CAN MOLECULAR-DYNAMICS SUPERVISION HELP PREDICT PROTEIN FUNCTIONS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein functions depend on conformational ensembles, but incorporating molecular-dynamics (MD) information into sequence-based prediction requires supervision that can be learned without trajectories at inference. We introduce ESMLace, an ESM-2 35M encoder trained to predict compressed MD targets: moments, wavelets, cosine coefficients, or discrete VQ-VAE tokens from 96-frame windows. We transfer the frozen encoders to mutation-effect prediction and immune-recognition benchmarks, comparing matched mean-pool and token-level readouts. Across ten ProteinGym assays, VQ supervision changes from -0.008 Spearman relative to Base ESM under mean pooling to $+0.166$ under a two-block token readout, giving a compression-by-readout interaction of $+0.174$ (95% assay bootstrap $[0.083, 0.271]$, family-wise $p = .031$). VQ supervision also improves BATCAVE global peptide-activity ranking from 0.495 to 0.557 Spearman, while antibody and TCR–pMHC results vary by benchmark. Replica and matched-probe analyses identify reproducible MD signals and increased target decodability, while waveform reconstructions distinguish temporal-shape recovery from amplitude recovery. MD supervision can therefore improve sequence-based function prediction, with its benefit determined jointly by temporal representation, readout capacity, and task-label alignment.

1 INTRODUCTION

Protein function depends on conformational populations, local flexibility, and transitions among functional states in addition to a static fold (Henzler-Wildman & Kern, 2007). This is especially tangible in immune recognition. T cell receptors (TCRs), peptides, and major histocompatibility complexes form dynamic interfaces whose loop and peptide motions can change recognition despite similar crystallographic snapshots (Buckle & Borg, 2018). Antibody complementarity-determining regions, particularly CDR-H3, likewise populate ensembles from which binding-competent conformations may be selected (Fernández-Quintero et al., 2019). Mutations can redistribute such populations without changing the broad fold. A representation that captures only sequence regularities or one structure may therefore omit information relevant to TCR–pMHC recognition, antibody affinity, and mutation-effect landscapes.

Molecular dynamics (MD) offers atomistic, time-ordered observations of these motions, but generating and storing long trajectories for every downstream sequence is impractical (Karplus & McCammon, 2002). NMR, HDX-MS, and cryo-EM provide complementary experimental views but differ in resolution, timescale, and throughput (Palmer, 2001; Masson et al., 2019; Frank, 2017). Recent generative models amortize part of this cost by sampling protein ensembles (Jing et al., 2024; Lewis et al., 2025), while SeqDance and ESMDance supervise language models with biophysical summaries (Hou et al., 2026). These advances motivate transferring MD supervision into sequence representations. Its functional benefit still depends on which trajectory targets the encoder learns and how a downstream readout uses its residue tokens.

We ask *can MD supervision help predict protein functions?* We address this question with ESMLace, an ESM-2 sequence encoder trained to predict compressed MD descriptors at global, residue, and residue-pair positions. Four target families—moments, wavelets, cosine coefficients, and discrete VQ-VAE tokens—provide different summaries of the same 96-frame windows. At transfer time, the MD heads are removed and the frozen encoder supplies residue tokens to a trainable downstream

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

model. This design isolates the choice of MD supervision under a shared backbone and requires no trajectory at inference.

The main evaluation tests mutation-effect prediction on ProteinGym and transfer to antibody affinity and TCR–pMHC benchmarks. Readout capacity is a central experimental factor: direct mean pooling and a two-block token readout expose different information from the same frozen encoder. ProteinGym shows a VQ-by-readout interaction, while the immune tasks reveal gains and losses under their respective split and label definitions. Independent MD replicas, static and sequence controls, fresh matched probes, and waveform reconstructions connect these transfer outcomes to the information retained by the targets and learned embeddings.

Our contributions are:

1. We train ESMLace with four compressed MD target families under a shared ESM-2 35M framework, and compare the resulting frozen encoders without MD at downstream inference (Section 3).
2. We identify a statistically supported VQ-by-readout interaction on ProteinGym and evaluate transfer to antibody and TCR–pMHC tasks, including positive BATCAVE global-rank transfer and benchmark-dependent gains and losses (Section 4).
3. We support the transfer evaluation with replica analysis on 363 DynaRepo complexes, matched encoder probes, and waveform audits that characterize retained information and reconstruction limits (Section 5 and Appendix D.1).

2 Related Work

Protein dynamics and learned ensembles. MD has long connected molecular structure to motion and energetics (Karplus & McCammon, 2002), but conventional simulations remain expensive at broad sequence scale. AlphaFlow adapts structure predictors using flow matching, and BioEmu learns sequence-conditioned equilibrium ensembles from MD, structures, and stability data (Jing et al., 2024; Lewis et al., 2025). These models generate conformations. Our target is different: a compact sequence encoder whose internal representation is supervised by time-ordered MD descriptors, followed by an audit of what information the representation retains.

Dynamics-supervised protein encoders. SeqDance and ESMDance predict mean, dispersion, interval, and mode-like properties derived from MD and normal-mode analysis; ESMDance adapts ESM-2 and reports zero-shot mutation-effect gains (Hou et al., 2026). DPLM instead aligns sequence embeddings with trajectory embeddings learned from residue-contact-map videos by contrastive learning and evaluates mutation effects, stability, and functional clustering (Jiang et al., 2026). These studies establish two complementary routes to dynamics-aware sequence representations: predicting predefined biophysical targets or aligning sequence and trajectory latent spaces. Our study focuses on the interaction between temporal targets and downstream readouts. Under one ESM-2 35M backbone and one 96-frame protocol, we compare moment, frequency, multiresolution, and discrete-motif supervision to determine how the resulting representations support protein-function prediction.

Representation comparison. CKA measures similarity between representation geometries and is invariant to isotropic scaling and orthogonal transformations (Kornblith et al., 2019). We use linear CKA across independent replicas, but first remove tested sequence and same-replica static components. Cross-replica retrieval provides a complementary instance-level measure: CKA can be high when global geometry agrees even if individual systems are not uniquely identifiable.

3 Dynamics-Supervised Protein Representations

3.1 Fixed-window MD targets

For protein sequence s and a 96-frame trajectory window $X \in \mathbb{R}^{96\times L\times d}$, we compute global, residue, and residue-pair descriptor waveforms. The complete target set contains 12 descriptors;

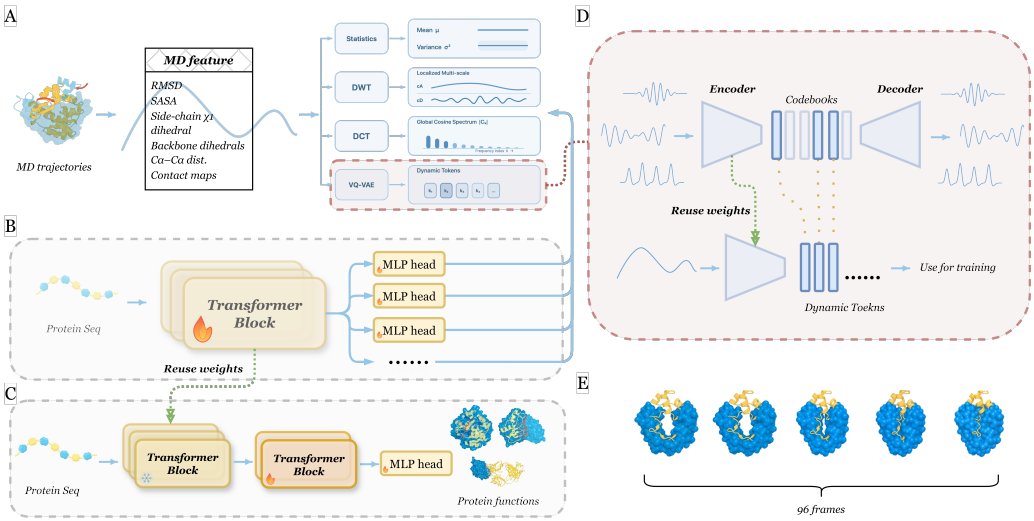

Figure 1: MD supervision is transferred through the sequence encoder. Waveforms are compressed by Stats, DWT, DCT, or a pretrained VQ-VAE, using either the complete descriptor set (All) or distances and dihedrals (D&D). The encoder and MD heads are trained jointly; MD heads are then removed, and the frozen encoder feeds a trainable downstream readout.

the geometry-focused D&D subset retains Cα distances and backbone dihedrals. We compare four temporal representations:

- **Stats**: mean and population variance over the window;
- **DWT**: six low-resolution Haar wavelet coefficients (Mallat, 1989);
- **DCT**: six low-frequency cosine coefficients (Ahmed et al., 1974);
- **VQ-VAE**: six categorical tokens from a frozen vector-quantized autoencoder trained to reconstruct normalized waveforms (van den Oord et al., 2017).

DWT, DCT, and VQ-VAE therefore use the same six-target temporal budget per channel; Stats uses two moments. Continuous targets use ordinary elementwise mean-squared error and VQ tokens use cross-entropy. Appendix A.2 gives descriptor shapes and checkpoint rules.

3.2 Sequence encoder and transfer

Every arm starts from the same ESM-2 35M backbone (Lin et al., 2023). We call the resulting model family ESMLace: an ESM-2 sequence encoder trained to predict MD-derived signals, rather than a trajectory generator. During MD supervision, the encoder and feature-specific prediction heads are optimized from sequence alone. Downstream experiments discard the MD heads and freeze the encoder. We compare Base ESM-2 against two supervision variants for each compression: All, using the complete descriptor set, and D&D, using distances and dihedrals. Checkpoint-selection details are given in Appendix A.2. No downstream test label selects an MD checkpoint.

For single-chain tasks, the downstream model either directly mean-pools residue tokens or applies two Transformer blocks followed by masked mean pooling. For multi-chain immune tasks, chains are encoded separately, concatenated with learned chain-identity and token-position embeddings, mixed by the same two-block readout, and masked-mean pooled. Thus MD is not supplied at inference time: any difference arises from how MD supervision changed the sequence representation.

4 When Does MD Supervision Help Function Prediction?

4.1 Evaluation protocol

We evaluate ten ProteinGym substitution assays spanning activity, binding, expression, organismal fitness, and stability (Notin et al., 2023); AbRank antibody–antigen affinity ranking (Liu et al., 2025); BATCAVE TCR peptide-activity ranking (Banerjee et al., 2025); ITRAP-derived TCR–pMHC labels (Montemurro et al., 2023); and the original TABR-BERT S1–S4 tests (Zhang et al., 2024). ProteinGym uses mutation-identity splits within each assay, so it measures within-landscape, assay-local mutation generalization rather than transfer to unseen proteins, families, or assay properties. All downstream arms use three matched random seeds and validation-selected downstream epochs. ProteinGym and BATCAVE use pairwise rank loss (Burges et al., 2005); AbRank uses exact-label MSE with a censor-aware margin term; and binary interaction tasks use weighted binary cross-entropy. Full splits, counts, optimization settings, readout definitions, and complete reported arm matrices are in Appendices B–C.

The held-out axis differs across benchmarks and bounds each conclusion. AbRank holds out antibody–antigen cluster combinations but not necessarily either component; BATCAVE holds out a source fold while retaining all 49 scan groups across partitions; ITRAP is peptide-disjoint but largely reuses TCR and MHC components; and the TABR-BERT sets mix peptide-novel and non-zero-shot source protocols. We therefore interpret each score under its own split and do not pool evidence across these generalization regimes.

4.2 ProteinGym exposes a representation–readout interaction

With direct mean pooling, VQ-VAE D&D is slightly worse than Base ESM (-0.008 macro assay Spearman). With the matched two-block token readout, VQ-VAE D&D reaches 0.558 versus 0.392 for Base ESM, a $+0.166$ difference (Fig. 2A–B). Nine of ten assays have a positive difference-in-differences between readouts (Table 4). The mean interaction is $+0.174$, and an assay bootstrap gives a 95% interval of $[0.083, 0.271]$. An exact shared-assay max-$|t|$ correction over the originally examined MD arms gives family-wise $p = .031$ for the interaction (Appendix A.2). The separately corrected two-block VQ gain has $p = .180$, placing the strongest statistical evidence on the representation–readout interaction.

This pattern suggests that MD supervision can create spatially distributed residue-level organization that direct global averaging obscures. The effect is not unique to VQ-VAE: continuous Stats and DWT encoders also improve materially when token mixing is introduced (Table 5). For VQ-VAE specifically, motif classification may reduce sensitivity to exact amplitude, but discreteness is entangled with per-waveform min–max normalization and a codebook pretrained on a separate BioEmu-derived corpus. The evidence therefore supports a broad representation–readout compatibility effect; it does not isolate discretization as its cause.

4.3 Transfer is not universal

Immune tasks expose benchmark-dependent transfer (Fig. 3). On BATCAVE, global RankNet training raises test global Spearman from 0.495 ± 0.010 for Base ESM to 0.557 ± 0.003 for VQ-VAE D&D, a mean difference of $+0.061$. Stats D&D reaches 0.532 ± 0.013, whereas DCT D&D reaches 0.505 ± 0.012. BATCAVE uses a held-out source fold and permits ranking pairs across scan groups; it is therefore a global ranking benchmark over source-normalized activity, not an unseen-TCR or component-OOD test. VQ also changes AbRank Spearman from -0.005 ± 0.002 to 0.122 ± 0.089 ($+0.127$ in the three-seed means); the large VQ seed standard deviation makes this result less stable than its mean difference alone suggests. Its result on the original TABR-BERT S1–S4 macro ROC-AUC is lower than Base ESM, whereas DCT has the best mean PR-AUC. ITRAP shows a large aggregate VQ gain ($+0.454$ ROC-AUC), but the Base ESM result is itself sub-random and variable (0.374 ± 0.062), 99.57% of test rows contain a training-seen TCR, every test MHC occurs in training, and class balance differs strongly by peptide. We retain the contrast as a benchmark association rather than attributing its full magnitude to general interaction recognition.

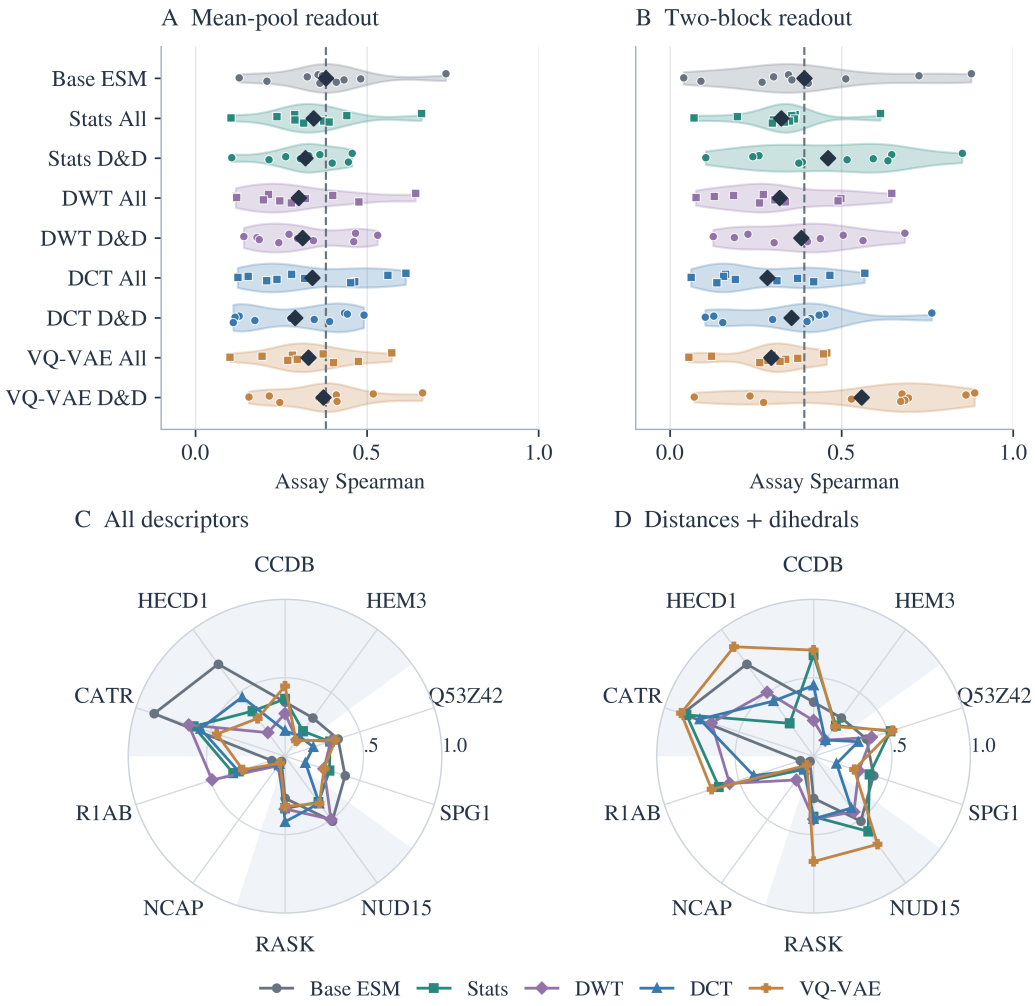

Figure 2: MD-transfer performance varies across encoders, readouts, and assays. **A–B**: absolute Spearman scores across ten assays, including Base ESM; points average three seeds within each assay, diamonds are assay means, and violins show descriptive densities. Dashed lines mark the corresponding Base ESM mean. **C–D**: two-block Spearman for Base ESM and All or D&D encoders on all ten assays. Adjacent pairs are CCDB/HEM3 (activity), Q53Z42/SPG1 (binding), NUD15/RASK (expression), NCAP/R1AB (organismal fitness), and CATR/HECD1 (stability); pairing does not combine assay values. Both radars use the same 0–1 scale; area is not an aggregate score.

5 Supporting Analysis of MD Supervision

5.1 DynaRepo protocol

DynaRepo provides all-atom trajectories for macromolecular complexes (Mokhtari et al., 2026). We use 363 protein–protein complexes with three independent 500-ns replicas. To hold the number of observations fixed, every condition contains 96 frames. Frame intervals of $0.1, 0.2, 0.5, 1, 2$, and 4 ns yield spans of $9.5, 19, 47.5, 95, 190$, and 380 ns.

Let $Z_{r,m,t}$ denote a system-level representation for replica r, method m, and span t. We subtract the representation obtained by repeating the first structure of the same replica, isolating the measured temporal component from a static control. We then fit a linear map from frozen ESM sequence embeddings to this dynamic representation using five-fold cross-fitting and analyze the residuals. Folds are assigned over complexes with a fixed shuffled K-fold split; they are not constructed from

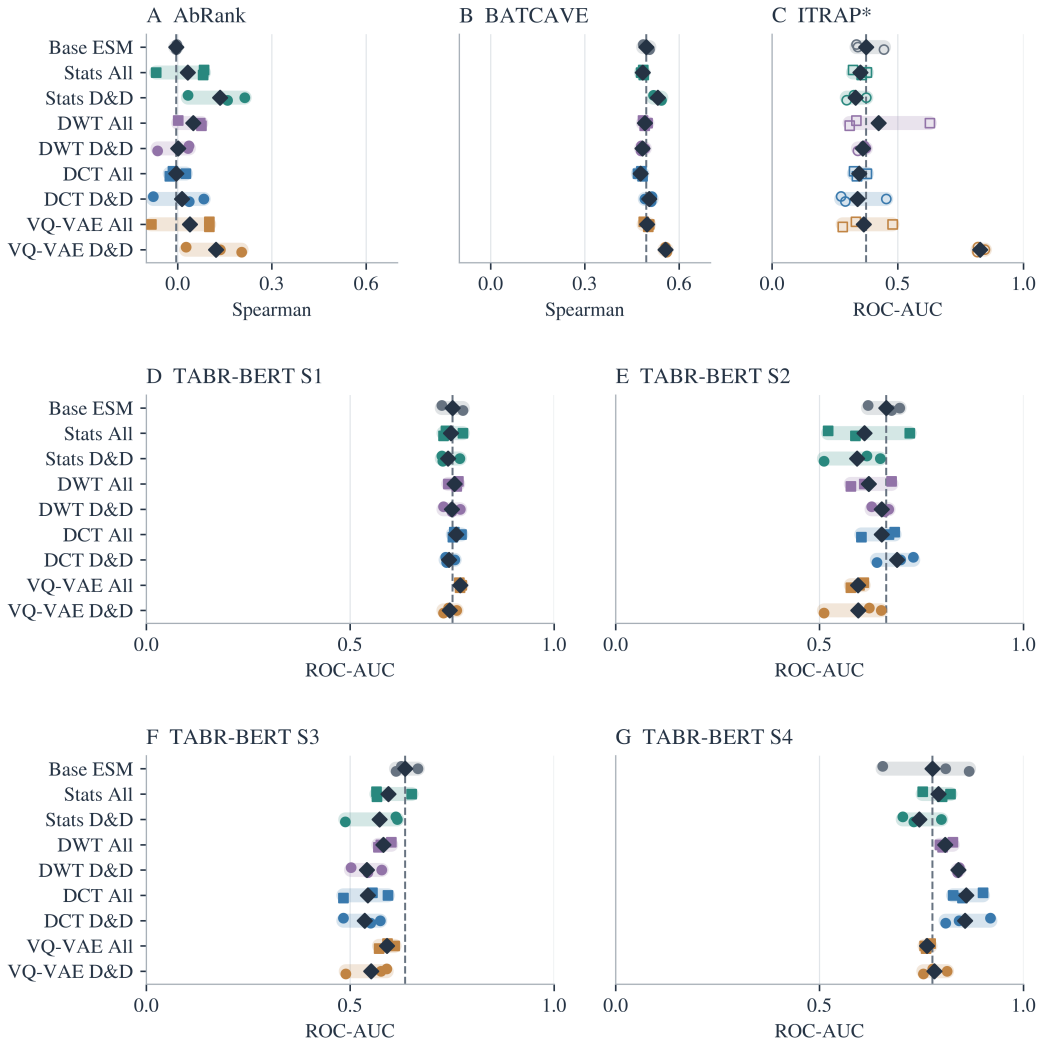

Figure 3: Absolute performance varies across immune tasks and MD encoders. Every panel includes Base ESM and eight MD encoders; points are three seeds, dark diamonds are means, and light bands show the observed seed range, not a confidence interval or density. Dashed lines mark each panel's Base ESM mean. **A–C**: AbRank and BATCAVE Spearman, and ITRAP ROC-AUC. **D–G**: ROC-AUC on the original TABR-BERT S1, S2, S3, and S4 tests separately, without averaging across sets. Open ITRAP markers retain the component-recurrence caveat. All ROC-AUC panels share a 0–1 scale; the Spearman panels use a separate common scale.

sequence-family clusters. The analysis therefore controls for out-of-fold linear sequence prediction at the complex level, without claiming homology-disjoint evaluation or removal of nonlinear sequence and family effects.

For each pair of replicas, we report linear CKA, Spearman correlation between representation dissimilarity matrices, and top-k nearest-neighbor retrieval of the same complex. CKA for centered Gram matrices K and L is

$$\mathrm{CKA}(K, L) = \frac{\langle K, L \rangle_F}{\|K\|_F \|L\|_F}.$$

At the longest span, 1,000 complex-label permutations test residual CKA; five compression/raw hypotheses are Bonferroni corrected. We additionally measure amplitude entropy, spectral entropy, lag-1 autocorrelation, VQ code entropy, and effective rank. Repeated-static and time-shuffled controls distinguish variability from ordered temporal dependence.

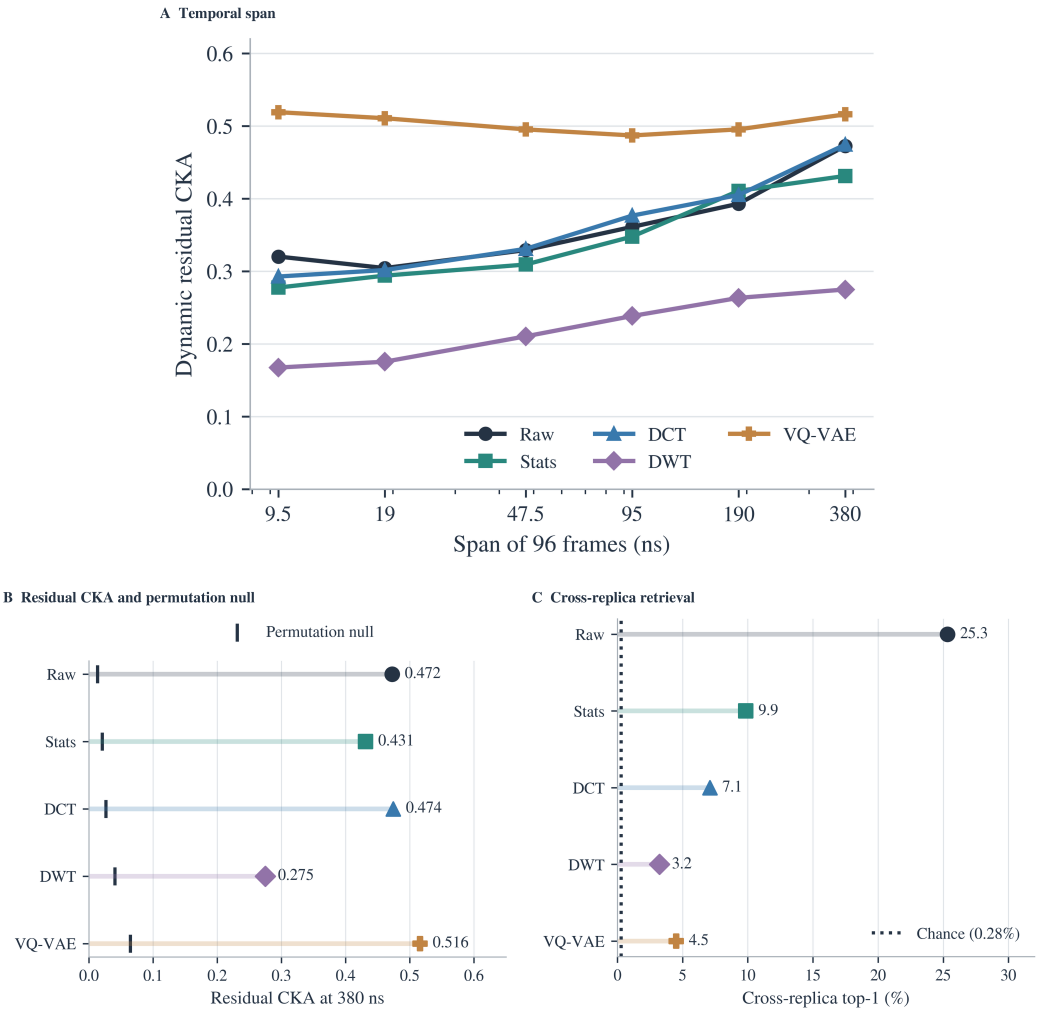

Figure 4: DynaRepo information analysis across three independent replicas. **A**: dynamic residual CKA as the interval between the 96 frames increases. **B**: observed residual CKA and permutation null (black tick) at 380 ns; all five corrected permutation tests give $p = .005$. **C**: cross-replica top-1 retrieval after static subtraction; dashed line is chance. Global alignment and instance retrieval need not rank compressions identically.

5.2 Independent replicas share residual dynamic structure

Figure 4A shows that residual alignment depends on both timescale and representation. At 380 ns, residual CKA is 0.472 for raw dynamics, 0.431 for Stats, 0.474 for DCT, 0.275 for DWT, and 0.516 for VQ-VAE, compared with permutation means of 0.013–0.065. Every observed value has permutation $p = 0.000999$ and Bonferroni-corrected $p = 0.004995$ (Fig. 4B). Because static and tested linear sequence components were removed, these results support reproducible dynamic information rather than identity through those controls.

The representations retain different kinds of information. Raw dynamics retrieves the same complex across replicas for 25.3% of systems at 380 ns, versus 9.9% for Stats, 7.1% for DCT, 3.2% for DWT, 4.5% for VQ-VAE, and 0.28% chance (Fig. 4C). VQ-VAE has the highest global CKA but lower instance retrieval, consistent with a shared motif geometry that is reproducible across proteins without uniquely fingerprinting each complex. DCT and Stats improve with longer spans; VQ alignment is already high at 9.5 ns and changes little, suggesting early saturation of its discretized motif vocabulary.

5.3 Entropy is informative only with temporal controls

Amplitude and spectral entropy change modestly across frame intervals and remain high after time shuffling, whereas shuffling collapses lag-1 autocorrelation to approximately zero. Repeating a static frame makes all temporal entropies and autocorrelation zero (Appendix Fig. 5). Hence "more entropy" is not by itself evidence of meaningful dynamics: unordered noise can have high entropy. Replica agreement, ordered controls, and retrieval supply the necessary complementary evidence.

Matched encoder and waveform checks. On a separate strict MD-target cohort, identical fresh probes on frozen encoders give an uncorrected paired 95% interval above zero for 12 of 16 encoder–feature comparisons, below zero for 3, and crossing zero for 1. DCT/DWT give the largest distance-target improvements, while VQ-VAE D&D gains are smaller but positive. Target decodability and downstream performance therefore favor different representations. Post-hoc waveform examples distinguish recovered temporal shape from attenuated amplitude. Appendix D.1 reports the full protocols, intervals, selection criteria, and results.

Additional VQ-VAE examples show why waveform shape and error should be examined separately: in four of six post-hoc audited cases, VQ has a higher prediction-to-raw correlation but a larger normalized RMSE than at least one continuous comparator. Native-scale and shape-normalized views of the two highest-correlation cases illustrate this distinction (Appendix Fig. 7); the appendix documents their selection and the full audit.

6 Discussion and Limitations

The downstream experiments show that the value of MD supervision depends on the readout and task. On ProteinGym, changing the readout reverses the VQ-minus-Base comparison, with a statistically supported interaction. BATCAVE shows improved global ranking, while the original TABR-BERT tests show a VQ loss. Thus the training target alone does not determine functional utility. Supporting replica and probe analyses show reproducible target structure and changes in decodability, but their representation rankings differ from downstream rankings. Waveform examples further distinguish recovery of temporal shape from recovery of amplitude.

Temporal detail beyond reconstruction error. Figure 6 suggests a further distinction between the compression families. In this example, decoding the true VQ-VAE codes retains finer, faster-varying waveform detail than the six low-frequency DCT coefficients or coarse DWT representation, despite lower reconstruction correlations in some comparisons. The sequence-predicted VQ waveform also shows local variation, although its amplitude is strongly attenuated. Thus a target can retain relatively high-frequency temporal structure without minimizing overall waveform error. We hypothesize that such local temporal motifs provide useful supervision for residue representations and contribute to VQ-VAE's gains with a token-mixing readout on ProteinGym and BATCAVE. The relevant frequencies are relative to the stored-frame sampling interval. This visual pattern motivates the mechanism; it does not yet quantify faithful high-frequency recovery or its contribution to downstream performance.

Label construction and the sequence baseline. Negative-label construction may partly explain the competitive sequence-only baseline and smaller MD benefit on TABR-BERT. Both its training protocol and released S1–S4 tests construct negatives by pairing pMHCs with randomly sampled TCRs from a healthy-donor repertoire (Zhang et al., 2024). Repertoire-source differences can introduce sequence-level cues, a documented concern in TCR-specificity benchmarks (Dens et al., 2023). We hypothesize that these cues help Base ESM separate the labels and reduce the incremental value of dynamics-oriented features. BATCAVE instead compiles measured peptide-mutant activity, including weak and non-activating variants (Banerjee et al., 2025); its improvement therefore provides a more direct test against experimentally measured functional variation. Our ITRAP benchmark also derives from single-cell pMHC measurements and filtering (Montemurro et al., 2023). In our supplied data construction, negatives are other pMHCs detected in the same cells, excluding known positive pairs (Appendix B). These assay-derived, nonassigned pairs differ from background-repertoire mismatches, although they are not independently validated nonbinders. This

label-construction hypothesis explains a possible benchmark dependence; a high baseline alone does not establish saturation or identify its cause.

Model scale and computational budget. Our computational budget restricts MD-supervised training to the relatively small ESM-2 35M backbone. Learning diverse residue- and pair-level dynamic targets within this capacity may limit how much MD information can be integrated alongside existing sequence features. A larger encoder could capture richer dependencies across residues and temporal motifs, retain more of the complementary sequence and dynamics information, and improve transfer and generalization. Model scaling is therefore a promising extension of ESMLace, rather than a change to its supervision principle. Whether these gains materialize remains an empirical question, since larger capacity also increases data and optimization requirements.

Scope of the evaluation. The replica audit removes the tested static and linear sequence components; nonlinear sequence effects and family overlap remain. Fixed 96-frame windows couple observation span to sampling interval. VQ-VAE combines discretization, per-waveform normalization, and an externally pretrained codebook, so these contributions remain coupled. Most transfer results use one MD-pretraining seed and three downstream seeds. ProteinGym measures within-assay mutation generalization, while immune tasks retain component recurrence and benchmark-specific negative labels.

The analysis also clarifies why more information may reduce downstream accuracy. An auxiliary signal can emphasize variation irrelevant to a label, displace useful evolutionary features in a finite-capacity encoder, or remain distributed across tokens that a readout averages away. Negative transfer is therefore compatible with genuine dynamic information. Future work should vary one axis at a time: matched optimization budgets across compressions, multiple dynamics-pretraining seeds, one-block/no-position readouts, experimentally anchored kinetic observables, and family-disjoint downstream tests.

7 Conclusion

MD supervision can improve protein-function prediction from sequence. ESMLace achieves this by training ESM-2 on compressed trajectory targets and transferring the resulting residue representations without MD at inference. The ProteinGym readout interaction and BATCAVE ranking gains show the importance of combining an informative target with a readout that can use it. Mixed transfer across immune benchmarks further identifies label construction and task alignment as key design factors. Replica, probe, and waveform analyses connect these outcomes to reproducible dynamic signals, target decodability, and temporal-shape recovery. Together, the results make temporal representation, token-level readout, and benchmark design joint choices when building dynamics-supervised protein models.

Reproducibility Statement

Appendix A specifies preprocessing, splits, targets, losses, checkpoint selection, statistical tests, and seed identities. Every main-text number is traced to the corresponding dataset-level evaluation record, and figures are generated directly from these results. The supplementary code includes deterministic figure generation and a claim-to-evidence map.

Ethics Statement

This study reuses public molecular simulation and protein benchmark data and does not collect new human-subject data. TCR benchmarks derive from previously released immune-repertoire studies; we use processed sequence/label tables only and make no attempt to re-identify donors. Predictive scores should not be interpreted as clinical decisions or experimental binding validation.

AI USE STATEMENT

Generative AI tools were used to provide feedback on experimental design, assist with analysis and plotting code, interpret validated outputs, identify and verify literature, and draft and edit manuscript text and figures. They were not used to generate synthetic experimental observations or theoretical proofs. Authors reviewed the AI-assisted work, tested executable code, traced numerical claims to saved artifacts, and manually checked cited sources. We take responsibility for the final content, claims, and artifacts in this work.

REFERENCES

Nasir Ahmed, T. Natarajan, and K. R. Rao. Discrete cosine transform. *IEEE Transactions on Computers*, C-23(1):90–93, 1974.

Amitava Banerjee, David J. Pattinson, Cornelia L. Wincek, Paul Bunk, Armend Axhemi, Sarah R. Chapin, Saket Navlakha, and Hannah V. Meyer. T cell receptor cross-reactivity prediction improved by a comprehensive mutational scan database. *Cell Systems*, 16(8), 2025.

Ashley M. Buckle and Natalie A. Borg. Integrating experiment and theory to understand TCR–pMHC dynamics. *Frontiers in Immunology*, 9:2898, 2018. doi: 10.3389/fimmu.2018.02898.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 89–96, 2005.

Ceder Dens, Kris Laukens, Wout Bittremieux, and Pieter Meysman. The pitfalls of negative data bias for the T-cell epitope specificity challenge. *Nature Machine Intelligence*, 5(10):1060–1062, 2023. doi: 10.1038/s42256-023-00727-0.

Monica L. Fernández-Quintero, Johannes Kraml, Guy Georges, and Klaus R. Liedl. CDR-H3 loop ensemble in solution—conformational selection upon antibody binding. *mAbs*, 11(6):1077–1088, 2019. doi: 10.1080/19420862.2019.1618676.

Joachim Frank. Advances in the field of single-particle cryo-electron microscopy over the last decade. *Nature Protocols*, 12:209–212, 2017.

GetContacts developers. GetContacts: Analysis of interatomic contacts from molecular-dynamics trajectories. `https://getcontacts.github.io/`, 2026. Open-source software, accessed 2026.

Katherine Henzler-Wildman and Dorothee Kern. Dynamic personalities of proteins. *Nature*, 450 (7172):964–972, 2007. doi: 10.1038/nature06522.

Chao Hou, Haiqing Zhao, and Yufeng Shen. Protein language models trained on biophysical dynamics inform mutation effects. *Proceedings of the National Academy of Sciences*, 123(4): e2530466123, 2026. doi: 10.1073/pnas.2530466123.

Yuexu Jiang, Duolin Wang, Ibrahim A. Imam, Dong Xu, and Qing Shao. DPLM: Dynamics-aware protein language model via contrastive learning between sequence and molecular dynamics simulation trajectory. *bioRxiv*, 2026. doi: 10.64898/2026.04.29.721692. Preprint.

Bowen Jing, Bonnie Berger, and Tommi Jaakkola. AlphaFold meets flow matching for generating protein ensembles. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 22277–22303, 2024.

Martin Karplus and J. Andrew McCammon. Molecular dynamics simulations of biomolecules. *Nature Structural Biology*, 9:646–652, 2002. doi: 10.1038/nsb0902-646.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 3519–3529, 2019.

S. Y. Lee, H. S. Cho, J. G. Pelton, D. Yan, E. A. Berry, and D. E. Wemmer. Crystal structure of activated CheY. comparison with other activated receiver domains. *Journal of Biological Chemistry*, 276:16425–16431, 2001. doi: 10.1074/jbc.M101002200.

Sarah Lewis, Tim Hempel, José Jiménez-Luna, Michael Gastegger, Yu Xie, Andrew Y. K. Foong, et al. Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science*, 389(6761):eadv9817, 2025. doi: 10.1126/science.adv9817.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023. doi: 10.1126/science.ade2574.

Chunan Liu, Aurelien Pelissier, Yanjun Shao, Lilian Denzler, Andrew C. R. Martin, Brooks Paige, and María Rodríguez Martínez. AbRank: A benchmark dataset and metric-learning framework for antibody–antigen affinity ranking. *arXiv preprint arXiv:2506.17857*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Stéphane G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989.

Glenn R. Masson, John E. Burke, Natalie G. Ahn, Ganesh S. Anand, Christoph Borchers, Séverine Brier, et al. Recommendations for performing, interpreting and reporting hydrogen deuterium exchange mass spectrometry experiments. *Nature Methods*, 16:595–602, 2019.

Robert T. McGibbon, Kyle A. Beauchamp, Matthew P. Harrigan, Christoph Klein, Jason M. Swails, Carlos X. Hernández, et al. MDTraj: A modern open library for the analysis of molecular dynamics trajectories. *Biophysical Journal*, 109:1528–1532, 2015.

Omid Mokhtari, Emmanuelle Bignon, Hamed Khakzad, and Yasaman Karami. DynaRepo: The repository of macromolecular conformational dynamics. *Nucleic Acids Research*, 54(D1):D393–D401, 2026. doi: 10.1093/nar/gkaf1130.

Alessandro Montemurro, Helle Rus Povlsen, Leon Eyrich Jessen, and Morten Nielsen. Benchmarking data-driven filtering for denoising of TCR–pMHC single-cell data. *Scientific Reports*, 13: 16147, 2023. doi: 10.1038/s41598-023-43048-3.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Hansen Spinner, et al. ProteinGym: Large-scale benchmarks for protein fitness prediction and design. In *Advances in Neural Information Processing Systems*, volume 36, pp. 64331–64379, 2023.

Arthur G. Palmer. NMR probes of molecular dynamics: Overview and comparison with other techniques. *Annual Review of Biophysics and Biomolecular Structure*, 30:129–155, 2001.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

K. Volz and P. Matsumura. Crystal structure of Escherichia coli CheY refined at 1.7-å resolution. *Journal of Biological Chemistry*, 266:15511–15519, 1991. URL `https://www.rcsb.org/structure/3CHY`.

Jiawei Zhang, Wang Ma, and Hui Yao. Accurate TCR–pMHC interaction prediction using a BERT-based transfer learning method. *Briefings in Bioinformatics*, 25(1):bbad436, 2024. doi: 10.1093/bib/bbad436.

A Reproducibility Protocol

A.1 Scope and excluded experiments

The manuscript evaluates encoders pretrained to predict compressed MD targets and used without MD at downstream inference. We intentionally exclude (i) multi-task downstream models that predict MD as an auxiliary output, (ii) raw or spatiotemporally compressed MD supplied directly to an affinity model, (iii) MISATO/DynaRepo affinity training, and (iv) generative mediation. These are different interventions and would confound the representation question studied here.

A.2 MD corpus, preprocessing, and pretraining

Source corpus and split. The MD-pretraining corpus contains MDTraj and GetContacts arrays derived from BioEmu CATH-like folded-domain, mutant/all-atom, and octapeptide sources (Lewis et al., 2025; McGibbon et al., 2015; GetContacts developers, 2026). It contains 23,922 parsed protein entries, 23,812 unique amino-acid sequences, and 23,919,436 extracted frames. The deterministic split contains 67,479 protein–run groups: 53,983 training, 6,747 validation, and 6,749 test groups. A group is the Cartesian key of protein identifier and run index. Groups are shuffled with seed 42 and assigned in an 80/10/10 ratio. All descriptors and windows from one protein–run group remain in one partition. Distinct runs of the same protein can occur in different partitions, so the MD-validation split is run-disjoint but not necessarily sequence-disjoint or family-disjoint. Downstream test labels are not used in this split or in MD checkpoint selection.

Trajectory alignment and descriptors. Trajectories are unwrapped when required and superposed to frame zero before computing descriptors. MDTraj supplies RMSD, solvent-accessible surface area (SASA), backbone dihedrals, and side-chain χ_1 (McGibbon et al., 2015). GetContacts supplies seven binary residue-pair contact maps: hydrogen bond, salt bridge, π–cation, π–stacking, T-shaped stacking, van der Waals, and hydrophobic contacts (GetContacts developers, 2026). Cα–Cα distances are extracted from the corresponding stored frames in Å. If the MDTraj and contact arrays contain different numbers of frames, distances are truncated to the contact-map time axis; unavailable contact types are represented by zeros. The All target family contains RMSD, SASA, backbone dihedrals, χ_1, Cα distances, and the seven contact descriptors. The geometry-focused D&D family retains only backbone dihedrals and Cα distances.

Window construction and numerical handling. Every accepted array is divided into $T = 96$ stored-frame windows with a stride of 50 stored frames. Starts are $0, 50, 100, \ldots$; a final start at $n_{\mathrm{frames}} - 96$ is added only if it is not already present, retaining the trajectory tail without duplicating a window. A nonempty trajectory shorter than 96 frames is right-padded by repeating its final stored frame, and an empty array is represented by zeros. No temporal loss mask is applied to this repeated tail, so padded frames contribute to Stats and temporal transforms; this is a preprocessing limitation for short trajectories. Feature arrays are reshaped into independent scalar waveforms at global, residue, or residue-pair locations. Non-finite floating targets are replaced by zero after target construction and are not separately masked in the loss. Because source trajectories do not share one physical saving interval, the transform axis denotes stored-frame index rather than a universal frequency in ps^{-1}.

Temporal targets. For a scalar waveform x, Stats predicts $(\mathrm{nanmean}_t\, x, \mathrm{nanvar}_t\, x)$, with population variance ($\mathrm{ddof} = 0$), and uses MSE. DWT applies a level-4 Haar decomposition with symmetric boundary extension and retains the six level-4 approximation coefficients; discarded detail coefficients are not supervised (Mallat, 1989). DCT applies an orthonormal type-II transform and retains coefficient indices $0, \ldots, 5$ without min–max normalization (Ahmed et al., 1974). DWT and DCT use MSE on six real-valued coefficients per scalar waveform.

VQ-VAE independently min–max normalizes each finite scalar waveform to $[0, 1]$; non-finite entries are filled with the finite minimum, while constant and all-non-finite waveforms become zero. Its encoder contains four dilated one-dimensional convolutions with kernel sizes $7/5/5/5$ and dilations $1/2/4/8$, batch normalization, LeakyReLU after the first three convolutions, and adaptive average pooling to six positions. The resulting six 32-dimensional vectors are quantized against an EMA-updated 256-entry codebook (van den Oord et al., 2017). The decoder linearly upsamples the six

quantized vectors to 96 frames and applies the mirrored convolutional stack. It is trained with waveform-reconstruction MSE and the EMA vector-quantization objective. The frozen encoder and codebook then convert each MD waveform into six code indices. ESMLace predicts six independent 256-way categorical distributions per channel with cross-entropy averaged over all spatial locations and tokens; it does not reconstruct trajectories during ESM supervision.

Target shapes. Let N be residue count after feature extraction and C the descriptor-channel count. Input sequences are capped at 1,022 residues for ESM-2 35M. RMSD is global with stored shape $[T]$ and $C = 1$. SASA has shape $[T, N]$ and $C = 1$. Side-chain χ_1 has shape $[T, N, 2]$ for sine and cosine components. Backbone dihedrals have shape $[T, N, 4]$ for $\sin\phi$, $\cos\phi$, $\sin\psi$, and $\cos\psi$. Cα distance and each contact descriptor have shape $[T, N, N]$ and $C = 1$. At each location, Stats, DWT, DCT, and VQ-VAE contain $2C$, $6C$, $6C$, and $6C$ supervised targets, respectively. Global targets are predicted from the ESM beginning-of-sequence token; residue targets use the corresponding residue token. For a pair (i, j), the pair head forms the symmetric elementwise product $h_i \odot h_j$ and passes it through the same layer-normalized two-layer MLP used at every pair, evaluated in chunks to limit memory. The VQ classifier expands each discrete target into 256 logits, producing head widths of 1,536, 3,072, and 6,144 for $C = 1, 2$, and 4; these logits are classification outputs, not additional target features.

Optimization and checkpoints. All feature losses have unit weight. Within one sample, continuous heads use the elementwise MSE averaged over every output element, VQ heads use cross-entropy averaged over spatial locations and temporal tokens, and the resulting head losses are averaged across the available features. There is no additional target-variance or feature-range normalization, so objective values are comparable only within a target family. Every arm starts from the same ESM-2 35M weights, and the ESM encoder and feature heads are optimized jointly with AdamW at learning rate 10^{-4}, weight decay 10^{-2}, and gradient clipping at 1.0 (Loshchilov & Hutter, 2019). ReduceLROnPlateau uses factor 0.5 and patience 2. At epoch 3, batch sizes are 16 for All Stats/DWT/DCT, 4 for All VQ-VAE, 64 for D&D Stats/DWT/DCT, and 32 for D&D VQ-VAE. The All variant uses the epoch-3 checkpoint. The D&D runs continue for up to 20 epochs; MD-validation-loss minima select the reported D&D checkpoints at Stats epoch 18, DWT epoch 19, VQ-VAE epoch 14, and DCT epoch 19. The VQ-VAE itself is trained for three epochs on training groups with Adam at learning rate 10^{-3}, batch size 10,240, embedding width 32, hidden width 64, commitment cost 0.25, and EMA decay 0.99 (Kingma & Ba, 2015). Its epoch-3 encoder and codebook are frozen for the All and D&D ESM-supervision arms. No downstream validation or test label selects an MD checkpoint.

Reporting variants and multiplicity. The reported comparison contains eight MD-supervised encoders (four compressions, each with All and D&D) plus Base ESM. All and D&D differ in descriptor set, batch schedule, and checkpoint-selection rule; their contrast does not isolate the effect of descriptor choice alone. The ProteinGym family-wise tests retain the originally screened family of 12 MD arms, including four intermediate checkpoints omitted from the current presentation. Thus reducing the displayed set does not retrospectively relax multiplicity correction. Matched-probe descriptive counts refer only to the eight reported encoders and their 16 encoder–feature comparisons.

A.3 DynaRepo information analysis

We select 363 complexes with three complete 500-ns replicas. The native saved frame interval is 0.1 ns. Strides $\{1, 2, 5, 10, 20, 40\}$ preserve exactly 96 frames and yield the spans reported in the main text. Static controls repeat the first sampled structure 96 times. Time-shuffled controls permute frame order within each system and replica.

For each compression and replica, we first construct fixed-width fingerprints separately for intra-chain distances, interchain distances, and backbone dihedrals, then concatenate them. For Stats, DCT, and DWT, each temporal coefficient is reduced over the variable number of spatial locations by eight statistics: mean, standard deviation, mean absolute value, and the 0.10, 0.25, 0.50, 0.75, and 0.90 quantiles. Raw trajectories use the same eight-statistic reduction for temporal standard deviation, range, amplitude entropy, spectral entropy, and lag-1/lag-5 autocorrelation. VQ indices are never treated as ordered numbers: every one of the six token positions is mapped to its 32-dimensional codebook embedding, for which spatial mean and standard deviation are concatenated



Figure 5: Temporal controls for intrachain signals at a 1-ns frame interval. Time shuffling preserves amplitude entropy but removes lag-1 correlation; repeating one static frame removes all temporal variation. Values are normalized within their respective metric definitions.

together with normalized code entropy. These reductions yield one fixed-dimensional system vector despite variable chain lengths.

Vectors from the three replicas are pooled for standardization and randomized PCA to at most 64 dimensions (seed 20260916). Dynamic features subtract the matched repeated-static representation before sequence residualization. Because per-waveform min–max normalization maps every constant waveform to zero, the VQ repeated-static fingerprint is the codebook representation of the universal zero waveform rather than a complex-specific absolute-geometry baseline; VQ static subtraction therefore removes that common constant-code offset and is not equivalent to the continuous static control. The sequence covariate concatenates the across-chain mean and standard deviation of mean-pooled frozen ESM embeddings with chain count and total, mean, and standard deviation of chain length. Both covariates and targets are standardized. Ridge regression with $\alpha = 10$ is fit in five shuffled complex-level folds (seed 20260916), and held-out residuals are concatenated before replica comparison. The folds contain disjoint complexes but are not sequence-family clustered; the analysis removes tested out-of-fold linear sequence components, not every nonlinear or homology-associated effect. Linear CKA, dissimilarity-matrix Spearman, top-1/top-5 retrieval, adjusted Rand index over $k \in \{8, 16, 32\}$, and effective rank are averaged over the three replica pairs. Longest-span permutation tests jointly shuffle system correspondence and repeat the complete CKA calculation 1,000 times with a fixed random seed.

B Downstream Datasets, Splits, and Optimization

B.1 ProteinGym mutation-effect ranking

We use ten experimental single-mutant assays selected before model comparison. Candidates have a wild-type sequence of at most 512 residues, at least 1,000 annotated single mutants, and a UniProt identifier not used by another selected assay. The retained assays contain 34,935 variants in total. For each assay, a deterministic mutation-identity partition assigns 80/10/10% of variants to training, validation, and test. Exact substitutions therefore do not recur across partitions, but all three partitions share one wild-type protein and one experimental assay. This protocol measures within-landscape interpolation to held-out mutations; it is not protein-, family-, or property-out-of-distribution evaluation. Each model receives the mutated amino-acid sequence only and predicts the direction-normalized experimental DMS score. Batch-all-pairs RankNet with temperature 1.0 omits tied pairs (Burges et al., 2005). Validation Spearman selects the downstream epoch. Test Spearman is first averaged over the ten assays with equal assay weight and then summarized over seeds 42, 43, and 44.

Table 1: ProteinGym assays fixed before model comparison. Variant count is the total retained count before the assay-local 80/10/10 mutation split.

Property	ProteinGym assay identifier	Variants
Activity	CCDB_ECOLI_Adkar_2012	1,176
Activity	HEM3_HUMAN_Loggerenberg_2023	5,689
Binding	Q53Z42_HUMAN_McShan_2019_binding-TAPBPR	3,344
Binding	SPG1_STRSG_Olson_2014	1,045
Expression	NUD15_HUMAN_Suiter_2020	2,844
Expression	RASK_HUMAN_Weng_2022_abundance	3,066
Organismal fitness	NCAP_I34A1_Doud_2015	9,462
Organismal fitness	R1AB_SARS2_Flynn_2022	5,725
Stability	CATR_CHLRE_Tsuboyama_2023_2AMI	1,340
Stability	HECD1_HUMAN_Tsuboyama_2023_3DKM	1,244

B.2 Antibody and TCR–pMHC benchmarks

AbRank. The raw table contains 342,356 rows (Liu et al., 2025). We retain 75,483 rows with a finite numeric fitness or log-affinity label, nonempty ESM-compatible heavy-chain, light-chain, and antigen sequences, and length at most 1,022 for every sequence field. After required-field filtering, 20,501 invalid-token rows and 73,254 over-length rows are removed. Inputs are ordered heavy-chain, light-chain, and antigen triplets. The split unit is a joint Ab25–Ag25 cluster key. Deterministic size-matched holdouts contain 60,387/7,548/7,548 rows and 751/5/6 joint keys. An observed antibody-cluster–antigen-cluster combination cannot recur at evaluation, but either component can occur with a different partner. The split audit finds no exact concatenated heavy–light–antigen input overlap, while 99.9% of unique test antigen sequences also occur in training and six unique test heavy-chain sequences recur. This is joint-combination transfer rather than strict antibody- or antigen-component OOD evaluation. Exact labels use MSE. Left- and right-censored labels use a squared margin penalty only when a prediction violates the reported bound. Spearman and RMSE are evaluated on the 4,741 exact-label test rows.

BATCAVE global rank. We use 8,186 BATCAVE rows with reconstructed full TCRα, full TCRβ, peptide, and full MHC amino-acid sequences (Banerjee et al., 2025); 172 rows originating from the separate Knapp source are excluded. Each sequence is limited to 512 residues and the concatenated input to 1,022 residues. Source-provided fold 0 is held out, and 10% of the remaining rows is selected for validation with seed 42, yielding 5,796/642/1,748 train/validation/test rows. Every partition contains all 49 `TCR|index-peptide|MHC|assay` scan groups, so the source fold is a row-level activity holdout rather than a scan-group holdout. The continuous target is source-normalized peptide activity. RankNet pairs may cross the 49 scan groups, so global test Spearman is the primary metric. We additionally report the unweighted mean of per-scan Spearman values as a diagnostic, but do not train a separate scan-restricted objective. The fold split is not a claim of unseen-TCR, unseen-peptide, or component-disjoint transfer.

ITRAP. The ITRAP subset of our TCR–pMHC benchmark collection contains 23,235 binary-labeled receptor–pMHC pairs after background-repertoire filtering. The underlying experiment measures single-cell pMHC binding, with ITRAP-based filtering (Montemurro et al., 2023). Dataset construction uses TCRα/TCRβ/pMHC UMI thresholds of $(1, 1, 5)$ for positive assignment and labels other pMHCs detected in the same cells as negative, after excluding known positive pairs. Removing TCRs present in the background repertoire gives 3,244 positives and 19,991 negatives. These labels are assay-derived nonassignments, not independently validated nonbinding measurements; this reconstruction differs from the source paper's separate synthetic-negative evaluation sets. Our downstream processing preserves these labels and does not rerun denoising or generate negatives. Inputs are ordered TCRβ CDR3, peptide, and full-MHC sequences. A peptide-disjoint split yields 18,587/2,324/2,324 rows and 41/4/5 train/validation/test peptides. Weighted BCE-with-logits uses class weights computed from training data. The test set contains 492 positives and 1,832 negatives. Although peptide identities are disjoint, 99.57% of test rows contain a training-seen TCR and every test MHC occurs in training; this is peptide-held-out evaluation with recurring receptor components.

TABR-BERT TCR–pMHC. The release contains a 70,423-row positive training pool and four balanced tests (Zhang et al., 2024). A seed-42 row split produces 62,598/7,825 source positives for training and validation without peptide-, TCR-, or MHC-disjointness. Five healthy-background CDR3 sequences are paired with every positive while peptide and HLA remain fixed, giving 375,588/46,950 training/validation rows.

The original test labels are preserved. In all four released tests, experimentally determined positives are paired with an equal number of negatives formed by randomly matching pMHCs to healthy-repertoire TCRs (Zhang et al., 2024). S1 is a 2,826-row master-set subset (1,413 positives; 497 epitopes) and is zero-shot for TABR-BERT in the source protocol, though not for every comparator. S2 contains 1,236 pMTnet-derived rows (618 positives; 222 epitopes) and is non-zero-shot. S3 contains 794 PanPep-derived rows (397 positives; 266 epitopes) and is source-defined zero-shot; S4 is a 676-row filtered combination of S1–S3 (338 positives; 183 epitopes) and is also source-defined zero-shot. Against our actual training partition, the split audit finds no peptide overlap for S1, S3, or S4, while S2 shares 30 peptide sequences with training. TCR and MHC components can recur, so peptide novelty is not full complex novelty. Existing validation-selected checkpoints are evaluated on each released set; ROC-AUC and PR-AUC are reported separately and then summarized by an unweighted macro mean. These tests assess discrimination against healthy-background negatives under the stated peptide-overlap conditions.

B.3 Frozen encoder and downstream readouts

The frozen ESM-2 35M encoder contains 12 Transformer blocks, embedding width 480, 20 attention heads with width 24 per head, and feed-forward width 1,920 (Lin et al., 2023). The token readout adds two newly initialized pre-normalized Transformer blocks with the same width, head count, and feed-forward width, GELU activation, and dropout 0.1 in attention, feed-forward, and residual paths (Vaswani et al., 2017). ProteinGym uses learned 480-dimensional token-position embeddings for at most 512 tokens. Multi-chain tasks additionally use learned protein-identity embeddings and permit a maximum concatenated length of 2,044 tokens. Masked mean pooling is followed by layer normalization at width 480, a 480-to-512 linear projection, GELU, dropout 0.2, and a scalar output. The direct ProteinGym ablation removes both aggregation blocks and learned positions, mean-pools frozen ESM tokens, and retains the identical 512-hidden-unit prediction head.

ProteinGym uses batch size 32, AdamW at learning rate 10^{-4} and weight decay 10^{-2}, and at most ten epochs. The four multi-chain benchmark families use batch size 8, the same optimizer, gradient clipping at 1.0, ReduceLROnPlateau with factor 0.5 and patience 2, and validation-loss early stopping with patience 3. All downstream results use seeds 42, 43, and 44 and report the validation-selected epoch. The encoder remains frozen in every reported downstream arm.

C Complete Numerical Results

C.1 DynaRepo longest-span summary

Table 2 reports the longest-span values underlying the main-text comparison.

Table 2: DynaRepo dynamic representation metrics at a 380-ns span. Retrieval is cross-replica top-1; chance is 0.28%. Null is the mean residual CKA under permutation.

Representation	Residual CKA	Null	Corrected p	Retrieval (%)
Raw	0.472	0.013	0.005	25.3
Stats	0.431	0.021	0.005	9.9
DCT	0.474	0.027	0.005	7.1
DWT	0.275	0.040	0.005	3.2
VQ-VAE	**0.516**	0.065	0.005	4.5

C.2 PROTEINGYM SUMMARY

Table 3: ProteinGym macro Spearman over ten assays. Values are mean ± standard deviation across three matched fine-tuning seeds.

Frozen encoder	Mean-pool MLP	Two-block token readout
Base ESM-2 35M	0.381 ± 0.004	0.392 ± 0.017
Stats D&D	0.321 ± 0.005	0.461 ± 0.008
DWT D&D	0.312 ± 0.010	0.382 ± 0.053
DCT D&D	0.291 ± 0.004	0.354 ± 0.024
VQ-VAE D&D	0.373 ± 0.003	$\mathbf{0.558 \pm 0.020}$

Table 4: Assay-level VQ-minus-Base differences and readout interaction. Values average the three matched seeds within each assay.

Property	Assay	Mean pool	Two-block	Interaction
Activity	CCDB	+0.161	+0.331	+0.170
Activity	HEM3	-0.110	-0.069	+0.041
Binding	Q53Z42	-0.067	+0.173	+0.240
Binding	SPG1	-0.117	-0.130	-0.013
Expression	NUD15	-0.122	+0.183	+0.305
Expression	RASK	+0.002	+0.404	+0.402
Organismal fitness	NCAP	+0.028	+0.031	+0.003
Organismal fitness	R1AB	+0.173	+0.594	+0.421
Stability	CATR	-0.069	+0.010	+0.078
Stability	HECD1	+0.041	+0.137	+0.097

C.3 COMPLETE PROTEINGYM COMPRESSION–READOUT MATRIX

Table 5 reports all nine frozen-encoder initializations under both readouts. The property rows are descriptive averages of two assays and are not treated as independent replications.

Table 5: ProteinGym pairwise-rank transfer (Notin et al., 2023). Values are assay-equal test Spearman mean ± s.d. over three fine-tuning seeds (ten assays; two assays for each of five assay categories). All initializations, assay-local splits, seeds, and readout hyperparameters are matched within a column. Bold denotes the largest reported mean in a column, not a test-selected checkpoint.

Initialization	Mean-pool Spearman ↑	Two-block Spearman ↑
Base ESM-2 35M	**0.381 ± 0.004**	0.392 ± 0.017
Stats All	0.344 ± 0.007	0.324 ± 0.022
Stats D&D	0.321 ± 0.005	0.461 ± 0.008
DWT All	0.301 ± 0.002	0.320 ± 0.017
DWT D&D	0.312 ± 0.010	0.382 ± 0.053
VQ-VAE All	0.330 ± 0.006	0.295 ± 0.012
VQ-VAE D&D	0.373 ± 0.003	**0.558 ± 0.020**
DCT All	0.341 ± 0.007	0.284 ± 0.004
DCT D&D	0.291 ± 0.004	0.354 ± 0.024

Table 6: Two-block ProteinGym outcome by official coarse selection type (Notin et al., 2023). Each property is an assay-equal mean of two selected assays, then averaged over three seeds; categories are descriptive strata, not independent replications.

Property	Base ESM-2 35M	VQ-VAE D&D	Difference
Activity	0.323	0.455	+0.131
Binding	0.379	0.401	+0.022
Expression	0.390	0.683	+0.293
Stability	0.802	0.875	+0.073
Organismal fitness	0.065	0.377	+0.312

C.4 IMMUNE-TASK BOUNDARIES

Table 7: VQ-VAE D&D relative to Base ESM. These are descriptive three-seed means; metrics are task-specific and should not be compared in absolute magnitude across rows.

Dataset	Metric	Base	VQ-VAE	Difference
BATCAVE global rank	Spearman	0.495	0.557	+0.061
AbRank	Spearman	-0.005	0.122	+0.127
TCR–pMHC S1–S4	Macro ROC-AUC	0.707	0.668	-0.039
ITRAP*	ROC-AUC	0.374	0.828	+0.454

BATCAVE is trained with one global ranking objective across source-normalized activity values; the secondary scan-macro score in Table 8 is diagnostic only. The focused audit's combined TCR table uses a separately generated combined-test protocol and gives a VQ-minus-Base change of $+0.018$. Table 7 instead reports the original released S1–S4 tests, which we regard as the primary protocol. The two TCR values are not pooled. *ITRAP's aggregate association is subject to the recurrence and peptide-prevalence caveats described above.

C.5 COMPLETE IMMUNE-TASK MATRICES

Table 8 reports the complete nine-arm BATCAVE global-rank matrix, including DCT. Global Spearman is the prespecified primary metric; scan-macro Spearman is shown only to expose heterogeneity across scan groups.

Initialization	Global Spearman	Scan-macro Spearman	Pearson
Base ESM	0.495 ± 0.010	0.214 ± 0.008	0.422 ± 0.013
Stats All	0.484 ± 0.005	0.159 ± 0.029	0.424 ± 0.010
Stats D&D	0.532 ± 0.013	0.289 ± 0.015	0.481 ± 0.006
DWT All	0.491 ± 0.007	0.151 ± 0.027	0.426 ± 0.010
DWT D&D	0.483 ± 0.007	0.173 ± 0.014	0.410 ± 0.010
DCT All	0.477 ± 0.007	0.106 ± 0.053	0.412 ± 0.003
DCT D&D	0.505 ± 0.012	0.180 ± 0.009	0.434 ± 0.013
VQ-VAE All	0.498 ± 0.009	0.242 ± 0.025	0.426 ± 0.018
VQ-VAE D&D	0.557 ± 0.003	0.367 ± 0.020	0.497 ± 0.014

Table 8: Complete BATCAVE global-rank matrix. Values are mean $\pm$ sample SD over seeds 42–44. The primary metric is global Spearman over the 1,748-example held-out source fold. Scan-macro Spearman is a secondary diagnostic averaged over 49 scan groups; it is not a separately trained scan-restricted objective.

Tables 9–11 report every reported initialization for AbRank, ITRAP, and the macro average of the original TABR-BERT tests. Tables 12–13 then resolve S1–S4 separately so that differences among their source and novelty definitions are not hidden by the macro average.

Table 9: Complete AbRank transfer matrix (Liu et al., 2025): mean ± s.d. over three downstream seeds on the joint antibody–antigen cluster-combination holdout. All encoders are frozen; checkpoints are selected without downstream test values.

Initialization	Spearman ↑	RMSE ↓
Base ESM-2 35M	-0.005 ± 0.002	1.723 ± 0.071
Stats All	0.032 ± 0.087	1.603 ± 0.081
Stats D&D	0.135 ± 0.093	1.590 ± 0.210
DWT All	0.050 ± 0.042	1.528 ± 0.134
DWT D&D	0.002 ± 0.056	1.714 ± 0.149
VQ-VAE All	0.039 ± 0.107	1.746 ± 0.194
VQ-VAE D&D	0.122 ± 0.089	1.477 ± 0.174
DCT All	-0.005 ± 0.027	1.819 ± 0.109
DCT D&D	0.014 ± 0.083	1.564 ± 0.113

Table 10: Complete ITRAP transfer matrix (Montemurro et al., 2023): mean ± s.d. over three downstream seeds on the peptide-disjoint split. TCR and MHC component recurrence limits interpretation (Appendix B).

Initialization	ROC-AUC ↑	PR-AUC ↑
Base ESM-2 35M	0.374 ± 0.062	0.170 ± 0.017
Stats All	0.352 ± 0.028	0.159 ± 0.008
Stats D&D	0.333 ± 0.040	0.156 ± 0.011
DWT All	0.424 ± 0.177	0.185 ± 0.060
DWT D&D	0.360 ± 0.017	0.159 ± 0.004
VQ-VAE All	0.364 ± 0.103	0.163 ± 0.023
VQ-VAE D&D	0.828 ± 0.017	0.458 ± 0.046
DCT All	0.347 ± 0.027	0.158 ± 0.003
DCT D&D	0.340 ± 0.099	0.160 ± 0.019

Table 11: Complete TCR–pMHC macro transfer matrix (Zhang et al., 2024): mean ± s.d. over three downstream seeds. Each metric is an unweighted mean across the original balanced S1–S4 tests. The backbone is frozen and checkpoint-selection rules are given in Appendix A.2.

Initialization	Macro ROC-AUC ↑	Macro PR-AUC ↑
Base ESM-2 35M	0.707 ± 0.044	0.765 ± 0.031
Stats All	0.686 ± 0.051	0.724 ± 0.056
Stats D&D	0.662 ± 0.047	0.731 ± 0.033
DWT All	0.692 ± 0.023	0.746 ± 0.018
DWT D&D	0.696 ± 0.019	0.752 ± 0.007
VQ-VAE All	0.680 ± 0.009	0.732 ± 0.005
VQ-VAE D&D	0.668 ± 0.042	0.731 ± 0.030
DCT All	0.704 ± 0.028	0.762 ± 0.021
DCT D&D	0.706 ± 0.019	0.769 ± 0.019

Table 12: TCR–pMHC results on the original balanced TABR-BERT S1 and S2 tests (Zhang et al., 2024) (mean ± s.d.; seeds 42–44). Released labels are unchanged and no test negatives are resampled. All checkpoints are validation-selected.

Initialization	S1 ROC-AUC	S1 PR-AUC	S2 ROC-AUC	S2 PR-AUC
Base ESM-2 35M	0.751 ± 0.026	0.807 ± 0.021	0.664 ± 0.040	0.707 ± 0.039
Stats All	0.747 ± 0.026	0.797 ± 0.026	0.610 ± 0.102	0.605 ± 0.118
Stats D&D	0.740 ± 0.025	0.790 ± 0.022	0.592 ± 0.072	0.650 ± 0.049
DWT All	0.756 ± 0.013	0.806 ± 0.010	0.621 ± 0.051	0.656 ± 0.051
DWT D&D	0.750 ± 0.020	0.800 ± 0.017	0.652 ± 0.022	0.676 ± 0.015
VQ-VAE All	0.770 ± 0.003	0.819 ± 0.005	0.594 ± 0.016	0.620 ± 0.003
VQ-VAE D&D	0.744 ± 0.016	0.792 ± 0.016	0.595 ± 0.074	0.644 ± 0.047
DCT All	0.760 ± 0.011	0.807 ± 0.017	0.652 ± 0.044	0.689 ± 0.030
DCT D&D	0.742 ± 0.012	0.791 ± 0.017	0.690 ± 0.045	0.741 ± 0.048

Table 13: TCR–pMHC results on the original balanced TABR-BERT S3 and S4 tests (Zhang et al., 2024), with the same reporting and checkpoint protocol as Table 12.

Initialization	S3 ROC-AUC	S3 PR-AUC	S4 ROC-AUC	S4 PR-AUC
Base ESM-2 35M	0.635 ± 0.028	0.724 ± 0.021	0.777 ± 0.110	0.823 ± 0.056
Stats All	0.594 ± 0.049	0.692 ± 0.037	0.791 ± 0.035	0.804 ± 0.046
Stats D&D	0.572 ± 0.073	0.690 ± 0.039	0.745 ± 0.049	0.794 ± 0.031
DWT All	0.582 ± 0.017	0.698 ± 0.002	0.808 ± 0.017	0.822 ± 0.015
DWT D&D	0.541 ± 0.038	0.677 ± 0.021	0.840 ± 0.002	0.855 ± 0.012
VQ-VAE All	0.591 ± 0.019	0.699 ± 0.018	0.764 ± 0.007	0.789 ± 0.004
VQ-VAE D&D	0.552 ± 0.054	0.675 ± 0.031	0.782 ± 0.029	0.813 ± 0.030
DCT All	0.543 ± 0.055	0.678 ± 0.033	0.860 ± 0.038	0.874 ± 0.032
DCT D&D	0.536 ± 0.047	0.673 ± 0.022	0.857 ± 0.056	0.871 ± 0.058

D Additional Checks and Failure Modes

D.1 Does MD supervision add decodable trajectory information?

The DynaRepo analysis establishes that several 96-frame target descriptions are reproducible across independent replicas, but reproducibility alone does not show that MD supervision changed the sequence representation. We therefore separate three claims. First, replica agreement tests whether a target contains system-specific signal. Second, the original-head audit below tests whether an end-to-end MD predictor beats a simple train-only sequence descriptor prior. Third, the matched-probe audit discards every original MD head and asks whether the frozen MD-supervised encoder makes the same held-out targets more decodable than Base ESM under an identical newly initialized readout.

Existing-head audit. For each of 2,094 strict protein–run test groups, the previously trained MD head is compared with a descriptor prior fitted on 2,048 original-training groups. The distance prior uses unordered amino-acid pair and binned sequence separation; the dihedral prior uses amino-acid identity, relative position, and protein-length bin. Relative skill is $1 - L_{\mathrm{model}}/L_{\mathrm{prior}}$. Its role is limited: it tests the complete trained predictor against that particular prior, not the MD encoder against Base ESM.

Table 14: Existing-head strict-test skill relative to a train-only sequence descriptor prior (mean [95% group-bootstrap interval]). Values are comparable only within a target family and feature.

Checkpoint	Distance	Dihedral
Stats All	0.698 [0.611, 0.773]	0.784 [0.780, 0.787]
Stats D&D	0.065 [−0.020, 0.144]	0.665 [0.661, 0.669]
DWT All	0.920 [0.912, 0.927]	0.759 [0.755, 0.763]
DWT D&D	0.889 [0.881, 0.896]	0.795 [0.792, 0.799]
DCT All	0.923 [0.914, 0.929]	0.772 [0.768, 0.775]
DCT D&D	0.892 [0.884, 0.899]	0.800 [0.797, 0.804]
VQ-VAE All	−0.011 [−0.014, −0.009]	0.054 [0.053, 0.055]
VQ-VAE D&D	0.030 [0.029, 0.031]	0.112 [0.111, 0.113]

Matched frozen-encoder audit. We freeze Base ESM-2 35M or one MD-supervised encoder, discard its old heads, and fit the same $\mathrm{LayerNorm} \to 480{\times}256 \to \mathrm{GELU} \to \mathrm{dropout} \to 256{\times}d$ probe. Each target family uses the same 2,048 training protein–run groups, a 512-group protein-ID-disjoint validation subset, and the locked 2,094-group strict test cohort, whose protein IDs occur in neither the original MD-training nor MD-validation split (exact protein-ID disjointness, not homology-level separation). We use the first 96 stored frames, all usable dihedral positions, and deterministic non-self upper-triangle distance pairs (at most 256 per training group and 512 per validation/test group). Regression targets use dimension-standardized training objectives and are evaluated by native-space MSE; VQ-VAE codes use token cross-entropy. Three fresh-probe seeds are averaged before a paired 2,000-resample test-group bootstrap. Relative skill is $1-L_{\mathrm{MD\ encoder}}/L_{\mathrm{Base\ ESM}}$, so positive values directly indicate additional matched-readout-accessible MD-target information under the matched nonlinear readout.

The two tables answer different questions and should not be expected to have identical rankings. Table 14 includes the head learned jointly during MD supervision and compares it with a hand-designed descriptor prior. Table 15 removes that head advantage and compares encoders with identical capacity, data, optimization, and sampled positions. A positive matched-probe interval is evidence that MD supervision placed target-relevant information into the frozen embedding beyond Base ESM; a null or negative interval means that this protocol does not establish such an increment, even if the original predictor reconstructs its target. The strongest matched increments are for DCT/DWT distance targets (0.893–0.918 across the tested checkpoints). The validation-selected D&D encoders also improve dihedral decodability for DWT (0.174) and DCT (0.157). VQ-VAE D&D gives much smaller but positive increments (0.011 distance and 0.024 dihedral). This ordering differs from downstream transfer, where VQ-VAE can be preferable with a token-mixing readout.

Table 15: Fresh matched-probe relative skill over Base ESM (mean [95% paired group-bootstrap interval]). Positive values favor the MD-supervised encoder. Comparisons are only within a target family and feature.

Frozen encoder	Distance	Dihedral
Stats All	0.232 [0.200, 0.265]	0.001 [−0.017, 0.019]
Stats D&D	−0.343 [−0.493, −0.212]	−0.548 [−0.581, −0.515]
DWT All	0.913 [0.907, 0.918]	0.066 [0.049, 0.083]
DWT D&D	0.893 [0.886, 0.900]	0.174 [0.156, 0.191]
DCT All	0.918 [0.913, 0.924]	0.139 [0.125, 0.154]
DCT D&D	0.900 [0.893, 0.906]	0.157 [0.140, 0.173]
VQ-VAE All	0.005 [0.005, 0.005]	−0.006 [−0.006, −0.005]
VQ-VAE D&D	0.011 [0.010, 0.012]	0.024 [0.023, 0.026]

Direct reconstruction skill, representation usability, and biological-label alignment are therefore separate properties. The 16 matched intervals are descriptive: they average three fresh-probe seeds but do not vary the original MD-pretraining seed, and no family-wise correction is applied across the 16 comparisons. Neither table proves physical trajectory reconstruction or downstream usefulness by itself; those are assessed separately by the plausibility, waveform, and transfer analyses.

D.2 Illustrative paired distance reconstruction across compression families

Figure 6 shows one paired Cα–Cα distance example across DCT, DWT, and VQ-VAE. All panels use the same strict-test group, first 96-frame window, and residue pair 31–70, which had the highest joint score among the six pairs retained for this group in the existing audit. Candidate groups were ranked by the mean percentile of their existing complete-group distance losses; the group and candidate pairs were chosen jointly rather than separately by method. The visualization is post-hoc and illustrative; aggregate reconstruction and matched-probe tables, not this example, support population-level claims. Each panel separates compression fidelity from prediction fidelity: original MD is compared with the inverse transform or decoder output of the true six-element representation, and then with reconstruction from the sequence-predicted coefficients or codes. The curves retain the affine scale defined by the raw distance trace and expose amplitude attenuation. VQ-VAE uses its native per-waveform normalized decoder space, so its panel does not establish absolute-distance recovery. The larger heat-map atlas is retained in the artifact but omitted here to avoid repeated illustrations.

D.3 Additional VQ-VAE temporal-shape examples

Temporal shape and waveform error can favor different predictions. We revisit the existing 40-group, 13,804-pair audit and select additional illustrations using only VQ-VAE prediction-to-raw Pearson correlation > 0.5, compression-to-raw correlation > 0.5, raw standard deviation > 0.4968 Å, and raw range ≥ 0.5 Å. These gates retain 12 pairs. Ranking them by VQ prediction-to-raw correlation and allowing at most one pair per protein–run group gives six audited groups; we display only the top two, while retaining all six in the artifact. DCT and DWT scores are not selection criteria; selection on VQ quality makes these examples illustrative rather than an unbiased comparison of methods. Several audited groups are variants of the same protein and are not independent family-level replications. Across the six audited examples, VQ-VAE prediction-to-raw correlations range from 0.521 to 0.563; the two displayed cases have correlations 0.563 and 0.556. Its raw-window-normalized RMSE is larger than at least one continuous comparator in 4/6 cases and larger than both in 3/6. Correlation is insensitive to an additive offset and positive amplitude scaling, whereas squared error penalizes both; a better-shaped curve can therefore have a larger error. This comparison does not establish that VQ-VAE universally preserves shape better or that discretization alone causes the difference. The native panels retain the original distance window's affine scale for MD and continuous predictions and use the VQ decoder's native normalized output. The shape panels independently min–max normalize each trace to expose its timing and relative profile. Neither view shifts time, aligns peaks, nor smooths the original MD waveform. Independent normalization can magnify a weak-amplitude prediction, so native and shape views should be read together. The normalized RMSE is not an independently predicted absolute-distance error in Å.

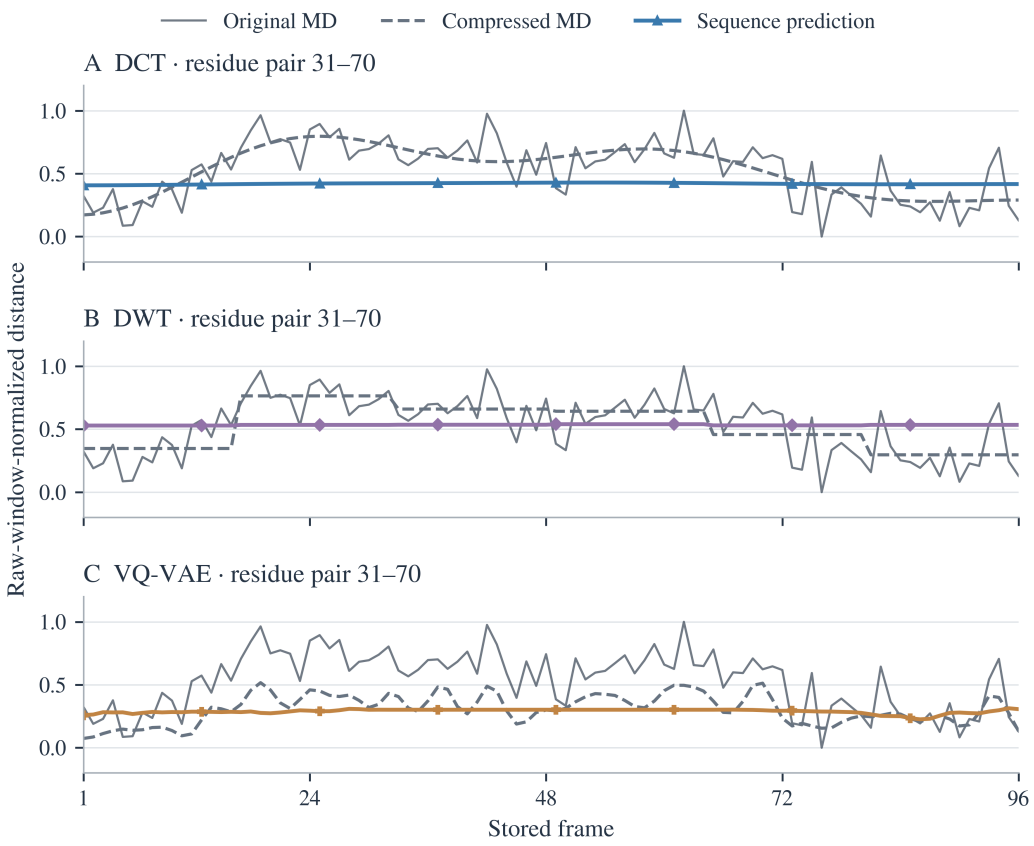

Figure 6: A selected distance waveform separates compression fidelity from prediction fidelity. Panels show original MD, compressed MD reconstruction, and sequence-predicted reconstruction for the same pair 31–70 in strict-test group 1UBQ_L67S__G53N&000000. DCT has $r_c = 0.79$ and $r_p = 0.76$, DWT $r_c = 0.73$ and $r_p = 0.53$, and VQ-VAE $r_c = 0.77$ and $r_p = 0.44$, where r_c compares compression with raw MD and r_p compares prediction with compression. The raw-derived normalized scale preserves amplitude attenuation; VQ remains in its native normalized decoder space. The fixed joint shape gates retain 25/13804 candidate pairs, so this selected illustration is not an estimate of typical recovery.

Table 16: The two displayed temporal-shape cases, ranked by VQ prediction-to-raw correlation. Entries give prediction-to-raw Pearson correlation / raw-window-normalized RMSE. Selection is VQ-focused and post-hoc; all six audited cases remain in the artifact.

Case	Residue pair	DCT	DWT	VQ-VAE
01	11–42	0.008 / 0.269	0.102 / 0.242	0.563 / 0.232
02	41–70	0.426 / 0.266	0.306 / 0.281	0.556 / 0.343

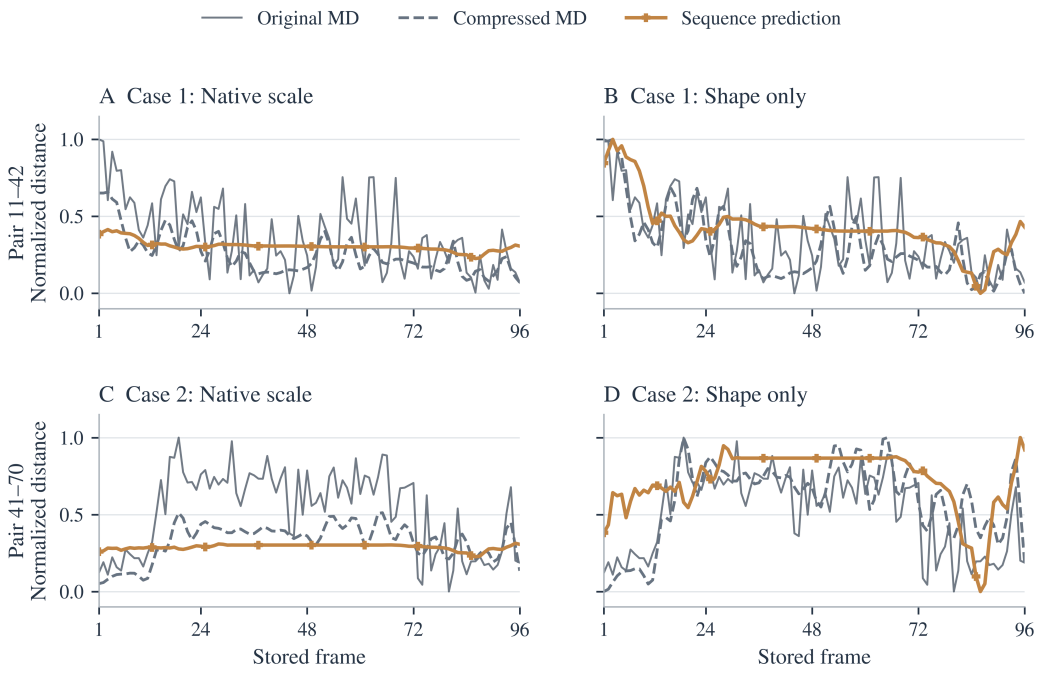

Figure 7: Two selected VQ-VAE examples illustrate temporal shape despite amplitude attenuation. Rows show case 01 (2M2L__E3H_K24E&000000, pair 11–42) and case 02 (1UBQ_L67S__G53N&000000, pair 41–70). Left panels retain the native normalized scale; right panels independently normalize each trace to display shape. Every panel contains original MD, true compressed reconstruction, and sequence-predicted reconstruction; the two cases have the highest VQ prediction-to-raw correlations among the six audited groups. Selection and error definitions are given in Table 16.

D.4 CheY: a sequence-only illustration of coordinated waveform directions

We examine whether a single sequence produces decoded waveforms whose coordinated changes resemble the directions observed between two experimental conformations. CheY provides an illustrative comparison between the inactive structure 3CHY and the BeF_3^--activated structure 1FQW, using chain A in both entries (Volz & Matsumura, 1991; Lee et al., 2001). Their deposited sequences are identical and all 128 Cα positions are available. The cartoons in Figure 8 show these experimental structures after rigid Cα alignment; they are not structures generated from the model's predictions. The model receives only the sequence, without either reference structure, an initial conformation, or the activating ligand.

Pair selection and decoding. We reuse the frozen VQ-VAE D&D distance head and codebook decoder to produce 96 values per residue pair. The reference structures identify 86 changing pairs with sequence separation greater than three residues, an absolute Cα distance difference of at least 1 Å, and distance at most 12 Å in at least one state. Each predicted waveform is centered and divided by its temporal standard deviation, then signed by the experimental activated-minus-inactive distance change. A positive signed deviation therefore indicates motion in the activated-state direction, not agreement with its absolute distance. With a fixed seed (20260925), 43 pairs select the minimum and maximum of their mean signed waveform; the remaining 43 pairs evaluate the distance-change directions between those same two frames. The selection gives frame indices 28 and 86. We do not choose a separate frame for each pair, and frame indices are not calibrated physical times.

Observed pattern and its limits. On the held-out pairs, 32/43 (74.4%) predicted changes agree in direction with the experimental comparison; the cosine between standardized predicted changes and experimental direction signs is 0.395. The all-pair signed-waveform span exceeds controls based on independent per-pair circular shifts, independent time permutations, and direction-sign permutations within 2-Å bins of inactive-state distance. Using 4,999 randomizations and Holm adjustment across the 30 testable case–model combinations, the corresponding adjusted p values are 0.006, 0.006, and 0.026. However, the held-out cosine does not exceed the spatial direction-sign control (adjusted $p = 1.000$). Thus the directional agreement is a descriptive case pattern, not an independently validated recovery of the activated state. Pairs may also share residues, so they are not independent biological replications.

This illustration was selected after inspecting six models (DCT, DWT, and VQ-VAE, each with All and D&D supervision) on four functional-state comparisons and two same-state structural controls. It is not an unbiased estimate of success across proteins. VQ preprocessing removes each waveform's absolute offset and range, which cannot be reconstructed from the predicted codes alone; we do not use the experimental endpoint distances to supply those missing scales. Consequently, the decoded curves cannot establish absolute contact formation, a physically valid three-dimensional trajectory, transition kinetics, or ligand-dependent state populations. The case illustrates partial coordination of relative waveform directions, while retaining the distinction between that pattern and complete functional-state recovery.

D.5 Why residual CKA and retrieval can disagree

CKA compares the geometry of all systems: it is high if pairwise variation spans similar directions across replicas. Retrieval asks whether each individual system is the nearest match to itself. A quantized representation can preserve a stable coarse geometry while mapping many systems to similar motif-frequency profiles, which can produce high CKA and lower retrieval. Conversely, raw continuous trajectories can retain idiosyncratic amplitudes that improve instance identity without being ideal for transfer.

D.6 Why VQ-VAE may favor token readout

Each VQ target assigns local waveforms to discrete motifs. Cross-entropy encourages boundaries between motif identities rather than minimizing every amplitude error. The supervision is applied at residue or residue-pair positions, so the resulting signal may remain spatially distributed. Mean pooling removes order and can cancel local changes; two token-mixing blocks can condition one position on others before pooling. Continuous Stats and DWT encoders also benefit from this readout,

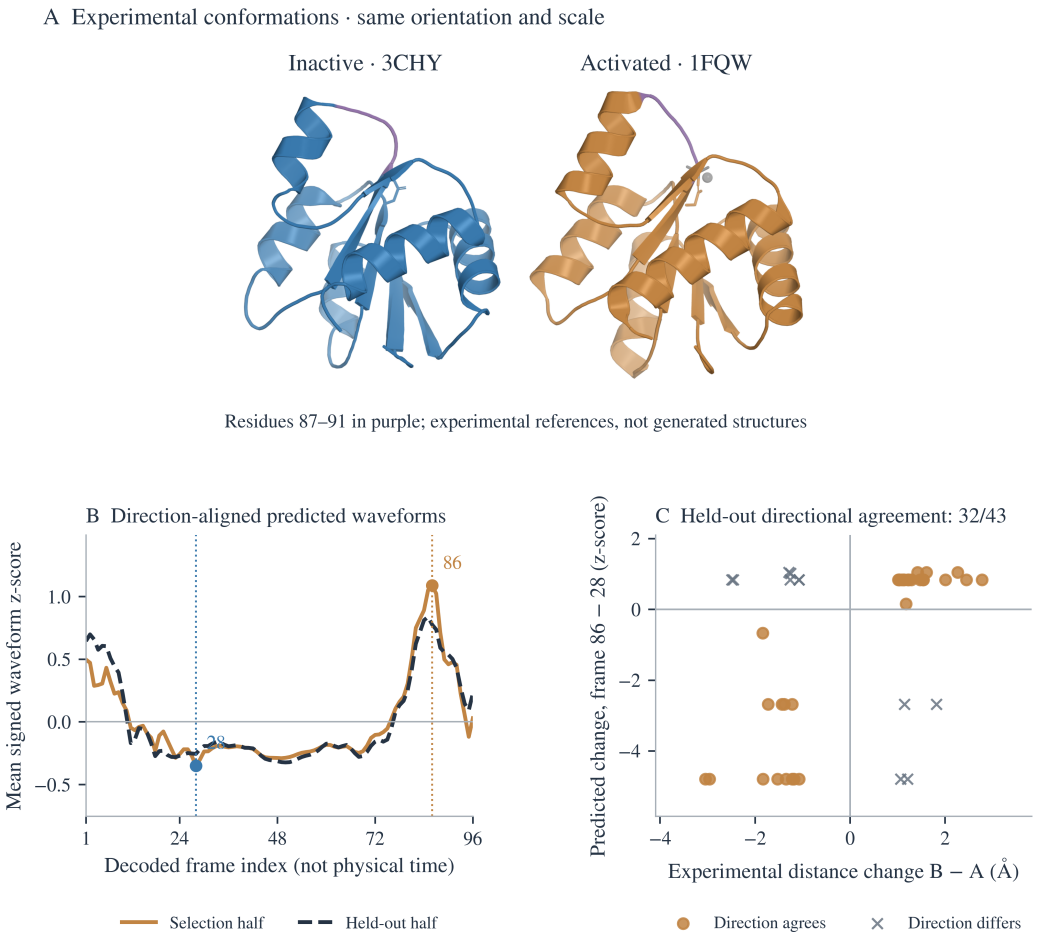

Figure 8: CheY VQ-VAE D&D waveforms show partial agreement with experimental state-change directions, without establishing complete state recovery. **A**: adjacent, separate experimental inactive (3CHY, blue) and BeF_3^--activated (1FQW, ochre) cartoons in the same orientation and scale after rigid Cα alignment. Residues 87–91 are purple; sticks mark Asp57, and the activated reference also displays BeF_3^- and Mn. **B**: mean signed, per-pair standardized waveforms for the frame-selection and held-out halves; positive deviations point toward the activated-state distance-change direction, and vertical lines mark selected frames 28 and 86. **C**: experimental distance changes versus predicted standardized changes on all 43 held-out pairs, with agreement for 32 pairs (74.4%). These are normalized waveform directions, not predicted distances in Å; the held-out spatial-control test is not significant, as detailed in the text.

so spatial distribution is the broader explanation supported by the matrix. VQ-specific explanations remain unresolved because discreteness, per-waveform normalization, and the externally pretrained codebook change together.

D.7 Physical plausibility

Predicted MD summaries preserve some expected correlations between flexibility, compactness, and geometry on held-out and external trajectories. They also violate some exact geometric relationships because each head predicts a summary independently. We therefore treat the encoder as a learned representation, not a simulator and not a source of physically valid reconstructed trajectories.

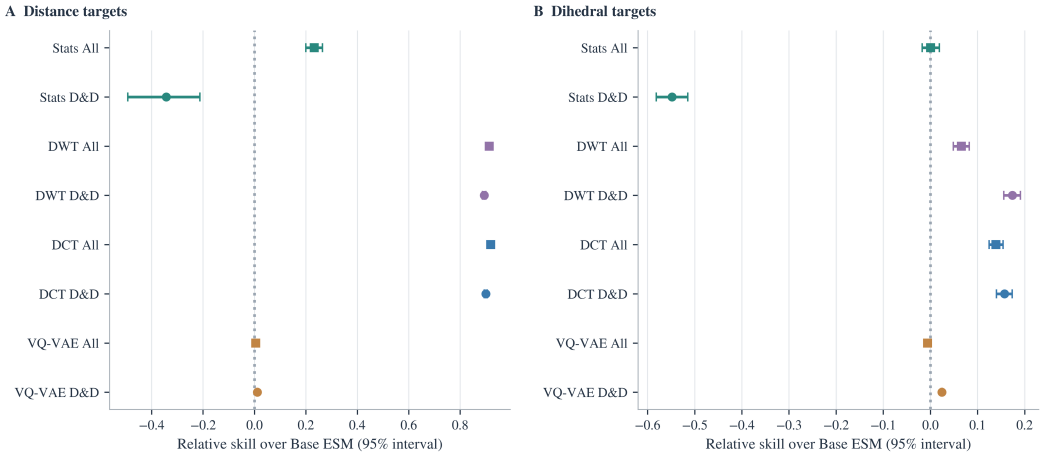

Figure 9: The retained encoders differ in matched-probe decodability. Points and error bars reproduce the relative skills and paired 95% test-group bootstrap intervals in Table 15. Squares denote All and circles D&D; colors denote compression. Zero denotes equal loss to the matched Base ESM probe. Comparisons remain target-family-specific.

E Artifact Provenance

The numerical figure panels are produced by deterministic rendering scripts with a shared palette and typography configuration. Results are organized by analysis: DynaRepo representation, entropy, and permutation tests; ProteinGym assay- and seed-level evaluations; immune benchmarks; matched probes; and selected waveform reconstructions. Reconstruction data retain the fixed checkpoints and original case selections. A machine-readable figure index records the corresponding data sources and panel positions. Source run identifiers remain unchanged while displayed names are All and D&D. Figures are rendered from recorded results without retraining models.

The separate CheY illustration uses cached sequence-only predictions and public experimental coordinates. Its rendering script records the coordinate and prediction checksums, shared cartoon view, fixed pair split, and selected frame indices. Experimental cartoons and decoded waveform panels are explicitly distinguished; no predicted three-dimensional coordinates or target-derived VQ distance scales are introduced.