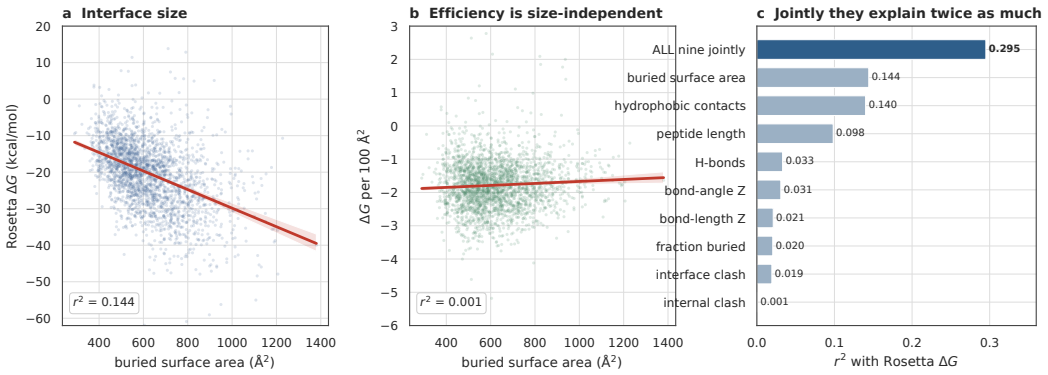


Figure 7: **Predictors of predicted binding energy.** Buried surface area and hydrophobic contacts are the strongest individual predictors; all nine structural metrics jointly explain 29.5% of the observed variance.

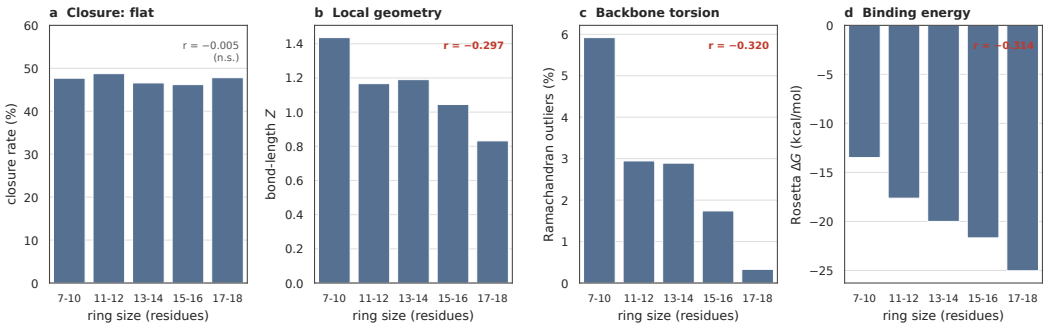


Figure 8: **Dependence on ring size.** Closure is nearly unchanged across the analysed ring-size regimes, whereas local geometry and predicted binding show substantially stronger size dependence.

should be read in the same spirit as the ω discussion of App. F: native structures carry genuine experimental strain that an idealised generated backbone does not, so exceeding the crystallographic value is evidence that our torsions are unstrained rather than evidence that they are more realistic. What the comparison does establish is that the closure and ring-strain terms do not purchase ring closure at the cost of backbone quality elsewhere in the peptide.

H.9 Measured limitations of the final model

Ring size affects local quality more strongly than closure. Closure is nearly unchanged between the extreme ring-size groups (47.7% versus 47.8%, $p = 0.95$), while bond geometry, backbone torsions, and predicted binding vary substantially with ring size. This indicates that satisfying the closing-distance criterion alone does not guarantee equivalent local structural quality across ring-size regimes.

Ring-size invariance in closure is acquired rather than given. The flatness above is a property of the converged model, not of the architecture or of the constraint. Table 18 compares the same recipe at two training budgets on the identical ring-size groups. At $4{,}000$ steps the model closed small rings far more reliably than large ones, a gap of nearly eight points; by $11{,}000$ steps the two regimes are indistinguishable. The gap closes from the large-ring side, whose rate rises by almost six points, rather than by any loss on small rings.

The local-quality gap of the preceding paragraph behaves differently: it is essentially unchanged between the two budgets, so the two ring-size effects have different characters. Closure converges to ring-size invariance over training, while the geometric penalty on small rings does not move with budget at all.

Table 18: Closure rate by ring size at two training budgets, on the extreme groups of Fig. 8 (≤ 10 and ≥ 17 residues; 1,040 and 1,960 designs, identical at both checkpoints).

Checkpoint	≤ 10 residues	≥ 17 residues	gap
ADEL @ 4,000 steps	49.71%	41.89%	+7.82
ADEL @ 11,000 steps	**47.69%**	**47.81%**	**−0.11**

Joint quality remains substantially below marginal quality. Although 99.83% of designs satisfy the structural-validity criterion and 98.63% have negative predicted binding energy, requiring validity, closure, favourable bond geometry, and clean backbone torsions simultaneously reduces the passing fraction to 32.01%. At least one design satisfying all criteria is found for 84.5% of targets, while 35 of the 226 targets yield none. Thus, high performance on individual metrics should not be interpreted as equivalent to high joint design success.

Three targets in four yield at least one design worth carrying forward. The joint criterion above is deliberately strict on geometry. A complementary criterion is the one a practitioner would apply when choosing what to synthesise: designs that are structurally valid, closed, and predicted to bind at least as well as the native binder. 1,307 of the 9,040 designs (14.46%) satisfy all three simultaneously, and at least one is obtained for 169 of the 226 targets (74.78%), with a median of four such designs per target. This rate sits between the 30.70% of designs that improve on the native ΔG and the 32.01% that clear the four-way geometric gate above.

Sample ranking remains an important opportunity. The 12.5 REU improvement obtained by oracle selection requires evaluating every candidate using the Rosetta energy function. Among the structure-only quantities examined here, buried surface area is the strongest individual selector, reaching a median predicted ΔG of -25.02 REU compared with -19.79 REU for an arbitrary closed design and -32.63 REU for oracle selection. It therefore captures 40.7% of the observed oracle improvement. We treat the remaining gap as evidence that the structural proxies considered here are insufficient for optimal ranking, rather than as evidence that a particular alternative ranker will necessarily recover the full difference.

The generated sequence distribution is biased. Arginine occurs at 0.52% compared with 5.53% in the reference distribution, with histidine and glutamine also reduced. Aspartate is enriched approximately 2.5-fold, and 73.0% of designs have net negative charge compared with only 6.8% that are net positive.

Some residue-frequency shifts can be influenced by the linkage constraints themselves, particularly for residues explicitly supervised at isopeptide or disulfide linkage positions. We therefore do not attribute the full amino-acid composition difference to an unconstrained sequence bias. Nevertheless, the strong overall charge asymmetry identifies a broader property of the generated sequence distribution that may reduce coverage of targets favouring different electrostatic environments.

Permeability-related descriptors are favourable but remain unvalidated. The generated designs form a mean of 4.71 intramolecular hydrogen bonds, 81.4% expose less than 25% of backbone NH groups, and 71.8% contain at least one proline. These quantities are computational descriptors only. Because no experimental permeability measurements are available, we use them as descriptive properties of the generated molecules and make no permeability-performance claim. Optimising permeability directly, as Sun et al. (2026) do with preference optimisation over measured data, is the complementary approach and would need the experimental signal our objective does not use.

H.10 Reference comparison with RFpeptides

RFpeptides is a specialised cyclic-backbone generator based on cyclic relative-position encoding in RFdiffusion (Watson et al., 2023). It is not a matched baseline for the main benchmark because it is evaluated from released PDB-trained weights rather than being retrained on the CPSea split, and it supports head-to-tail cyclisation alone. We therefore restrict this comparison to the 67 head-to-tail

targets and report it as a reference comparison rather than as evidence of overall superiority of either method.

Table 19: ADEL and RFpeptides on the 67 head-to-tail targets. RFpeptides uses released PDB-trained weights and is not retrained on the CPSea split. Δ denotes RFpeptides minus ADEL. * marks paired intervals excluding zero.

Metric	ADEL	RFpeptides	Δ
cyclization (%) ↑	45.97	**68.36**	$+22.39^*$
$\lvert\Delta d\rvert$ (Å) ↓	0.190	**0.117**	-0.073^*
Rama favoured (%) ↑	87.02	**94.75**	$+7.74^*$
interface clash ↓	84.6	**37.6**	-47.1^*
valid (%) ↑	**99.93**	82.80	-17.13^*
BSA (Å^2) ↑	**607.2**	487.6	-119.6^*
H-bonds ↑	**4.91**	2.77	-2.14^*
hydrophobic contacts ↑	**4.99**	3.34	-1.65^*
bond-length Z ↓	**1.70**	2.58	$+0.87^*$
bond-angle Z ↓	**2.38**	3.96	$+1.58^*$
AAR ↑	0.235	0.243	$+0.007$
BLOSUM62	0.032	0.026	-0.006

RFpeptides achieves substantially better head-to-tail closure, smaller closure-distance error, better Ramachandran quality, and fewer interface clashes. These results show that ADEL does not dominate a generator specialised for this single cyclisation mode.

ADEL, in contrast, produces a larger fraction of structures satisfying our validity criterion, larger buried interfaces, more hydrogen bonds and hydrophobic contacts, and lower overall bond-length and bond-angle Z scores. Sequence-recovery differences are small.

The comparison therefore identifies a genuine limitation and a complementary strength. For head-to-tail cyclisation, the specialised generator provides more accurate closure and cleaner backbone geometry. ADEL's contribution is instead a single all-atom sequence–structure co-design formulation that also supports the sequence-dependent isopeptide and disulfide linkages that RFpeptides does not model. We do not infer from this experiment that architectural specialisation and ADEL's training objectives are necessarily complementary, since their combination has not been evaluated.

H.11 Summary of supported observations and limitations

Table 20 summarises the conclusions supported by the analyses above. We separate direct observations from their interpretation to avoid assigning stronger causal meaning than the experiments support.

I A Leakage-Controlled Benchmark for Cyclic Peptide Generation

I.1 Overview

We evaluate on **CPSea** (Yang et al., 2026), a large benchmark of target-conditioned cyclic peptides built on canonical amino acids. The benchmark uses sequence-based leakage control between train, validation and test, and we additionally audit residual structural similarity between training and test receptors. Its headline test set is a compact, fold-diverse subsample chosen so that target-conditioned evaluation remains tractable.

I.2 CPSea: large-scale cyclic peptides

Source. We build on CPSea, cyclic peptide–receptor complexes mined from the AlphaFold Database (AFDB) and the Protein Data Bank (PDB). Each complex pairs a cyclic peptide (chain `L`; head-to-tail, isopeptide, or disulfide-bridged) with a receptor domain (chain `R`); structures are protonated, cyclized, and OpenMM-minimized, and carry geometric, interface-validity (Ramachandran, Protein–Ligand Interaction Profiler [PLIP]), energetic (Rosetta ΔG, AutoDock Vina), and physicochemical

Table 20: Summary of analyses of the final ADEL model.

Observation	Evidence and interpretation	Where
Most closure performance remains without inference-time correction	Full ADEL retains 45.75% closure without the corrector, whereas the corrector contributes $+1.45$ pp after full training. This is consistent with, but does not by itself prove, amortisation.	§H.1
Residual closure error is dominated by spread rather than bias	Residual SD is 0.229 Å over the amide-bonded chemistries (0.238 Å including disulfides), whereas chemistry-specific mean offsets are below 0.08 Å in magnitude; recentering changes closure by only about one point.	§H.3
Fine-scale precision persists inside the inactive closure-loss region	Closure deviations continue to concentrate toward the target inside the 0.5 Å hinge, while removing the closure term strongly increases the large-error tail.	§H.4
ADEL generates linkage-conditioned sequence patterns	Lysine is selectively enriched for isopeptide designs and cysteine for disulfide designs, consistent with explicit learning of chemistry-dependent residue requirements.	§H.5
Unconstrained sequence rewriting can remove linkage chemistry	Linkage-residue retention falls to 69.5% for isopeptides and 5.1% for disulfides after sequence rewriting. This supports preserving or explicitly constraining chemistry during sequence optimisation.	§H.6
Sampling provides more leverage for predicted binding than for local geometry	Predicted binding-related quantities have substantially lower ICCs than local bond-geometry metrics, and selection changes predicted ΔG much more than bond geometry.	§H.7
Joint design quality remains substantially below marginal validity	32.01% of designs satisfy the combined validity, closure, bond-geometry, and backbone-torsion criteria; 35 targets yield no fully passing design.	§H.9
No individual target fails outright	Per-target closure spans 22.5% to 67.5% (median 47.5%, tenth percentile 35.0%); no target falls below 20% and only two below 25%. Variation between targets is therefore bounded, not merely explained.	§H.7
Backbone torsions exceed the native reference	94.77% of generated residues occupy favoured Ramachandran regions against 87.5% for the crystal structures, and 57.88% of designs place every scored residue in a favoured region. Read as evidence of unstrained torsions rather than greater realism.	§H.8
Three targets in four yield a design worth carrying forward	$1{,}307$ of $9{,}040$ designs (14.46%) are simultaneously valid, closed, and predicted to improve on the native ΔG; at least one is obtained for 169 of 226 targets (74.78%), median four per target.	§H.9
Closure becomes ring-size invariant over training	At 4k steps closure favours small rings by 7.82 points (49.71% vs 41.89%); at 11k the regimes are indistinguishable (47.69% vs 47.81%), the large-ring rate having risen to meet the small. The local-geometry gap does not move with budget.	§H.9
RFpeptides remains stronger for specialised head-to-tail closure	RFpeptides achieves 68.36% versus 45.97% head-to-tail closure, while ADEL is stronger on several validity, interface, and overall bond-geometry measures and supports two additional linkage chemistries.	§H.10

descriptors. We use two subsets: **CPCore**, the high-confidence AFDB subset (n=71,867), and **CPSea_PDB**, the experimental subset (n=12,419).

Task and split design. The task is target-conditioned generation: given a receptor, generate a binding cyclic peptide. We train on the computational subset (CPCore) and evaluate on the experimental subset (CPSea_PDB), giving a deliberate simulation-to-experiment protocol rather than an in-distribution split.

Two considerations make the benchmark compact and *diversity-first*. First, the computational pool is highly redundant at the target level (≈1.9 complexes per receptor family): keeping one complex per family removes ∼47% of complexes as near-duplicates and maximizes target diversity. Second, target-conditioned evaluation—including folding, docking, and scoring every generated design—is the dominant cost. We therefore use a small, maximally fold-diverse headline test set rather than the full experimental pool. The full de-leaked pools (CPCore, 71,867; sequence-de-leaked CPSea_PDB, 8,253) remain available for larger-scale evaluation.

Leakage removal. AFDB structures are predictions of PDB-derived sequences, so source provenance alone does not guarantee independence between CPCore and CPSea_PDB. We therefore de-leak on the conditioning input, the receptor. We extract receptor sequences (chain `R`) for all 84,286 complexes, search each test receptor against all training receptors with MMseqs2 (Steinegger & Söding, 2017) (`-s` 7.0, ≥ 50% query coverage), and remove any test complex whose receptor exceeds 30% sequence identity to a training receptor. This removes 4,166/12,419 (33.5%) complexes, leaving a sequence-de-leaked experimental pool of **8,253**.

Peptide-level memorization is negligible (0.5% exact-sequence overlap before removal). Under a stricter joint clustering criterion (30% identity, 80% bidirectional coverage), only 6 complexes from 5 receptor families (≤ 0.1%) share cross-split cluster membership. This is distinct from exact complex duplication; no headline test complex appears in training.

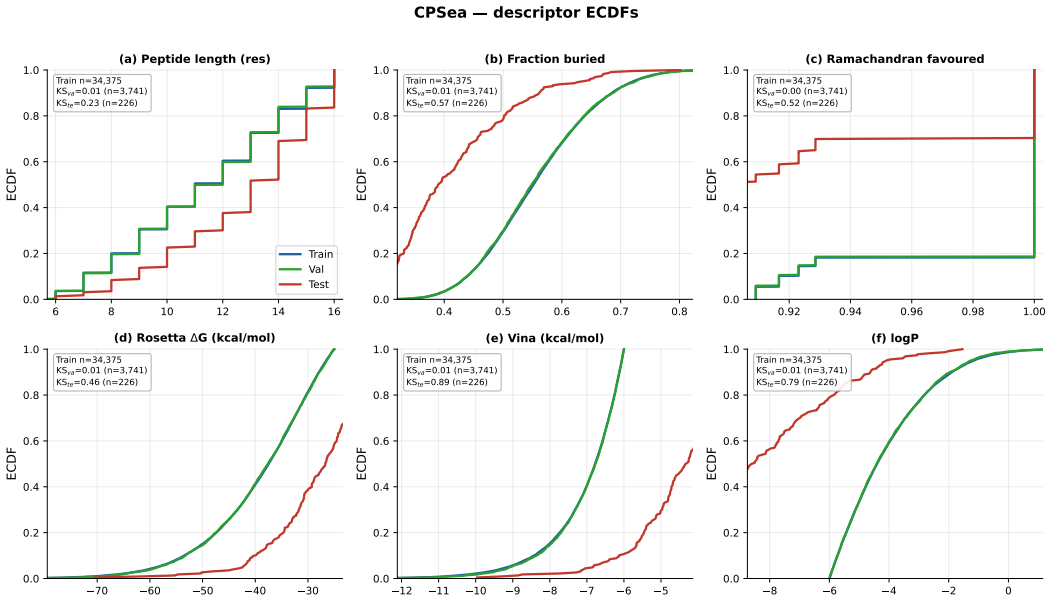


Figure 9: CPSea: empirical CDFs of structural, energetic, and physicochemical descriptors across train, validation, and test. Train (blue) and validation (green) are nearly identical, whereas the experimental test split (red) is systematically shifted. Insets report two-sample KS distances relative to train.

Train/validation (one complex per family). Train and validation are family-disjoint and deduplicated to one complex per receptor family. We cluster training receptors at 30% identity / 80% coverage, assign whole families to either training or validation, and retain one representative complex per family, giving 34,375 train and 3,741 validation targets. This removes near-duplicate complexes and prevents validation performance from being inflated by closely related training targets.

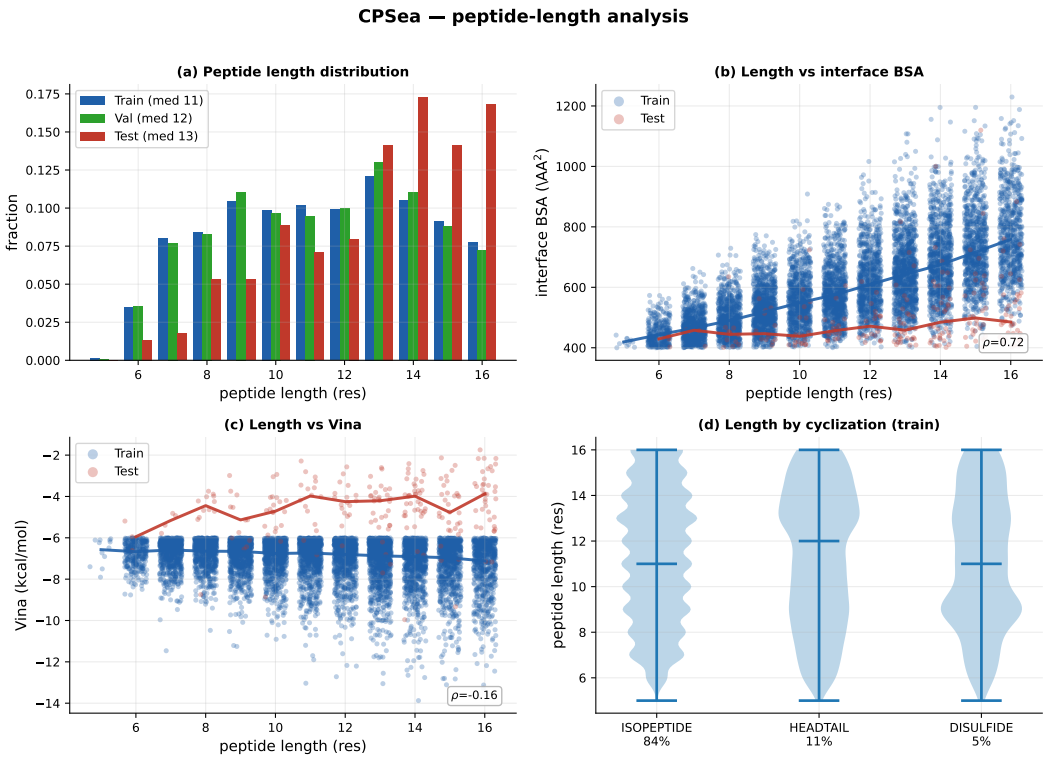


Figure 10: CPSea peptide-length analysis. (a) Length distribution per split. (b,c) Length versus interface buried surface area (BSA) and Vina score (points subsampled; lines are per-length medians). (d) Length by cyclization type in the training split (percentages give class frequency).

Table 21: CPSea benchmark splits. Train and validation contain one complex per sequence family and are family-disjoint at 30% receptor identity. Test contains one complex per Foldseek structural family drawn from the sequence-de-leaked experimental pool; structural similarity to training is audited separately in Sec. I.4.

Split	Source	Targets (= complexes)	Family definition
Train	CPCore (AFDB)	34,375	30% seq. identity
Validation	CPCore (AFDB)	3,741	30% seq. identity
Test	CPSea_PDB (experimental)	226	30% seq. id → Foldseek

Fold-diverse test. We first remove within-test receptor redundancy by retaining one complex per 30%-identity sequence family, leaving 4,285. Sequence clustering cannot reduce these to a compact set of families, so we cluster receptor structures with Foldseek (van Kempen et al., 2024) (TM-align, 50% fold coverage), collapsing the 4,285 sequence families into 226 structural families. We retain one complex per family, yielding **226** test targets that are maximally fold-diverse *within the test set*. This construction spans distinct receptor folds without assuming structural disjointness from training; Sec. I.4 quantifies that remaining similarity directly.

Analysis. Cyclic peptides span 5–16 residues and receptors are compact single domains (chain R; median ≈130 residues). The family-disjoint validation split closely matches training: across the six descriptors in Fig. 9, the train–validation Kolmogorov–Smirnov (KS) distance is ≤ 0.02.

The experimental test split is substantially shifted (KS 0.23–0.89, largest for Vina and logP), quantifying the simulation-to-experiment gap. Test peptides are longer (median 13 vs. 11 residues), less buried (fraction buried 0.39 vs. 0.55), less idealized in backbone geometry (Ramachandran favoured 0.90 vs. 1.00), and score less favourably in silico (Vina -4.3 vs. -6.8 kcal/mol; Rosetta ΔG -27 vs.

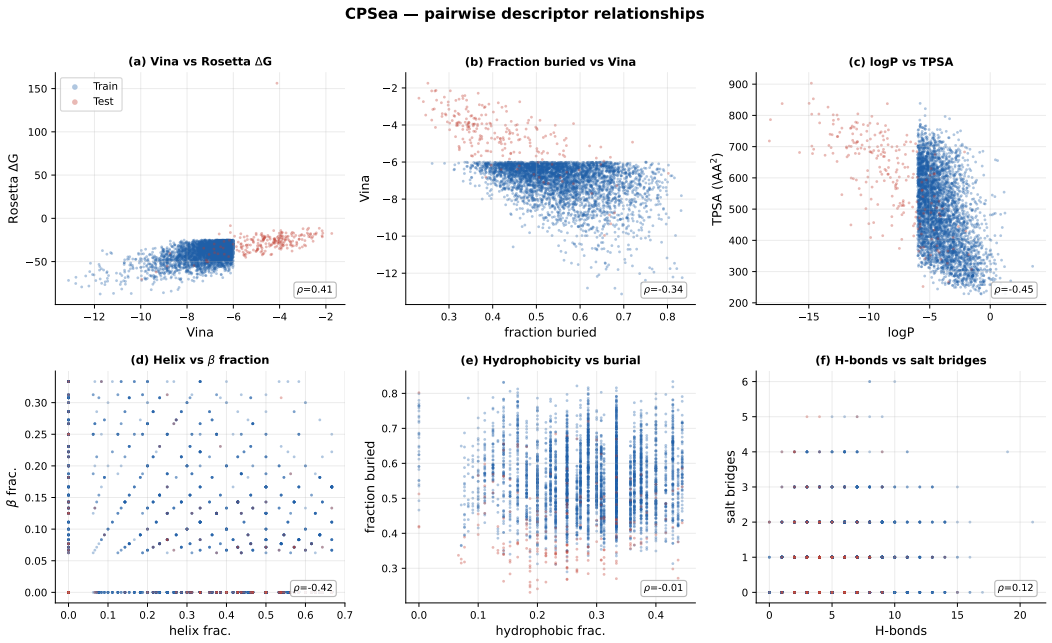


Figure 11: CPSea pairwise descriptor relationships (train blue, experimental test red; subsampled). Titles report the training-split Spearman correlation. (a) Agreement between the two in-silico binding oracles; (b) interface burial vs. Vina; (c) physicochemical space; (d) secondary structure; (e) hydrophobicity vs. burial; (f) interface hydrogen bonds vs. salt bridges.

-37 REU). Peptide length grows with interface area (Spearman ρ_{train}=0.72; Fig. 10) but not with docking score, and is similar across cyclization chemistries. The two in-silico binding oracles agree only moderately (Vina vs. Rosetta ΔG, ρ=0.41 train and 0.51 test; Fig. 11), motivating reporting both.

I.3 Evaluation protocol and limitations

We recommend the sequence-de-leaked experimental test set as the headline evaluation, with in-silico oracles (AutoDock Vina, Rosetta ΔG, Ramachandran/PLIP validity, and a folding-based designability check) as secondary metrics. CPCore is systematically more idealized than the experimental test distribution: Ramachandran favoured is $\approx$1.0 versus 0.90 (Cliff's δ=0.60), interfaces are more buried (δ=0.69), and Vina scores are $\approx$2.5 kcal/mol more favourable (δ=0.89). Because these quantities also participate in filtering and evaluation, scoring only against training-distribution statistics would overestimate success; the experimental test set and an orthogonal folding check partly mitigate this circularity.

I.4 Structural train–test leakage audit

A 30% sequence-identity screen cannot detect remote structural homology. We therefore compare every one of the 226 test receptors against all 34,375 training receptors with Foldseek in TM-align mode, giving 7,768,750 alignments. For each test receptor we report the query-normalised TM-score of its single best training match. As a reference for the structural diversity already present within the test set, we repeat the same best-match calculation against the other test receptors.

Sequence de-leaking does not imply structural disjointness. The median sequence identity of the 226 best train–test structural matches is 4.9%, and exactly one exceeds the 30% sequence threshold. Thus, 225 best structural matches lie below the sequence screen, and 141 of the 226 test targets still have a training neighbour with TM-score > 0.5. Such remote fold reuse is expected when AFDB and PDB are sampled from the same protein universe unless structural similarity is explicitly used as a cross-split exclusion criterion.

Table 22: Structural similarity of the test set. Query-normalised TM-score of each test receptor's best match against training and against the remaining test receptors. The test-vs-test column measures within-test structural diversity rather than leakage.

	vs. train (34,375)		vs. test (reference)	
Best-match TM	targets	share	targets	share
> 0.4	208	92.0%	116	51.3%
> 0.5	141	62.4%	35	15.5%
> 0.6	80	35.4%	10	4.4%
> 0.7	34	15.0%	1	0.4%
> 0.8	11	4.9%	0	0.0%
median best-match TM	0.545		0.403	

Closure shows no detectable dependence on train–test fold similarity. The audit lets us test whether this residual structural similarity is associated with our headline closure result. It is not: per-target closure is essentially uncorrelated with best structural similarity to training ($r^2 = 0.003$). On the 85 targets with no training neighbour above TM-score 0.5, ADEL closes 46.6% of designs, compared with 47.6% on the 141 targets that do have such a neighbour. The estimated difference is $+0.98$ percentage points, with a 95% bootstrap confidence interval of $[-1.4, +3.4]$ percentage points. Covalent-geometry metrics likewise show no detectable shift between the two groups.

No test complex appears in training. The six most structurally similar train–test pairs reach TM-score 0.85–0.90 but only 21–28% sequence identity, and correspond to AFDB predictions from different organisms matched to experimental PDB entries.

Two secondary quantities do improve on fold-similar targets. Cα-RMSD and binding energy per unit buried surface are better in this subset, but PepFlow and DiffPepBuilder show similar shifts, suggesting that these receptors are intrinsically easier for multiple methods rather than uniquely memorized by ADEL. Sequence recovery is less clear: BLOSUM62 is 0.031 higher on fold-similar targets for ADEL and 0.025 higher for DiffPepBuilder, whereas the other baselines show little effect. Structural similarity explains about 5% of the variation in this quantity, so we treat it as a caveat on sequence-recovery claims rather than on closure.

Limits of the audit. We did not compute the full structural similarity distribution within the 34,375-target training set; an exhaustive all-against-all search would require approximately 1.18 billion alignments. We instead report test-vs-test similarity as a reference for the diversity enforced by our Foldseek subsampling. We also audit whole-receptor folds rather than binding-site geometry; a pocket-level alignment would provide a sharper measure of local structural reuse. These remain limitations of the audit, but they do not change the empirical observation that closure shows no detectable dependence on receptor-level train–test fold similarity.

What the benchmark does and does not measure. (i) CPSea complexes excise a peptide segment and its receptor from the same parent chain, so the benchmark evaluates bound-state geometry and interface recovery against a known reference rather than true co-folding.

(ii) The peptides contain canonical L-amino acids and occupy a polar, non-permeability-optimized chemical regime; non-canonical and lipophilic macrocycles require a different corpus.

(iii) Cyclization is isopeptide-dominated (per family 84.0%; head-to-tail 10.6%; disulfide 5.4%), so all closure results are reported separately by chemistry as well as in aggregate.

(iv) The train–test exclusion criterion is 30% receptor sequence identity. This is sufficient for sequence de-leaking but not structural disjointness; Sec. I.4 quantifies the remaining remote structural similarity and finds no detectable dependence of the headline closure result on it.

(v) The headline test comprises 226 fold-diverse targets because target-conditioned evaluation is expensive. This increases uncertainty relative to the full pool, so comparisons are paired and bootstrapped over targets. The complete 8,253-complex sequence-de-leaked experimental pool remains available for larger-scale evaluation.

J ARCHITECTURE OF OUR MODEL

This appendix states each contribution as mathematics rather than description. We give three results that we consider load-bearing and prove each: that distilling the constraint projection alone is provably vacuous (Lemma 1), that our truth anchor cannot select a direction away from the native structure (Lemma 2), and that adding teacher heads can only enlarge the set of supervised atoms (Proposition J.4).

J.1 PRELIMINARIES

Backbone architecture. We inherit the backbone unchanged, and state its structure here because our distillation term interacts with the split between its two halves (Fig. 12). A *trunk* builds the conditioning representation: an input embedder (atom encoder, 3 blocks, 4 heads) that carries our constraint flags for cyclicity, the design mask and the binding site; a template module (2 blocks, weights frozen); a token-distance module (2 blocks); an MSA module (3 blocks), which is inactive here because cyclic peptide design runs single-sequence; and a stack of 12 *Miniformer* blocks with 16 heads. The trunk emits a single representation $s_{\mathrm{trunk}} \in \mathbb{R}^{384}$ and a pair representation $z_{\mathrm{pair}} \in \mathbb{R}^{128}$.

Each Miniformer block updates the pair stream and then the single stream, all residually: a triangular update on z_{pair}, a $4\times$-expansion transition on z_{pair}, a pair-biased attention that reads z_{pair} to update s_{trunk}, and a $4\times$-expansion transition on s_{trunk}. It is a lighter block than a full Pairformer, which applies four triangle operations per block rather than one; the configuration selects it through `use_miniformer`. Each layer of the diffusion module's token transformer is likewise residual: an adaptive layer norm conditioned on s_{trunk}, a pair-biased attention, and a conditioned transition (Fig. 12, bottom).

A *diffusion module*, the score model F_θ of Eq. 14, then maps noised coordinates to a coordinate update conditioned on $(s_{\mathrm{trunk}}, z_{\mathrm{pair}})$: an atom encoder (3 blocks, 4 heads, $a_s = 128$, $a_z = 16$), a token transformer (8 blocks, 16 heads) and an atom decoder (3 blocks). Auxiliary heads predict residue type, which is what makes the model a sequence-structure co-design model and is where our chemistry term attaches, alongside distogram, confidence and b-factor heads. Inputs are padded to 256 tokens and 2048 atoms.

Why the split matters for us. The trunk is evaluated once per training step and its output is reused for every subsequent call. Our teacher (Sec. J, Eq. 21) passes the same conditioning tensors to a second invocation of the score model, so its marginal cost is one diffusion-module forward rather than one full-model forward: the 12-block Miniformer stack is not recomputed. This is what makes a second-denoising-call teacher affordable enough to run during training at all.

Diffusion parameterisation. We inherit the EDM preconditioning of Karras et al. (2022) unchanged. Writing σ for the noise level, $\sigma_d = 16$ for the data scale and F_θ for the raw network, the denoised prediction is

$$\hat{x}_0 = c_{\text{skip}}(\sigma)\, x_\sigma + c_{\text{out}}(\sigma)\, F_\theta\big(c_{\text{in}}(\sigma)\, x_\sigma,\ c_{\text{noise}}(\sigma)\big), \tag{14}$$

$$c_{\text{skip}} = \frac{\sigma_d^2}{\sigma^2 + \sigma_d^2}, \quad c_{\text{out}} = \frac{\sigma\,\sigma_d}{\sqrt{\sigma^2 + \sigma_d^2}}, \quad c_{\text{in}} = \frac{1}{\sqrt{\sigma^2 + \sigma_d^2}}, \quad c_{\text{noise}} = \tfrac{1}{4}\log\frac{\sigma}{\sigma_d}, \tag{15}$$

with training noise drawn log-normally, $\sigma = \sigma_d \exp(P_{\text{mean}} + P_{\text{std}}\,\xi)$, $\xi \sim \mathcal{N}(0,1)$, $P_{\text{mean}} = -1.2$, $P_{\text{std}} = 1.5$. We write $D_\theta(x_\sigma, \sigma)$ for the map in Eq. 14.

Macrocyclisation constraints. A constraint is a tuple $s = (i, j, d^*, \tau, c)$: two atom indices that must form a bond, the ideal bond length, a tolerance, and a chemistry type $c \in \{\text{HT}, \text{ISO}, \text{SS}\}$ for head-to-tail, isopeptide and disulfide. We use $d^* = 1.33$ Å for the two amide chemistries, the Engh-Huber peptide C–N length (Engh & Huber, 1991), and $d^* = 2.05$ Å for disulfide, with $\tau = 0.5$ Å throughout training. Each constraint defines an *admissible set*

$$\mathcal{C}_s \;=\; \big\{\, x \in \mathbb{R}^{N\times 3} \;:\; \big|\, \|x_i - x_j\| - d^* \big| \le \tau \,\big\}, \tag{16}$$

a closed, non-convex set (it is the preimage of an annulus under the pair distance). We write x^* for the native structure, rigid-aligned to the prediction, and $e(x) = \max\big(|\,\|x_i - x_j\| - d^*| - \tau,\, 0\big)$ for the constraint violation, so $x \in \mathcal{C}_s \iff e(x) = 0$.

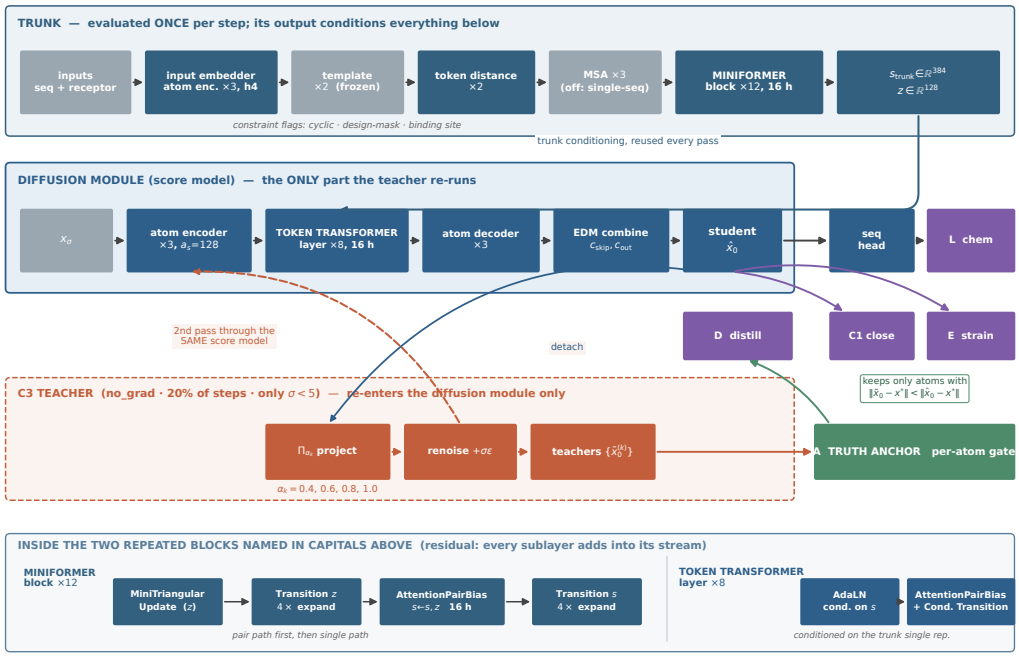


Figure 12: **The training graph of our model.** The backbone splits into a *trunk* (top), evaluated once per step, and a *diffusion module* (middle) conditioned on its output. We change neither. Our four terms attach at three places: two geometric penalties on $\hat{x}_0$, one cross-entropy on the residue-type head, and one distillation term whose target comes from the teacher (bottom). The teacher runs under `no_grad` on 20% of steps and only where $\sigma < 5$, and it re-enters *only* the diffusion module, reusing s_{trunk} and z_{pair}; the 12-block Miniformer stack is never recomputed. The bottom strip expands the two repeated blocks. The truth anchor sits between teacher and loss and decides, per atom, whether the teacher may teach at all. Greyed blocks are inactive or frozen in our setting.

Total objective. Figure 12 shows where each term attaches. Our four terms are added to the backbone loss $\mathcal{L}_{\text{base}}$ (diffusion, distogram, confidence, b-factor and residue-type):

$$\mathcal{L} = \mathcal{L}_{\text{base}} + \underbrace{\lambda_{\text{close}}\mathcal{L}_{\text{close}} + \lambda_{\text{E}}\mathcal{L}_{\text{strain}}}_{\text{closure constraints}} + \lambda_{\text{D}}\mathcal{L}_{\text{distill}} + \lambda_{\text{L}}\mathcal{L}_{\text{chem}}, \tag{17}$$

with $(\lambda_{\text{close}}, \lambda_{\text{E}}, \lambda_{\text{D}}, \lambda_{\text{L}}) = (1.0,\ 0.2,\ 0.3,\ 1.0)$. Only the four terms after $\mathcal{L}_{\text{base}}$ in Eq. 17 are ours.

The closure term. $\mathcal{L}_{\text{close}}$ is a squared hinge on the violation,

$$\mathcal{L}_{\text{close}}(x) \;=\; e(x)^2 \;=\; \Big(\max\big(\,|\,\|x_i - x_j\| - d^*|\,-\tau,\ 0\,\big)\Big)^2, \tag{18}$$

zero on $\mathcal{C}_s$ and quadratic outside. It constrains exactly one scalar functional of the structure. Everything we add exists because one scalar is not enough, and Lemma 1 makes that statement precise.

A note on the tolerance. Sec. 2 explains why the training band is wider than the evaluation window: the penalty acts on $\hat{x}_0$ at every noise level, and a 0.15 Å demand on a conditional mean fights the score at large σ. The implementation also supports making that explicit through a noise-dependent band $\tau_{\text{eff}}(\sigma) = \tau + \kappa\sigma$, which would tighten the constraint as σ falls instead of holding one compromise value throughout. We report all results at $\kappa = 0$, recovering the fixed band exactly, and note the mechanism because it is the natural route to closing the train/evaluation gap.

J.2 D: DISTILLING THE PREDICT-PROJECT-RENOISE OPERATOR

The operator distilled here is the published predict-project-renoise corrector of Huang et al. (2024). What is ours is the distillation, the truth-anchoring and the multi-head schedule, not the corrector itself.

The learned proximal operator we did not ship. The released model contains no corrector parameters: the projection is analytic, and the module implementing the learned alternative is never constructed, so its auxiliary loss is inert and absent from every logged run in the shipped lineage. The learned operator (a small MLP over constraint-relative per-atom features, with a Fourier embedding of $\log \sigma$ and a zero-initialised output layer, so that it begins as the identity) was trained on predicted-to-native pairs and abandoned for the reason given in Sec. 2: applied at every gated sampling step it compounds its own error, and the metrics moved adversely on every axis. We record the outcome as a design rationale rather than as a measured ablation, since we did not preserve that run's per-design scores; a reader who wants the comparison should treat it as motivation for distillation and not as evidence about learned correctors in general.

The training-free corrector applies a projection that moves only the two constrained atoms. With $m = \frac{1}{2}(x_i + x_j)$, $u = (x_i - x_j)/\|x_i - x_j\|$ and $t = \mathrm{clip}(\|x_i - x_j\|,\, d^* - \tau,\, d^* + \tau)$, define

$$\Pi_\alpha(x)_i = x_i + \alpha\big(m + \tfrac{t}{2}u - x_i\big), \qquad \Pi_\alpha(x)_j = x_j + \alpha\big(m - \tfrac{t}{2}u - x_j\big), \tag{19}$$

leaving all other atoms fixed, with strength $\alpha \in (0, 1]$. Proposition J.2 states the sense in which this is the right map, and Eq. 25 below is the loss it ultimately feeds.

[Minimal displacement] $\Pi_1(x)$ is the Euclidean projection of x onto $\mathcal{C}_s$ among maps that move only atoms i and j, and $\Pi_1(x) = x$ whenever $x \in \mathcal{C}_s$.

Proof. Any admissible relocation must set the pair distance to some $t' \in [d^* - \tau, d^* + \tau]$. For fixed t', the displacement $\|x_i - y_i\|^2 + \|x_j - y_j\|^2$ subject to $\|y_i - y_j\| = t'$ is minimised by moving both atoms symmetrically along u about their midpoint, giving $\frac{1}{2}(t' - d)^2$ with $d = \|x_i - x_j\|$. Minimising over $t' \in [d^* - \tau, d^* + \tau]$ gives $t' = \mathrm{clip}(d, d^* - \tau, d^* + \tau) = t$, which is Eq. 19 at $\alpha = 1$. When $x \in \mathcal{C}_s$ we have $t = d$ and the map is the identity. □

Distilling the projection would teach nothing new. It is tempting to supervise the model directly on the projection, that is on $\|\hat{x}_0 - \Pi_\alpha(\hat{x}_0)\|^2$. The following shows that this objective is not merely similar to the closure term but exactly proportional to it.

Lemma 1 (Vacuity of projection distillation)**.** *For every x and every $\alpha \in (0, 1]$,*

$$\big\|x - \Pi_\alpha(x)\big\|^2 \;=\; \frac{\alpha^2}{2}\, e(x)^2 \;=\; \frac{\alpha^2}{2}\, \mathcal{L}_{\text{close}}(x). \tag{20}$$

The two objectives therefore have identical gradients up to the constant $\alpha^2/2$, and a model trained on the left-hand side learns nothing that $\mathcal{L}_{\text{close}}$ does not already supply.

Proof. Write $d = \|x_i - x_j\|$, so $x_i = m + \frac{d}{2}u$ and $x_j = m - \frac{d}{2}u$. From Eq. 19, $\Pi_\alpha(x)_i - x_i = \frac{\alpha}{2}(t - d)\, u$ and $\Pi_\alpha(x)_j - x_j = -\frac{\alpha}{2}(t - d)\, u$. Since $\|u\| = 1$ and no other atom moves, $\|x - \Pi_\alpha(x)\|^2 = 2 \cdot \frac{\alpha^2}{4}(t - d)^2 = \frac{\alpha^2}{2}(t - d)^2$. It remains to show $|t - d| = e(x)$. If $d > d^* + \tau$ then $t = d^* + \tau$ and $|t - d| = d - d^* - \tau = |d - d^*| - \tau = e(x)$. If $d < d^* - \tau$ then $t = d^* - \tau$ and $|t - d| = d^* - \tau - d = |d - d^*| - \tau = e(x)$. Otherwise $t = d$ and both sides vanish. □

What we distil instead. Lemma 1 tells us the information is not in the projection but in what the network does *after* being restarted from a projected structure. This matches the corrector's measured behaviour, which is global rather than local: it improves clash by 25% and bond-angle Z by 31% against a closure gain of 4.9 percentage points. Our teacher is therefore the second denoising call,

$$\tilde{x}_0^{(\alpha)} \;=\; D_\theta\Big(\Pi_\alpha(\hat{x}_0) + \sigma\varepsilon,\; \sigma\Big), \qquad \varepsilon \sim \mathcal{N}(0, I), \qquad \text{under } \texttt{no_grad}. \tag{21}$$

Stochastic firing and the effective weight. The teacher is a second pass through the diffusion module, cheaper than a full forward because the trunk conditioning is reused, but not free; we fire it with probability $p = 0.2$ using a rank-synchronised Bernoulli draw, ensuring every data-parallel worker makes the same decision and the accumulated gradient stays well defined. We additionally gate to $\sigma < \sigma_{\max} = 5$, the regime in which the sampler applies the corrector. The contribution to the expected objective is therefore

$$\mathbb{E}\big[\lambda_{\text{D}}\mathcal{L}_{\text{distill}}\big] = \lambda_{\text{D}} \cdot p \cdot \mathbb{E}\big[\mathcal{L}_{\text{distill}} \,\big|\, \text{fired}\big], \tag{22}$$

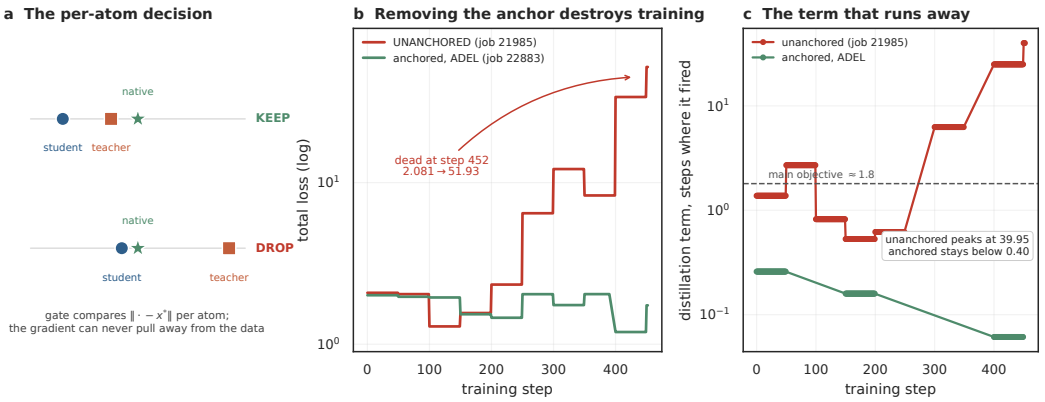


Figure 13: **The truth anchor is a stability requirement, not a refinement.** *(a)* A per-atom comparison against the native structure; the term survives only where the teacher is strictly closer to the truth. *(b, c)* Real training curves. Without the anchor the objective has no fixed point and the run dies at step 452; the identical recipe with the gate is stable. Panel (c) shows only steps on which the teacher fired.

so the nominal $\lambda_{\mathrm{D}} = 0.3$ acts as an effective weight of 0.06. Student and teacher arise from the same network on different inputs and may differ by a rigid transform, so we rigid-align $\tilde{x}_0$ and x^* into the frame of $\hat{x}_0$ before any distance is evaluated; without this the loss would measure a rigid offset rather than geometry.

J.3 A: THE TRUTH ANCHOR

The failure without an anchor. Equation 21 defines a target produced by the model being trained. Consider the naive objective $\|\hat{x}_0 - \mathrm{sg}[T(\hat{x}_0)]\|^2$ where T is the teacher map and sg is the stop-gradient. Every fixed point of T is a global minimiser, and the set of such points is not related to the data in any way: the student can reduce the objective to zero by drifting to any configuration on which projection followed by renoising is idempotent. The system has no attractor at x^*. We observe this at two weights, and the severity scales with the weight. At $\lambda_{\mathrm{D}} = 1.0$ an unanchored run was healthy through its 200 warmup steps at weight zero, then within 200 steps of the weight engaging the total loss went $2.081 \to 51.930$ while the distillation term alone reached 39.949 against a main objective near 1.8; it was dead at step 452 (Fig. 13b, c). At the weight we actually use, $\lambda_{\mathrm{D}} = 0.3$, an otherwise identical unanchored run completes $4{,}000$ steps, but the auxiliary term grows until it constitutes 96.7% of the training objective against 3.0% for the anchored model, and the loss trend reverses from decreasing to increasing. The anchor is therefore what keeps the term an auxiliary one rather than the objective.

The gate. For each atom a we compare both candidates against the native structure and keep the term only where the teacher is strictly better:

$$g_a \;=\; \mathbb{1}\big[\, \big\|\tilde{x}_{0,a} - x^*_a\big\| \;<\; \big\|\hat{x}_{0,a} - x^*_a\big\| \,\big], \qquad \text{evaluated under } \texttt{no_grad}. \tag{23}$$

The gate converts a self-referential objective into an anchored one, and the property it secures is stated below: at every supervised atom, a step toward the teacher reduces the distance to the native structure. The result is elementary by design. We chose the gate so that this holds by construction rather than by tuning, which is what allows the term to be added to an existing objective without a stability analysis of its own. Its scope is coordinate space: it characterises the direction of an idealised step at a single atom. How that composes into a parameter update, under shared weights, an aggregated Huber loss and gradient clipping, is a separate and empirical question, and Fig. 13 answers it.

Lemma 2 (The gate cannot select an ascent direction)**.** *Fix an atom a with $g_a = 1$ and write $\mathbf{a} = \hat{x}_{0,a} - x^*_a$, $\mathbf{b} = \tilde{x}_{0,a} - x^*_a$, so that $\|\mathbf{b}\| < \|\mathbf{a}\|$. Let $f(\eta) = \big\|\hat{x}_{0,a} + \eta(\tilde{x}_{0,a} - \hat{x}_{0,a}) - x^*_a\big\|^2$ be the squared distance to the truth after a step of size η toward the teacher. Then*

$$f'(0) = 2\big(\langle \mathbf{a}, \mathbf{b}\rangle - \|\mathbf{a}\|^2\big) < 0, \qquad \textit{and} \qquad \eta^* = \frac{\|\mathbf{a}\|^2 - \langle \mathbf{a}, \mathbf{b}\rangle}{\|\mathbf{b} - \mathbf{a}\|^2} \;\geq\; \tfrac{1}{2}, \tag{24}$$

where η^* minimises f. Every gated atom therefore admits a strict descent direction toward the native structure, and the minimiser lies at least halfway to the teacher. The argument is a convexity one, and the bound on η^* follows from the same inequality as the sign of $f'(0)$. We state the result explicitly because it delimits the gate exactly: it fixes the sign of the coordinate-space step, which is the property the anchor was introduced to provide.

Proof. $f(\eta) = \|\mathbf{a}+\eta(\mathbf{b}-\mathbf{a})\|^2$ is a convex quadratic with $f'(0) = 2\langle \mathbf{a}, \mathbf{b}-\mathbf{a}\rangle = 2(\langle \mathbf{a}, \mathbf{b}\rangle - \|\mathbf{a}\|^2)$. By Cauchy-Schwarz and the gate condition, $\langle \mathbf{a}, \mathbf{b}\rangle \le \|\mathbf{a}\|\|\mathbf{b}\| < \|\mathbf{a}\|^2$, so $f'(0) < 0$. Setting $f'(\eta) = 0$ gives the stated η^*, which is positive by the same inequality. For the bound, $\eta^* \ge \frac{1}{2}$ is equivalent to $2(\|\mathbf{a}\|^2 - \langle \mathbf{a}, \mathbf{b}\rangle) \ge \|\mathbf{b}\|^2 - 2\langle \mathbf{a}, \mathbf{b}\rangle + \|\mathbf{a}\|^2$, that is to $\|\mathbf{a}\|^2 \ge \|\mathbf{b}\|^2$, which is the gate condition. □

The gate is not a heuristic filter. By Lemma 2 the term applies only at atoms where movement toward the target reduces the distance to the native structure, so no supervised atom contributes a direction pointing away from the data. The effect of that on the trained model is measured rather than assumed: Fig. 13 shows the objective with and without the anchor, and the difference is unambiguous. The converse is equally concrete. Among atom pairs where the teacher is *worse* than the student, which the gate discards, an infinitesimal step toward the teacher increases the distance to truth for a substantial fraction of configurations; in a numerical check over 10^5 random instances this occurred 18.2% of the time. Those are precisely the updates the anchor removes.

Form of the term. Let $M_a = g_a \cdot m_a^{\text{atom}} \cdot m_a^{\text{design}} \cdot \mathbb{1}[\sigma < \sigma_{\max}]$ be the full mask. The loss is a Huber function of the per-atom displacement, averaged over atoms:

$$\mathcal{L}_{\text{distill}} = \frac{\sum_a M_a \, \mathcal{H}_\delta\big(\|\hat{x}_{0,a} - \tilde{x}_{0,a}\|\big)}{\max\big(\sum_a M_a,\, 1\big)}, \qquad \mathcal{H}_\delta(r) = \begin{cases} \frac{1}{2}r^2, & r \le \delta, \\ \delta\big(r - \frac{1}{2}\delta\big), & r > \delta, \end{cases} \tag{25}$$

with $\delta = 1.0$. Three choices here each repair a failure we encountered.

(i) Huber rather than a clamp. Writing r_a for the per-atom distance,

$$\Big\| \frac{\partial \mathcal{H}_\delta(r_a)}{\partial \hat{x}_{0,a}} \Big\| = \min(r_a, \delta) \;\le\; \delta, \qquad \text{whereas} \qquad \frac{\partial}{\partial \hat{x}_{0,a}} \min\big(\mathcal{L}, c\big) = 0 \;\text{ when } \mathcal{L} > c. \tag{26}$$

A hard clamp on the mean has *exactly zero* gradient above the cap, so it switches the term off precisely in the regime where it is needed; Huber bounds the per-atom gradient magnitude at δ while preserving its direction. Bounding the gradient, not the loss, is the actual requirement: both divergences we recorded came from gradient magnitude.

(ii) The denominator counts atoms. Normalising by $\sum_a M_a$ rather than by the sample count keeps the term scale-free in the number of supervised atoms. An earlier term that normalised by samples reached 1.25×10^{16}.

(iii) The design mask. Restricting to macrocycle atoms prevents the receptor, which the corrector never touches, from dominating the average.

J.4 Multi-head teachers: a measured trade-off

By Lemma 1 the displacement induced by Π_α scales as $\alpha^2 e(x)^2$, so a single α fixes one displacement schedule for every structure at every noise level. That is an ill-suited prior: a large α rescues a badly open ring and over-corrects a nearly closed one. We therefore run K teachers on a spread of strengths,

$$\alpha_k = \alpha_{\text{base}} \left(0.4 + 0.6\,\frac{k}{K-1}\right), \quad k = 0, \ldots, K{-}1 \qquad \Longrightarrow \qquad \alpha \in \{0.4,\, 0.6,\, 0.8,\, 1.0\} \;\text{ at } K = 4, \tag{27}$$

and select per atom the head closest to the native structure,

$$k_a^* = \arg\min_k \big\|\tilde{x}_{0,a}^{(k)} - x_a^*\big\|, \qquad \tilde{x}_{0,a} = \tilde{x}_{0,a}^{(k_a^*)}. \tag{28}$$

[Monotone coverage in K] Let $G_K = \{\, a : g_a = 1 \,\}$ be the set of supervised atoms when the target is formed by Eq. 28 over heads $\{1, \ldots, K\}$. Then $G_K = \bigcup_{k \le K} G^{(k)}$, where $G^{(k)}$ is the gated set of

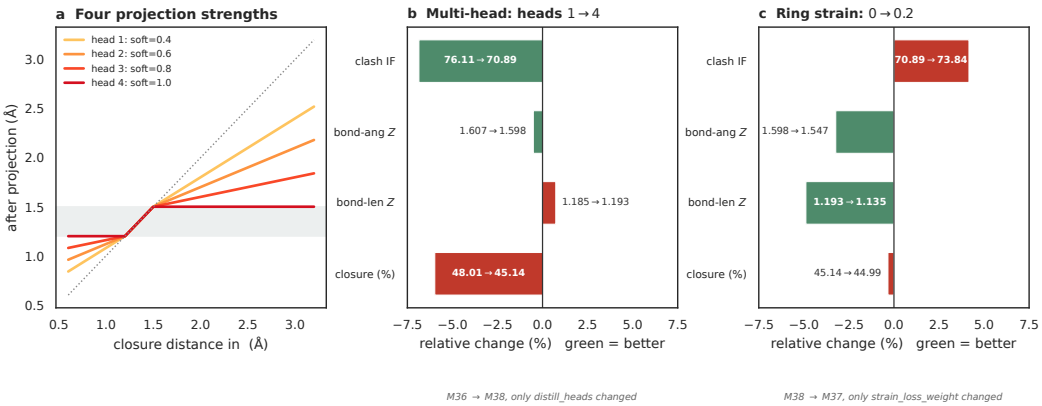


Figure 14: **Projection strength and what changing it buys.** *(a)* The four heads apply Eq. 19 at $\alpha = 0.4, 0.6, 0.8, 1.0$; the shaded band is $\mathcal{C}_s$. *(b, c)* Single-variable comparisons at a matched 4,000-step budget from a common parent.

head k alone. Consequently $G_K \supseteq G_{K'}$ for $K \geq K'$: adding heads can only enlarge the supervised set, and can never introduce an atom whose target is worse than the student's own prediction.

Proof. Selection and gating use the same functional $\|\cdot - x_a^*\|$. Hence $g_a = \mathbb{1}\big[\min_{k\leq K} \|\tilde{x}_{0,a}^{(k)} - x_a^*\| < \|\hat{x}_{0,a} - x_a^*\|\big] = \max_{k\leq K} \mathbb{1}\big[\|\tilde{x}_{0,a}^{(k)} - x_a^*\| < \|\hat{x}_{0,a} - x_a^*\|\big]$, which is the indicator of $\bigcup_{k\leq K} G^{(k)}$. Monotonicity in K is immediate, and every $a \in G_K$ satisfies the gate by construction, so Lemma 2 applies unchanged. □

Proposition J.4 is why reusing the anchor's criterion for selection is safe rather than merely convenient: it is what makes extra heads monotone. Note also that selection occurs after rigid alignment, so Eq. 28 compares geometry and not rigid offsets, and that when the anchor is disabled the minimisation in Eq. 28 has no reference; the implementation then falls back to the mean $\tilde{x}_{0,a} = K^{-1} \sum_k \tilde{x}_{0,a}^{(k)}$.

The head count is a measured trade-off, not a uniform gain. Figure 14b isolates K at fixed settings by changing only the head count from 1 to 4. Multi-head supervision improves angular geometry and interface packing while reducing closure. Because ring strain acts on related geometry, we read that arm together with Fig. 14c, which changes only λ_{E} from the same four-head configuration; the three arms share a parent and a budget, so each adjacent pair separates one of the two. The remaining cell of that grid, one head with the strain term on, was not run, so we separate the two effects without estimating their interaction. We therefore treat $K = 4$ as a measured geometry–closure trade-off rather than a uniformly improving component.

J.5 Closure constraints, part two: Engh-Huber-informed ring strain

By Lemma 1, $\mathcal{L}_{\text{close}}$ controls exactly one scalar, the pair distance d. The fibre $\{x : d(x) = d^*\}$ is enormous, and contains configurations whose local geometry is chemically implausible. We constrain two further functionals: the bond angles flanking the closing bond (Fig. 15a),

$$\theta_1 = \angle\big(\mathrm{CA}_i,\, \ell_i,\, \ell_j\big), \qquad \theta_2 = \angle\big(\ell_i,\, \ell_j,\, \mathrm{CA}_j\big), \qquad \angle(a, o, c) = \arccos \frac{\langle a - o,\; c - o\rangle}{\|a - o\|\,\|c - o\|}, \tag{29}$$

penalised by the same squared-hinge form as Eq. 18, evaluated in radians:

$$\mathcal{L}_{\text{strain}}(x) = \sum_{n\in\{1,2\}} \Big(\max\big(|\theta_n - \theta_0| - \tau_\theta,\; 0\big)\Big)^2, \qquad \theta_0 = 111^\circ \text{ (Engh \& Huber, 1991)}, \quad \tau_\theta = 25^\circ. \tag{30}$$

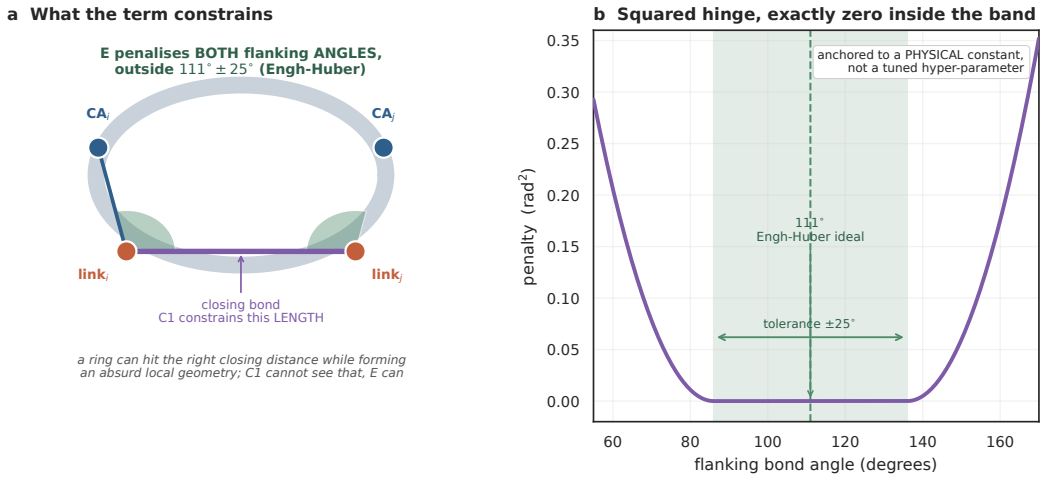


Figure 15: **Ring strain constrains geometry the closure term cannot see.** *(a)* The two bond angles flanking the closing bond. *(b)* A squared hinge in radians, exactly zero inside $111° \pm 25°$, anchored to a tetrahedral reference informed by Engh-Huber rather than a swept threshold.

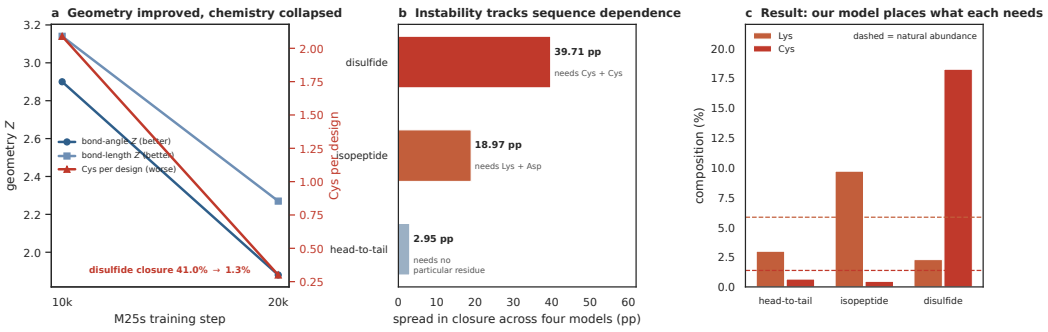


Figure 16: **Why a chemistry term is necessary, and what it produces.** *(a)* In an earlier run every geometry metric improved between step 10,000 and 20,000 while cysteine content collapsed, because only geometry carried a gradient. *(b)* Closure instability tracks how much sequence information each linkage requires. *(c)* Our final model places the residue each chemistry needs, and places it only where it is needed.

Numerical conditioning. The derivative $\mathrm{d}\arccos(z)/\mathrm{d}z = -(1-z^2)^{-1/2}$ is unbounded as $z \to \pm 1$, that is at collinear configurations. We clamp the cosine to $|z| \le 1 - 10^{-4}$, which bounds the angular gradient by $(1 - 0.9999^2)^{-1/2} \approx 70.7$ and keeps the term finite for degenerate geometries. The norms in the denominator are floored at 10^{-6} for the same reason. The term returns zero where the flanking CA atoms are unavailable, so it is safe to evaluate on any constraint.

θ_0 is tabulated, not tuned. The value $111°$ is the Engh-Huber tetrahedral reference for a backbone carbon, taken from small-molecule crystallography and used here as a generic centre rather than a per-linkage ideal. What matters for our purposes is not that it is exact for any one linkage, which it is not, but that it was fixed before we saw a result: it is a tabulated value rather than a knob turned toward our evaluation metric. Changing only λ_{E} from 0 to 0.2, with parent checkpoint, budget, batch size and every other key held fixed, improves bond-length Z by 0.058 and bond-angle Z by 0.052, which are $7\times$ and $4\times$ our replicate error bars, at a closure change of 0.15 percentage points that lies inside noise (Fig. 14c). These are exactly the two quantities Eq. 30 constrains.

J.6 L: LINKAGE CHEMISTRY

Every term above is a functional of coordinates alone. But the admissible set of Eq. 16 is defined on *atoms*, and for two of the three chemistries the identity of those atoms is determined by the residue type: an isopeptide bond requires (LYS, ASP) and a disulfide requires (CYS, CYS). Formally, $\mathcal{C}_s$ is well posed only once the sequence supplies atoms i and j, and no coordinate objective supplies that.

The measured consequence. In an earlier run, between step 10,000 and 20,000, bond-length Z improved $3.14 \rightarrow 2.27$, bond-angle Z improved $2.90 \rightarrow 1.88$ and interface clash improved $92.4 \rightarrow 71.1$, while cysteine content fell $2.09 \rightarrow 0.30$ per design and disulfide closure collapsed from 41.03% to 1.32%. Notably 39% of those designs still placed atoms inside the sulphur-sulphur distance window: the right shape with the wrong element, which is exactly the failure mode a coordinate-only objective cannot see. Geometry and chemistry moved in opposite directions in the same run because only one had a gradient. Across four of our models, closure spread is 2.95 percentage points for head-to-tail, which requires no particular residue, against 18.97 for isopeptide and 39.71 for disulfide (Fig. 16b): variance tracks sequence dependence.

The term. We place a cross-entropy on residue identity at the two linkage positions, read from the sequence head the backbone already carries. Let t_i, t_j be the token indices of the linkage residues and $(r_i(c), r_j(c))$ the pair demanded by chemistry c:

$$\mathcal{L}_{\text{chem}} = \tfrac{1}{2}\Big[\text{CE}\big(z_{t_i},\, r_i(c)\big) + \text{CE}\big(z_{t_j},\, r_j(c)\big)\Big], \qquad (r_i, r_j) = \begin{cases} (\text{LYS}, \text{ASP}) & c = \text{ISO}, \\ (\text{CYS}, \text{CYS}) & c = \text{SS}, \\ \varnothing & c = \text{HT}, \end{cases} \tag{31}$$

where z_t are the residue-type logits. Head-to-tail constraints map to $\varnothing$ and are dropped through `ignore_index`, so they contribute nothing by construction; the term is identically zero when no sample in a batch carries a chemistry-constrained linkage. Because the sequence head runs once per diffusion sample, the targets are repeated along the multiplicity axis to match.

Ordering matters, and so does the guard. We apply Eq. 31 as a *warm start* rather than jointly. Training it concurrently with the geometric terms from a common initialisation left it permanently inert, whereas sequencing it does not. This is a property of the recipe, not of the loss, and it has a consequence we state plainly: once our model is warm-started from a chemistry-corrected parent, the term is numerically negligible during the second phase, peaking at 1.1×10^{-5} against 0.42 in the run where chemistry was the objective. It behaves as a guard that is a no-op when the prediction is already admissible. The leave-one-out arm that isolates the guard shows it is nonetheless load-bearing. Removing it while keeping the chemistry-corrected parent degrades bond-angle Z on 88.1% of held-out targets, together with interface clash and ω deviation (Sec. E). The ordering across three arms is what identifies the mechanism: the unguarded arm is worse than *both* our model and an arm in which the linkage chemistry was never learned at all. A teaching signal would rank these monotonically in how much chemistry each arm received. A constraint predicts exactly what we observe, because a model that never acquired the chemistry has nothing to lose, whereas one that acquired it and then trains under the geometric objectives with nothing holding it in place drifts away from it. The rarity with which the term fires is therefore a property of a guard doing its job, not evidence that it is inert.

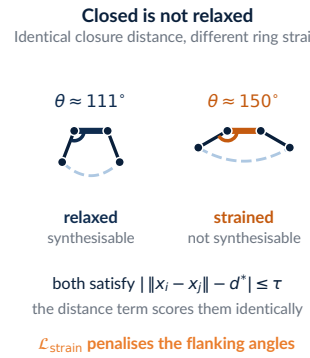


Figure 17: A distance constraint cannot separate a relaxed ring from a strained one.

What our model produces. Lysine reaches 9.70% in isopeptide designs against 2.99% and 2.29% where it is not required and 5.84% in natural proteins; cysteine reaches 18.24% in disulfide designs against 0.65% and 0.46% elsewhere and 1.38% naturally (Fig. 16c). The enrichment is specific to the chemistry that requires it.

Table 23: Our three contributions, their components, the mechanism of each, the formal guarantee where one exists, and the comparison that isolates it. No layer, attention pattern or parameter count changes, so these components add no model-side inference cost.

Contribution	Component	Mechanism	Guarantee	Isolating evidence
1. Truth-anchored multi-head distillation	**A** truth anchor	Eq. 23	Lemma 2	5 metrics worse; bond-len Z on 95.1% of targets
	D corrector distillation	Eq. 21	Lemma 1	bond-ang Z -0.110, better on 89.8% of targets
	multi-head (K=4)	Eq. 28	Prop. J.4	K:1→4 alone: bond-ang Z -0.068, closure -2.44 pp
2. Closure constraints	closure, bond length	Eq. 18	n/a	closure $39.53 \rightarrow 34.04\%$ without it
	E ring strain	Eq. 30	n/a	bond-len Z -0.058, better on 85.0% of targets
3. Linkage chemistry	**L** residue identity	Eq. 31	n/a	bond-ang Z -0.083, better on 88.1% of targets

J.7 The corrector at sampling time

Everything above treats Π_α as a teacher. It is also the operator the sampler runs, and since the two uses share more than notation it is worth setting down how it behaves at inference and why we configured it as we did.

The rule itself is short. Whenever a denoising step falls below $\sigma_{\max} = 5$ we take the denoised prediction $\hat{x}_0$, project the two designated atoms using Eq. 19 at strength $\alpha = 0.5$, and renoise back to the current level:

$$x_{t-1} \;=\; \Pi_\alpha\big(\hat{x}_0\big) \;+\; \sigma_t\,\varepsilon, \qquad \varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{32}$$

Three choices in that sentence deserve a reason.

The projection replaces the update rather than following it. Equation 32 stands in for the ordinary EDM step; it is not applied on top of one. The alternative costs a forward pass: if the corrector ran after the usual update, the network would have to denoise the corrected state before sampling could continue, and an operator that fires across most of the late trajectory would nearly double the cost of that stretch. Substituting leaves the budget where it was. A gated step performs one denoiser call, exactly as an ungated one does, and the only arithmetic we add moves two atoms of an all-atom structure. This is also why we can call the corrector free in practice while Sec. 2 still motivates distillation by the cost of the published variant, which adds the pass we avoid.

The strength is $\alpha = 0.5$ rather than a hard snap. Choosing $\alpha = 1$ lands on $\mathcal{C}_s$ in a single step, which sounds preferable until one remembers that the projection moves two atoms and leaves every neighbour where it was. Early in the gated range the required displacement can be large, and taking all of it at once pulls the closing atoms away from bond geometry that was otherwise sound. A half step is the conservative reading of Eq. 4: each firing removes the fraction α of whatever violation remains, so across many gated steps the sequence still contracts toward the admissible band, but gradually enough that the surrounding structure relaxes around the constraint instead of being dragged by it.

The gate is deliberately late. Above $\sigma_{\max}$ the denoised prediction is close to a posterior mean over structures that have not yet settled into a fold, and projecting such a prediction pins a distance that carries little information about the final geometry. We let the global fold form first and intervene only once the prediction is specific enough for the closing distance to mean something. The same threshold gates the teacher during training, which is deliberate rather than incidental: distilling a correction the sampler would never apply teaches the model about a regime it does not encounter.

One interaction is worth recording, because it is easy to get wrong. Since Eq. 32 substitutes for the update, a corrected $\hat{x}_0$ that were also fed back as self-conditioning would end up conditioning its own successor, and nothing in the loop damps that path. Self-conditioning is therefore off in every reported run, and our implementation refuses the combination outright rather than letting it degrade quietly.

J.8 Summary

Table 23 collects the three contributions and the components that make them up. Two of the three carry formal guarantees, and every component is isolated by a comparison that changes it alone; where a single comparison would have moved two settings at once, we trained the intermediate arm that separates them rather than reporting the pair jointly. Two entries are reported as costs rather than gains: the truth anchor is a stability requirement rather than a performance contribution, and the multi-head teacher is a measured trade-off that buys backbone geometry and interface packing at the price of closure. We would rather name both as such than fold them into a claim of uniform improvement. All evidence in the final column is a paired per-target comparison with family-wise correction over the full set of tests (Sec. E).

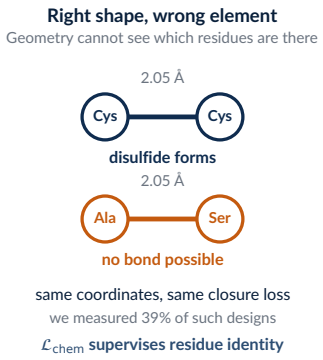


Figure 18: Identical coordinates, identical closure loss, and only one pair can bond.

Monotonicity of the multi-head supervised set. Selection reuses the gate's own criterion, and this is what makes extra teachers safe rather than merely convenient. Because $\min_k \|\tilde{x}_{0,a}^{(k)} - x_a^*\| < \|\hat{x}_{0,a} - x_a^*\|$ holds exactly when some individual head satisfies the gate, the supervised set is a union

$$G_K = \bigcup_{k \le K} G^{(k)}, \qquad \text{hence} \qquad G_K \supseteq G_{K'} \text{ for } K \ge K', \tag{33}$$

so adding teachers can only enlarge the supervised set and can never introduce an atom whose target is worse than the model's own prediction (Proposition J.4).

J.9 The learned corrector we trained and abandoned

Why not learn a better corrector. The obvious alternative is to keep the corrector and improve it: replace the analytic clamp, which knows the constraint and nothing about the data manifold, with a small network predicting a displacement toward the nearest point that is both feasible and plausible, trained on predicted-to-native pairs. We built and trained this operator before arriving at the objective below, and report the outcome because it shaped the design. It fails for a reason specific to the setting rather than to the parameterisation. A corrector fires at roughly 300 steps of a single trajectory, and each application conditions the next, so its error does not average out but accumulates along the path; a learned operator has error to accumulate where a projection onto a fixed distance has almost none. Every axis we tracked degraded. The operator also inherits its own training pathology: it is a refinement module, so it must map $\hat{x}_0$ to the native structure, and early in training $\hat{x}_0$ is noise and the displacement it is asked to predict is unbounded, which we could stabilise only by holding its weight at zero until the backbone was usable and clamping the term so that an outlying batch could not dominate the shared gradient-norm clip.

The failure is informative, and it is what selects the objective we do use. A component evaluated once per sample can afford to be learned; a component evaluated three hundred times per sample cannot, unless its error is bounded by construction, which is what the analytic projection buys at the price of knowing no chemistry. The way to get a learned operator's behaviour without its accumulation is to stop applying it at inference and to put it in the weights, where it is expressed once through the network that already runs. That is the distillation term: we distil the corrector rather than deploy it.

K Hyperparameters and Reproduction

We list every setting we changed and every setting a reader would need to reproduce our model. Values inherited unchanged from the backbone architecture and training recipe are included too, because "unchanged" is only useful information if the reader can see what it was.

Stage 3 additionally warms the distillation term over its first 200 steps and gates the teacher below $\sigma = 5$, firing on 20% of steps; the chemistry term has no warmup in either stage that uses it. We state the stages because Eq. 2 is the objective at convergence rather than a recipe for a single run: reproducing the reported model means running the three stages in this order, not optimising the sum from one initialisation.

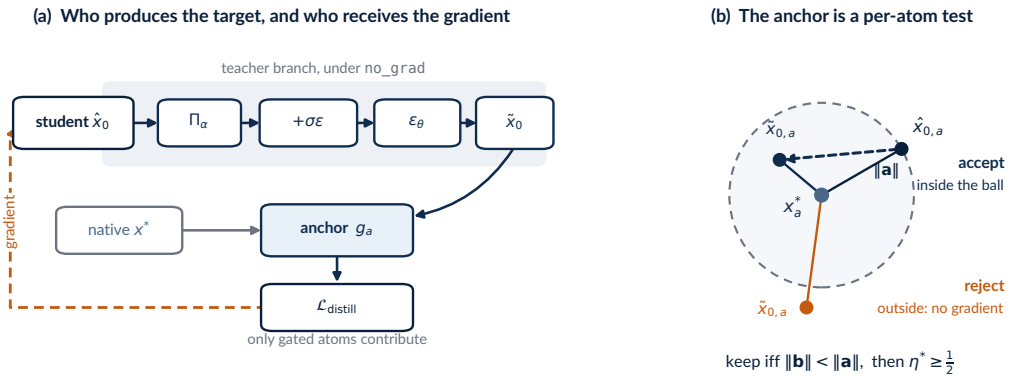


Figure 19: The student, the teacher and the anchor. **(a)** The teacher is the student's own network applied a second time: the prediction is projected by Π_α, renoised, and denoised again. That entire branch runs under `no_grad`, so the teacher supplies a target and never a gradient, and the gradient of $\mathcal{L}_{\text{distill}}$ returns only to the student. The anchor sits between the teacher and the loss and consults the native structure, which the teacher itself never sees. **(b)** What the anchor decides, for one atom. The dashed ball is centred on the native position with radius $\|\mathbf{a}\|$, the student's current error. A teacher landing inside the ball is closer to the truth than the student, so the atom is supervised and Lemma 2 gives a descent direction with $\eta^* \geq \frac{1}{2}$. A teacher landing outside is discarded and contributes nothing. On 10^5 sampled instances, 18.2% of atoms fall outside.

Table 24: The three training stages. Each initialises from the previous checkpoint and does not carry over optimiser state. All stages share the AF3 schedule with max lr $= 2 \times 10^{-4}$, 500 warmup steps, weight decay 0.003 and an effective batch of 96; they differ only in which terms are active and for how long. The base model inherits the BoltzGen backbone, so no stage trains from random initialisation.

Stage	Active terms	Steps	Role
1. Base	$\mathcal{L}_{\text{base}} + \mathcal{L}_{\text{close}}$	26,317	closure only, from the pretrained backbone
2. Chemistry	$+ \mathcal{L}_{\text{chem}}$ (λ_{L}=1.0)	4,000	corrects linkage residues before geometry
3. ADEL	$+ \mathcal{L}_{\text{distill}}$ (λ_{D}=0.3) $+ \mathcal{L}_{\text{strain}}$ (λ_{E}=0.2)	11,000	the model we report

K.1 Our four terms

Table 25 is the complete specification of our contribution: ten values, against the 226 leaf keys of the full configuration. Two deserve a note. The distillation weight of 0.3 is a nominal value: because the teacher fires on a fraction $p = 0.2$ of steps, its expected contribution is 0.06. And `distill_truth_anchor` does not appear in the released configuration files, because it defaults to `true`; we added the flag solely so that the ablation of Sec. E could switch the anchor off, and every run predating it is bit-identical.

K.2 Optimisation

Table 26 lists the optimisation settings. The effective batch of 96 is the product of devices and gradient accumulation, and every ablation arm preserves it: the arms that ran on 4 GPUs use `accumulate_grad_batches` $= 24$. Because the learning-rate decay is scheduled to begin at step 50,000 and no run here exceeds 15,000, all of our training happens at the constant peak rate after warmup. The ADEL stage took approximately 2.4 days of wall-clock on one 8× A100 node.

K.3 Architecture and diffusion process

Table 27 gives the architecture and diffusion process. The design task is conditioned through embedder flags for cyclicity, the design mask, the binding-site specification and secondary structure, and the

Table 25: The settings that define ADEL. These ten values are the entire difference between our model and the chemistry-corrected parent it starts from.

Component	Key	Value	Notes
closure	`closure_loss_weight`	1.0	squared hinge, Eq. 18
	`closure_tol` (τ)	0.5 Å	class default, not overridden
D distillation	`distill_c3`	`true`	teacher enabled
	`distill_loss_weight`	0.3	effective 0.06 after firing prob.
	`distill_prob`	0.2	rank-synchronised Bernoulli
	`distill_below_sigma`	5.0	gate matching the sampler
multi-head teacher	`distill_heads` (K)	4	$\alpha \in \{0.4, 0.6, 0.8, 1.0\}$
	`distill_soft` (α_{base})	1.0	ladder scales from this
A truth anchor	`distill_truth_anchor`	`true`	default; `false` diverges
E ring strain	`strain_loss_weight`	0.2	$111^\circ \pm 25^\circ$, Eq. 30
L linkage chemistry	`chem_loss`	`true`	guard during geometry training
	`chem_loss_weight`	1.0	inherited from the parent stage
warmups	`distill_warmup_steps`	200	weight held at 0 before this
	`chem_warmup_steps`, `warmup_steps`	0	closure and chemistry already trained in

Table 26: Optimisation and trainer settings for the ADEL stage. All values are inherited from the backbone recipe; we changed none of them. §: the reported run uses seed 42; App. E.1 retrains the same recipe under two further seeds and reports the run-to-run spread.

Setting	Value	Setting	Value
optimiser	Adam	devices	8× A100
β_1, β_2	0.9, 0.95	`batch_size`	1
ϵ	10^{-8}	`accumulate_grad_batches`	12
peak learning rate	2×10^{-4}	**effective batch**	**96**
warmup steps	500	`diffusion_multiplicity`	12
LR schedule	AF3	precision	`bf16-mixed`
decay starts at step	50,000 (never reached)	gradient clip	10.0
decay factor / period	0.95 / 50,000	EMA decay	0.999
weight decay	0.003 (excl. norms/bias)	random seed	42§
train diffusion steps	20	recycling steps	3

data selector applies structure conditioning with probability 0.4, secondary-structure conditioning with probability 0.1, and coordinate noise of 1 Å standard deviation to conditioned structures.

K.4 Data

Table 28 gives the partition. Sampling uses a cluster sampler over 100,000 samples per epoch, a multimer cropper, and a size filter admitting 1 to 8 chains. MSAs are disabled and all training is single-sequence, which is appropriate for short macrocyclic peptides and is why the template and MSA modules are inert.

K.5 Evaluation protocol

Table 29 reports the sampling budget and inference settings used for ADEL. Every method is evaluated on the same 226 held-out targets with 40 generated designs per target, giving 9,040 designs per model, and all outputs are rescored with the same evaluation pipeline. Each baseline otherwise uses the sampling procedure specified by its released implementation.

Scoring. Metrics are computed in two passes: a fast pass covering cyclisation, geometry, validity and co-design quality, and a Rosetta pass for interface energy. Designs are chains A and B while native references are chains L and R, so the binder-target chain mapping must be supplied explicitly

Table 27: Architecture and EDM settings, all inherited unchanged. Of 142.20M parameters, 142.06M (99.90%) receive gradients; only the 0.15M-parameter template module is frozen, and it is inactive here because design runs single-sequence.

Setting	Value	Setting	Value
`token_s` / `token_z`	384 / 128	σ_{data}	16.0
`atom_s` / `atom_z`	128 / 16	$\sigma_{\min}$, $\sigma_{\max}$	4×10^{-4}, 160
input embedder (atom enc.)	3 blocks, 4 h	ρ	7
template module	2 blocks (frozen)	P_{mean}, P_{std}	-1.2, 1.5
token-distance module	2 blocks	γ_0, $\gamma_{\min}$	0.8, 1.0
MSA module	3 blocks (inactive)	noise / step scale	1.0 / 1.0
Miniformer trunk	**12 blocks, 16 h**	coordinate augmentation	`true`
token transformer	8 blocks, 16 h	rotational alignment	`true`
atom decoder	3 blocks, 4 h	`max_tokens`	256
Fourier dim	256	`max_atoms`	2048

Backbone loss weights (unchanged): diffusion 4.0, distogram 3×10^{-2}, residue type 3×10^{-2}, b-factor 10^{-3}, confidence 10^{-4}; smooth-lDDT enabled, explicit bond loss disabled.

Table 28: Dataset partition. Splits are disjoint by target and the test split was never seen during any stage of the lineage.

Split	Targets	Use
train	34,375	all three training stages
validation	3,741	monitoring only
test	**226**	every number we report
total	38,342	

or every native-dependent metric silently voids. Closure is judged against a type-aware window: 1.20 to 1.50 Å for amide linkages and the corresponding sulphur-sulphur window for disulfides.

K.6 Software and reproduction

Our runs used PyTorch 2.9.0 with CUDA 12.8, PyTorch Lightning 2.6.4 and Python 3.12.13, on A100 GPUs in `bf16-mixed` precision with fused kernels enabled.

K.7 Ordering and weighting of the loss terms

Ordering matters, weighting does not. Once the sequence head places the required residues, Eq. 12 is satisfied and stays satisfied, so its loss mass is concentrated in the first few hundred steps. Raising λ_{L} therefore achieves nothing, and applying the term from a common initialisation leaves it inert, because the sequence head is still unformed when its window of influence passes. ADEL applies it as a warm start, which is a property of the schedule rather than of the loss.

K.8 Baseline protocol

The comparison is between models trained on the same data, not between released checkpoints of unknown provenance. Every baseline was trained with the recipe its own repository recommends, on the identical CPSea train/validation/test partition used for our model, so no baseline saw a test target during training and none is handicapped by a split it was not built for. Where a repository's recipe draws on an external augmentation corpus — ProtFrag for the PepGLAD lineage, which CP-Composer also inherits — that corpus is omitted, since including it would train one baseline on data no other model sees and break the same-split premise. The one exception to retraining is RFpeptides, which we ran from released weights and therefore report in Sec. H.10 as a reference point rather than as a member of this comparison.

We do not evaluate GeoCycler (Zhang et al., 2026) because no public implementation is available for retraining on our controlled CPSea split.

Table 29: Sampling budget and inference settings for ADEL. All methods are evaluated on the same 226 targets with 40 designs per target and are rescored using the same evaluator. Baselines otherwise follow their respective sampling procedures.

Setting	Value
designs per target	40
targets	226
designs total	**9,040**
sampling steps	200
recycling steps	3
devices	2× A100
wall-clock	3 h 12 m

What each model is given at inference. The receptor is supplied as the carved domain; the epitope is the set of receptor residues within 8 Å of the native peptide; and the peptide length is the native length. These three are identical across models. Nothing else from the native structure is provided, and in particular no native peptide coordinates or sequence are given to any model. Each model produces 40 designs for each of the 226 targets.

The linkage specification is not consumed by every model. The linkage chemistry and the two linkage positions come from the native annotation, but supplying a specification is not the same as conditioning on it, and the models differ in whether they have an interface that accepts one. BoltzGen does: its input schema carries a `cyclic` period and an explicit `bonds` list, and the per-target specifications we pass populate them. UniMoMo does not: its released code contains no cyclisation or linkage-chemistry path, and its generation entry point accepts a config, a checkpoint and a device, with no argument through which a bond could be named. We were unable to establish the position of the remaining baselines from their released code with the same confidence, and we state that rather than assume either answer.

This matters for how Table 1 should be read. A model with no interface for a side-chain linkage is not attempting the constraint and failing it; it is producing an unconditioned macrocycle that we then score against a constraint it was never told about. For the two sequence-dependent chemistries, which are 70.3% of the test set, the comparison is therefore a measure of benchmark coverage rather than of performance on a shared task. The 67 head-to-tail targets are the subset every model can express, and App. H.2 reports that subset separately for exactly this reason.

Harmonisation, checked rather than assumed. Every model but one returns 9,040 designs over 226 targets at a mean peptide length of 12.80 residues spanning 6 to 16, matching our model exactly. Two departures are worth recording. UniMoMo returns 9,032 designs, eight having failed to write, and its length column reads 2 residues higher on the 159 targets whose natives carry ACE and NME caps, because it counts the caps as residues; on the 67 uncapped head-to-tail targets it matches exactly. The underlying peptides are the same length in both cases. BoltzGen is additionally reported in two configurations, the design module alone and the full pipeline with inverse folding, folding and filtering, since the pipeline is what its authors recommend for use.

Scoring. All models are scored by one program on one machine, with the same chain mapping, the same cap-aware alignment and the same acceptance windows. No baseline is scored by numbers taken from its own paper.

L METRIC DEFINITIONS

Every metric below is computed by a single evaluator, applied identically to all models, against the same 226 native complexes. Unless stated otherwise a metric is first averaged over the 40 designs for a target and then over targets, so that a target with many failures cannot dominate; reported uncertainties are the standard deviation across targets.

Table 30: Every metric, its direction, and the constant that defines it. Rates are averaged over the 40 designs of a target and then over the 226 targets; ± in the main text is the standard deviation across targets.

Metric	Better	Range	Defining constant
Closure and chemistry			
linkage chemistry	↑	$[0.296, 1]$	required pair: Lys–Asp, Cys–Cys
closure	↑	$[0, 1]$	1.336 ± 0.15 Å; 2.05 ± 0.25 Å (SS)
closure deviation	↓	≥ 0	distance from d^*, reported whenever resolved
closed and valid	↑	$[0, 1]$	conjunction of the two above
Physical validity			
VALID	↑	$[0, 1]$	well-formed ∧ stereo
stereo	↑	$[0, 1]$	every residue L or Gly
fused atoms	↓	≥ 0	non-bonded pair < 1.0 Å
Geometry			
bond-length Z	↓	≥ 0	Engh–Huber, Eq. 35
bond-angle Z	↓	≥ 0	Engh–Huber, Eq. 35
internal clash	↓	≥ 0	per 1,000 atoms, $\delta = 0.4$ Å
interface clash	↓	≥ 0	per 1,000 atoms, $\delta = 0.4$ Å
Ramachandran favoured	↑	$[0, 100]$	98% contour, receptor-derived
Ramachandran outlier	↓	$[0, 100]$	outside 99.95% contour
ω deviation	↓	$[0, 90]$	departure from planarity, degrees
cis non-Pro	↓	$[0, 100]$	$\lvert\omega\rvert < 30^\circ$
twisted ω	↓	$[0, 100]$	$30^\circ \leq \lvert\omega\rvert \leq 150^\circ$
Sequence			
AAR	↑	$[0, 1]$	per-residue identity
BLOSUM62	↑	$\mathbb{R}$	best over circular rotations
Energetics and accuracy			
ΔG	↓	$\mathbb{R}$	REU, `InterfaceAnalyzerMover`
ΔG/100 Å^2	↓	$\mathbb{R}$	normalised by buried SASA
$\Delta\Delta G$	↓	$\mathbb{R}$	vs. crystal rescored, same build
beats native	↑	$[0, 1]$	fraction with $\Delta G <$ native
Cα-RMSD	↓	≥ 0	after optimal superposition

L.1 Closure and linkage chemistry

A macrocyclisation constraint $s = (i, j, d^*, \tau, c)$ is read from the native: the two atoms that bond, the ideal bond length, the tolerance, and the chemistry type. The design is then asked the same question the native answers.

Linkage chemistry. Whether the residues the bond requires are present at the two linkage positions: lysine and aspartate for isopeptide, two cysteines for disulfide. Head-to-tail closes through the backbone termini and requires no particular residue, so it passes by construction. Head-to-tail is 67 of 226 targets, which puts a **floor of** 0.296 on this metric over the full set; a score near 0.30 means no constrained site was satisfied, not partial credit. We therefore also report the rate over the 159 constrained targets alone.

Closure. Writing x_i, x_j for the two designated atoms, the design is closed when

$$\big|\, \|x_i - x_j\| - d^* \big| \leq \tau, \qquad (d^*, \tau) = \begin{cases} (1.336,\ 0.15)\ \text{Å} & c \in \{\text{HT}, \text{ISO}\}, \\ (2.05,\ 0.25)\ \text{Å} & c = \text{SS}. \end{cases} \tag{34}$$

The amide reference is the same 1.336 Å used as the C–N bond length in Eq. 35, so a closed head-to-tail ring is one whose closing bond is a peptide bond by the same standard applied to every other bond in the chain. *Closure deviation* is $\big|\, \|x_i - x_j\| - d^* \big|$ reported for every design whose closing pair resolves, passing or not, and is the quantity that shows how far a failure actually falls.

Closed and valid. The conjunction of closure and physical validity, per design. A ring that shuts inside a structure with broken covalent geometry or inverted stereocentres is not a usable design, and this joint rate is the metric we regard as decisive.

Alignment. Linkage positions are indices into the native chain, and all 159 constrained-chemistry natives carry ACE/NME terminal caps while the 67 head-to-tail natives do not. Models differ in whether they emit caps, so indices are aligned cap-aware before any linkage-dependent quantity is evaluated. Comparing by raw index shifts the lookup by one position for any uncapped model and scores present residues as absent.

L.2 Physical validity and geometry

VALID. The conjunction of *well-formed* and *stereochemically correct*. Well-formed requires finite coordinates, no fused atoms (no non-bonded heavy-atom pair closer than 1.0 Å), and bond-length and bond-angle Z below 10. Stereochemically correct requires every residue to be L or glycine.

Bond-length and bond-angle Z. For each covalent bond and each backbone angle we take the deviation from its Engh–Huber reference (Engh & Huber, 1991) in units of that reference's standard deviation, and report the root-mean-square over all such terms in the peptide:

$$Z_{\text{len}} = \sqrt{\frac{1}{|B|}\sum_{b\in B}\left(\frac{\ell_b-\ell_b^0}{\varsigma_b}\right)^2}, \qquad Z_{\text{ang}} = \sqrt{\frac{1}{|A|}\sum_{a\in A}\left(\frac{\theta_a-\theta_a^0}{\varsigma_a}\right)^2}. \tag{35}$$

The references (ℓ^0, ς) in Å are N–Cα $(1.459, 0.020)$, Cα–C $(1.525, 0.021)$, C–O $(1.229, 0.019)$ and the peptide bond C–N $(1.336, 0.023)$. The angle references (θ^0, ς) in degrees are N–Cα–C $(111.0, 2.7)$, Cα–C–O $(120.8, 1.7)$, Cα–C–N $(116.2, 2.0)$, C–N–Cα $(121.7, 2.5)$ and O–C–N $(123.0, 1.6)$. Because these are population Z-scores, a value near 1.0 is what a well-resolved crystal structure gives, and the scale is comparable across the two quantities.

Clash. A clash is a heavy-atom pair whose separation falls below the sum of van der Waals radii by more than a tolerance: $r_i + r_j - d_{ij} > \delta_{\text{clash}}$, with $\delta_{\text{clash}} = 0.4$ Å and radii C 1.70, N 1.55, O 1.52, S 1.80, P 1.80, Se 1.90 Å. Covalently bonded pairs and 1–3 neighbours are excluded. Counts are normalised per 1,000 peptide atoms so that peptides of different length are comparable. *Internal* clashes are pairs within the designed peptide; *interface* clashes are peptide–receptor pairs.

Torsional statistics. For each peptide bond we compute $\omega = \text{dihedral}(\text{C}\alpha_k, \text{C}_k, \text{N}_{k+1}, \text{C}\alpha_{k+1})$ and report the mean planarity deviation $\min(|\,|\omega| - 180^\circ|,\ |\omega|)$. A bond with $|\omega| < 30^\circ$ is *cis* and is counted when the following residue is not proline; a bond with $30^\circ \le |\omega| \le 150^\circ$ is *twisted*. The 30° cut matters: taking the naive 90° threshold reports every twisted bond as cis and inflates the cis rate several fold.

Ramachandran. Rather than import external contours we build empirical ϕ/ψ densities from the *receptor* chains of the 226 native complexes, which are real PDB protein and are independent of any designed peptide being scored. Densities are binned at 8° and split into the four standard classes of Kleywegt & Jones (1996): glycine, proline, pre-proline and general. A residue is *favoured* if it falls inside the smallest contour containing 98% of the reference mass and an *outlier* if it falls outside the 99.95% contour.

L.3 Sequence

AAR. Per-residue identity between the designed and native peptide sequences, aligned positionally, averaged over residues.

BLOSUM62. Global pairwise alignment of the designed sequence against the native under the BLOSUM62 substitution matrix. Because the peptides are cyclic, the alignment is evaluated over all circular rotations of the design and the best score is taken; scoring a single arbitrary rotation would penalise a correct sequence that happens to start at a different residue. We report the raw score and its

per-residue normalisation. The sign carries meaning: a negative score means the design aligns worse than an unrelated sequence would, which is a qualitatively different outcome from a small positive one, and three of the models we compare against sit there.

Why both AAR and BLOSUM62. AAR is a hard identity test and treats a lysine-for-arginine substitution as a complete failure, though it conserves charge and often function. BLOSUM62 credits such substitutions. Reporting both separates "the model recovered the native residue" from "the model recovered the native chemistry at that position", and the two orderings are not identical across the models we compare.

L.4 Binding energetics

All energetic quantities come from Rosetta's interface analyser applied to the design and its receptor, using one PyRosetta build for every model and for the native structures.

ΔG. Rosetta interface energy in REU, computed with `InterfaceAnalyzerMover` under the full-atom score function with the peptide and receptor declared as the two docking partners. More negative is tighter. REU are Rosetta score units and are not calibrated against experimental binding free energies; we report differences in REU throughout and do not convert them to kcal/mol, because a score difference of a given size carries no guaranteed thermodynamic interpretation. One PyRosetta build is used for every model *and* for the native structures, so all energies are on a single scale; energies from different Rosetta versions are not comparable and we do not mix them.

ΔG per 100 Å^2. ΔG normalised by buried solvent-accessible surface, so that a larger interface cannot inflate the score. We report this alongside raw ΔG because peptide length is not fixed across models.

Rosetta scoring coverage. The Rosetta interface score requires canonical residue types and complete side chains. Two baselines do not always supply them: UniMoMo places D-residues at roughly half its positions, and BoltzGen's full pipeline emits backbone coordinates only. Scoring therefore succeeds on 30.4% and 7.0% of their designs respectively, and their energetic columns are computed over those subsets. The scored and unscored subsets are close on every quantity we can measure for both, differing by under two points in validity and well-formedness, so we do not read the loss as selecting the better designs; it is nonetheless a smaller sample than the other columns rest on. Every other model, including ours and the crystal structures, scores at 100% coverage.

$\Delta\Delta G$ and beats-native. The crystal structures are rescored with the same build, giving a native ΔG per target on a common scale. $\Delta\Delta G$ is the design's energy minus that native value, and *beats native* is the fraction of designs whose ΔG falls below it. Beating the native is a high bar by construction: the reference is an experimentally determined complex.

L.5 Structural accuracy

Cα-RMSD. Root-mean-square deviation of peptide Cα atoms from the native after optimal superposition, so it measures shape rather than placement. We report the superposed form because several methods do not place the complex in the native frame, which makes the unsuperposed value uninformative.

L.6 Metrics we compute but do not report

Two quantities the evaluator produces are omitted from the tables, and it is worth saying why rather than leaving them out silently.

Side-chain completeness counts residues missing an expected heavy atom. Residues of unknown type carry no expectation and are skipped, so a model emitting them scores *higher*; the native crystal structures themselves score only 0.37 because real structures have unresolved side chains. The metric rewards the wrong behaviour and we do not use it.

Forms-any-macrocycle fires whenever any two atoms in the peptide come within bonding distance, irrespective of which atoms. Collapsed geometry satisfies it almost everywhere, so it reads near 1.0 exactly when a structure has failed. Closure against the native's designated atom pair is the meaningful version and is what we report.