Base-and-Sugar Dual-Frame Flow Matching for RNA Co-Design

Junzhe Li [*1 2] **Lijian Peng** [*2] **Yuhao Li** [1 2] **Yize Zhou** [1 2] **Hanqun Cao** [2] **Cheng Tan** [3] **Shengchao Liu** [1 2]

Abstract

Current frame-based RNA generators typically represent each nucleotide with a single coordinate frame. However, our representation analysis across $11{,}497$ static chains and $31{,}432$ multi-state relation groups shows that a single frame is insufficient to simultaneously capture base-mediated interactions and atom-level reconstruction. Motivated by this observation, we introduce **DuetRNA**, a dual-frame model for joint RNA sequence-structure generation. Each nucleotide carries two coupled $\mathrm{SE}(3)$ frames: a base-centered frame anchored at the glycosidic nitrogen for base mediated relationship, and a Gram-Schmidt sugar-centered frame for sugar-backbone reconstruction. DuetRNA learns their joint distribution with $\mathrm{SE}(3)$ flow matching and co-generates nucleotide identities in the same forward pass, avoiding post-hoc sequence recovery by an external inverse-folding model. Our experiments across two de novo generation protocols show that DuetRNA produces RNA structures with strong folding-based self-consistency. When trained on RNASolo, DuetRNA achieves 48.67% scTM-validity under inverse-folded sequence evaluation and 38.83% RhoFold-based scTM-validity under direct joint generation. These results demonstrate that explicitly factorizing RNA residues into base-centered and sugar-backbone-centered frames is a strong representation choice for RNA sequence–structure generation.

1. Introduction

RNA molecules perform diverse biological functions through their three-dimensional structures, which are shaped not only by chain connectivity but also by rich tertiary interactions such as base pairing, base stacking, and long-range packing (Saenger, 1984; Leontis & Westhof, 2001). Recent progress in RNA machine learning has substantially improved *sequence-conditioned* tertiary-structure prediction, with models such as DeepFoldRNA, DRfold, RhoFold+, and NuFold learning increasingly accurate maps from sequence to structure (Pearce et al., 2022; Li et al., 2023; Shen et al., 2024; Kagaya et al., 2025). A complementary line of work has begun to address *de novo* RNA generation, including structure-first generative models such as RNA-FrameFlow and emerging joint sequence-structure generators such as RiboGen and RiboFlow (Anand et al., 2025; Rubin et al., 2025; Ma et al., 2025). These advances suggest that RNA generative modeling is becoming feasible, but they also expose a central difficulty that is specific to RNA geometry: *what is the right generative object for a nucleotide?*

For proteins, residue-level rigid-body representations are often reasonably aligned with backbone-centered organization (Chothia, 1984). However, RNA has its own characteristics: its tertiary organization is dominated by nucleobase-mediated interactions (Leontis & Westhof, 2001), whereas its sugar-phosphate backbone remains substantially more flexible (Saenger, 1984). This asymmetry creates a representation bottleneck for the generation of single-frame residues.

An interesting observation on frame representation analysis. To make this bottleneck concrete, we systematically analyze seven single-frame candidates on $11{,}497$ static chains and $31{,}432$ multi-state relation groups. As detailed in Section 5.3, the tested single-frame representations do not simultaneously provide stable base-mediated inter-residue poses and accurate sugar-backbone reconstruction. A complementary dual-frame reconstruction analysis further shows that assigning base and sugar geometry to separate coordinate systems substantially reduces atom reconstruction error relative to the tested single-frame alternatives. Figure 1(b) visualizes the atom-anchored dual-frame construction deployed in DuetRNA: a sugar Gram–Schmidt frame combined with a chemically anchored nucleobase-plane frame.

Our contributions. Motivated by this observation, we propose DuetRNA, a physically informed dual-frame framework for joint RNA sequence-structure generation. DuetRNA is organized around a *base-centered frame* that

[*]Equal contribution [1]Wave Intelligence Lab [2]The Chinese University of Hong Kong, Hong Kong, China [3]Shanghai Artificial Intelligence Laboratory, Shanghai, China. Correspondence to: Junzhe Li <tpmam@connect.hku.hk>, Lijian Peng <1155211002@link.cuhk.edu.hk>.

Accepted at the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026)

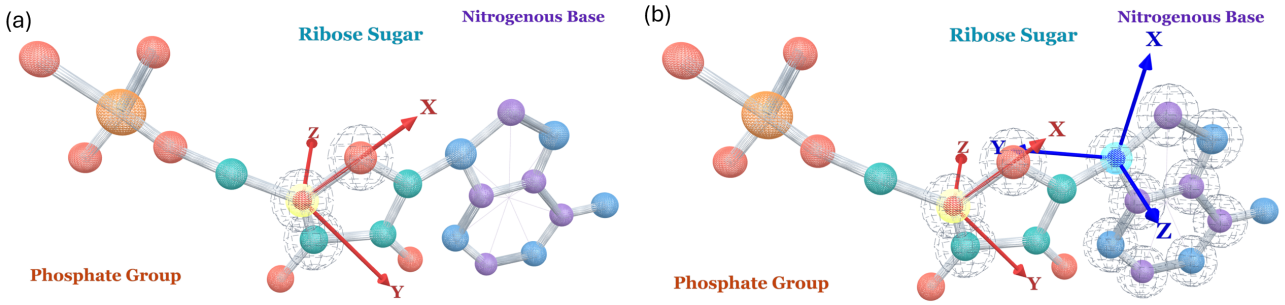

Figure 1. Comparison of RNA frame representations. Figure 1(a) builds a single a sugar Gram-Schmidt frame on $(O4', C4', C3')$. Figure 1(b) illustrates the dual-frame representation deployed by DuetRNA: the combination of a sugar Gram-Schmidt frame and a base frame which is chemically anchored and nucleobase-plane oriented.

captures residue-level RNA organization and a *sugar-centered frame* that supports torsion-conditioned sugar-backbone reconstruction. Each nucleotide is therefore represented by two coupled $\mathrm{SE}(3)$ frames that are jointly modeled by flow matching. We further introduce a reparameterized dual-frame constraint during training to enforce their within-residue coupling: after the network predicts the terminal base and sugar frames, we supervise the composed base-to-sugar transform derived from those predictions. This keeps the two frames chemically coupled rather than allowing them to drift as independent generated objects. With these generated frames, together with predicted sequence and intra-residue torsions, DuetRNA completes RNA atom-level reconstruction for fine-grained joint RNA sequence-structure generation.

To evaluate DuetRNA, we consider two complementary de novo generation settings. The first evaluates backbone designability through an inverse-folded sequence protocol, where generated structures are assigned candidate sequences and then independently forward-folded. The second evaluates direct joint generation, where the sequence produced by the model is directly folded by independent structure predictors, thereby testing whether the generated sequence and structure are mutually compatible without an external inverse-folding post-processing step.

Across these settings, DuetRNA produces RNA designs with strong folding-based self-consistency. When trained on the same RNASolo corpus used by RNA-FrameFlow, DuetRNA achieves 48.67% IF scTM-validity, compared with the published RNA-FrameFlow result of 41.00% under the same gRNAde-to-RhoFold evaluation protocol. The same RNASolo-trained checkpoint achieves 38.83% scTM-validity when evaluated directly with its generated sequences, without inverse-folding post-processing. These results indicate that a substantial fraction of the generated RNA sequence–structure pairs remain compatible under independent forward-folding evaluation. Together with our frame representation analysis, they support dual-frame factorization as an effective inductive bias for RNA generative modeling: base-centered frames capture nucleobase-mediated organization, while sugar-centered frames preserve the intra-residue geometry needed for backbone reconstruction.

2. Related Work

Sequence-conditioned RNA tertiary-structure prediction. Most recent deep-learning systems for RNA 3D modeling follow a sequence-conditioned forward-folding paradigm: given an RNA sequence $\mathbf{s}$, they predict a tertiary structure $\mathbf{x}$. Representative RNA-specific predictors include DeepFoldRNA, DRfold, RhoFold+, and NuFold (Pearce et al., 2022; Li et al., 2023; Shen et al., 2024; Kagaya et al., 2025), while broader biomolecular models such as RoseTTAFoldNA and AlphaFold3 extend AlphaFold-style architectures to nucleic acids and protein-nucleic-acid complexes (Baek et al., 2024; Abramson et al., 2024). These models are valuable as forward oracles in RNA design: once a candidate sequence has been proposed, they provide a practical estimate of whether it is compatible with a target fold. However, they do not directly sample RNA structures or sequence-structure pairs.

From a representation perspective, these predictors also reveal the importance of RNA-specific geometric parameterizations. DRfold learns nucleotide-wise local frames and inter-nucleotide geometric restraints, but uses a coarse-grained representation based on P, $C4'$, and glycosidic N atoms, with full-atom coordinates recovered afterward (Li et al., 2023). NuFold is particularly relevant to our motivation: it adopts a flexible nucleobase-centered representation with explicit torsional modeling, and shows improved local geometry as well as accurate reconstruction of both $C3'$-endo and $C2'$-endo sugar puckers (Kagaya et al., 2025). These results suggest that base orientation and sugar-backbone conformation play distinct geometric roles in RNA modeling. At the same time, systematic bench-

marks show that current prediction methods still struggle on orphan RNAs and non-Watson-Crick interactions, with local interaction fidelity remaining a major bottleneck (Bahai et al., 2024). This limitation is especially relevant for generation, where the model must construct plausible global folds together with fine-grained base-mediated interactions.

Structure-first *de novo* RNA generation. A complementary line of work learns a prior over RNA structural space directly, rather than only a sequence-conditioned folding map. RNA-FrameFlow introduces SE(3) flow matching for *de novo* RNA backbone generation by representing each nucleotide as a rigid-body frame and predicting the remaining backbone atoms through torsional degrees of freedom (Anand et al., 2025). Its evaluation protocol further frames generation as a design problem: sampled backbones are passed through inverse folding to obtain candidate sequences and then through forward folding to measure self-consistency (Anand et al., 2025; Joshi et al., 2025). This is a substantial step beyond forward prediction, because it learns a generative prior over RNA backbone geometry.

However, structure-first generation still separates scaffold generation from sequence compatibility. RNA-FrameFlow outputs backbone scaffolds, while compatible sequences are assigned downstream by a separate inverse-design model (Anand et al., 2025; Joshi et al., 2025). Recent generalized biopolymer design frameworks such as RFDpoly further indicate that structure-first generation for RNA, DNA, proteins, and mixed assemblies is becoming increasingly practical (Favor et al., 2025). Nevertheless, these pipelines do not directly model sequence and structure as coupled generated variables. Moreover, their generative states remain organized primarily around backbone or residue-level geometry, leaving base organization to be recovered indirectly.

Joint sequence-structure generation and representation gap. RiboGen moves toward direct modeling of $p(\mathbf{s}, \mathbf{x})$ by combining continuous flow matching for geometry with discrete flow matching for sequence (Rubin et al., 2025), demonstrating that RNA co-generation is feasible. However, its frameless parameterization offers excessive modeling freedom. With no rigid-body scaffold to regularize the generation and with a training loss that does not explicitly enforce bond lengths, bond angles, or torsional preferences, the model has difficulty learning the physical constraints inside RNA, resulting in lower validity (Rubin et al., 2025). RiboFlow (Ma et al., 2025) addresses the physical-plausibility gap by retaining a single residue-level SE(3) frame (as in RNA-FrameFlow) and reconstructing atoms from backbone frame and predicted torsions, which imposes implicit geometric constraints on the output. Nevertheless, as our frame-representation analysis shows (§B), single-frame construction can not capture base-mediated interaction geometry and backbone reconstruction simultaneously.

Taken together, these create a methodological opportunity. RNA-FrameFlow establishes that SE(3) flow matching is a viable framework for RNA backbone generation (Anand et al., 2025). NuFold suggests that nucleobase-centered representations with explicit torsional structure can improve RNA local geometry (Kagaya et al., 2025). RiboGen shows that joint RNA sequence-structure generation is possible (Rubin et al., 2025). RiboFlow (Ma et al., 2025) addresses the physical-plausibility gap left by RiboGen by retaining a single rigid frame.

Our work addresses the remaining representation gap. We formulate RNA generation around a base-sugar dual-frame state space: a base-centered frame captures inter-residue organization such as pairing, stacking, and tertiary packing, while a sugar frame captures sugar-backbone geometry and chain realization. This factorization is designed to avoid forcing high-level base organization and local backbone flexibility into a single nucleotide-level rigid body, and provides a physically informed state space for joint RNA sequence-structure generation with SE(3) flow matching.

3. Preliminary

3.1. Rigid-Body Motions and $\mathrm{SE}(3)$

The Rotation Group $\mathrm{SO}(3)$. A rotation of a rigid body is represented by a 3×3 matrix R that preserves distances and orientations. All such matrices form the *special orthogonal group*

$$\mathrm{SO}(3) = \{\, R \in \mathbb{R}^{3\times 3} \mid R^\top R = I,\ \det(R) = 1 \,\}. \quad (1)$$

$\mathrm{SO}(3)$ is a three-dimensional *Lie group* and can be viewed as a Riemannian manifold once equipped with a suitable metric. Since $\mathrm{SO}(3)$ is not a Euclidean vector space, operations such as interpolation, averaging, and regression must respect its manifold structure.

The Special Euclidean Group $\mathrm{SE}(3)$. A full rigid-body pose adds a translation $t \in \mathbb{R}^3$ to the orientation. The pair belongs to the *special Euclidean group* (Barfoot, 2017)

$$\mathrm{SE}(3) = \left\{ \begin{pmatrix} R & t \\ \mathbf{0}^\top & 1 \end{pmatrix} \,\middle|\, R \in \mathrm{SO}(3),\ t \in \mathbb{R}^3 \right\}. \quad (2)$$

In the text, we often denote an element compactly as $g = (R, t)$. Composition and inversion follow the familiar rules

$$g_1 \circ g_2 = (R_1 R_2,\ R_1 t_2 + t_1), g^{-1} = (R^\top,\ -R^\top t). \quad (3)$$

$\mathrm{SE}(3)$ is also a Lie group. Its tangent space at the identity, the Lie algebra $\mathfrak{se}(3)$, is a six-dimensional vector space An element of $\mathfrak{se}(3)$ is parameterized by a six-dimensional coordinate $\xi = (\omega, v)$, with ω encoding the angular component and v the translational component. The exponential and

logarithmic maps

$$\mathrm{Exp} : \mathfrak{se}(3) \to \mathrm{SE}(3), \qquad \mathrm{Log} : \mathrm{SE}(3) \to \mathfrak{se}(3), \quad (4)$$

provide a local correspondence between the Lie algebra and the group. These maps enable local computations in the Lie algebra while representing configurations as points on $\mathrm{SE}(3)$.

3.2. Flow Matching On Riemannian Manifolds

Let $\mathcal{M}$ be a Riemannian manifold and let p_{data} denote the unknown data distribution on $\mathcal{M}$. Flow matching (Lipman et al., 2023; Chen & Lipman, 2024) learns a time-dependent vector field whose induced flow transports a simple prior distribution p_{prior} on $\mathcal{M}$ to the data distribution.

Probability flow ODE. A time-dependent vector field $v_t : \mathcal{M} \to T\mathcal{M}$ defines an ordinary differential equation as

$$\frac{\mathrm{d}}{\mathrm{d}t} z_t = v_t(z_t), \quad z_0 \sim p_{\mathrm{prior}}. \quad (5)$$

Here, $T\mathcal{M}$ denotes the tangent bundle of $\mathcal{M}$, and $v_t(z) \in T_z\mathcal{M}$ for each $z \in \mathcal{M}$. Integrating this ODE from $t = 0$ to $t = 1$ yields a flow ψ_t such that $z_t = \psi_t(z_0)$.

Interpolation construction. Since the marginal probability path $p_t = (\psi_t)_{\#} p_{\mathrm{prior}}$ is intractable, we adopt a simulation-free training strategy. For each data point $z_1 \sim p_{\mathrm{data}}$, we draw a random noise sample $z_0 \sim p_{\mathrm{prior}}$ and construct an interpolation z_t between them. A natural choice on $\mathcal{M}$ is the geodesic path:

$$z_t = \exp_{z_0}\big(t \log_{z_0}(z_1)\big), \quad t \in [0, 1], \quad (6)$$

where $\exp_z$ and $\log_z$ denote the Riemannian exponential and logarithmic maps at point z.

Training via Endpoint parameterization. We adopt an *endpoint parameterization* (Lipman et al., 2023), in which the network predicts the terminal point $\hat{z}_1 = h_\theta(t, z_t) \in \mathcal{M}$ from interpolation point z_t. This parameterization allows the velocity to be constructed geometrically through the logarithmic map. The learning objective is then written in the tangent space at z_t as

$$\mathcal{L}(\theta) = \mathbb{E}_{t, z_0 \sim p_{\mathrm{prior}}, z_1 \sim p_{\mathrm{data}}} \big\| \log_{z_t}(\hat{z}_1) - \log_{z_t}(z_1) \big\|^2. \quad (7)$$

Inference. Once trained, the model generates samples by drawing $z_0 \sim p_{\mathrm{prior}}$ and numerically integrating the learned ODE from $t = 0$ to $t = 1$. At each integration step, the network outputs an endpoint estimate $\hat{z}_1$, and the vector field is reconstructed via $v_\theta(t, z_t) = \log_{z_t}(\hat{z}_1)/(1-t)$. The ODE flow $\frac{\mathrm{d}}{\mathrm{d}t} z_t = v_\theta(t, z_t)$, where $z_0 \sim p_{\mathrm{prior}}$, is then solved forward in time. The specific form of the tangent-space distance for our product manifold setting is given in Section 4.

4. Method

We propose DuetRNA, a base-and-sugar dual-frame generative framework for joint RNA sequence-structure modeling. As shown in Figure 2, each nucleotide is described by two physically meaningful rigid objects: a nucleobase-centered frame for residue-level organization and a sugar-centered frame for sugar-backbone realization. The model learns a joint SE(3) flow over such two frames, predicts nucleotide identity and intra-residue torsional variables from coupled geometric features, and decodes the final state into an atom23 heavy-atom representation.

Problem formulation. Let $\mathcal{C} = \{\mathrm{A}, \mathrm{U}, \mathrm{G}, \mathrm{C}\}$ denote the RNA nucleotide alphabet, $s = (s_1, \ldots, s_N) \in \mathcal{C}^N$ the nucleotide sequence, and $\phi = (\phi_1, \ldots, \phi_N)$ the backbone torsional variables. Our model jointly designs the structure and sequence of RNA by learning $p_\theta(T^{\mathrm{base}}, T^{\mathrm{sugar}}, s, \phi)$, which is decomposed into:

$$\begin{aligned} p_\theta(T^{\mathrm{base}}, T^{\mathrm{sugar}}, s, \phi) = {} & p_\theta^{\mathrm{frame}}(T^{\mathrm{base}}, T^{\mathrm{sugar}}) \\ & \cdot p_\theta^{\mathrm{seq}}(s \mid T^{\mathrm{base}}, T^{\mathrm{sugar}}) \\ & \cdot p_\theta^{\mathrm{loc}}(\phi \mid T^{\mathrm{base}}, T^{\mathrm{sugar}}, s). \end{aligned} \quad (8)$$

The following subsections mirror our generative pipeline. Section 4.1 first transforms raw RNA data into a dual-frame representation. Section 4.2 estimates the joint distribution over base and sugar frames. Section 4.3 and Section 4.4 describe sequence prediction and intra-residue torsion estimation for atom23 completion. Section 4.5 defines the deterministic atom23 completion path, and Section 4.6 summarizes the training objective and model architecture.

4.1. Dual-Frame Representation of RNA

Systematic analysis of residue-level coordinate frames (details in Appendix B) indicates that no single frame simultaneously captures base-mediated inter-residue geometry and intra-residue backbone reconstruction geometry. Base frames preserve the inter-residue relationship more faithfully, while sugar-centered frames capture the relationship between local residue atoms required for reconstruction. To reconcile this trade-off, we adopt a dual-frame representation: a base frame T_i^{base} captures base-mediated inter-residue geometry, and a sugar frame T_i^{sugar} captures the intra-residue torsional geometry.

The construction of sugar frame and base frame deployed in DuetRNA. As shown in figure Figure 1, DuetRNA constructed one sugar frame and one base frame for each nuleotide. Specifically, the sugar frame T^{sugar} uses **Sugar-GS**, a three-atom Gram-Schmidt frame anchored at $C4'$ and built from $(O4', C4', C3')$. The base frame T^{base} uses **Base-Plane**, a chemically anchored nucleobase-plane frame whose origin is the glycosidic connection atom ($N9$

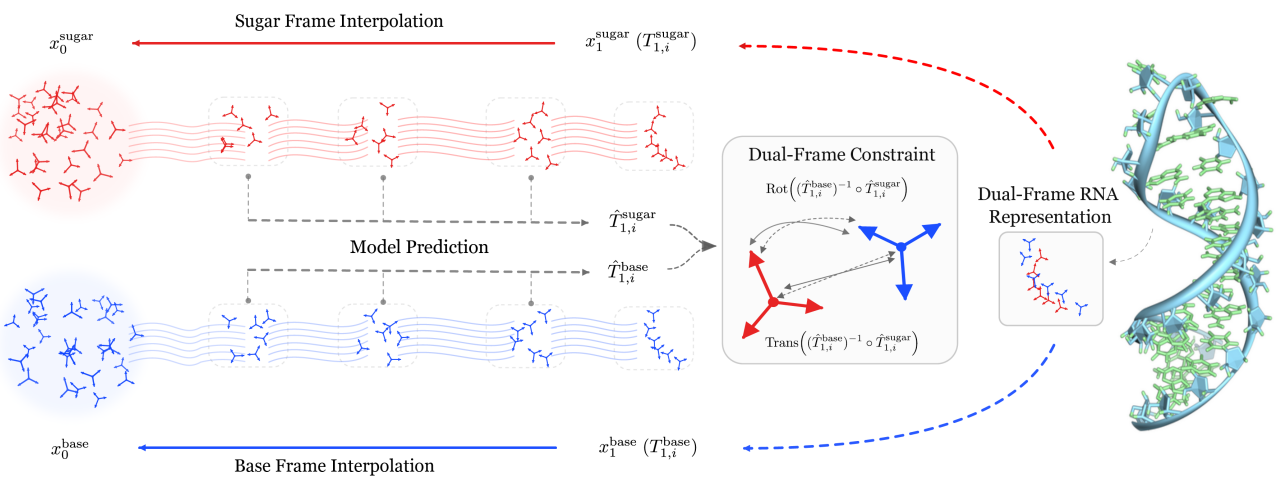

Figure 2. Overview of DuetRNA. In DuetRNA, each nucleotide is represented by two coupled SE(3) frames: a sugar frame and a base frame. The model jointly predicts terminal sugar and base frames from noisy interpolants, while a dual-frame constraint matches the predicted base-to-sugar relative translation and rotation to their ground-truth relative pose. This couples the two generated frames into a physically consistent dual-frame RNA representation for downstream sequence and atom-level structure reconstruction.

for purines and $N1$ for pyrimidines) and whose orientation is defined by the fitted base plane and an in-plane chemical axis. The exact coordinate definitions of these two deployed frames are given in Appendix B.

Dual-Frame representation for RNA. For a chain of length N, the RNA geometric state is thus described by

$$\begin{aligned} T^{\text{base}} &= (T_1^{\text{base}}, \ldots, T_N^{\text{base}}), \\ T^{\text{sugar}} &= (T_1^{\text{sugar}}, \ldots, T_N^{\text{sugar}}), \end{aligned} \tag{9}$$

4.2. Dual-Frame Flow Matching

For an RNA molecule with N residues, the learning target for continuous geometry is the pair of frames

$$X_t = \left(T_t^{\text{base}}, T_t^{\text{sugar}}\right) \in (\text{SE}(3) \times \text{SE}(3))^N. \tag{10}$$

Source endpoints and Dual-Frame interpolation. For $f \in \{\text{base}, \text{sugar}\}$, we sample a source endpoint $T_0^f = (R_0^f, x_0^f)$ from an unconditional source distribution with Gaussian translations and uniform rotations on SO(3) as the prior distribution p_0, and take a clean data frame $T_1^f = (R_1^f, x_1^f)$ as the target distribution p_1. We then define the interpolation path $T_t^f = (R_t^f, x_t^f)$ at time t as:

$$\begin{aligned} x_t^f &= (1-t)x_0^f + t x_1^f, \\ R_t^f &= \text{Exp}_{R_0^f}\left(t \, \text{Log}_{R_0^f}(R_1^f)\right). \end{aligned} \tag{11}$$

Reparameterized vector field estimation. Given T_t^{f} and time t, the model estimates the vector field by predicting terminal frame state $\hat{T}_1^{\text{f}}$. The implied translation and rotation vector fields from this reparameterization are

$$u_t^{\text{trans},f} = \frac{\hat{x}_1^f - x_t^f}{1-t}, \qquad u_t^{\text{rot},f} = \text{Log}_{R_t^f}(\hat{R}_1^f). \tag{12}$$

Learning objectives for vector field estimation. The estimation objective designed for Dual-Frame flow matching is

$$\mathcal{L}_{\text{Dual-Frame}} = \sum_{f \in \{\text{base},\text{sugar}\}} \mathcal{L}_{\text{frame}}^f + \mathcal{L}_{\text{constraint}} \tag{13}$$

Here the rigid term is the combined SE(3) flow matching loss described in RNA-FrameFlow:

$$\begin{aligned} \mathcal{L}_{\text{frame}}^f = &\frac{1}{N}\sum_{i=1}^{N} \frac{\left\|(\hat{t}_{1,i}^f - t_{1,i}^f)\right\|_2^2}{(1-\bar{t})^2} \\ &+ \frac{1}{N}\sum_{i=1}^{N} \frac{\left\|\text{Log}_{R_{t,i}^f}(\hat{R}_{1,i}^f) - \text{Log}_{R_{t,i}^f}(R_{1,i}^f)\right\|_2^2}{(1-\bar{t})^2} \end{aligned} \tag{14}$$

To ensure the base and sugar frames preserve the within-residue relationship during the flow matching process, we further introduce the reparameterized Dual-Frame constraint term in Equation (13) which supervises terminal relative translation and rotation at the predicted terminal state ($t = 1$).

$$\begin{aligned} \mathcal{L}_{\text{constraint}} = &\frac{1}{N}\sum_{i=1}^{N} \Big\| \text{Trans}\left((\hat{T}_{1,i}^{\text{base}})^{-1} \circ \hat{T}_{1,i}^{\text{sugar}}\right) \\ &- \text{Trans}\left((T_{1,i}^{\text{base}})^{-1} \circ T_{1,i}^{\text{sugar}}\right) \Big\|_2^2 \\ &+ \frac{1}{N}\sum_{i=1}^{N} \Big\| \log\Big(\text{Rot}\left((\hat{T}_{1,i}^{\text{base}})^{-1} \circ \hat{T}_{1,i}^{\text{sugar}}\right)^\top \\ &\cdot \text{Rot}\left((T_{1,i}^{\text{base}})^{-1} \circ T_{1,i}^{\text{sugar}}\right)\Big)\Big\|_2^2. \end{aligned} \tag{15}$$

Here $\mathrm{Trans}(\cdot)$ and $\mathrm{Rot}(\cdot)$ extract the translation and rotation of an $\mathrm{SE}(3)$ transform, respectively.

4.3. Sequence Prediction

The sequence head estimates $p_\theta^{\mathrm{seq}}(s \mid T^{\mathrm{base}}, T^{\mathrm{sugar}})$ as a residue-wise categorical distribution. Let $\pi_i(c)$ be the predicted probability of nucleotide $c \in \{A, U, G, C\}$ at residue i:

$$p_\theta^{\mathrm{seq}}(s \mid T^{\mathrm{base}}, T^{\mathrm{sugar}}) = \prod_{i=1}^{N} \pi_i(s_i).$$

The maximum-likelihood objective is the residue-wise negative log-likelihood

$$\mathcal{L}_{\mathrm{Seq}} = -\frac{1}{N} \sum_{i=1}^{N} \log \pi_i(s_i^{\mathrm{gt}}). \tag{16}$$

4.4. Intra-residue Torsion Estimation

The intra-residue torsional variables are the torsion angles required by atom23 completion:

$$p_\theta^{\mathrm{loc}}(\phi \mid T^{\mathrm{base}}, T^{\mathrm{sugar}}, s). \tag{17}$$

Let $K = 8$ be the number of torsion angles used by the completion map. These angles are predicted from the coupled geometric features by

$$\hat{u}_i^{\phi} = g_\theta^{\mathrm{tor}}(T^{\mathrm{base}}, T^{\mathrm{sugar}}, s)_i \in \mathbb{R}^{K \times 2} \tag{18}$$

During training, these internal-coordinate variables are supervised both directly and through the decoded atom23 structure described in Section Section 4.5. The main local objective is

$$\mathcal{L}_{\mathrm{Loc}} = \mathcal{L}_{\mathrm{tor}} + \mathcal{L}_{\mathrm{atom}}. \tag{19}$$

The torsion supervision uses a sine-cosine parameterization:

$$\mathcal{L}_{\mathrm{tor}} = \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=1}^{K} \left\| \hat{u}_{i,k}^{\phi} - u(\phi_{i,k}^{\mathrm{gt}}) \right\|_2^2. \tag{20}$$

where $u(\phi_{i,k}) = (\sin \phi_{i,k}, \cos \phi_{i,k}) \in S^1$.

Let $\hat{X}^{\mathrm{atom23}}$ denote the completed atom23 structure obtained from the deterministic completion map in Section 4.5. The atom reconstruction loss is

$$\mathcal{L}_{\mathrm{atom}} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|\mathcal{A}_i|} \sum_{a \in \mathcal{A}_i} \left\| \hat{x}_{i,a} - x_{i,a}^{\mathrm{gt}} \right\|_2^2, \tag{21}$$

where $\mathcal{A}_i$ is the valid heavy-atom subset for residue i under the atom23 mask. Additional decoded-geometry supervision terms used in ablations are summarized in Appendix C.

4.5. Atom23 Completion

Atom23 denotes a fixed-width heavy-atom representation with up to 23 nucleotide-specific atom slots. Residue-specific absent atoms are masked, and hydrogens are excluded. Atom23 completion is a deterministic mapping from the predicted geometric variables to this compact heavy-atom representation:

$$\hat{X}^{\mathrm{atom23}} = \mathcal{A}\Big(\hat{T}^{\mathrm{base}}, \hat{T}^{\mathrm{sugar}}, \hat{\phi}, s\Big), \tag{22}$$

The sugar/backbone atoms are generated from the predicted sugar frame $\hat{T}_i^{\mathrm{sugar}}$ together with the predicted torsions, whereas the base atoms are placed from canonical nucleotide-specific templates attached to the predicted base frame $\hat{T}_i^{\mathrm{base}}$. The final output is a compact atom23 heavy-atom representation rather than a hydrogen-complete all-atom model.

4.6. Training Objectives And Model Architecture

Objective. Overall, the training loss is the sum of three components from Sections 4.2 to 4.4:

$$\mathcal{L} = \mathcal{L}_{\mathrm{Dual-Frame}} + \lambda_{\mathrm{Seq}} \mathcal{L}_{\mathrm{Seq}} + \mathbf{1}[t > \tau] \lambda_{\mathrm{Loc}} \mathcal{L}_{\mathrm{Loc}}, \tag{23}$$

where $\mathcal{L}_{\mathrm{Dual-Frame}}$, $\mathcal{L}_{\mathrm{Seq}}$ and $\mathcal{L}_{\mathrm{Loc}}$ are described in Equations (13), (16) and (19), respectively. We use $\tau = 0.25$ to gate $\mathcal{L}_{\mathrm{Loc}}$ to delay decoded-geometry supervision until the noised frames have moved beyond the earliest part of the interpolation path. The loss weights and their ablation ranges are listed in Appendix C.

Architecture. The network architecture contains a coupled dual-frame geometric trunk and two output heads for sequence estimation and torsion angle estimation. The trunk takes the base and sugar frames as input, refines their coupled representations across depth, and exposes final features to the estimation heads described above. A more detailed discussion is provided in Appendix A.

5. Experiments

5.1. Experimental Setup

Datasets. We train DuetRNA on RNASolo (Adamczyk et al., 2022) following the choice of RNA-FrameFlow, RiboGen, and RiboFlow (Anand et al., 2025; Rubin et al., 2025; Ma et al., 2025). As another version, we also train DuetRNA on RNA3DB (Szikszai et al., 2024), a PDB-derived RNA structure collection curated for deep-learning benchmarks with sequence- and structure-aware nonredundant splits. From both datasets, we retain RNA-only homomeric chains with no protein chains and modeled lengths between 40 and 150 nucleotides. Since released training code and checkpoints are not available for RiboGen and RiboFlow at the time of writing, we therefore report their published results

Table 1. Results of de novo RNA generation (600 samples per method unless noted). IF assigns eight gRNAde sequences to each generated structure and evaluates the best candidate with RhoFold. GS directly evaluates the model-generated sequence.

Model	Train data	Protocol	Folder	N_T/α	Val. scTM ↑	Val. scRMSD ↑	Div. ↑	Nov. ↓
MMDiff (Anand et al., 2025)	–	IF	RhoFold	100/–	0.00	–	–	–
RNA-FrameFlow (paper) (Anand et al., 2025)	RNASolo	IF	RhoFold	50/–	41.00	–	**0.610**	**0.540**
DuetRNA (Ours)	**RNASolo**	**IF**	**RhoFold**	**100/20**	**48.67**	33.83	0.268	0.632
DuetRNA (Ours)	RNA3DB	IF	RhoFold	100/20	44.00	**37.00**	0.353	0.586
RiboGen (Rubin et al., 2025)	RNASolo	GS	Boltz-1	100/–	27.17	–	**0.604**	–
RiboGen (Rubin et al., 2025)	RNASolo	GS	Boltz-1	200/–	32.17	–	0.553	–
RiboGen (Rubin et al., 2025)	RNASolo	GS	Boltz-1	300/–	34.17	–	0.585	–
DuetRNA (Ours)	**RNA3DB**	**GS**	**Boltz-1**	**100/20**	**38.50**	26.40	0.353	0.600
RiboFlow (Ma et al., 2025)	RNASolo	GS	RhoFold	50/–	34.70	–	**0.577**	**0.562**
DuetRNA (Ours)	**RNASolo**	**GS**	**RhoFold**	**100/20**	**38.83**	33.17	0.263	0.629
DuetRNA (Ours)	RNA3DB	GS	RhoFold	100/20	37.83	**33.50**	0.353	0.600

as contextual references under the corresponding generated-sequence protocols.

Generation benchmark. Following RNA-FrameFlow (Anand et al., 2025), we evaluate *de novo* generation across lengths $L \in \{40, 50, \ldots, 150\}$, sampling 50 independent structures per length and therefore 600 structures per full run. We compare against MMDiff, RNA-FrameFlow, RiboFlow, and RiboGen under this length-controlled benchmark.

Sequence-source protocols. Because existing baselines differ in whether they generate nucleotide sequences, we score generated structures under two complementary protocols. **IF** denotes the inverse-folded sequence protocol: each generated backbone is assigned $N_{\text{seq}} = 8$ candidate sequences from gRNAde (Joshi et al., 2025), folded with RhoFold (Shen et al., 2024), and scored by the best self-consistency TM-score across the eight candidates. This protocol measures backbone designability independent of the model's own sequence channel. **GS** denotes the generated-sequence protocol: the model-generated sequence is folded directly, without an external inverse-folding model. We evaluate self-consistency TM-score with two independent structure predictors: RhoFold (Shen et al., 2024) (to align with RiboFlow's protocol) and Boltz-1 (to align with RiboGen's protocol) (Rubin et al., 2025; Wohlwend et al., 2024).

Metrics. A structure is considered valid when its self-consistency TM-score satisfies $\text{scTM} \geq 0.45$. We additionally report scRMSD-validity with threshold $\text{scRMSD} \leq 4.3$ Å, qTMclust diversity, and train-set pdbTM novelty, where lower pdbTM indicates higher novelty. Unless otherwise specified, DuetRNA results use the 180K checkpoint sampled with $N_T = 100$ and rotation exponent $\alpha = 20$.

5.2. Main Results

Table 1 reports the performance of the dual-frame state under both backbone-designability and one-shot sequence-structure generation protocols. Under IF, where sequence recovery is supplied externally by gRNAde, DuetRNA reaches 48.67% scTM-validity on RNASolo dataset. This is above the published RNA-FrameFlow reference value under a corpus- and evaluation-route-matched inverse-folding protocol. The model also reaches 33.83% scRMSD-validity. However, its qTMclust diversity is 0.268, compared with 0.610 for RNA-FrameFlow, and its train-reference pdbTM novelty score is 0.632, compared with 0.540 for RNA-FrameFlow. These result indicates that the generated dual-frame geometry remains highly designable when evaluated by an established backbone-first pipeline, with a potentially stronger concentration on a smaller and less novel set of structural modes.

The same checkpoint also remains effective when its own generated sequence is used. Under GS with RhoFold folding, DuetRNA reaches 38.83% scTM-validity on RNASolo without any external inverse-folding step. The RNA3DB-trained checkpoint obtains similar direct validity—37.83% with RhoFold and 38.50% with Boltz-1. Together with the within-paper frame analysis, these results support the intended role of the dual-frame formulation: it improves backbone-level designability while preserving enough base-mediated information to generate compatible nucleotide identities in a single pass.

5.3. Frame Representation Analysis

The frame-representation analysis is separate from the de novo generation benchmark and does not use the RNA3DB training split. Its static relation pool is built from $11{,}497$ processed RNASolo structure units (Adamczyk et al., 2022), while the multi-state pool is built from 523 single-chain NMR RNA entries comprising $31{,}432$ relation groups. We use these pools to motivate the dual-frame factorization with two empirical findings measured on raw RNA structures before model training.

First, base-anchored coordinates are better aligned with inter-residue RNA organization. On the multi-state pool,

Table 2. Effect of the inference sampling schedule on the 120K model under IF and GS. All rows use RhoFold, 600 generated structures, and the same model weights.

		IF		GS	
N_T	α	Val. scTM ↑	Div. ↑	Val. scTM ↑	Div. ↑
50	10	32.00	**0.585**	24.67	**0.587**
100	20	**40.67**	0.443	**31.17**	0.442
200	10	37.83	0.487	27.00	0.485
200	20	38.00	0.477	28.50	0.485
300	10	36.00	0.485	27.50	0.485
300	20	38.17	0.433	28.83	0.438

the deployed Base-Plane frame reduces canonical-pair drift from $13.35°/1.70$ Å under Sugar-GS to $7.78°/0.61$ Å. Base-Center gives slightly lower drift ($7.76°/0.47$ Å) and also performs best in the offline base-reconstruction diagnostic. We retain Base-Center as an analysis control rather than the deployed frame because its translation origin is a virtual, residue-dependent nucleobase centroid. In contrast, the current DuetRNA endpoint parameterization and atom23 template attachment use an explicit chemical atom as the frame origin. We therefore deploy Base-Plane, whose origin is the glycosidic N9/N1 atom and whose orientation is defined by the nucleobase plane.

Second, the base and sugar frames carry complementary information rather than defining interchangeable coordinate systems. In the reconstruction-path sweep over $51{,}013$ interaction pairs, the deployed Base-Plane + Sugar-GS dual frame reduces base-channel RMSD from 6.21 Å under Sugar-GS alone to 1.58 Å, while preserving the Sugar-GS backbone and bridge errors (1.88 Å and 2.84 Å). The Base-Center + Sugar-GS analysis control gives the lowest base-channel RMSD (1.09 Å), further supporting the benefit of separating base and sugar roles. The full candidate-frame catalog, static relation-separation analysis, multi-state breakdown, and reconstruction tables are reported in Appendix B.

5.4. Ablation Studies

5.4.1. Inference Sampling Schedule

Holding the 120K model, loss function, and evaluation pipeline fixed, we ablate the inference schedule by varying the integration steps N_T and rotation exponent α. Table 2 shows that $N_T = 100, \alpha = 20$ gives the strongest validity under both protocols, raising IF scTM-validity from 32.00% to 40.67% and GS scTM-validity from 24.67% to 31.17% relative to $N_T = 50, \alpha = 10$. Longer trajectories do not improve validity, while $N_T = 50, \alpha = 10$ retains higher diversity.

5.4.2. Training Budget Ablation

The 180K checkpoint gives the highest observed validity under the fixed $N_T = 100, \alpha = 20$ sampler. With architecture, objective, sampler, and evaluation fixed, Table 3 shows that increasing training from 120K to 180K raises IF scTM-validity from 40.67% to 44.00% and GS scTM-validity from 31.17% to 37.83%. Further training to 190K or 200K reduces both metrics, so we use the 180K model for the main comparison.

Table 3. Training-budget ablation of the RNA3DB-trained model under the fixed $N_T = 100, \alpha = 20$ sampler. The 180K model is selected for the main result; further training to 190K and 200K regresses. All rows use 600 generated structures.

	IF		GS	
Training steps	Val. scTM ↑	Val. scRMSD ↑	Val. scTM ↑	Val. scRMSD ↑
120K	40.67	**38.83**	31.17	32.00
180K	**44.00**	37.00	**37.83**	**33.50**
190K	38.33	31.67	29.17	26.67
200K	36.67	31.33	31.00	26.00

Table 4. Loss ablation study on RNA3DB. We remove each geometric loss term individually and report IF TM-validity (threshold 0.45) and diversity. The full model includes all loss terms.

Model	Validity (TM≥0.45)	Div. ↑
DuetRNA	44.00%	0.353
no_constraint	30.00%	0.55
no_torsion	27.92%	0.64
no_atom	24.83%	0.55
no_frame	6.87%	0.87

5.4.3. Loss Ablation

The complete loss consists of five parts, including constraint loss, torsion loss, atom loss, frame loss and sequence loss. When sequence loss is ablated, DuetRNA is unable predict the sequence, thus unable to reconstruct the RNA coordinates. Hence we only ablate below four geometric loss terms: the constraint loss ("no_constraint"), the frame loss ("no_frame"), the torsion loss ("no_torsion"), and the atom coordinate loss ("no_atom").The full model, which includes all four loss terms, achieves the highest IF TM-validity of 44.00% on 600 samples. Removing the constraint loss (no_constraint) drops validity to 30.00%, confirming its positive contribution. Removing frame loss leads to collapses and validity drops to only 6.87% , highlighting the essential role of SE(3) frame restriction in generating valid structures. Dropping torsion or atom losses also reduces performance, to 27.92% and 24.83% , respectively, while significantly improve the diversity. These results establish the importance of inter-frame coupling and geometric supervision, and the core frame loss is indispensable for structural validity.

6. Conclusion

In this work, we introduce DuetRNA, a dual-frame $\mathrm{SE}(3)$ flow matching framework for RNA sequence–structure co-generation. Motivated by a systematic analysis of seven single-frame candidates across $11{,}497$ static chains and $31{,}432$ multi-state relation groups, DuetRNA rep-

resents each nucleotide with coupled base- and sugar-centered frames, enabling residue-level organization and intra-residue backbone reconstruction geometry to be generated within a unified flow.

Our results show that DuetRNA with this factorization yields strong folding-based self-consistency, achieving 48.67% scTM-validity under inverse-folded sequence evaluation and 38.83% RhoFold-based TM-validity under direct joint generation on RNASolo dataset. More broadly, our analysis supports a geometric view of RNA generation in which base-centered frames capture nucleobase-mediated pairing, stacking, and long-range organization, while sugar-centered frames preserve the degrees of freedom needed for atomistic backbone reconstruction.

Current experiments focus on medium-length RNAs of 40–150 nt from RNASolo and RNA3DB, and evaluating dual-frame generation across shorter fragments, longer structured RNAs, and broader benchmark settings remains an important next step. Although modeling $(\mathrm{SE}(3) \times \mathrm{SE}(3))^N$ provides greater representational capacity than single-frame formulations, it also introduces additional computational cost. Meanwhile, the validity gains are accompanied by reduced diversity and novelty, identifying mode coverage as an important remaining limitation. Future work will extend DuetRNA to conditional and functional design tasks, such as ligand-binding aptamer and riboswitch design, and develop more efficient inter-frame coupling schemes and architectural optimizations.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning, specifically within the domain of structural bioinformatics and geometric modeling. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630: 493–500, 2024. doi: 10.1038/s41586-024-07487-w.

Adamczyk, B., Antczak, M., and Szachniuk, M. Rnasolo: a repository of cleaned pdb-derived rna 3d structures. *Bioinformatics*, 38(14):3668–3670, 2022. doi: 10.1093/bioinformatics/btac386.

Anand, R., Joshi, C. K., Morehead, A., Jamasb, A. R., Harris, C., Mathis, S. V., Didi, K., Ying, R., Hooi, B., and Liò, P. Rna-frameflow: Flow matching for de novo 3d rna backbone design. *Transactions on Machine Learning Research*, 2025.

Baek, M., McHugh, R., Anishchenko, I., Jiang, H., Baker, D., and DiMaio, F. Accurate prediction of protein–nucleic acid complexes using rosettafoldna. *Nature Methods*, 21: 117–121, 2024. doi: 10.1038/s41592-023-02086-5.

Bahai, A., Kwoh, C. K., Mu, Y., and Li, Y. Systematic benchmarking of deep-learning methods for tertiary rna structure prediction. *PLOS Computational Biology*, 20(12):e1012715, 2024. doi: 10.1371/journal.pcbi.1012715.

Barfoot, T. D. *State Estimation for Robotics*. Cambridge University Press, Cambridge, UK, 2017.

Chen, R. T. Q. and Lipman, Y. Flow matching on general geometries. In *International Conference on Learning Representations (ICLR)*, 2024.

Chothia, C. Principles that determine the structure of proteins. *Annual Review of Biochemistry*, 53:537–572, 1984. doi: 10.1146/annurev.bi.53.070184.002541.

Favor, A., Quijano, R., Chernova, E., Kubaney, A., Weidle, C., Esler, M. A., McHugh, L., Carr, A., Hsia, Y., Juergens, D., et al. De novo design of rna and nucleoprotein complexes. *bioRxiv*, 2025. doi: 10.1101/2025.10.01.679929.

Joshi, C., Jamasb, A., Viñas, R., Harris, C., Mathis, S., Morehead, A., Anand, R., and Liò, P. grnade: Geometric deep learning for 3d rna inverse design. In *International Conference on Learning Representations*, 2025.

Kagaya, Y., Zhang, Z., Ibtehaz, N., Wang, X., Nakamura, T., Punuru, P. D., and Kihara, D. Nufold: end-to-end approach for rna tertiary structure prediction with flexible nucleobase center representation. *Nature Communications*, 16:881, 2025. doi: 10.1038/s41467-025-56261-7.

Leontis, N. B. and Westhof, E. Geometric nomenclature and classification of RNA base pairs. *RNA*, 7(4):499–512, 2001. doi: 10.1017/S1355838201002515.

Li, Y., Zhang, C., Feng, C., Pearce, R., Freddolino, P. L., and Zhang, Y. Integrating end-to-end learning with deep geometrical potentials for ab initio rna structure prediction. *Nature Communications*, 14:5745, 2023. doi: 10.1038/s41467-023-41303-9.

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Ma, R., Zhang, Z., Wang, Z., Hua, C., Rao, J., Zhou, Z., and Zheng, S. RiboFlow: Conditional De Novo RNA co-design via synergistic flow matching. In *The Thirty-Ninth*

Annual Conference on Neural Information Processing Systems, volume 38, 2025.

Pearce, R., Omenn, G. S., and Zhang, Y. De novo rna tertiary structure prediction at atomic resolution using geometric potentials from deep learning. *bioRxiv*, 2022. doi: 10.1101/2022.05.15.491755.

Rubin, D., dos Santos Costa, A., Ponnapati, M., and Jacobson, J. Ribogen: Rna sequence and structure co-generation with equivariant multiflow, 2025.

Saenger, W. *Principles of Nucleic Acid Structure*. Springer-Verlag, New York, 1984. ISBN 0-387-90761-0. doi: 10.1007/978-1-4612-5190-3.

Shen, T., Hu, Z., Sun, S., Liu, D., Wong, F., Wang, J., Chen, J., Wang, Y., Hong, L., Xiao, J., Zheng, L., Krishnamoorthi, T., King, I., Wang, S., Yin, P., Collins, J. J., and Li, Y. Accurate rna 3d structure prediction using a language model-based deep learning approach. *Nature Methods*, 21:2287–2298, 2024. doi: 10.1038/s41592-024-02487-0.

Szikszai, M., Magnus, M., Sanghi, S., Kadyan, S., Bouatta, N., and Rivas, E. Rna3db: A structurally-dissimilar dataset split for training and benchmarking deep learning models for rna structure prediction. *Journal of Molecular Biology*, 436(17):168552, 2024. doi: 10.1016/j.jmb.2024.168552.

Wohlwend, J., Corso, G., Passaro, S., Getz, N., Reveiz, M., Leidal, K., Swiderski, W., Atkinson, L., Portnoi, T., Chinn, I., Silterra, J., Jaakkola, T., and Barzilay, R. Boltz-1: Democratizing biomolecular interaction modeling. *bioRxiv*, 2024. doi: 10.1101/2024.11.19.624167. Preprint.

A. Method Details

A.1. Model Architecture

A.1.1. Dual-Frame IPA-Transformer Trunk

Input conditioning. Each residue first receives a shared node embedding from timestep and positional encodings. Frame-specific pose conditioning is then added for the current base and sugar poses together with a frame-type embedding that distinguishes base tokens from sugar tokens.

Shared pair representation. The trunk constructs three pair channels: a base-view pair representation from base-node features and base translations, a sugar-view pair representation from sugar-node features and sugar translations, and a bridge-view pair representation from a per-residue summary of the base and sugar states together with their translation difference. These channels are fused into a shared pair tensor used by both geometric streams.

Dual geometric streams. At block k, the trunk maintains $T_i^{\text{base},(k)}$ and $T_i^{\text{sugar},(k)}$ together with their corresponding feature streams. Two invariant point attention modules process the current base and sugar poses in parallel, followed by per-stream transitions and rigid updates

$$T_i^{\text{base},(k+1)} = T_i^{\text{base},(k)} \circ \Delta T_i^{\text{base},(k)}, \qquad T_i^{\text{sugar},(k+1)} = T_i^{\text{sugar},(k)} \circ \Delta T_i^{\text{sugar},(k)}. \tag{24}$$

Object-token mixing and dynamic edge refinement. To enable explicit cross-object communication, the residue-wise base and sugar tokens are periodically concatenated into a sequence of $2L$ object tokens and processed by a joint transformer encoder. Between blocks, residue-level summaries from the current base and sugar streams are written back into the shared pair tensor through an edge transition, so the pair representation evolves jointly with trunk depth.

Self-conditioning. The pair stack also receives distograms computed from self-conditioned translations. During training, with probability 0.5, a stop-gradient forward pass provides predicted base and sugar translations that are fed back into the pair embedders. During sampling, the previous-step predictions play the same role.

Sequence and Torsion Heads.

Let H_i^{base} and H_i^{sugar} denote the final residue features of the two streams. The sequence head reads the final base token together with pooled pair context from the base-view, sugar-view, and fused pair representations. The torsion head reads the final sugar feature together with the initial sugar embedding.

A.2. Auxiliary Loss Definitions

Soft Base-Template Loss. For each candidate nucleotide $c \in \mathcal{C}$, let $X_i^{\text{base}}(c)$ be the base-template atomic coordinates of nucleotide c placed in the predicted base frame at residue i. Given predicted categorical probabilities $\pi_i(c)$, the probability-weighted base-template coordinates are

$$\tilde{X}_i^{\text{base}} = \sum_{c \in \mathcal{C}} \pi_i(c)\, X_i^{\text{base}}(c). \tag{25}$$

The corresponding auxiliary loss is

$$\mathcal{L}_{\text{soft-base}} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|\mathcal{B}_i^{\text{atom}}|} \sum_{a \in \mathcal{B}_i^{\text{atom}}} \left\| \tilde{x}_{i,a}^{\text{base}} - x_{i,a}^{\text{base,gt}} \right\|_2^2, \tag{26}$$

where $\mathcal{B}_i^{\text{atom}}$ is the atom23 base-atom subset for residue i. This term gives the sequence probabilities structural feedback through base-template coordinates; it does not replace the categorical cross-entropy objective.

Chain and Clash Auxiliary Losses. Let $\mathcal{E}$ denote the valid adjacent-residue set for which both $O3'_i$ and P_{i+1} are present. The chain-continuity auxiliary supervises the decoded adjacent $O3'_i$-P_{i+1} bridge:

$$\mathcal{L}_{\text{chain}} = \frac{1}{|\mathcal{E}|} \sum_{i \in \mathcal{E}} \left(\|\hat{x}_{i,O3'} - \hat{x}_{i+1,P}\|_2 - \left\| x_{i,O3'}^{\text{gt}} - x_{i+1,P}^{\text{gt}} \right\|_2 \right)^2. \tag{27}$$

A.3. Inference

Inference keeps two rigid objects for every residue throughout the whole trajectory. We first sample independent source chains $T_0^{\text{base}} \sim p_0^{\text{base}}$ and $T_0^{\text{sugar}} \sim p_0^{\text{sugar}}$. At step t, one coupled model evaluation takes $(T_t^{\text{base}}, T_t^{\text{sugar}})$ as input and predicts terminal frames $(\hat{T}_1^{\text{base}}, \hat{T}_1^{\text{sugar}})$. Both channels are then updated in parallel:

$$x_{t+\Delta t}^{\text{base}} = x_t^{\text{base}} + \Delta t \, \frac{\hat{x}_1^{\text{base}} - x_t^{\text{base}}}{1-t}, \qquad R_{t+\Delta t}^{\text{base}} = \text{Exp}_{R_t^{\text{base}}}\Big(\alpha(t)\Delta t \, \text{Log}_{R_t^{\text{base}}}(\hat{R}_1^{\text{base}})\Big), \tag{28}$$

$$x_{t+\Delta t}^{\text{sugar}} = x_t^{\text{sugar}} + \Delta t \, \frac{\hat{x}_1^{\text{sugar}} - x_t^{\text{sugar}}}{1-t}, \qquad R_{t+\Delta t}^{\text{sugar}} = \text{Exp}_{R_t^{\text{sugar}}}\Big(\alpha(t)\Delta t \, \text{Log}_{R_t^{\text{sugar}}}(\hat{R}_1^{\text{sugar}})\Big). \tag{29}$$

If self-conditioning is enabled, the previous-step predicted base and sugar translations are fed back into the pair representation. After the final integration step, the model produces sequence probabilities and intra-residue torsion estimates from the terminal base and sugar features. Then we choose the sequence by

$$\hat{s}_i = \arg\max_{c \in \mathcal{C}} p_\theta(s_i = c \mid T^{\text{base}}, T^{\text{sugar}}) \tag{30}$$

and complete atom23 coordinates from the terminal pair of frames:

$$\hat{X}^{\text{atom23}} = \mathcal{A}\Big(\hat{T}^{\text{base}}, \hat{T}^{\text{sugar}}, \hat{\phi}, \hat{s}\Big) \tag{31}$$

to obtain the final atom23 structure.

B. Frame Representation Analysis: Extended Results

This appendix provides the extended diagnostics underlying the frame representation analysis of §5.3. The main text reports the two core findings: base anchoring better captures inter-residue relation geometry, while a separate sugar/backbone frame is needed for local reconstruction. Here we report the candidate-frame catalog and deployed coordinate definitions, the canonicalization and continuity screen (§B.2), the static relation-separation analysis (§B.3), the per-relation multi-state breakdown including noncanonical pairs and stacking (§B.4), and the single-frame and dual-frame reconstruction sweep (§B.5).

B.1. Candidate Frame Catalog

We audit seven residue-level single-frame candidates spanning sugar/backbone and base-anchored constructions. The display names below are used consistently in the paper; the corresponding implementation identifiers are shown in parentheses.

Base family.

- **Base-Plane** (`base_plane_anchor`): a chemically anchored base-plane frame. The origin is the glycosidic connection atom ($N9$ for purines and $N1$ for pyrimidines). The orientation is defined by the fitted nucleobase plane normal and an in-plane chemical axis ($C4 \rightarrow C8$ for purines, $C4 \rightarrow C2$ for pyrimidines).
- **Base-Inertial** (`base_inertial`): an inertial frame over all nucleobase heavy atoms, with residue-type-specific sign and quadrant disambiguation. The origin remains the glycosidic connection atom.
- **Base-Center** (`base_center_standard`): the base-origin control. It uses the same inertial orientation family as Base-Inertial, but moves the origin to the mass centroid of the nucleobase heavy atoms.

Sugar/backbone family.

- **Sugar-GS** (`baseline_geom`): the legacy three-atom Gram-Schmidt frame built from $(O4', C4', C3')$ with origin at $C4'$, following the axis convention used in RNA-FrameFlow (Anand et al., 2025).
- **Sugar-Inertial-4** (`sugar4_inertial`): a four-atom mass-weighted inertial frame over $(O4', C4', C3', C5')$ with origin at $C4'$ and anchor-based sign disambiguation.

- **Sugar-Inertial-6** (`sugar6_inertial`): a six-atom inertial variant over $(C1', C2', C3', C4', O4', C5')$ with origin at $C4'$.
- **Sugar-Ring** (`sugar5_ring_axis`): a five-membered sugar-ring frame over $(O4', C1', C2', C3', C4')$, with ring-plane orientation and a sugar-ring centroid origin.

B.1.1. Deployed Frame Definitions

The deployed dual-frame model uses Sugar-GS for $T_i^{\text{sugar}} = (R_i^{\text{sugar}}, t_i^{\text{sugar}})$ and Base-Plane for $T_i^{\text{base}} = (R_i^{\text{base}}, t_i^{\text{base}})$. We give their exact coordinate definitions here.

Sugar-GS. Let $x_{O4'}, x_{C4'}, x_{C3'} \in \mathbb{R}^3$ be the corresponding atom coordinates. We set the origin at $C4'$,

$$t_i^{\text{sugar}} = x_{C4'}, \qquad e_{1,i}^{\text{sugar}} = \frac{x_{O4'} - x_{C4'}}{\|x_{O4'} - x_{C4'}\|_2}, \tag{32}$$

orthogonalize the $C3'$ direction against $e_{1,i}^{\text{sugar}}$,

$$\tilde{e}_{2,i}^{\text{sugar}} = (x_{C3'} - x_{C4'}) - \langle x_{C3'} - x_{C4'}, e_{1,i}^{\text{sugar}} \rangle e_{1,i}^{\text{sugar}}, \qquad e_{2,i}^{\text{sugar}} = \frac{\tilde{e}_{2,i}^{\text{sugar}}}{\|\tilde{e}_{2,i}^{\text{sugar}}\|_2}, \tag{33}$$

and complete a right-handed frame,

$$e_{3,i}^{\text{sugar}} = e_{1,i}^{\text{sugar}} \times e_{2,i}^{\text{sugar}}, \qquad R_i^{\text{sugar}} = \left[e_{1,i}^{\text{sugar}}, e_{2,i}^{\text{sugar}}, e_{3,i}^{\text{sugar}}\right]. \tag{34}$$

Base-Plane. Let $\mathcal{B}_i$ denote the heavy-atom set of the nucleobase, excluding sugar atoms, and let $a_i^{\text{link}} \in \{N9, N1\}$ be the glycosidic connection atom. The origin is

$$t_i^{\text{base}} = x_{a_i^{\text{link}}}. \tag{35}$$

We fit a plane to $\{x_a : a \in \mathcal{B}_i\}$ and denote its unit normal by $\tilde{n}_i$. To resolve the normal sign, we use the residue-type-specific reference normal

$$n_i^{\text{ref}} = \begin{cases} \left(x_{C4} - x_{C8}\right) \times \left(x_{C2} - x_{N9}\right), & \text{purine}, \\ \left(x_{C4} - x_{C2}\right) \times \left(x_{C6} - x_{N1}\right), & \text{pyrimidine}, \end{cases} \tag{36}$$

and set

$$e_{3,i}^{\text{base}} = \text{sign}\big(\langle \tilde{n}_i, n_i^{\text{ref}} \rangle\big) \tilde{n}_i. \tag{37}$$

The in-plane chemical axis is

$$u_i = \begin{cases} x_{C4} - x_{C8}, & \text{purine}, \\ x_{C4} - x_{C2}, & \text{pyrimidine}, \end{cases} \qquad \tilde{e}_{1,i}^{\text{base}} = u_i - \langle u_i, e_{3,i}^{\text{base}} \rangle e_{3,i}^{\text{base}}, \tag{38}$$

which gives

$$e_{1,i}^{\text{base}} = \frac{\tilde{e}_{1,i}^{\text{base}}}{\|\tilde{e}_{1,i}^{\text{base}}\|_2}, \qquad e_{2,i}^{\text{base}} = e_{3,i}^{\text{base}} \times e_{1,i}^{\text{base}}, \qquad R_i^{\text{base}} = \left[e_{1,i}^{\text{base}}, e_{2,i}^{\text{base}}, e_{3,i}^{\text{base}}\right]. \tag{39}$$

B.2. Frame Canonicalization and Continuity Screening

Before comparing downstream relation or reconstruction metrics, each frame must provide a usable $\mathrm{SE}(3)$ supervision target. We therefore first apply a canonicalization and continuity gate. A candidate passes only if its valid rate is at least 0.95, the maximum deterministic rebuild error is at most 10^{-5} degrees, and the fraction of perturbation cases with a rotation jump above 150° is zero. The perturbation suite includes micro sugar-pucker, χ-rotation, phosphate-swing, and bridge-motion probes.

Sugar-Inertial-4 is therefore not used as a formal relation-stability frame in the main evidence chain. Its average continuity is not poor, but the small nonzero catastrophic-jump tail makes it unsafe as a smooth flow-matching label. We retain it only as an exploratory reconstruction diagnostic when explicitly marked as such.

Table 5. Frame canonicalization and continuity screening. The gate is strict on catastrophic rotation jumps: any nonzero $> 150^\circ$ jump rate fails the candidate.

Frame	Valid rate	Median cont. ($^\circ$)	P95 cont. ($^\circ$)	Jump_{150}	Gate
Sugar-GS (`baseline_geom`)	1.00000	1.88	5.00	0	pass
Sugar-Inertial-4 (`sugar4_inertial`)	0.99982	1.16	22.36	9.8×10^{-4}	fail
Sugar-Inertial-6 (`sugar6_inertial`)	0.99969	0.59	5.00	0	pass
Sugar-Ring (`sugar5_ring_axis`)	0.99987	0.89	5.00	0	pass
Base-Plane (`base_plane_anchor`)	0.99584	0.00	1.7×10^{-6}	0	pass
Base-Inertial (`base_inertial`)	0.99584	0.00	1.7×10^{-6}	0	pass
Base-Center (`base_center_standard`)	0.99584	0.00	1.7×10^{-6}	0	pass

B.3. Static Relation Separation

On the full static pool of 11,497 processed RNASolo structure units, 1,765 nonredundant representatives, and 1,045 motif-rich chains, we examine which frame family more readily separates relation types and which more tightly clusters identical relation instances. We run 20 independent sub-experiments, structured as one full set and three filtered subsets (representative, motif-rich, stem-like), across five relation categories (canonical cWW pairs, noncanonical pairs, stacking, base-phosphate, and base-backbone). Canonical cWW pairs include Watson-Crick AU/UA and GC/CG pairs plus GU/UG wobble pairs. In 11 of the 20 subset-relation combinations, the frame with the best inter-group/intra-group distance ratio belongs to the sugar/backbone family while the frame with the best 1-NN retrieval accuracy belongs to the base family. This result implies that frames rooted in sugar/backbone atoms spread different relation categories farther apart, whereas frames rooted in base atoms group geometrically identical instances closer together. No single frame family achieves the best result in every channel. Table Table 6 reports the family-level split pattern.

For a residue pair (i, j), the inter-residue relation is represented by

$$R_{ij} = R_i^\top R_j, \qquad t_{ij} = R_i^\top (t_j - t_i), \tag{40}$$

where R_i and t_i are the rotation and translation of residue i.

Table 6. Family-level split pattern on the static pool. Each row reports the number of subset-relation combinations (out of 20) in which the indicated pattern holds: "sep" is the family that better separates relation types, and "ret" is the family that more tightly retrieves identical relation instances.

Channel	Split pattern (sep | ret)	# Combinations	Fraction
joint pose	sugar/backbone | base	11	0.55
joint pose	sugar/backbone | sugar/backbone	7	0.35
joint pose	base | base	1	0.05
joint pose	base | sugar/backbone	1	0.05
translation	sugar/backbone | base	13	0.65
translation	sugar/backbone | sugar/backbone	5	0.25
translation	base | base	1	0.05
translation	base | sugar/backbone	1	0.05

These counts are descriptive evidence of two distinct representational roles in RNA residue geometry, not formal statistical claims.

B.4. Multi-State Stability: Per-Relation Breakdown

Due to the flexibility of RNA, the same molecule may adopt different conformations. A frame that separates relation types in a static snapshot does not necessarily keep the same relation stable across conformational states. NMR and other multi-model entries therefore provide a natural test bed for measuring drift across plausible states of the same molecule.

Here we extend the multi-state analysis of §5.3 to the five paper-facing frames used in the completed external run: Sugar-GS, Sugar-Inertial-6, Sugar-Ring, Base-Plane, and Base-Center. The pool consists of 523 single-chain multi-model entries (496 tier-A NMR plus 27 other multi-model structures), comprising 31,432 relation groups, 984,382 normalized relation rows, and 336 high-coverage relation keys. Variation reflects a combination of conformational flexibility and experimental uncertainty.

For each key, *orientation drift* ΔR (median geodesic angle from the medoid rotation) and *translation drift* ΔT (median Euclidean distance from the medoid translation) are computed under each candidate frame. A key is *base-like* in a channel if the best base-family frame outperforms all sugar/backbone-family frames.

Table 7. Aggregate multi-state stability over all 31,432 relation groups. Lower is better.

Frame	Family	R^{rel} drift (°) ↓	t^{rel} drift (Å) ↓
Sugar-GS (`baseline_geom`)	sugar/Gram-Schmidt	15.21	1.45
Sugar-Inertial-6 (`sugar6_inertial`)	sugar/inertial	14.60	1.40
Sugar-Ring (`sugar5_ring_axis`)	sugar/ring	14.44	1.31
Base-Plane (`base_plane_anchor`)	base/plane	12.26	1.05
Base-Center (`base_center_standard`)	base/center	**12.26**	**1.03**

Table 8. Per-relation multi-state stability across the five external-run frames. Lower is better. "cP" = canonical cWW pairs ($n = 2{,}334$), "ncP" = noncanonical pairs ($n = 2{,}125$), and "stk" = stacking ($n = 26{,}973$).

	cP ($n = 2{,}334$)		ncP ($n = 2{,}125$)		stk ($n = 26{,}973$)	
Frame	R°	t Å	R°	t Å	R°	t Å
Sugar-GS	13.35	1.70	19.03	2.48	15.27	1.36
Sugar-Inertial-6	12.32	1.52	18.71	2.42	14.70	1.33
Sugar-Ring	11.85	1.31	18.72	2.30	14.58	1.25
Base-Plane	7.78	0.61	16.42	1.85	12.67	1.07
Base-Center	**7.76**	**0.47**	**16.41**	**1.79**	**12.67**	**1.06**

Canonical cWW pairs strongly favor base-anchored frames in both rotation and translation channels. Noncanonical pairs favor base-anchored frames at the aggregate level, but remain key-heterogeneous: among 191 high-coverage noncanonical relation keys, 76 keys exhibit a sugar/backbone-like rotation pattern with a base-like translation pattern, while by evaluation weight 74% of pairs are base-like in both channels. Stacking is the most heterogeneous: by evaluation weight, 93.6% of stacking pairs favor base-like rotation stability, while 44.6% of stacking translations remain sugar/backbone-like; by key count, 72.7% of stacking keys are base-like in rotation. The multi-state evidence therefore identifies canonical cWW pairing as cleanly base-governed and stacking as a relation in which orientation and translation behave differently.

Table Table 9 reports, for each category, the fraction of keys whose lowest drift comes from a base-family frame. For canonical cWW pairing, a base-family frame achieves the lowest drift for every key in both channels, consistent with the physical organization of Watson-Crick and wobble pairing at the nucleobase level. Noncanonical pairing and stacking are more heterogeneous: only 38% of noncanonical keys have their lowest orientation drift from a base-family frame, yet 68% have their lowest translation drift from a base-family frame; stacking shows 73% of keys favoring a base-family frame in orientation and 68% in translation. Thus, no single frame family yields the lowest drift in both channels for every relation type.

B.5. Reconstruction Path: Full Single-Frame and Dual-Frame Sweep

The main text reports the core reconstruction tradeoff. Here we report the full reconstruction-path sweep on 51,013 interaction pairs for the gate-passing single frames supported by this path and the three dual configurations that pair a base-anchored frame with the Gram-Schmidt Sugar-GS frame. Sugar-Inertial-4 is excluded from this formal table because it failed the screening gate in Table Table 5.

Among single-frame choices, base-anchored frames preserve the relation geometry identified in the preceding sections but suffer from poor local sugar/backbone reconstruction; sugar/backbone frames give better local reconstruction but cannot accurately recover base atoms. The dual configurations decouple these roles: replacing the global frame with a base-anchored channel substantially reduces base-channel RMSD while retaining the Sugar-GS local reconstruction path.

Overall, the three analyses converge on the same conclusion: selecting any one single frame as a universal representative is inadequate for RNA. Base-anchored frames best capture inter-residue relation geometry (§B.3, §B.4) but do not by themselves solve local sugar/backbone reconstruction (§B.5); sugar/backbone-anchored frames show the complementary behavior. The dual-frame design resolves this by assigning each role to a dedicated channel: a global base frame for relation-bearing geometry and a sugar frame for reconstruction-bearing geometry (§4.1).

Table 9. Fraction of high-coverage relation keys whose lowest drift is achieved by a base-family frame.

Relation category	Keys	Orientation	Translation
Canonical cWW pairs	6	100%	100%
Noncanonical pairs	191	38%	68%
Stacking	139	73%	68%

Table 10. Reconstruction-path comparison on 51,013 interaction pairs. $\mathrm{RMSD}_{\mathrm{base}}$ measures atoms reconstructed in the base channel; $\mathrm{RMSD}_{\mathrm{bb}}$ and $\mathrm{RMSD}_{\mathrm{bridge}}$ measure atoms reconstructed in the sugar/backbone channel and the next-residue $O3' \rightarrow P$ bridge.

Configuration	$\mathrm{RMSD}_{\mathrm{base}}$ (Å) ↓	$\mathrm{RMSD}_{\mathrm{bb}}$ (Å) ↓	$\mathrm{RMSD}_{\mathrm{bridge}}$ (Å) ↓	Clash rate ↓
Single-frame configurations				
Sugar-GS	6.21	1.88	2.84	4.9×10^{-4}
Sugar-Ring	5.37	1.85	2.75	5.5×10^{-2}
Sugar-Inertial-6	7.12	1.96	2.30	8.3×10^{-4}
Base-Plane	5.85	2.98	4.06	4.9×10^{-4}
Base-Inertial	5.60	2.34	2.91	2.1×10^{-3}
Base-Center	5.61	3.20	3.74	2.4×10^{-3}
Dual-frame configurations with Sugar-GS local frame				
Base-Plane + Sugar-GS	1.58	1.88	2.84	4.8×10^{-4}
Base-Inertial + Sugar-GS	1.52	1.88	2.84	4.8×10^{-4}
Base-Center + Sugar-GS	1.09	1.88	2.84	4.9×10^{-4}

C. Experiment Details and Ablation Studies

C.1. Auxiliary Losses, Ablations, and Hyperparameters

The auxiliary losses are defined in Appendix A.2. This appendix reports their empirical ablation and the hyperparameter ranges used in the paper.

C.1.1. Auxiliary Objective Ablation

We ablate auxiliary objectives beyond the core frame, sequence, torsion, and atom23 reconstruction losses. Each variant uses the same architecture, the same 120K training budget, and the same fixed sampler $(N_T = 50, \alpha = 10)$. The auxiliary losses are the soft base-template loss, chain-continuity loss, and steric clash loss defined above.

Table 11. Auxiliary objective ablation at fixed sampling $(N_T = 50, \alpha = 10)$. The listed weights correspond to auxiliary terms only; all rows keep the core training objective unchanged and use 600 generated structures.

Auxiliary setting	λ_{soft}	λ_{chain}	λ_{clash}	IF scTM ↑	IF scRMSD ↑	GS scTM ↑	GS scRMSD ↑
none	0.0	0.0	0.0	**34.33**	28.83	**26.00**	22.33
soft	0.5	0.0	0.0	25.67	23.50	20.17	19.17
soft + chain	0.5	1.0	0.0	32.33	27.83	23.17	22.17
soft + chain + clash	0.5	1.0	5.0	33.50	**32.33**	25.67	**27.17**

At the fixed sampler, the core objective without auxiliary terms gives the best scTM-validity under both IF and GS. The soft base-template loss alone substantially reduces validity, suggesting that this auxiliary can over-constrain the sequence-structure channel when used without additional geometric regularization. Adding the chain-continuity term partially recovers performance, and adding the steric clash term gives the best scRMSD-validity, consistent with its role as an atom-level geometric regularizer.

C.1.2. Hyperparameter Ranges

Table Table 12 lists the objective weights that appear in the main method, together with the auxiliary-loss ranges used or reserved for ablations.

Table 12. Core objective weights and auxiliary-loss ablation ranges.

Symbol	Reported value(s)	Ablation range	Role
τ	0.25	$[0.1, 0.5]$	time gate for decoded-geometry supervision
λ_{Seq}	1.0	$[0.5, 2.0]$	nucleotide cross-entropy
λ_{Loc}	1.0	$[0.25, 2.0]$	torsion and decoded atom23 reconstruction objective
λ_{soft}	0 or 0.5	$[0, 1.0]$	soft base-template auxiliary
λ_{chain}	0 or 1.0	$[0, 2.0]$	adjacent $O3'$-P bridge continuity
λ_{clash}	0	$[0, 1.0]$	optional steric clash auxiliary

Table 13. Augmented Table 2 with Wilson 95% CIs and significance vs. $(N_T{=}100, \alpha{=}20)$.

		IF		GS	
N_T	α	Val. scTM % [95% CI]	vs. ref.	Val. scTM % [95% CI]	vs. ref.
50	10	32.00 [28.39, 35.84]	**	24.67 [21.39, 28.27]	*
100	**20**	**40.67 [36.81, 44.64]**	—	**31.17 [27.59, 34.98]**	—
200	10	37.83 [34.04, 41.78]	ns	27.00 [23.60, 30.69]	ns
200	20	38.00 [34.20, 41.95]	ns	28.50 [25.03, 32.24]	ns
300	10	36.00 [32.26, 39.92]	ns	27.50 [24.08, 31.21]	ns
300	20	38.17 [34.37, 42.12]	ns	28.83 [25.35, 32.58]	ns

C.1.3. Significance Analysis

All validity metrics are binomial proportions over $n = 600$ generated structures. We report Wilson 95% confidence intervals (CI) and assess pairwise differences with a two-proportion z-test. Significance against the reference row of each table is marked as $^{*}p{<}0.05$, $^{**}p{<}0.01$, $^{***}p{<}0.001$; ns: $p \geq 0.05$.

Table 13: only $(50, 10)$ differs significantly from the chosen $(100, 20)$; higher budgets show diminishing returns. Table 14: 190K/200K are significantly worse than 180K, supporting the post-180K regression.

Table 14. Augmented Table 3 with Wilson 95% CIs and significance vs. the 180K checkpoint.

	IF Val. scTM %		GS Val. scTM %	
Steps	Value [95% CI]	vs. 180K	Value [95% CI]	vs. 180K
120K	40.67 [36.81, 44.64]	ns	31.17 [27.59, 34.98]	*
180K	**44.00 [40.08, 48.00]**	—	**37.83 [34.04, 41.78]**	—
190K	38.33 [34.53, 42.29]	*	29.17 [25.67, 32.93]	**
200K	36.67 [32.91, 40.60]	**	31.00 [27.43, 34.81]	*