PHYSICS-ALIGNED DECODING (PAD) FOR DISCRETE PROTEIN STRUCTURE REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Discrete representations learned by deep autoencoders are increasingly reused as intermediate state spaces in generative, conditional, and autoregressive models. In this work, we empirically identify an objective-level failure mode in discrete protein structure tokenizers trained with reconstruction-aligned losses: despite low global reconstruction error, learned tokens encode locally unphysical geometry, including covalent distortions and steric clashes. We show that these violations are deterministic and persistent under reuse. We test the hypothesis that this behavior arises from objective misspecification rather than architectural limitations, and introduce Physics-Aligned Decoding (PAD), a minimal intervention that augments reconstruction objectives with differentiable physical priors. Without changing architecture or regenerating the codebook, PAD reshapes token semantics and restores physical validity while preserving reconstruction accuracy. Our results highlight how loss geometry determines representation semantics, and demonstrate the importance of objective alignment when discrete representations are reused beyond static reconstruction.

1 INTRODUCTION

Deep representation learning has traditionally emphasized reconstruction fidelity as a proxy for representation quality. In vector-quantized autoencoders (VQ-VAEs) (van den Oord et al., 2017; Razavi et al., 2019), this paradigm enables compression of high-dimensional inputs into discrete latent codes that can be reused for downstream modeling. Recently, this approach has been extended to protein structure, where discrete structural tokens serve as interfaces to generative models, conditional editors, and large language models (Gao et al., 2024; Lin et al., 2025; Yuan et al., 2025).

In these settings, discrete tokens are no longer treated as passive compression artifacts. Instead, they define a latent *state space* over which downstream models operate. This shift raises a critical question: what semantic content do these discrete states encode? When learned representations are reused as intermediate states, errors that are benign under static reconstruction may become amplified or irreparable.

Protein structures provide a stringent test case. Valid conformations lie on a highly constrained physical manifold defined by covalent geometry, steric exclusion, and torsional regularities (Engh & Huber, 1991; Ramachandran et al., 1963; Dunbrack, 2002). Yet most protein tokenizers are trained using reconstruction-aligned objectives—including coordinate-level L_2 losses or frame-aligned variants such as FAPE (Jumper et al., 2021)—that are insensitive to localized physical violations.

In this work, we show that this mismatch leads to a systematic failure mode: discrete tokens encode unphysical micro-geometry despite achieving low global reconstruction error. We argue that this behavior reflects objective misspecification rather than architectural deficiency, and we test this hypothesis through controlled intervention.

2 RECONSTRUCTION-ALIGNED DECODING AND ITS BLIND SPOTS

Reconstruction-aligned objectives minimize additive losses over geometric features such as Cartesian coordinates, distances, or frames (Ingraham et al., 2019; Gao et al., 2024). These losses induce

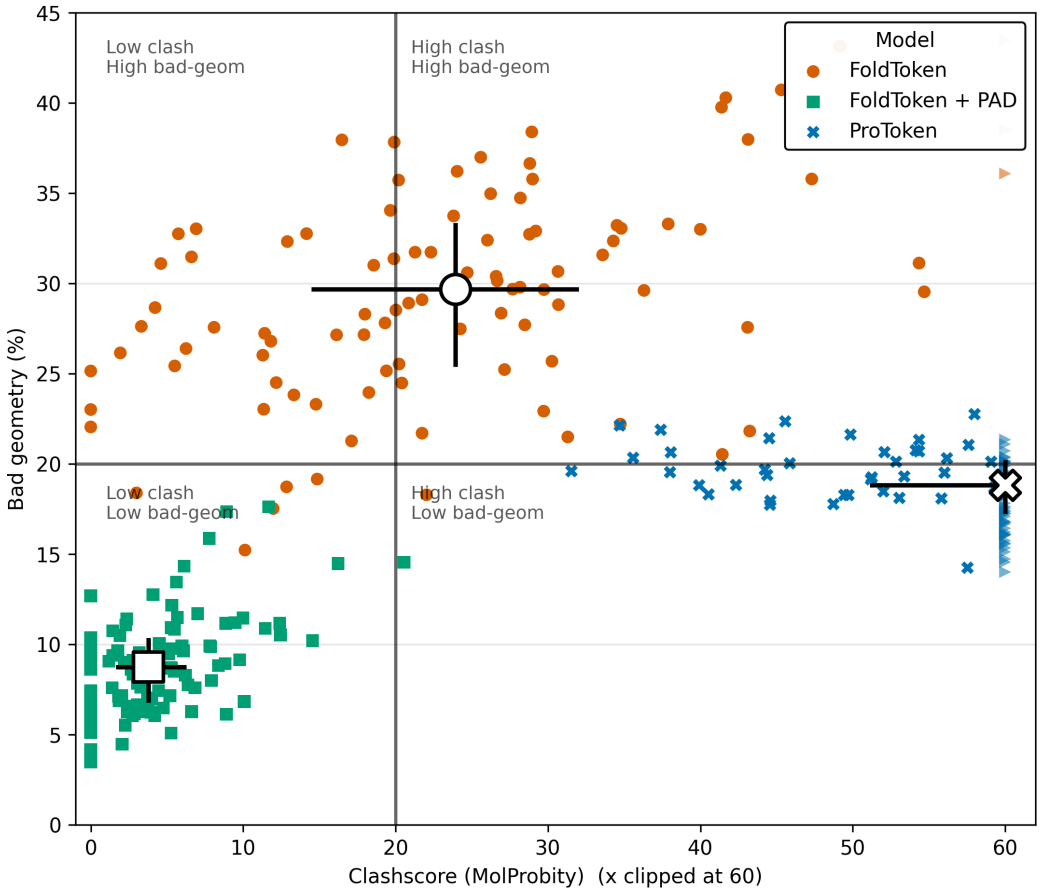


Figure 1: **Deterministic physical violations under reconstruction-aligned decoding.** Clashscore versus bad-geometry fraction for decoded structures on held-out conformations. Despite low global error, reconstruction-aligned tokenizers exhibit severe local violations. Physics-Aligned Decoding (PAD) restores physical validity without architectural changes.

smooth, symmetric penalties that treat all deviations as comparable, regardless of physical interpretation.

In physically constrained systems, this symmetry is problematic. A small displacement into empty space and an equally small displacement into another atom's excluded volume incur similar penalties under an L_2 loss, even though the latter configuration is physically forbidden (Bondi, 1964; Chen et al., 2010). As a result, the optimizer may reduce average reconstruction error by introducing localized covalent distortions or steric overlaps.

Figure 1 illustrates this failure mode across representative tokenizers. Notably, violations are reproducible under a single deterministic encode–decode pass, indicating that they reflect stable optima of the training objective rather than stochastic decoding noise.

3 Scientific-Method Framing

We cast this observation as a hypothesis-driven empirical study:

Observation. Reconstruction-aligned decoding produces discrete tokens whose decoded structures exhibit local physical violations.

Hypothesis. These violations arise from objective misspecification: symmetric reconstruction losses fail to encode asymmetric physical constraints such as steric exclusion.

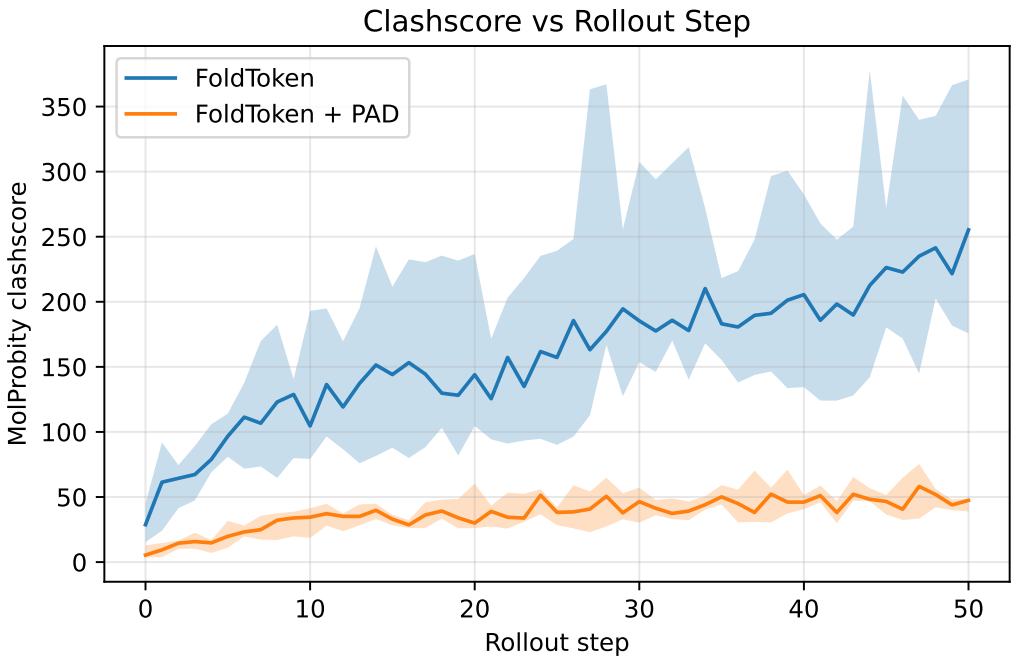


Figure 2: **Token reuse reveals semantic stability.** Median clashscore over multi-step autoregressive rollouts. Reconstruction-aligned tokens rapidly accumulate physical violations, whereas PAD tokens remain stable under reuse.

Prediction. If the hypothesis is correct, modifying only the decoding objective—while keeping architecture and codebook fixed—should reshape token semantics and eliminate violations under both single-step decoding and downstream reuse.

Test. We implement Physics-Aligned Decoding (PAD) and evaluate physical validity under controlled conditions.

4 Physics-Aligned Decoding

Physics-Aligned Decoding augments the reconstruction objective with differentiable physical priors that encode local feasibility. The decoding loss takes the form

$$\mathcal{L}_{\text{PAD}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{geom}}\mathcal{L}_{\text{geom}} + \lambda_{\text{vdw}}\mathcal{L}_{\text{vdw}},$$

where $\mathcal{L}_{\text{geom}}$ enforces empirical covalent and torsional constraints (Engh & Huber, 1991; Shapovalov & Dunbrack, 2011) and $\mathcal{L}_{\text{vdw}}$ penalizes steric overlap using smooth van der Waals-inspired terms (Bondi, 1964). Full definition can be found in the appendix.

Crucially, PAD does not alter model architecture or regenerate the discrete codebook. We implement it via parameter-efficient fine-tuning using LoRA (Hu et al., 2021), ensuring that observed changes reflect objective-induced semantic reorganization rather than increased model capacity.

5 Results

Applying PAD shifts decoded structures into physically admissible regimes while preserving reconstruction fidelity (Fig. 1). To test whether this improvement reflects genuine semantic change rather than post hoc smoothing, we evaluate token reuse in downstream generative settings.

Figure 2 shows that under autoregressive reuse, reconstruction-aligned tokens accumulate steric clashes rapidly, whereas PAD tokens maintain bounded physical quality. Importantly, improvements

are visible at early rollout steps, indicating that unphysical micro-geometry is encoded directly into baseline token semantics.

6 Discussion

Our results demonstrate that reconstruction fidelity alone is an insufficient training signal when discrete representations are reused as state spaces. The geometry of the loss function determines which regions of representation space are accessible, and symmetric objectives may admit semantically invalid optima in constrained domains.

While our experiments focus on protein structure, the underlying principle is general. Any discrete representation learned on a constrained manifold—and reused for planning, rollouts, or conditional editing—may require objective terms that encode feasibility constraints rather than average reconstruction error.

7 Limitations

Physics-Aligned Decoding enforces local feasibility but does not model full molecular energetics or long-range interactions. More broadly, this work does not propose a universal objective, but illustrates how objective misspecification can be diagnosed and corrected using controlled empirical analysis.

References

A. Bondi. van der waals volumes and radii. *Journal of Physical Chemistry*, 68(3):441–451, 1964.

Vincent B. Chen, William B. Arendall, Jeffrey J. Headd, et al. Molprobity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography*, 66(1):12–21, 2010. doi: 10.1107/S0907444909042073.

Roland L Dunbrack. Rotamer libraries in the 21st century. *Current Opinion in Structural Biology*, 12(4):431–440, 2002. ISSN 0959-440X. doi: https://doi.org/10.1016/S0959-440X(02)00344-5.

R. A. Engh and R. Huber. Accurate bond and angle parameters for x-ray protein structure refinement. *Acta Crystallographica Section A*, 47:392–400, 1991.

Zhangyang Gao, Cheng Tan, Jue Wang, Yufei Huang, Lirong Wu, and Stan Z. Li. Foldtoken: Learning protein language via vector quantization and beyond. 2024. URL https://arxiv.org/abs/2403.09673.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. 2021. URL https://arxiv.org/abs/2106.09685.

John Ingraham, Vikas Garg, Regina Barzilay, and Tommi Jaakkola. Generative models for graph-based protein design. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Xiaohan Lin, Zhenyu Chen, Yanheng Li, et al. Unifying sequence-structure coding for advanced protein engineering via a multimodal diffusion transformer. *Chem. Sci.*, 16:11087–11102, 2025. doi: 10.1039/D5SC02055G. URL http://dx.doi.org/10.1039/D5SC02055G.

Christine A. Orengo, Andrew D. Michie, et al. CATH: a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108, 1997. doi: 10.1016/S0969-2126(97)00260-8.

G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7:95–99, 1963.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. 2019. URL https://arxiv.org/abs/1906.00446.

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

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NeurIPS*, 2017.

Yann Vander Meersche, Gabriel Cretin, Aria Gheeraert, Jean-Christophe Gelly, and Tatiana Galochkina. Atlas: protein flexibility description from atomistic molecular dynamics simulations. *Nucleic Acids Research*, 52(D1):D384–D392, 11 2023. ISSN 0305-1048. doi: 10.1093/nar/gkad1084. URL https://doi.org/10.1093/nar/gkad1084.

Xinyu Yuan, Zichen Wang, Marcus Collins, and Huzefa Rangwala. Protein structure tokenization: Benchmarking and new recipe. 2025. URL https://arxiv.org/abs/2503.00089.

A Detailed van der Waals and Steric Potentials

This appendix provides the explicit functional forms of the differentiable steric-exclusion and van der Waals terms used in Physics-Aligned Decoding (PAD). The forms are inspired by Rosetta-style Lennard–Jones potentials, but are modified to ensure smooth, numerically stable gradients for end-to-end training (soft-floored distances, a linear continuation at short range, and a smooth cutoff for attraction).

A.1 Soft-Floored Interatomic Distances

For atoms indexed by p, q with decoded coordinates $\hat{x}_p, \hat{x}_q$, define

$$d_{pq} = \|\hat{x}_p - \hat{x}_q\|.$$

To avoid unbounded gradients as $d_{pq} \to 0$, we use a soft-floored distance

$$\tilde{d}_{pq} = \sqrt{d_{pq}^2 + d_0^2}, \tag{1}$$

with small constant $d_0 > 0$. All Lennard–Jones terms below are evaluated using $\tilde{d}_{pq}$.

A.2 Differentiable Steric Clash Penalty

A coarse differentiable clash barrier is applied over backbone atoms and pseudo-C_β atoms with radii r_p. For atom pairs $p < q$, we penalize overlaps using

$$\mathcal{L}_{\text{clash}}(\hat{X}) = \mathbb{E}_{p<q}\Big[\big(\max\{0,\ (r_p + r_q + \delta) - \tilde{d}_{pq}\}\big)^2\Big], \tag{2}$$

where $\delta > 0$ is a small margin. Using $\tilde{d}_{pq}$ (rather than d_{pq}) ensures bounded gradients even under severe overlap. This definition is used consistently throughout the paper; the squared term applies to distance penetration quadratically.

A.3 Rosetta-Inspired van der Waals Energy

We include a smooth van der Waals loss $\mathcal{L}_{\text{vdw}}$ comprising a repulsive (fa_rep) and an attractive (fa_atr) component. For atom pairs p, q, let σ_{pq} and ϵ_{pq} denote Lennard–Jones radius and well-depth parameters. Define

$$x_{pq} = \frac{\sigma_{pq}}{\tilde{d}_{pq}}.$$

Shifted Lennard–Jones form. We use the shifted potential

$$E_{pq}^{\text{LJ+}}(\tilde{d}) = \epsilon_{pq}\big(x_{pq}^{12} - 2x_{pq}^{6} + 1\big), \tag{3}$$

which is zero at $\tilde{d} = \sigma_{pq}$ and increases sharply for $\tilde{d} < \sigma_{pq}$.

Repulsive term (fa_rep**).** Let $\tilde{d}_t = 0.6\sigma_{pq}$. For $\tilde{d} \le \tilde{d}_t$, we use a linear continuation

$$E_{pq}^{\text{lin}}(\tilde{d}) = a_{pq}\tilde{d} + b_{pq},$$

where (a_{pq}, b_{pq}) are chosen to match both the value and slope of $E_{pq}^{\text{LJ+}}$ at $\tilde{d}_t$. The repulsive energy is

$$E_{pq}^{\text{rep}}(\tilde{d}) = \begin{cases} E_{pq}^{\text{lin}}(\tilde{d}), & \tilde{d} \le \tilde{d}_t, \\ E_{pq}^{\text{LJ+}}(\tilde{d}), & \tilde{d}_t < \tilde{d} \le \sigma_{pq}, \\ 0, & \tilde{d} > \sigma_{pq}. \end{cases} \tag{4}$$

Attractive term (fa_atr)**.** Define the standard Lennard–Jones form

$$E_{pq}^{\mathrm{LJ}}(\tilde{d}) = \epsilon_{pq}\big(x_{pq}^{12} - 2x_{pq}^{6}\big).$$

We apply a smooth cutoff between an inner radius d_{in} and an outer cutoff d_{out} using a cubic Hermite switch

$$s(t) = 2t^3 - 3t^2 + 1, \qquad t = \frac{\tilde{d} - d_{\mathrm{in}}}{d_{\mathrm{out}} - d_{\mathrm{in}}}. \tag{5}$$

The attractive energy is

$$E_{pq}^{\mathrm{atr}}(\tilde{d}) = \begin{cases} -\epsilon_{pq}, & \tilde{d} \le \sigma_{pq}, \\ E_{pq}^{\mathrm{LJ}}(\tilde{d}), & \sigma_{pq} < \tilde{d} \le d_{\mathrm{in}}, \\ E_{pq}^{\mathrm{LJ}}(\tilde{d})\, s(t), & d_{\mathrm{in}} < \tilde{d} \le d_{\mathrm{out}}, \\ 0, & \tilde{d} > d_{\mathrm{out}}. \end{cases} \tag{6}$$

In our implementation, $(d_{\mathrm{in}}, d_{\mathrm{out}})$ are chosen as fixed multiples of σ_{pq}, and the switch ensures continuous energy and gradients at the cutoff.

Total van der Waals loss. The full van der Waals loss sums over valid atom pairs, excluding bonded and near-bonded interactions via an exclusion mask $\chi_{pq} \in \{0, 1\}$:

$$\mathcal{L}_{\mathrm{vdw}}(\hat{X}) = w_{\mathrm{rep}} \sum_{p<q} \chi_{pq} E_{pq}^{\mathrm{rep}}(\tilde{d}_{pq}) + w_{\mathrm{atr}} \sum_{p<q} \chi_{pq} E_{pq}^{\mathrm{atr}}(\tilde{d}_{pq}). \tag{7}$$

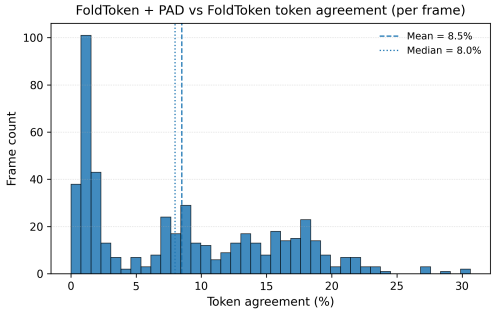
Together, $\mathcal{L}_{\mathrm{clash}}$ and $\mathcal{L}_{\mathrm{vdw}}$ provide complementary shaping: the clash barrier suppresses forbidden overlaps, while the van der Waals term provides a smooth repulsive/attractive landscape that guides decoded structures toward physically plausible local minima under differentiable optimization.

B AUTOREGRESSIVE ROLLOUT MODEL.

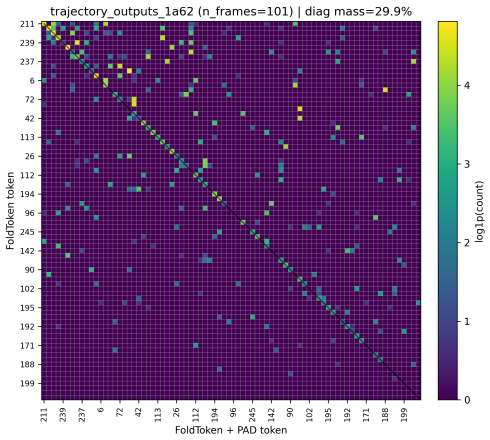
For the downstream multi-step reuse stress test (Sec. **??**), we train a lightweight Transformer encoder that predicts the next-frame token at each residue given the current-frame token sequence. Concretely, the model embeds per-residue discrete tokens with an embedding dimension of $d = 256$, adds sinusoidal positional encodings over residue index, and applies $L = 4$ TransformerEncoder layers with $H = 8$ attention heads and feedforward width $4d$, followed by a linear classifier to K token logits per residue. The model is trained with cross-entropy on next-frame prediction pairs (z_t, z_{t+1}) extracted from tokenized MD trajectories, where $z_t \in \{0, \ldots, K-1\}^L$ is the full-length residue token vector for a frame. We use an $80/20$ split over trajectory frames (first $0.8T$ frames for training, remaining $0.2T$ frames for seeding held-out rollouts), AdamW optimization with learning rate 3×10^{-4} and weight decay 10^{-2}, batch size 16, gradient clipping at 1.0, and train for 10 epochs (dropout 0.1; GELU; pre-norm). The vocabulary size K is inferred from the maximum token id observed in the training frames ($K = \max(z_{\le 0.8T}) + 1$) unless provided explicitly. All rollout experiments use identical model architecture, training procedure, and sampling hyperparameters for FoldToken and FoldToken+PAD, so differences in physical validity arise from token semantics rather than downstream model differences.

C TOKEN-LEVEL EFFECTS OF PHYSICS-ALIGNED DECODING

This appendix provides a detailed, token-level analysis of the semantic changes induced by Physics-Aligned Decoding (PAD), complementing the aggregate statistics reported in the main text. While the main paper focuses on per-frame agreement and summary metrics, the visualizations below reveal how these changes arise from structured transitions between discrete token states and from localized, position-dependent effects along the sequence. Together, these analyses clarify that PAD induces controlled semantic reorganization rather than diffuse or random perturbations.



(a) Per-frame token agreement between FoldToken and FoldToken+PAD, defined as the fraction of residues whose discrete token assignment is unchanged. Dashed and dotted lines indicate the mean (8.5%) and median (8.0%) agreement, respectively.



(b) Aggregated token transition matrix between FoldToken and FoldToken+PAD over 101 frames. Each entry counts how often a residue assigned token i under the baseline model is reassigned to token j under PAD. Color indicates $\log(1 + \text{count})$.

Figure 3: Aggregate and token-level views of token reassignment under Physics-Aligned Decoding. Left: histogram of per-frame token agreement between FoldToken and FoldToken+PAD, defined as the fraction of residues whose discrete token assignment is unchanged. Right: aggregated token transition matrix between FoldToken and FoldToken+PAD over 101 frames, restricted to the top-k most frequently used tokens ($k = 20$). Each entry (i, j) counts the number of residue positions for which token i under the baseline FoldToken model is reassigned to token j under PAD; color indicates $\log(1 + \text{count})$.

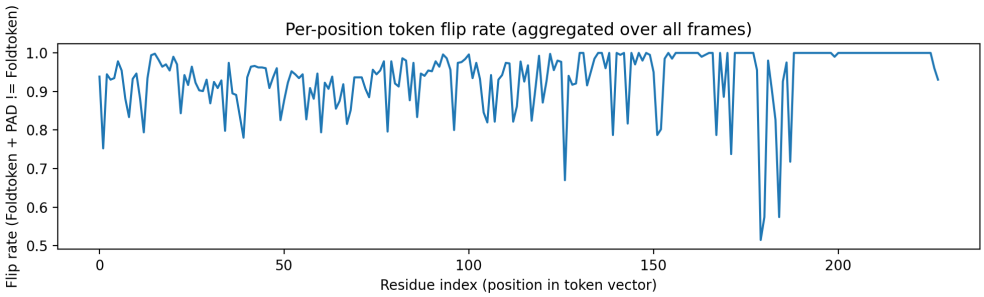


Figure 4: Per-position token flip rate aggregated across all frames, defined as the fraction of frames in which the FoldToken+PAD assignment differs from the baseline FoldToken assignment at a given residue index. While most positions exhibit high stability, several localized regions show consistently elevated flip rates, indicating position-dependent semantic reorganization under PAD.

D LoRA FINE-TUNING CONFIGURATION

This appendix summarizes the LoRA configuration used for the geometry-push experiments reported in the main text. We list only the hyperparameters and architectural targets that materially affect the learned representation; standard training, logging, and data-loading details are omitted for brevity.

LoRA hyperparameters. We use LoRA with rank $r = 32$ and scaling factor $\alpha = 64$, with dropout 0.1. The base FoldToken model is frozen during fine-tuning, and no bias parameters are trained. LoRA adapters are trained in feature-extraction mode.

Targeted modules. LoRA adapters are applied broadly across the encoder and decoder, including attention projections, feed-forward layers, and geometric feature pathways, while preserving the original coordinate prediction heads. Specifically, adapters are attached to:

- the VQ projection layers and embedding MLPs,
- encoder attention (W_Q, W_K, W_V, W_O), feed-forward, and geometric feature modules,
- decoder GNN attention, feed-forward, and geometric feature modules,
- decoder coordinate and quaternion prediction projections.

Optimization and geometry losses. Fine-tuning uses a learning rate of 1×10^{-4} with cosine scheduling and a warmup of 100 steps. Gradients are clipped to a norm of 2.0 and accumulated over 12 steps. Geometry-aware losses are enabled throughout training, including bond length, bond angle, Ramachandran, and ω-torsion penalties (each with weight 0.5), with an additional peptide planarity term (weight 0.05). A van der Waals clash loss is enabled with a small weight (0.01), ramped during training.

E TRAINING DATA, SPLITS, AND PROCEDURE

Training data. LoRA fine-tuning for Physics-Aligned Decoding is performed on molecular dynamics (MD) trajectory frames derived from the CATH domain dataset (Orengo et al., 1997). Training frames are extracted from equilibrium MD simulations of CATH domains, providing physically realistic intermediate conformations beyond static experimental structures. Only backbone atoms (N, C_α, C, O) and pseudo-C_β atoms are used during training, consistent with the representation described in the main text.

Data splits and coverage. Trajectories are split at the protein level to avoid information leakage across train and evaluation sets. Domains used for PAD fine-tuning are disjoint from those used in all static reconstruction, trajectory-level, and autoregressive rollout evaluations reported in Section 4. For downstream MD-based evaluations, additional out-of-distribution trajectories are drawn from the ATLAS dataset (Vander Meersche et al., 2023), which is never used during fine-tuning.

Frame sampling. From each training trajectory, frames are subsampled uniformly in time to reduce temporal correlation and to expose the decoder to a diverse range of physically valid conformations. No frame-level augmentation or artificial perturbation is applied. Each training batch consists of independently sampled frames rather than contiguous trajectory segments.

Training procedure. Fine-tuning is performed for a fixed number of epochs over the training frames using the AdamW optimizer. The pretrained FoldToken encoder, decoder, and codebook remain frozen except for the LoRA adapters described above. All reconstruction losses used in the base tokenizer are retained unchanged, with additional physics-aligned terms applied only at the decoding stage. Gradients are propagated through the straight-through estimator to the encoder, allowing token assignment boundaries to adapt without regenerating the discrete codebook.

Computational cost. Because PAD is implemented via parameter-efficient fine-tuning, training requires substantially less computation than full retraining. Inference-time decoding incurs only a small constant-factor overhead from evaluating the additional geometry-aware loss terms, with no change to model architecture or codebook size.

F SCIENCE OF DL IMPROVEMENT CHALLENGE SUBMISSION

F.1 WHAT MODEL ARE YOU TARGETING?

Provide a summary of the problem the deep net model is designed to solve. Good summaries should outline the state of the literature, provide an overview that domain experts would consider reasonable, and cite relevant sources.

We target deep discrete representation models trained with reconstruction-aligned objectives, with a particular focus on vector-quantized autoencoders (VQ-VAEs) used for protein structure tokenization. Such models map high-dimensional geometric inputs to a finite set of discrete latent states, enabling downstream generative, conditional, and autoregressive modeling.

In the protein domain, discrete structure tokenizers such as FoldToken and ProToken have been proposed to serve as compact interfaces between three-dimensional molecular structure and sequence-based models, including transformers and large language models. These models are typically trained using reconstruction-aligned losses (e.g., coordinate-level L_2 loss or frame-aligned variants such as FAPE) that emphasize global geometric fidelity.

While effective for static reconstruction benchmarks, these models are increasingly reused as latent state spaces for downstream tasks, including conditional editing, autoregressive rollouts, and latent dynamics modeling. In these regimes, discrete tokens no longer function merely as compression artifacts, but as semantic states over which learning and inference are performed. Our work targets this class of models and interrogates whether reconstruction-aligned training objectives produce representations whose discrete states are semantically valid when reused beyond reconstruction.

F.2 HOW DO YOUR RESULTS CONTRIBUTE—OR COULD POTENTIALLY CONTRIBUTE—TO UNDERSTANDING THESE MODELS?

What aspects of the models become better understood thanks to your work?

Our results clarify how the geometry of the training objective shapes the semantic content of discrete representations. We show that reconstruction-aligned objectives can admit stable optima in which discrete latent states encode locally invalid configurations, even when global reconstruction error is low. These violations are deterministic, reproducible, and persist under reuse, indicating that they arise from objective-level properties rather than stochastic decoding noise or insufficient model capacity.

By isolating objective effects through a controlled intervention that modifies only the decoding loss while keeping architecture and codebook fixed, we demonstrate that unphysical behavior is not an inherent limitation of discrete autoencoders, but a consequence of symmetric reconstruction losses that fail to encode asymmetric feasibility constraints. This provides empirical evidence that representation failure modes can arise from loss misspecification even when models appear successful under standard evaluation metrics.

More broadly, our work highlights that when learned representations are reused as state spaces, reconstruction fidelity alone is an insufficient proxy for representational validity. Understanding deep models in such settings requires analyzing not only architecture and capacity, but also how objective geometry constrains which regions of representation space are accessible.

F.3 HOW DO YOU EXPECT YOUR SUBMISSION TO INFLUENCE FUTURE WORK?

Propose ways in which your insights, findings, or methodologies could shape subsequent research directions, model design choices, or scientific applications.

We expect this work to encourage more systematic investigation of objective-induced failure modes in deep representation learning, particularly in settings where representations are reused as intermediate states. Our findings suggest that future model design should explicitly consider whether training objectives encode the feasibility constraints of the underlying data manifold, rather than relying solely on symmetric reconstruction losses.

Methodologically, we hope this work promotes the use of hypothesis-driven, controlled interventions to disentangle the roles of architecture, capacity, and objective geometry in learned representations. Substantively, the Physics-Aligned Decoding framework illustrates that minimal objective-level modifications can reshape representation semantics without architectural changes, offering a practical template for diagnosing and correcting misaligned representations.

While our experiments focus on protein structure, the underlying insights apply to any domain involving constrained state spaces, such as robotics, physical simulation, planning, or discrete latent dynamics. We anticipate that future work will extend these ideas to other modalities and develop principled objective designs that ensure learned representations remain semantically valid under reuse.