

Refined Topology-Pass Scheduling: Damped Trust-Region Compensation for Dynamic k -NN Graphs in SE(3)-Equivariant Structure Prediction

Jiankai Huang

huangj.k.08@gmail.com

Hanwen Shen

hs907@scarletmail.rutgers.edu

Abstract

Equivariant structure-prediction pipelines alternate coordinate updates with the construction of a k -nearest-neighbour (k NN) graph whose *membership* is a discrete, piecewise-constant function of the coordinates. Rebuilding that graph at every block boundary (PB) is the expensive part of an epoch, whereas reusing one frozen graph (F) is fast but leaves a topology gap. We study a two-pass design in which the cheap pass records a small amount of state (margins, k NN radii, measured per-boundary displacements) and a second pass *compensates conservatively*. We formulate the choice of the graph-construction basis as a damped fixed-point iteration whose unique fixed point is exactly PB, and derive (i) exact non-flip certificates with the two-sided constant $m_i/4$ and one-sided constant $m_i/2$, (ii) the minimum-variance (BLUE) replay gain γ^* and the resulting feasibility crossover, (iii) an ℓ_∞ trust-region projection that preserves convergence, and (iv) an analytic near-tie spacing law $E[m] \approx r_k/(3k)$ that matches measurements to 1%. Three compensation levels are implemented in C++ (L₁: frozen set with live geometry; L₂: margin-gated local repair; L₃: damped, projected basis) and validated under a matched-conditions protocol (identical seed, MSA depth and trajectory family, only the topology mode differs). L₂ reproduces the PB edge set exactly ($\mathcal{D} = 0.0000$ at every boundary), L₁ stays within 0.6% and L₃ interpolates monotonically: L₂ reproduces the PB edge set exactly when the trust region is removed, and L₃ interpolates between the frozen and the PB topology. The recorded trajectory itself yields a measured $1.5\times$ Pass-1 and $1.76\times$ epoch speed-up with no measurable loss gap. An open-source implementation of the runtime and of all the experiments reported here is available at <https://github.com/AlbertHuang96/ProteinPredictMachineLearning>.

Index Terms

dynamic k NN graphs, SE(3)-equivariant networks, topology pass scheduling, trust-region projection, fixed-point iteration, protein structure prediction

1. INTRODUCTION

1.1 Previous work

All-atom protein structure prediction is dominated by three-track (1D/2D/3D) architectures: AlphaFold2 established the MSA/pair/coordinate design with invariant point attention [1], AlphaFold3 extended it to a diffusion module over all chemical components [2], and RoseTTAFold All-Atom (RFAA) combined an SE(3)-equivariant track with all-atom residue frames [3]. The 3D track itself rests on equivariant message passing: tensor-field networks [4] and SE(3)-Transformers [5] build relative-position features from spherical harmonics and Clebsch–Gordan coefficients, i.e. from a *graph* of residue pairs. Graph construction (a k NN search) is therefore on the critical path of every iteration.

Two largely separate literatures address the consequences. (i) *Dynamic graph maintenance*: temporal GNNs update graph structure as data streams in [6], and approximate nearest-neighbour indexes such as HNSW support fast incremental queries [7]; spectral sparsification asks which edges even matter [8], and graph rewiring improves message passing by changing connectivity on purpose [9]. (ii) *Relaxing discrete selections*: straight-through estimators pass gradients through non-differentiable choices [10] and differentiable top- k operators replace a hard selection by a continuous relaxation [11]. Our setting is an instance of both: the graph is a discrete degree of freedom whose gradient vanishes almost everywhere, and it must be maintained under continuous drift.

1.2 Background and problem statement

The training pipeline we study comes from RFAA. What is specific here is the implementation route: our runtime builds the forward graph explicitly with ggml before execution, so the topology of every block — including the k -nearest-neighbour (k NN) graph construction — is bound at build time, while the coordinates that define that topology keep changing during training. This mismatch is the coordinate-optimisation problem addressed in this paper: the graph is a function of the coordinates, yet once the forward graph has been built it can only be reused as it stands, so one must decide *when* to rebuild it and *what* to reuse in the meantime. Concretely, the pipeline scans T block boundaries per iteration. At boundary t the graph

$$\mathcal{G}(x) = (E(x), D(x), W(x)) \quad (1)$$

is built by `make_graph` from coordinates x : a *directed* k NN edge set $E(x) \subset [L] \times [L]$, in which each residue keeps its $k = \min(64, L - 1)$ nearest neighbours, together with the edge geometry $d_{ij} = x_j - x_i$ and scalar edge features W .

Four terms are used throughout. F is the plain schedule in which the graph is built once, from the initial coordinates x^0 , and then reused unchanged for every later boundary, so that the network runs feed-forward. PB is the opposite schedule: the graph is rebuilt from the live coordinates at every boundary, which is the accuracy reference and the expensive option. A *topology pass* is one execution of the block sequence with a graph that is fixed in advance and reused, i.e. a feed-forward pass whose trajectory can be recorded. The *two-pass variant* combines a topology pass (Pass-1), which records the trajectory, with a second pass (Pass-2) that re-runs the same blocks and consumes the recorded state, so that only the graph step differs between the two passes. Measurements on the original pipeline summarise the trade-off:

- cost: PB needs ≈ 34.1 s per epoch versus 19.4 s for the two-pass variant ($1.76\times$); the k NN rebuild dominates (≈ 1.8 s *per block*), whereas refreshing the geometry over the *frozen* edge set $\mathcal{E}$ of (1) is $O(|\mathcal{E}|)$, with $|\mathcal{E}| \approx 7,000$ edges, and is negligible, and a local rebuild costs $O(|\Omega|k)$;
- fidelity: at $L = 103$ the first boundary replaces $\approx 25\%$ of the directed edges (100% of residues touched) and steady-state boundaries replace 2–3%; the common-edge geometry differs by only 0.29 Å;
- accuracy: no measurable loss gap between F and PB (15.49 vs. 15.91, 3 seeds, spread ± 1.3) — i.e. the benefit ceiling of *any* hybrid is still unknown.

Two questions follow. **(Q1)** Is a schedule that keeps the cheap F-style topology pass for most blocks feasible, and what is the cheapest useful hybrid? **(Q2)** Can such a hybrid — one that records only a little state and compensates *conservatively* — reduce the topology gap to PB beyond what the plain F schedule already achieves? This paper answers both with a fixed-point formulation, proved certificates, an implementation, and a matched-conditions experimental study.

1.3 Contributions

- 1) A fixed-point formulation of basis selection: the compensated pipeline is $x^{t+1} = \Phi_t(x^t, \mathcal{G}(B^t))$ and the damped iteration on B^t has PB as its *unique* fixed point (Proposition 1); Pass-1 supplies the first iteration for free.
- 2) Exact non-flip certificates: $4\Delta < m_i$ guarantees edge-wise invariance of the k NN set (Theorem 1); if only one endpoint moves, $2\Delta_i < m_i$ suffices (Theorem 2). The two constants give two *different* semantics for the same safety factor $\kappa_i = \sigma m_i / 2$.
- 3) Optimality of the replay gain: γ^* is the minimum-variance (Gauss–Markov/BBLUE) linear combination of the frozen prior and the recorded predictor, with residual variance $\sigma_d^2 \sigma_p^2 / (\sigma_d^2 + \sigma_p^2)$; the feasibility crossover $\rho^* \approx 0.25$ is the $\sigma_p = \sigma_d$ boundary.
- 4) A trust-region view: the ℓ_∞ box projection Π_κ is firmly non-expansive, so composing it with the damped iteration preserves convergence, and the projected step cannot leave the region in which the current graph is provably valid.
- 5) Quantified edge flapping: $E[m] \approx r_k / (3k)$ for a local 3D Poisson process, matching the measured margin median to 1% (0.078 vs. 0.0787 Å) and showing that flapping is intrinsic to top- k graphs, not an implementation defect.
- 6) An implementation of three compensation levels with matched-conditions validation: L_2 reproduces PB *exactly* ($\mathcal{D} = 0.0000$), L_1 is within 0.57–0.68%, and L_3 interpolates monotonically between them.

2. LOGICAL ROADMAP AND PROOF OBJECTIVES

2.1 Central question and logical structure

The questions (Q1)–(Q2) left by Section 1 are not whether a frozen graph is always identical to a graph rebuilt at every block. They are whether a cheap topology pass can record enough information to support conservative compensation towards the

fully coupled reference PB, and under what conditions that compensation is both useful and inexpensive, i.e. when it reduces the topology gap to PB beyond what the plain F schedule already achieves. The argument therefore separates three issues: the mathematical target of refinement, the conditions under which the proposed updates are justified, and the experimental evidence that the implementation improves topology fidelity at an acceptable cost.

The logical chain is

identify the reference target $\longrightarrow$ derive stability and replay conditions $\longrightarrow$ construct the compensation levels
 $\longrightarrow$ audit and measure the implementation $\longrightarrow$ state the supported conclusion and its limits.

Section 3 develops the mathematical and algorithmic part of this chain; Section 4 tests its observable consequences; Section 5 combines the two. A proof of a local property, an approximation under a statistical model, and an observation on the recorded samples play different roles: none is used here as a substitute for the others.

2.2 Algorithm chapter: what must be established, and in what order

Section 3.1: define a common problem before comparing schedules. The live coordinates x^t and the graph-construction basis B^t are separated in one closed-loop recursion. The frozen schedule F, the fully coupled schedule PB, and the refined schedule (this work) then become different choices within the same formulation. The margins m_i , radii κ_i , and distances $\mathcal{D}, \mathcal{D}_w$ specify, respectively, when topology is fragile, how compensation is constrained, and what improvement means. This subsection supplies definitions rather than a separate convergence claim.

Section 3.2: identify what refinement is trying to reach. Proposition 1 proves a consistency statement: if the damped coupling converges with $\lambda > 0$, its limiting basis equals its limiting live coordinates, so the limiting recursion is exactly PB. Proposition 2 then states the local convergence argument under assumptions (A1)–(A3), while Proposition 3 describes the role of ties, path noise, and the risk set in the residual topology gap. The dependency is important: identifying the fixed-point target does not itself prove that every trajectory reaches it. The following two subsections supply the geometric and projection ingredients used by the conditional convergence argument.

Section 3.3: determine when reusing topology is certified. The analysis first locates membership changes on equal-distance surfaces, explaining why smooth coordinate motion can produce a discrete edge change. Lemma 1 bounds distance perturbations; Theorems 1 and 2 then prove the sufficient non-flip conditions $4\Delta < m_i$ for two-sided motion and $2\Delta_i < m_i$ when only node i moves. These results explain the constant-edge-set premise in the local argument and motivate margin-based gating. Their purpose is to certify reuse where the inequalities hold, not to assert that all observed displacements satisfy them. This also introduces the distinction between preventing flips and tracking flips that actually occur.

Section 3.4: decide how much to replay and how to constrain it. Under the stated drift-and-predictor covariance model, the minimum-variance linear replay gain in (13)–(15) answers how strongly the recorded displacement should influence the basis. The comparison with $\gamma = 1$ identifies the $\sigma_p = \sigma_d$ boundary for naive replay versus freezing. The trust-region part answers a different question: how to constrain the proposed basis at low cost. Proposition 4 proves that the box projection is firmly non-expansive; the composition argument uses non-expansiveness together with $q < 1$ and a feasible reference fixed point. The subsequent analysis of flips, two-cycles, and chattering explains why a local contraction statement is not an unconditional global convergence guarantee across topology changes. Thus damping, replay gain, and projection have distinct logical jobs rather than being interchangeable tuning parameters.

Section 3.5: quantify why the stability conditions can be restrictive. The local 3D Poisson model gives the spacing law $\mathbb{E}[m] \approx r_k/(3k)$, and the crossing-rate model relates flapping to trajectory speed and near-tie density. These are model-based quantitative explanations of why even small coordinate drift can compete with the neighbour margin. Their role is to motivate measurement and calibration of the gate, not to replace the exact certificates. The comparison with measured margins is made in Section 4.3.

Sections 3.6 and 3.7: turn the preceding conditions into an executable schedule. The three levels act on different objects: L_1 refreshes geometry on frozen edges, L_2 repairs local neighbour sets, and L_3 changes the basis through damped, projected replay. The procedures connect these choices to Pass-1 recording, offline planning, and Pass-2 consumption. Measured displacements, clamp flags, neighbour closure of the risk set, and the acceptance metrics make the earlier reasoning operational. The result of the chapter is therefore a specified family of algorithms with explicit assumptions and tests, not just an abstract fixed-point formulation.

2.3 Experimental chapter: which claims the measurements test

Sections 4.1 and 4.2: establish that the comparison is meaningful, then test the implementation. The family-reconstruction audit checks that the offline graph builder reproduces the production edge sets, including tie-breaking. The matched-conditions protocol holds the sample, seed, MSA depth, and block schedule fixed. Only after these controls does the correctness study compare the compensation levels with PB. The observed exact agreement of L_2 , small residual of L_1 , and intermediate result of L_3 are implementation evidence on the tested boundaries, rather than a universal theorem about all samples or trajectories.

Section 4.3: check whether the theoretical regime describes the data. Sample characterisation separates non-degenerate structures from coincident-coordinate and complete-graph cases, then compares measured margins and drift with the certificates and the spacing law. This connects the mathematics to the actual operating regime: a certificate can be correct while its sufficient condition is too restrictive for a recorded trajectory.

Section 4.4: test the cost side of the original question. Wall-time measurements identify the benefit of the recorded two-pass design, while the shared cost model distinguishes inexpensive frozen-edge geometry updates from local repair work that depends on $|\Omega_t|$. This separation prevents the measured two-pass speed-up from being read as an unconditional cost guarantee for every compensation setting.

Section 4.5: test how much topology fidelity compensation can recover. Offline oracle replay evaluates the potential of compensation on recorded coordinates, and parameter sweeps expose the difference between a conservative stability radius and a radius large enough to track the observed drift. The synthetic predictor study then examines how that benefit depends on predictor quality and step saturation. These tests address the practical usefulness of the replay-and-gating mechanism; they do not assume that a noisy predictor inherits the oracle result.

Section 4.6: delimit what the experiments establish. The discussion distinguishes the edge-weight proxy from loss sensitivity, same-family runtime comparisons from same-coordinate offline replay, and topology agreement from prediction accuracy. Sample degeneracy and trajectory divergence remain relevant limitations. In particular, the reported loss comparison does not resolve whether recovering the remaining topology gap improves the training objective.

2.4 Conclusion chapter: the final claim and its boundary

Section 5 joins the two branches of the argument. The theoretical branch identifies PB as the fixed-point target of the damped coupling, supplies sufficient local non-flip certificates, derives the replay rule under its covariance model together with the $\sigma_p = \sigma_d$ feasibility boundary, quantifies edge flapping through an analytic near-tie spacing law that matches measurement to 1%, and analyses projection and switching. The experimental branch establishes that the proposed levels are implemented and can improve measured topology fidelity, while the recorded two-pass trajectory provides the reported wall-time benefit.

The final supported claim is therefore a *theory-guided and experimentally validated family of topology-pass schedules*, with identifiable conditions for stable reuse and useful compensation. It is not a proof that arbitrary discrete switching converges globally, that finite-pass compensation always reproduces PB, or that reducing $\mathcal{D}$ or $\mathcal{D}_w$ necessarily reduces prediction loss. The logical endpoint is controlled topology fidelity together with measured computational benefit; the accuracy value of the recovered edges remains the separate question identified in the conclusion.

3. ALGORITHM: FORMULATION, DERIVATIONS AND PROCEDURES

3.1 Notation and the closed loop

Let $x^t \in \mathbb{R}^{3L}$ collect the C_α coordinates at boundary t ($t = 0, \dots, T$) and let $B^t \in \mathbb{R}^{3L}$ be the *topology basis*, i.e. the coordinates used to build the graph at boundary t (B^t need not equal x^t). The block advance is

$$\Phi_t : x^{t+1} = \Phi_t(x^t, \mathcal{G}(B^t)), \quad t = 0, \dots, T-1, \quad (2)$$

with per-step raw offsets u_t and projection $\Pi(u)_i = \text{clamp}_{[-s, s]}(\alpha u_i)$ ($\alpha = 0.03$, $s = 3 \text{ \AA}$). For residue i let $d_{(k)}(i)$ be its k -th neighbour distance and define

$$\rho_i = d_{(k)}(i), \quad m_i = d_{(k+1)}(i) - d_{(k)}(i), \quad \kappa_i = \sigma \frac{m_i}{2}. \quad (3)$$

We call m_i the *margin* of residue i : the gap between its k -th and $(k+1)$ -th neighbour, i.e. how far that residue is from a tie at the top- k boundary. A small margin means a fragile membership, and a flip — a change of membership — is exactly the

event $m_i = 0$. Equation (2) is a *closed loop* (coordinates determine the graph, the graph determines coordinates) with three instances:

- F/topology pass: $B^t \equiv x^0$ (pure feed-forward, no feedback);
- PB: $B^t = x^t$ (fully coupled reference);
- refined (this work): $B^t = \Pi_\kappa(x^0 + \sum_{s < t} \gamma_s \Delta^s)$, an interpolation driven by *measured* Pass-1 displacements $\Delta^s = x^{s+1} - x^s$.

Topology distance is measured by

$$\mathcal{D} = 1 - \frac{|\mathcal{E} \cap \mathcal{E}'|}{|\mathcal{E} \cup \mathcal{E}'|}, \quad \mathcal{D}_w = \frac{\sum_{e \in \mathcal{E} \Delta \mathcal{E}'} w_e}{\sum_e w_e}, \quad (4)$$

with attention-mass proxy w_e (in this work $w_e = 1/(1 + \|d_e\|)$; see §4.6).

3.2 Fixed-point structure

Proposition 1. (Unique fixed point is PB) *Consider the damped coupling (a fixed-point iteration in the classical sense, cf. [12])*

$$B_{(n+1)}^t = (1 - \lambda)B_{(n)}^t + \lambda x_{(n)}^t, \quad x_{(n)}^{t+1} = \Phi_t(x_{(n)}^t, \mathcal{G}(B_{(n)}^t)). \quad (5)$$

If (5) converges to $(\bar{B}, \bar{x})$ with $\lambda > 0$, then $\bar{B}^t = \bar{x}^t$ for all t , hence the limit is the PB trajectory.

Proof. At the fixed point $B^t = (1 - \lambda)B^t + \lambda x^t$, so $\lambda(B^t - x^t) = 0$ and, since $\lambda > 0$, $B^t = x^t$; then (2) coincides with the PB recursion by definition. $\square$

Proposition 1 fixes the meaning of “conservative compensation”: the basis is gradually pulled towards the live structure, and the only fixed point of that process is PB itself. Because Pass-1 already starts at $B_{(0)}^t = x^0$ and records $\{\Delta^s\}$, one iteration comes for free.

Proposition 2. (Bounded, monotone refinement) *Assume (A1) Φ_t is L_Φ -Lipschitz in x ($L_\Phi \approx 1$ for residual blocks); (A2) inside the per-residue margin ball $\|B_i^t - x_i^t\| < m_i/2$ the edge set of i is constant, so the graph contributes no extra Lipschitz growth; (A3) Π is non-expansive. Then one full sweep is a contraction with modulus $q = (1 - \lambda) + \lambda L_\Phi$ and $\|z_{(n)} - z^*\| \leq q^n \|z_{(0)} - z^*\|$ for the trajectory z , hence convergence for $\lambda < 2/(1 + L_\Phi)$ — satisfied by any $\lambda \in (0, 1)$ when $L_\Phi \approx 1$.*

Proof (sketch). Compose the damped update (5) with Φ_t : the linear part contributes $(1 - \lambda)$ and the model part λL_Φ ; the graph term is Lipschitz-free inside the margin ball by (A2), and Π does not increase distances by (A3). The composition is a contraction, so the Picard–Banach fixed-point principle [13] together with the linear-convergence argument of [14] yields the geometric bound. $\square$

Where (A2) fails, $\mathcal{G}$ is discontinuous and the map is only piecewise contractive: a trajectory can cross the same flip surface twice ($A \rightarrow B \rightarrow A$), i.e. *edge flapping*, and chattering can destroy the fixed point. This is the reason $\kappa_i = \sigma m_i/2$ is not optional: it confines every step to the region where the current graph is provably valid, so discrete changes are *discovered* rather than *manufactured*.

Proposition 3. (What refinement can and cannot reach) *The residual with respect to PB has exactly two sources: (i) genuine ties ($m_i \approx 0$, unsolvable by any rule) and (ii) pass-to-pass model/path noise. Hence $\mathcal{D}_w$ has a floor and the attainable goal is $\mathcal{D}_w(\text{refined}) \leq 1.1 \mathcal{D}_w(\text{PB})$, not 0. The provable upper bound, computable offline, is*

$$\mathcal{D}(\text{refined vs. PB}) \leq \frac{\sum_{i \in \Omega} k + |\text{ties}|}{|\mathcal{E}|} \approx \frac{|\Omega|}{L} + \frac{|\text{ties}|}{|\mathcal{E}|}, \quad (6)$$

i.e. only residues that left their margin ball can contribute.

3.3 Discrete topology and exact non-flip certificates

Each edge indicator is $1[j \in S_i(x)] = 1[d_{ij}(x) \leq d_{i,(k)}(x)]$ with $d_{ij} = \|p_i - p_j\|$. The distance is smooth but the indicator jumps on the codimension-1 manifolds $\Sigma_{i,jl} = \{x : d_{ij} = d_{il}\}$; in generic position these are smooth and of measure zero, yet trajectories cross them. Three elementary steps give the measure-zero statement.

Smoothness and a local graph. Let $g_{ijl}(x) = d_{ij}(x) - d_{il}(x)$; the norm is smooth off the origin, so g_{ijl} is smooth wherever no two atoms coincide, and $\nabla g_{ijl} \neq 0$ on the surface. Fix a point with $g_{ijl}(x^*) = 0$ and write $x = (u, v)$ with $\partial g_{ijl}/\partial v(x^*) \neq 0$; the map $F(u, v) = (u, g_{ijl}(u, v))$ then has $\det DF(x^*) \neq 0$, so the inverse function theorem [12] inverts it locally and the level set is locally a graph,

$$\Sigma_{i,jl} \cap U(x^*) = \{(u, \varphi(u)) : u \in U'\}, \quad (7)$$

i.e. the surface is parametrised by $3L - 1$ free coordinates, the remaining one being determined by φ .

Countably many graphs suffice. The neighbourhood $U(x^*)$ depends on x^* , so read literally the construction gives uncountably many pieces. Since $\mathbb{R}^{3L}$ has a countable base of open balls $\{O_k\}$, one can choose for every point of the surface a base element contained in its own neighbourhood, so that

$$\Sigma_{i,jl} \subset \bigcup_{k \in K} \{(u, \varphi_k(u)) : u \in U_k\}, \quad K \text{ countable}, \quad (8)$$

and $\mathbb{R}^{3L} = \bigcup_k \bar{B}(0, k)$ reduces everything to a compact piece.

Compactness and uniform continuity. On the compact piece $\Sigma_{i,jl} \cap \bar{B}(0, k)$ finitely many of the graphs already cover, and each domain U_ℓ can be taken closed and bounded, hence compact; by Heine–Cantor φ_ℓ is then *uniformly* continuous,

$$\forall \varepsilon > 0 \quad \exists \delta > 0 : \quad \|u - u'\|_\infty < \delta \implies |\varphi_\ell(u) - \varphi_\ell(u')| \leq \varepsilon. \quad (9)$$

Partitioning U_ℓ into finitely many cubes of side $< \delta$ places the graph inside finitely many slabs of thickness 2ε ; writing $|\cdot|$ for the $3L$ -dimensional volume (Lebesgue measure) of a set,

$$|\Sigma_{i,jl} \cap B| \leq 2\varepsilon |B| \quad \text{for every } \varepsilon > 0, \quad (10)$$

which forces $|\Sigma_{i,jl} \cap B| = 0$, i.e. the indicator vanishes almost everywhere, $\int_{\mathbb{R}^{3L}} 1_{\Sigma_{i,jl}}(x) dx = 0$; finite unions of such sets remain of measure zero.¹ Almost every configuration therefore has all margins positive — the k NN set is well defined — and a C^1 trajectory crosses the surfaces at isolated times only. Consequently the distributional derivative of the edge set vanishes off Σ : *the gradient carries no information about who is a neighbour*, and topology is a genuine discrete degree of freedom. L_1 – L_3 are therefore relaxations, not solutions.

Order statistics are Lipschitz, $|d_{i,(k)}(x + \delta) - d_{i,(k)}(x)| \leq 2\Delta$ with $\Delta = \max_j \|\delta_j\|$, but m_i — being a difference of two Lipschitz functions — has *no positive lower bound*, so flips are generic rather than a bug.

Lemma 1. (Distance perturbation bound) $|d_{ij}(x + \delta) - d_{ij}(x)| \leq \|\delta_i\| + \|\delta_j\| \leq 2\Delta$.

Proof. $||a| - |b|| \leq \|a - b\|$ (reverse triangle inequality), and $(p_i - p_j) - (q_i - q_j) = \delta_j - \delta_i$. □

Theorem 1. (Two-sided non-flip certificate) *If $4\Delta < m_i$ then $S_i(x + \delta) = S_i(x)$ edge-wise.*

Proof. For members $j \in S_i$: $d'_{ij} \leq d_{ij} + 2\Delta \leq d_{i,(k)} + 2\Delta$. For non-members $l \notin S_i$: $d'_{il} \geq d_{il} - 2\Delta \geq d_{i,(k+1)} - 2\Delta = d_{i,(k)} + m_i - 2\Delta$. If $d_{i,(k)} + 2\Delta < d_{i,(k)} + m_i - 2\Delta$, every member is closer than every non-member, so the top- k set is unchanged; the inequality is equivalent to $4\Delta < m_i$. □

Theorem 2. (One-sided certificate) *If only node i moves, $2\Delta_i < m_i$ implies $S_i(x + \delta) = S_i(x)$.*

Proof. With $\delta_j \equiv 0$ for $j \neq i$, Lemma 1 gives $|d'_{ij} - d_{ij}| \leq \Delta_i$; replacing 2Δ by Δ_i in the previous argument yields the claim. □

The design parameter appears through $\kappa_i = \sigma m_i/2$, which yields two *different* readings of the same knob:

¹Classical ingredients: the finite-covering lemma (Borel–Lebesgue/Heine–Borel) and the Cantor–Heine theorem on uniform continuity ([12], Vol. I, secs. 2.3.2 and 4.2; countable base: $\mathbb{Q}^{3L}$ is countable, sec. 2.4.1); countable unions of measure-zero sets and the graph of a continuous function ([13], Vol. II, sec. 11.1.2, whose argument (10) reproduces).

- stability (provable non-flip): the two-sided certificate needs $\kappa_i \leq m_i/4$, i.e. $\sigma \leq 1/2$ — this is the default of the design, and it is exactly the critical value;
- tracking (reproducing real flips): needs $\kappa_i \gtrsim$ the actual drift, i.e. $\sigma \approx 2\text{--}4$; no proof, but statistically optimal in the sense of §3.4.

Both statements are consistent; they optimise different objectives (Proposition 2 versus Proposition 3). Finally, the hazard form of the certificates, $\Omega_{\text{risk}}(t) = \{i : m_i(x^t) \leq 4\Delta_t\}$, with the number of flips $N_{\text{flip}} \leq |\Omega_{\text{risk}}|$, identifies Ω_t as an event-driven adaptive update set with cost $O(|\Omega|)$.

3.4 Optimal replay, trust-region projection, and stability under flips

Write the true drift as $s = x^t - x^0$ (prior: $s \approx 0$ with covariance Σ_d) and the recorded predictor as $p = \hat{z} - x^0 = s + e$, with $E[e] = 0$, $\text{Cov}(e) = \Sigma_p$ and $\text{Cov}(s, e) = 0$. For the linear estimator $B - x^0 = \gamma p$ the error is $\varepsilon(\gamma) := s - \gamma p = (I - \gamma)s - \gamma e$, which is mean-zero ($E[\varepsilon] = (I - \gamma)E[s] - \gamma E[e] = 0$), so its second moment is its covariance. Expanding the outer product and taking expectations term by term (constant matrices pass through $E[\cdot]$),

$$\text{MSE}(\gamma) = \text{tr } E[\varepsilon \varepsilon^\top] = \text{tr}[(I - \gamma)\Sigma_d(I - \gamma)^\top + \gamma\Sigma_p\gamma^\top], \quad (11)$$

where the two cross terms $-(I - \gamma)E[se^\top]\gamma^\top$ and $-\gamma E[es^\top](I - \gamma)^\top$ vanish by $\text{Cov}(s, e) = 0$, and the trace turns the error covariance into a scalar (the total mean squared error over coordinates). Completing the square in γ , with $W := \Sigma_d + \Sigma_p$,

$$\text{MSE}(\gamma) = \text{tr}[(\Sigma_d^{-1} + \Sigma_p^{-1})^{-1}] + \text{tr}[(\gamma - \gamma^*)W(\gamma - \gamma^*)^\top], \quad \gamma^* = \Sigma_d W^{-1}, \quad (12)$$

so the second term is non-negative with curvature $2W \succ 0$ and vanishes at the unique minimiser $\gamma = \gamma^*$ (equivalently at the solution of $\partial \text{MSE}/\partial \gamma = 0$), while the minimum of (11) is $\text{tr}[(\Sigma_d^{-1} + \Sigma_p^{-1})^{-1}]$. This is the Gauss–Markov/BLEU solution [15]

$$\gamma^* = \Sigma_d(\Sigma_d + \Sigma_p)^{-1}, \quad \text{Cov}(x^t - B) = (\Sigma_d^{-1} + \Sigma_p^{-1})^{-1}, \quad (13)$$

Using $I - \gamma^* = \Sigma_p W^{-1}$ together with the identity $\Sigma_p W^{-1} \Sigma_d = \Sigma_d W^{-1} \Sigma_p = (\Sigma_d^{-1} + \Sigma_p^{-1})^{-1} =: P$, the residual covariance evaluates to $\text{Cov}(x^t - B) = \Sigma_p W^{-1} P + \Sigma_d W^{-1} P = P$, i.e. to the parenthesised second entry of (13). The minimiser is equivalently characterised by the orthogonality principle

$$\text{Cov}(x^t - B, p) = \Sigma_d - \gamma^*(\Sigma_d + \Sigma_p) = 0, \quad (14)$$

the normal equation of the minimisation. In the scalar case

$$\gamma^* = \frac{\sigma_d^2}{\sigma_d^2 + \sigma_p^2}, \quad \text{residual variance} = \frac{\sigma_d^2 \sigma_p^2}{\sigma_d^2 + \sigma_p^2} \quad (15)$$

(the familiar parallel-resistance form). Three consequences matter in practice: (i) replay with γ^* is *never* worse than freezing, unless $\sigma_p = \infty$; (ii) naive replay ($\gamma = 1$) beats freezing iff $\sigma_p < \sigma_d$, i.e. iff $\gamma^* > 1/2$ — the measured feasibility crossover $\rho^* \approx 0.25$ between predictor quality and ground truth is exactly the $\sigma_p = \sigma_d$ boundary; (iii) κ must be calibrated to the predictor error, which is the quantitative explanation of the empirically optimal $\sigma \approx 2\text{--}4$.

3.4.1 Which part of the LM method is used

Of the Levenberg–Marquardt (LM) trust-region method we use one facet only: the *projection* onto the trust region. The rest of the machinery is neither used nor needed — no normal-equation solve $(J^\top J + \mu I)p = -J^\top r$, no QR factorization of an augmented least-squares problem, and no root-finding for a damping parameter — because the residual of our problem is the *discrete* edge set $1[d_{ij} \leq d_{i,(k)}]$, which admits no Jacobian.

3.4.2 The projection, and which symbol is the output

The projection used by the refined basis is

$$\Pi_\kappa(u) = \arg \min_{\|y - x^0\|_\infty \leq \kappa} \frac{1}{2} \|y - u\|^2 = \text{clamp}(u, x^0 \pm \kappa), \quad (16)$$

where the input $u = x^0 + \sum_{s < t} \gamma_s \Delta^s$ is the unconstrained replayed position and is never modified, the optimisation variable is y , and the output is the minimiser of (16) — the new basis B^t , *not* u . Coordinate-wise that output is $y_i = \min\{\max\{u_i, x_i^0 - \kappa_i\}, x_i^0 + \kappa_i\}$: a coordinate passes through unchanged when u_i already lies in $[x_i^0 - \kappa_i, x_i^0 + \kappa_i]$, and is otherwise clipped onto a face of the box, which is the regime $\Omega_t = \{i : \|B_i^t - x_i^0\| \geq \kappa_i\}$ that triggers local repair. The constraint uses the ℓ_∞ norm while the objective uses the ℓ_2 norm; this is not a typo: the *shape* of the feasible set is dictated by the per-residue certificates of Theorems 1 and 2, whereas a projection is by definition the closest point in the Euclidean sense.

3.4.3 From the ellipsoidal region of LM to the box

LM admits a *scaled* region,

$$\min_p \frac{1}{2} \|J_k p + r_k\|_2^2 \quad \text{subject to} \quad \|D_k p\| \leq \Delta_k, \quad (17)$$

with $D_k = \text{diag}(d_i)$, $d_i > 0$, because least-squares variables can differ by many orders of magnitude; the lengths of the principal axes are tied to the typical values of the corresponding variables [14]. Writing $d_i = 1/\kappa_i$ and squaring the constraint,

$$\|D_k p\|_2 \leq \Delta_k \iff \sum_i (p_i/\kappa_i)^2 \leq \Delta_k^2, \quad (18)$$

so the semi-axis along coordinate i is $\Delta_k \kappa_i$: the diagonal of D_k is the reciprocal of the per-coordinate radius. Our design folds the scalar radius into the radii ($\Delta_k \equiv 1$), giving $\kappa_i = \sigma m_i/2$; with $\sigma \leq 1/2$ this box is precisely the ball of Theorem 1. The two regions are in fact members of one family: $\{\|D_k v\|_p \leq 1\}$ is an ellipsoid with semi-axes κ_i at $p = 2$, and since $\|v\|_\infty = \lim_{p \rightarrow \infty} \|v\|_p$ it converges to the box $\{|v_i| \leq \kappa_i\}$ as $p \rightarrow \infty$. We use that end of the family, for two reasons.

Tightness. The certificate is per residue, so the feasible set is a product of intervals. An ℓ_2 ellipsoid is a strict subset of that box — also safe, but strictly more conservative, since it forbids the simultaneous use of each coordinate’s own budget; note that the two regions agree per coordinate, as a single coordinate can reach κ_i in either. The box therefore uses exactly what the certificate permits, and it makes “which residues were clipped” a well-defined set.

Cost. Projection onto a box is the closed-form clamp in (16) ($O(L)$, no iteration). Projection onto an ellipsoid has no closed form: it requires a one-dimensional root-find for $\mu \geq 0$ with

$$p(\mu) = (I + \mu D_k^\top D_k)^{-1} u, \quad \|D_k p(\mu)\|_2 = \Delta_k \quad (\mu = 0 \text{ if } u \text{ is already feasible}), \quad (19)$$

the secular equation — the same inner iteration that LM performs on its damping parameter in order to meet a prescribed radius. The box projection removes that iterate entirely, which is why it is the form we use.

3.4.4 Non-expansiveness, and composition with the contraction

A map T is *non-expansive* (NE) if $\|Tu - Tv\| \leq \|u - v\|$, and *finitely non-expansive* (FNE) if

$$\langle Tu - Tv, u - v \rangle \geq \|Tu - Tv\|^2 \iff \|Tu - Tv\|^2 \leq \|u - v\|^2 - \|(u - Tu) - (v - Tv)\|^2, \quad (20)$$

equivalently $2T - I$ is non-expansive; FNE implies NE.

Proposition 4. (The box projection is finitely non-expansive) *The map Π_κ of (16) is finitely non-expansive.*

Proof. The box is a product of intervals and Π_κ acts coordinate-wise, so it suffices to prove the scalar statement for one clamp $P(t) = \min\{\max\{t, a\}, b\}$, $a < b$, and to sum the scalar inequalities over coordinates. If t_1, t_2 are both interior, P is the identity on them so both sides equal $(t_1 - t_2)^2$. If $t_1 > b$ and $t_2 \in [a, b]$, then $Pt_1 = b$ and $Pt_2 = t_2$, hence $\langle Pt_1 - Pt_2, t_1 - t_2 \rangle = (b - t_2)(t_1 - t_2) \geq (b - t_2)^2 = |Pt_1 - Pt_2|^2$ because $t_1 \geq b$. If $t_1 > b$ and $t_2 < a$, both sides reduce to $t_1 - t_2 \geq b - a$. If both arguments are clipped onto the same endpoint, both sides vanish; the remaining cases are symmetric. $\square$

Composing the projection with the damped update of Proposition 2 leaves the contraction modulus unchanged, because only non-expansiveness is needed (the requirement stated in (A3), verified for Π_κ in Proposition 4):

$$\|z_{n+1} - z^*\| = \|\Pi_\kappa(Fz_n) - \Pi_\kappa(Fz^*)\| \leq \|Fz_n - Fz^*\| \leq q \|z_n - z^*\|, \quad q = (1 - \lambda) + \lambda L_\Phi < 1. \quad (21)$$

Two remarks complete the argument. First, the step above presupposes that the fixed point is feasible, and it is: the box is centred at x^0 , hence $\Pi_\kappa(x^0) = x^0$, which is what makes the projection compatible with Proposition 1. Second, firmness is what keeps the constraint from generating oscillations of its own: on a face two distinct points can be mapped to the same point, and by (20) the difference of their displacements is annihilated as well, so the projection cannot manufacture the chattering discussed below.

We close this subsection with the discrete side. A *flip* is a change of membership in a residue’s k NN set (equivalently, as noted above, the crossing of a surface $\{m_i(x) = 0\}$), and we call its *size* the number of residues whose neighbour lists change. Let $J = \partial\Phi_t/\partial x$ be the block Jacobian of the sweep at a fixed graph and let

$$M_\lambda = (1 - \lambda)I + \lambda J \quad (22)$$

be the corresponding *linearised one-sweep operator*, i.e. the linearisation of the damped coupling (5) about a trajectory; the spectral radius $\rho(M_\lambda) < 1$ is the linear case of Proposition 2.

A single flip is a sparse perturbation: it changes only those rows of J (hence of M_λ) that belong to the residues whose neighbour lists changed. It therefore replaces M_λ by $M_\lambda + E$ with E non-zero on a few rows only, and the resulting eigenvalue drift is governed by the Bauer–Fike theorem [16]: every eigenvalue μ of $M_\lambda + E$ satisfies $\min_i |\mu - \mu_i(M_\lambda)| \leq \kappa_p(X) \|E\|_p$, where $\kappa_p(X) = \|X\|_p \|X^{-1}\|_p$ is the condition number of the eigenvector matrix X of M_λ and $p \in \{1, 2, \infty\}$; when M_λ is symmetric this factor is one and the bound reduces to Weyl’s. A single flip is therefore harmless as long as the spectrum keeps a margin below one.

Repetition is different. The sweep is not a single map: $\mathcal{G}(B)$ is piecewise constant in the coordinates, so on each cell $C_G = \{x : \mathcal{G}(B(x)) = G\}$ the sweep acts as one fixed map Ψ_G , and it is discontinuous across cells. Inside a cell, (A2)–(A3) make Ψ_G a contraction with modulus $q = (1 - \lambda) + \lambda L_\Phi < 1$, but globally there is no single $q < 1$: this is what *piecewise* contractive means. If a trajectory crosses the same flip surface repeatedly ($A \rightarrow B \rightarrow A$, edge flapping), the sweep alternates between two contractions Ψ_A, Ψ_B whose fixed points differ, $p_A \neq p_B$. The two-step map $\Psi_A \circ \Psi_B$ is still a contraction, so the orbit does converge — but to a fixed point of neither: the even and odd iterates approach the two ends of a *two-cycle*, whose amplitude is

$$a = \|p^* - \Psi_B(p^*)\|, \quad (23)$$

where p^* is the fixed point of $\Psi_A \circ \Psi_B$; for affine scalar maps $\Psi_A(x) = qx + (1 - q)p_A$, $\Psi_B(x) = qx + (1 - q)p_B$ it evaluates to $a = \frac{1-q}{1+q} \|p_A - p_B\|$.

The orbit therefore never settles. Since the even and odd iterates have different limits,

$$\lim_{n \rightarrow \infty} x_{2n} = p^* \neq \Psi_B(p^*) = \lim_{n \rightarrow \infty} x_{2n+1}, \quad \limsup_{n \rightarrow \infty} \|x_{n+1} - x_n\| \geq a > 0, \quad (24)$$

so the sequence $\{x_n\}$ has no limit and its ω -limit set is the two-point set $\{p^*, \Psi_B(p^*)\}$: a residual that never vanishes. This is the behaviour we call *chattering*.

The picture degrades further if crossing times accumulate, i.e. if

$$\exists T^* < \infty : \quad t_1 \leq t_2 \leq \dots, \quad t_k \rightarrow T^*, \quad N(t) := \text{card}\{k : t_k \leq t\} \rightarrow \infty \quad (t \rightarrow T^*), \quad (25)$$

equivalently if the switching intensity $\int |\dot{x}| \rho_{\text{near-tie}}(0; x) dt$ is not integrable on $[0, T^*]$. This is Zeno behaviour: what fails is precisely the hypothesis that crossing times are isolated, on which the piecewise argument rests.

The design limits this rather than removing it: the projected step keeps the trajectory inside the cell in which the current graph is provably valid, so flips are *discovered* rather than *manufactured* ((16) with $\kappa_i = \sigma m_i/2$, $\sigma \leq 1/2$); a flip that does occur is repaired locally by the gate rather than by a global rebuild; and L_3 follows the flipped graph instead of holding the stale one. Flapping itself cannot be eliminated, being intrinsic to top- k graphs, so damping, gating and tracking are the only responses available — and the quantitative question becomes how often flips occur, which is answered next.

3.5 How large is a flip in practice?

Near-tie density controls everything. Modelling neighbours locally as a 3D Poisson process of density ρ , the k -th neighbour sits at the radius fixed by

$$k = \frac{4}{3} \pi \rho r_k^3, \quad \frac{dk}{dr} = 4\pi \rho r^2, \quad (26)$$

so the expected spacing of the k -th and $(k+1)$ -th neighbour is the reciprocal intensity

$$E[m] \approx \frac{1}{4\pi \rho r_k^2} = \frac{r_k}{3k}. \quad (27)$$

A flip occurs exactly when a margin vanishes, i.e. on the surfaces $\{m_i(x) = 0\}$ where the k -th and $(k+1)$ -th neighbour of i exchange order. Here $x(t)$ is the *numerical* trajectory — the interpolation of the boundary states x^t recorded by the run — i.e. a deterministic curve, not a sample path of a stochastic process; accordingly the $\mathbb{E}[\cdot]$ of (28) and (33) is never taken over a distribution for x : it averages the displacement (margin) field — equivalently the run-to-run ensemble — at a fixed curve x . For a smooth $x(t)$ the expected number N of such crossings over an interval obeys the Kac–Rice form [19],

$$\mathbb{E}[N] \approx \int |\dot{x}(t)| \rho_{\text{near-tie}}(0; x(t)) dt, \quad (28)$$

i.e. the flapping rate is proportional to trajectory speed times the near-tie density $\rho_{\text{near-tie}}(0; \cdot)$, the density of the margins m_i at level zero, and the Poisson model above is its local form. At $k = 64$ ($r_k \approx 15 \text{ \AA}$) the predicted near-tie spacing is $0.078 \text{ \AA}$, within 1% of the measured margin median $0.0787 \text{ \AA}$ (Table 3). Three corollaries follow: larger k makes flips *more* likely ($m \propto 1/k$); for $k \geq L - 1$ (e.g. $L = 51$) the graph becomes complete and “margin” degenerates, so near-complete-graph samples cannot be used to study stability; and flapping is intrinsic to top- k graphs, hence can only be damped, gated or tracked, not eliminated.

3.5.1 Local Poisson model and the spacing law

Fix a residue and model the positions of its neighbours, in coordinates centred on it, as a homogeneous Poisson point process (PPP) of intensity ρ in $\mathbb{R}^3$ (a number density; the symbol ρ_i of (3) denotes a neighbour radius, not a density): the counts in disjoint Borel sets are independent Poisson variables whose means are ρ times the volume [17], so the number $N(r)$ of points within radius r satisfies

$$\mathbb{E}[N(r)] = \frac{4}{3}\pi\rho r^3, \quad \frac{d\mathbb{E}[N(r)]}{dr} = 4\pi\rho r^2. \quad (29)$$

Defining r_k by $\mathbb{E}[N(r_k)] = k$ gives $\rho = 3k/(4\pi r_k^3)$ and hence a radial rate $4\pi\rho r_k^2 = 3k/r_k$. The radii $\{|x_j|\}$ themselves form a PPP on $(0, \infty)$ with intensity $\lambda(r) = 4\pi\rho r^2$ (the image of a PPP under a measurable map is again a PPP, carrying the push-forward intensity [17]), so conditioned on a point at radius r the residual gap $r_{k+1} - r_k$ is exponential with that local rate:

$$\Pr(r_{k+1} - r_k > t \mid r_k = r) = \exp\left[-\frac{4\pi\rho}{3}((r+t)^3 - r^3)\right], \quad \mathbb{E}[m \mid r_k = r] = \frac{1 + O(\varepsilon_r)}{4\pi\rho r^2}, \quad (30)$$

where $\varepsilon_r = (4\pi\rho r^3)^{-1} = 1/(3k)$ measures how much the rate changes over one gap: at $k = 64$, $r_k \approx 15 \text{ \AA}$ this is $\varepsilon_r \approx 0.005$, i.e. the local rate $4\pi\rho r^2$ changes by $\approx 2/(3k) \approx 1\%$ across one gap. Substituting the radial rate into (30) gives (27), $\mathbb{E}[m] \approx (4\pi\rho r_k^2)^{-1} = r_k/(3k)$: the mean spacing is the reciprocal of the local rate. The exact moments give the same value,

$$\mathbb{E}[r_k] = \left(\frac{3}{4\pi\rho}\right)^{1/3} \frac{\Gamma(k + \frac{1}{3})}{\Gamma(k)}, \quad \mathbb{E}[r_{k+1}] - \mathbb{E}[r_k] \approx \frac{d\mathbb{E}[r_k]}{dk} \approx \frac{r_k}{3k}, \quad (31)$$

so (27) is exact to $O(1/k)$.

3.5.2 Crossing counts: the Kac–Rice form

A flip of residue i is exactly a crossing of the level set $\Sigma_i = \{m_i = 0\}$ by the trajectory, and 0 is a regular value of m_i wherever the k -th and $(k+1)$ -th neighbour are distinct, so Σ_i is a C^1 hypersurface by the implicit function theorem. The coarea formula, written here in its distributional (level-set) form, $\mathcal{H}^{n-1} \llcorner \{f = u\} = |\nabla f| \delta(f - u) dx$ [18], writes the level set as a volume integral; its one-dimensional case, applied along the curve $t \mapsto x(t)$, counts crossings, and only the velocity component normal to the level set contributes:

$$\text{card}\{t \in [0, T] : m_i(x(t)) = 0\} = \int_0^T |\langle \nabla m_i(x(t)), \dot{x}(t) \rangle| \delta(m_i(x(t))) dt. \quad (32)$$

Taking expectations and exchanging dt with $\mathbb{E}$ (Tonelli), and assuming the trajectory direction to be independent of ∇m_i and locally isotropic, gives

$$\mathbb{E}[N] = \int_0^T |\dot{x}(t)| \rho_{\text{near-tie}}(0; x(t)) dt, \quad \rho_{\text{near-tie}}(0; x) := \sum_i \mathbb{E}[|\partial_n m_i(x)| \delta(m_i(x))], \quad (33)$$

where ∂_n is the derivative along $\dot{x}$; the approximation sign in (28) therefore refers only to the local model for $\rho_{\text{near-tie}}(0)$ used next. This is the classical Kac–Rice formula for the expected number of level crossings in its total-crossings form [19], which is why no factor $1/2$ appears in (28). Its hypotheses are the standard ones — a.s. C^1 paths, transversal crossings, no multiple or accumulating zeros [19] — but they are not empirical claims about x : because $x(t)$ is a numerical trajectory interpolated between finitely many recorded boundaries, C^1 smoothness and the absence of accumulating zeros are properties of the interpolation by construction, and all expectations are taken over the margin field at that fixed curve. Only transversality is substantive — a margin that grazes zero without changing sign is not a flip, and the “no multiple zeros” clause of [19] is what excludes it.

TABLE 1
WHAT EACH LEVEL TOUCHES IN (35); COSTS FROM THE SHARED COST MODEL.

Level	Modified object	Cost / property
L ₁	only geometry: $D(B^t), W(B^t)$ (edge set stays $E(x^0)$)	$O(\mathcal{E}) \approx 7k \Rightarrow$ free; removes the 0.29–0.4 Å geometry term
L ₂	edge set locally: E_t in (35), evaluated inside C_i	$O(\sum_{i \in \Omega} C_i \log C_i)$ with $ C_i \leq 1 + k + k^2$ (two-hop); Ω closed under frozen neighbours
L ₃	the basis $B^t = \Pi_\kappa(x^0 + \sum_{s < t} \gamma_s \Delta^s)$ (predictor only)	storage $O(LT)$ ($T=8: \approx 2.5k$ floats); needs $\sigma_p < \sigma_d$
L ₄	full rebuild every boundary (PB)	1.8 s/block; reference

3.5.3 Local density and a numerical check

In the local Poisson model the margin is an exponential gap, so its density at level zero is the reciprocal mean spacing,

$$\rho_{\text{near-tie}}(0) \approx \sum_i \frac{1}{\mathbb{E}[m_i]} = \sum_i \frac{3k}{r_k} = \sum_i 4\pi\rho r_k^2; \quad (34)$$

this is the “local form” referred to above, and it is the same coarea-plus-conditional-density structure that yields the expected level-set measure of a stationary isotropic Gaussian field, $\nu(u) = p(u) \mathbb{E}|\nabla g|$ [19].

Numerically, $k = 64$ and $r_k \approx 15 \text{ Å}$ give $\rho = 3k/(4\pi r_k^3) \approx 4.5 \times 10^{-3} \text{ Å}^{-3}$ — consistent with a 142-residue chain of radius of gyration $\approx 20 \text{ Å}$ — and $4\pi\rho r_k^2 \approx 12.8 \text{ Å}^{-1}$, so the model predicts $\mathbb{E}[m] \approx 0.078 \text{ Å}$ against the measured margin median 0.0787 Å of Table 3. The same density predicts the topological churn as well: taking the measured per-boundary *step* of 0.02–0.06 Å as input, the model predicts 0.26–0.78 crossings per residue per boundary, whereas the measured churn corresponds to ≈ 0.64 crossings per residue per boundary ($\approx 1\%$ of the $\approx k$ out-edges of a residue are replaced). The two agree to within the accuracy of the local model, with no free parameter.

3.6 Three compensation levels

All levels share one update at boundary t ,

$$\mathcal{G}_t = (E_t, D(B^t), W(B^t)), \quad E_t = (E(x^0) \setminus \text{rows}_{\Omega_t}(x^0)) \cup \bigcup_{i \in \Omega_t} \{(i, j) : j \in \text{kNN}_k(i; C_i; B^t)\}, \quad (35)$$

where $\text{rows}_{\Omega_t}(x^0)$ are the outgoing rows of the active residues, *rebuilt* rather than added to $E(x^0)$ (rows outside Ω_t are copied verbatim, so no stale edge survives); the k neighbours of i are searched inside the local candidate set $C_i = E(x^0)_i \cup \{j : j \in \text{kNN}(x^0)_l, l \in \text{kNN}(x^0)_i\}$ evaluated on the basis B^t , sequence-local edges ($|i - j| < k_{\min}$) are kept as in $\mathcal{G}(x^0)$, and $\Omega_t = \{i : \|B_i^t - x_i^0\| \geq \kappa_i\}$.

Two practical rules complete the design. First, the recorded deltas must be *measured displacements*, never a rate: the per-block step clamps at exactly 3.000 Å on most boundaries, so extrapolating a saturated step is invalid. Second, Ω must be closed under neighbours in *both* directions, $\Omega \leftarrow \Omega \cup N_{E(x^0)}(\Omega)$: both endpoints of every frozen edge are tested, i.e. if either end of $(i, j) \in E(x^0)$ is in Ω then so is the other (one sweep of the edge list suffices). Otherwise a residue can flip because its *neighbour* moved and the repair misses it (offline: without closure L₂ scores 0.0886 against L₃’s 0.0670 at $\sigma = 4$).

3.7 Procedures

Algorithm 1–3 give the complete procedure. Only the graph-construction step of Pass-2 differs from the current pipeline.

3.7.1 Hyper-parameters

They are fixed by a 1-D *offline* sweep, without training: start at $\sigma = 1/2$, sweep $\gamma \in \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}$ and keep the best $\mathcal{D}_w$, then sweep $\sigma \in \{\frac{1}{4}, \frac{1}{2}, 1\}$ (the basis-iteration damping $\lambda = 1/2$ is only needed if the iteration (5) is actually run; Pass-1 already supplies iteration 0).

Algorithm 1 Pass-1: record (existing pass + a few arrays)

```
1:  $x \leftarrow x^0$ 
2: for  $t = 0, \dots, T - 1$  do
3:    $G_t \leftarrow \mathcal{G}(x)$  ▷ frozen behaviour,  $B \equiv x^0$ 
4:   if  $t = 0$  then record  $m_i, \rho_i$  from  $G_0$  ▷ 1 float per residue
5:   end if
6:    $(x_{\text{new}}, o_t) \leftarrow \Phi_t(x, G_t)$  ▷ new coords and block outputs  $o_t$ 
7:    $\Delta^t \leftarrow x_{\text{new}} - x$ ;  $\text{clamp}_i^t \leftarrow (|\Delta_i^t| \approx s)$ 
8:   record  $\Delta^t$ , clamp flags;  $x \leftarrow x_{\text{new}}$ 
9: end for
```

Algorithm 2 Offline plan (cheap; may be run once after the first epoch)

```
1:  $c^0 \leftarrow x^0$ ;  $c^{s+1} \leftarrow c^s + \Delta^s$  ▷ replay trajectory
2: for  $t = 1, \dots, T$  do
3:    $B^t \leftarrow \Pi_\kappa(x^0 + \sum_{s < t} \gamma_s \Delta^s)$  ▷  $L_3$ ,  $\kappa_i = \sigma m_i / 2$ 
4:   skip  $\gamma$ -extrapolation for clamp-flagged residues
5:    $\Omega_t \leftarrow \{i : \|B_i^t - x_i^0\| \geq \kappa_i\}$ 
6:    $\Omega_t \leftarrow \Omega_t \cup \{j \in E(x^0)_i, i \in \Omega_t\}$  ▷ neighbour closure
7:    $\text{plan}_t \leftarrow \{i \in \Omega_t : C_i\}$  ▷  $L_2$  local re-selection plan
8: end for
```

3.7.2 Falsifiable predictions

- (1) $\mathcal{D}_w$ decreases with γ and then flattens at the tie/noise floor; if it does not, (A1) fails and one should fall back to L_1 only.
- (2) $|\Omega_t|$ shrinks with t as the trajectory converges (cross-epoch topology is $\approx 95\%$ stable), so L_2 's cost decays during training.
- (3) If $\mathcal{D}_w(L_1 + L_2) \approx \mathcal{D}_w(\text{PB})$, the topology question is closed and only accuracy (E2, below) remains.

4. EXPERIMENTS

4.1 Protocol and implementation status

All three levels are implemented in the C++ runtime behind one switch (PPML_SE3_TOPO=fixed|per_block|l1|l2|l3) with two hyper-parameter knobs, PPML_TOPO_SIGMA (σ) and PPML_TOPO_GAMMA (γ); the graph builder constructs a host-side basis $B^t = x^0 + \text{clamp}(\gamma(x^t - x^0), \pm\kappa)$ only for L_3 and passes it to the edge-geometry kernel, so $L_1/L_2/L_3$ are bit-identical to the previous behaviour when disabled.

Two measurement rules turned out to be essential.

- **Family-reconstruction audit.** Before any comparison, the graph builder is re-implemented offline and required to reproduce the recorded dumps exactly. Agreement initially stalled at 44% because the production builder sorts candidate neighbours by the pair (distance, index), so ties are broken by ascending index, whereas a naive `argsort` picks tied members arbitrarily. With the lexicographic rule (and the production constants $k = \min(64, L - 1)$, $k_{\min} = 9$, float32 distances) the audit passes on *all* 18 recorded boundaries (Table 2).
- **Matched-conditions (MC) protocol.** Comparing two topology modes across two *different* trajectories is meaningless: an unseeded reference inflated the measured gap of L_1 by a factor of 2–3 ($\mathcal{D} = 0.0103/0.0194$ instead of $0.0057/0.0068$). Every comparison below therefore keeps seed, MSA depth, block schedule and sample fixed and switches *only* the topology mode.

4.2 Correctness

Table 2 collects the verification. The audit establishes that the topology question is being measured against a faithful reconstruction rather than against a re-implementation artefact. The runtime block then shows a clean ordering: frozen edges (L_1) miss 0.57–0.68% of the PB edge set, the damped projected basis of L_3 with $\gamma = 1/2$ reduces this to 0.44–0.59% (a strict interpolation between L_1 and PB), and L_2 reproduces the PB edge set *exactly* ($\mathcal{D} = 0$ at both boundaries, $|\mathcal{E}|$ identical, $J = 1.0000$).

Algorithm 3 Pass-2: consume (only the graph step changes)

```
1:  $x \leftarrow x^0$ 
2: for  $t = 0, \dots, T - 1$  do
3:    $B^t \leftarrow$  basis from plan (default  $x^0$ )
4:    $E_t \leftarrow E(x^0)$  ▷ L0 edge set
5:   if  $L_2$  enabled then  $E_t \leftarrow E_t \cup \bigcup_{i \in \Omega_t} \text{kNN}_k(i; B^t)$  inside  $C_i$ 
6:   end if
7:    $G_t \leftarrow (E_t, D(B^t), W(B^t))$  ▷  $L_1$  live geometry
8:    $x \leftarrow \Phi_t(x, G_t)$ 
9: end for
10: acceptance:  $\mathcal{D}_t = 1 - J(E_t, E_t^{\text{PB}})$ ,  $\mathcal{D}_w$ ; require  $\mathcal{D}_t(\text{refined}) < \mathcal{D}_t(\text{plain})$  and  $\mathcal{D}_w \leq 1.1 \mathcal{D}_w(\text{PB})$ 
```

TABLE 2

CORRECTNESS VERIFICATION. TOP: FAMILY-RECONSTRUCTION AUDIT (OFFLINE REBUILD OF `MAKE_GRAPH` FROM THE RECORDED COORDINATES MUST REPRODUCE THE DUMPED EDGE SET EXACTLY). BOTTOM: RUNTIME VERIFICATION IN THE MATCHED-CONDITIONS PROTOCOL (P69905, $L=142$, SEED 1234, MSA DEPTH 3, 3 BOUNDARIES; $\mathcal{D}$ MEASURED AGAINST A PB REFERENCE OF THE SAME FAMILY).

Family	L	boundaries	$ \mathcal{E} $	Jaccard
P62805 (degenerate)	103	6/6	7105–7107	1.0000
P62891	51	6/6	2550 (ident.)	1.0000
P69905	142	3/3	9089 = $64L+1$	1.0000
P62241	208	3/3	13588, 13595	1.0000
Runtime mode			$\mathcal{D}$ ($t=1$ / $t=2$)	$J(t, \text{frozen})$
L_1 (frozen set, live geometry)			0.0057 / 0.0068	1.0000
L_3 ($\sigma=4$, $\gamma=0.5$)			0.0044 / 0.0059	0.9985 / 0.9980
L_2 ($\sigma=0.5$ and 4)			0.0000 / 0.0000	0.9943 / 0.9932
L_3 $\gamma=0$ (anchor)			= L_1	—
L_3 $\kappa \rightarrow \infty$ (anchor)			= L_1	—

The two degenerate anchors behave as the theory requires: $\gamma = 0$ (basis = x^0) is identical to L_1 , and $\kappa \rightarrow \infty$ (empty risk set) is identical to L_1 as well. This is the strongest available statement of implementation correctness: the compensated pipeline collapses onto PB when the trust region is removed, and onto the frozen pipeline when the compensation is switched off.

4.3 Data characterisation and drift

Table 3 characterises the three samples. Here *drift* means the displacement of the recorded coordinates at a boundary relative to the basis x^0 ; it matters because the certificate $4\Delta < m_i$ is evaluated on the same Δ , so the measured drift supplies the empirical value of Δ and decides whether $4\Delta < m_i$ can hold at all. The $L = 103$ sample is a *degenerate limiting case*: its mapping covers only 24/103 residues, 86.8% of point pairs coincide, all 103 margins are exactly zero, and its mean C_α separation is 1.63 Å, so its topology is governed by tie noise and its absolute geometry is unreliable.

The two well-mapped samples are non-degenerate ($\leq 0.02\%$ coincident pairs, margin medians 0.0787 and 0.118 Å, no zero margins). Their coordinate drift is heavy-tailed and small in the bulk: at the last recorded boundary the *mean* displacement over residues is 0.04–0.07 Å while the C_α RMSD is 0.39–0.48 Å (Table 3), so a few residues carry most of the motion. The drift is nevertheless *not* small relative to the certificate: since $4\Delta < m_i$ must hold *per residue*, the certificate would require margins of order $4\Delta \approx 0.15$ – 0.27 Å, i.e. 1.9 – $3.4\times$ the median margin of P69905 (0.0787 Å) and 1.3 – $2.3\times$ that of P62241 (0.118 Å).

The two ratios are lower bounds, and by a wide margin: the certificate’s Δ is the per-boundary maximum $\max_j \|\delta_j\|$, whereas the values substituted above are means over residues. That maximum is itself set by the ± 3 Å step clamp rather than by the drift — on P69905 it is 5.196 Å = $3\sqrt{3}$, the corner of the projection box, and identical at both recorded boundaries — so the mean understates the violation by about two orders of magnitude ($78\times$ – $113\times$ on the two samples). The sufficient condition therefore fails on essentially every residue, consistent with the $\approx 1\%$ (unweighted) edge churn between PB and F.

For $k = 64$ and $r_k \approx 15$ Å the spacing law (27) predicts a median near-tie spacing of 0.078 Å, against a measured margin median of 0.0787 Å on P69905 — a 1% agreement. Two consequences follow. First, the model is accurate enough to be used quantitatively. Second, the spacing itself is far below the scale of the motion: the typical margin (≈ 0.08 Å) is comparable to

TABLE 3

SAMPLE CHARACTERISATION. COLLISIONS: FRACTION OF POINT PAIRS CLOSER THAN 10^{-3} Å. DRIFT: C_α RMSD OF THE LAST RECORDED BOUNDARY AGAINST x^0 (IN PARENTHESES: MEAN $|\Delta d|$ OVER THAT BOUNDARY’S OWN EDGE SET).

Sample	L	collisions	mean C_α dist.	margin median	drift (Å)
P69905	142	0.01%	19.2 Å	0.0787 Å	0.479 (0.074)
P62241	208	0.014%	29.6 Å	0.118 Å	0.387 (0.043)
P62805 (deg.)	103	86.8%	1.63 Å	$\equiv 0$ (103/103)	1.026 (0.263)

TABLE 4

MEASURED OPTIMISATIONS (FROM THE ORIGINAL PIPELINE AND FROM THIS IMPLEMENTATION). COSTS FOLLOW THE SHARED COST MODEL: FULL-SET REFRESH (KNN REBUILD) ≈ 1.8 s/BLOCK, FROZEN-SET GEOMETRY REFRESH $O(|\mathcal{E}|)$ (FREE), LOCAL REBUILD $O(\sum_{i \in \Omega} |C_i| \log |C_i|)$ WITH $|C_i| \leq 1 + k + k^2$ (THE IMPLEMENTATION USES `PARTIAL_SORT`, I.E. $O(\sum_{i \in \Omega} |C_i| \log k)$), EXTRA PASS ≈ 8 –11 s/EPOCH.

Item	baseline	optimised
epoch wall time (two-pass vs PB)	34.1 s	19.4 s (1.76 $\times$)
Pass-1 stage	—	1.5 $\times$
loss (F vs PB, 3 seeds, ± 1.3)	15.91	15.49
kNN rebuild per block	1.8 s	L_1 : 0; L_2/L_3 : $O(\Omega)$
geometry refresh (frozen set)	$O(\mathcal{E})$	free (held live)

the mean displacement relative to x^0 (0.02–0.07 Å), so margins inevitably cross zero and flips are intrinsic to top- k graphs rather than an implementation defect — which is why PB and F disagree on $\approx 1\%$ (unweighted) of edges.

4.4 Optimisations

Table 4 lists the measured optimisations. The free component is the *recording* that Pass-1 adds on top of the frozen-graph pass that the pipeline already runs: one margin m_i and one kNN radius ρ_i per residue, once at $t = 0$, plus the measured displacement Δ^t and the step-clamp flags at each boundary — $O(LT)$ storage, ≈ 2.5 k floats at $T = 8$. No graph is rebuilt, no neighbour is searched, and no computed quantity changes, so the recording itself costs neither measurable time nor accuracy. What it buys is the removal of the per-boundary kNN rebuild from the critical path, because the basis no longer has to wait for the live coordinates of the previous block: an epoch falls from 34.1 s (PB) to 19.4 s (two-pass), a factor 1.76, the recording pass itself was measured at 4.25 s against 6.40 s for the previous implementation of the same pass (1.5 $\times$), and the loss stays inside the seed spread (15.91 vs. 15.49, ± 1.3).

Within the compensation family the costs separate by orders of magnitude, and each one can be derived. Write L for the residues, $T = 8$ for the boundaries, $k = \min(64, L - 1)$ for the neighbour count and $|\mathcal{E}| \approx Lk \approx 7$ k for the directed edge set. The rebuild that the design removes ($L_4 = \text{PB}$) searches k neighbours for each of the L residues at every boundary, i.e. $O(L)$ distance evaluations per residue, $O(L^2)$ per boundary and $TO(L^2)$ per epoch, measured as 1.8 s per block. The three compensation levels avoid that search in different ways. L_1 recomputes only the edge distances and weights on the frozen edge index, $O(|\mathcal{E}|) = O(Lk)$ operations per boundary; the SE3 layers need those relative positions anyway, so the level is free, and it removes the 0.29–0.4 Å error of a stale basis geometry. L_2 keeps the only surviving neighbour search and confines it to the active residues: each $i \in \Omega_t$ is re-searched inside its two-hop candidate set C_i , $|C_i| \leq 1 + k + k^2$, which costs $O(|C_i| \log |C_i|)$ per residue — $O(|C_i| \log k)$ with the `partial_sort` of the implementation — and $O(\sum_{i \in \Omega_t} |C_i| \log |C_i|)$ per boundary, worst case $O(|\Omega_t| k^2 \log k)$; at $L = 103$ the two-hop set already covers most of the chain, so the practical cost is nearer $O(|\Omega_t| L \log L)$. L_3 adds the basis alone, an elementwise damp-and-clamp of $x^0 + \sum_{s < t} \gamma_s \Delta^s$, at $O(L)$ per boundary and $O(LT)$ storage, and no further search. The ladder is therefore $O(Lk) \ll O(L) \ll O(|\Omega_t| k^2 \log k) \ll TO(L^2)$, and L_2/L_3 stay cost-neutral exactly while the risk set Ω_t stays small — the quantity that the procedures subsection predicts shrinks with t .

4.5 Topology-gap reduction

Table 5 reports the offline (oracle) replay sweep: the basis is the live trajectory ($\gamma = 1$), so the numbers are an *upper bound* on what compensation can achieve. Two conclusions are stable across samples. First, the default $\sigma = 1/2$ is too tight: since $\kappa_i = \sigma m_i/2$, it fixes the trust radius at the two-sided non-flip threshold of Theorem 1, $\kappa_{\text{med}} = m_{\text{med}}/4$ — so the gate is saturated on 94% (P69905) and 46% (P62241) of the residues ($|\Omega| = 133/142$ and $96/208$ at $t = 1$, Table 5), and only 9–11% of $\mathcal{D}_w$ is removed. Second, $\mathcal{D}_w$ improves monotonically with σ up to -61% (P69905, $\sigma = 8$) and -41% (P62241, $\sigma = 8$), i.e. the design should calibrate κ to the motion (margin) scale rather than fix σ . The limit $\sigma \rightarrow \infty$, $\gamma = 1$ gives $\mathcal{D} = 0$ exactly,

TABLE 5

OFFLINE ORACLE REPLAY ($\gamma = 1$) ON BOTH SAMPLES: WEIGHTED TOPOLOGY DISTANCE $\mathcal{D}_w$ AGAINST THE PB EDGE SET AT THE SAME BOUNDARY, AND ITS REDUCTION AGAINST THE FROZEN BASIS ($\gamma = 0, B \equiv x^0$). $|\Omega|$ IS THE RISK-SET SIZE AT $t = 1$ FOR BOTH SAMPLES, OUT OF L RESIDUES (IDENTICALLY 0 WHEN $B \equiv x^0$); THE FIVE-POINT σ SWEEP IS COMPLETE ON BOTH.

Sample	setting	κ_{med} (Å)	$\mathcal{D}_w$	reduction	$ \Omega $
P69905	frozen ($\gamma=0$)	—	0.0578	—	0
P69905	$\sigma=0.5$	0.020	0.0515	−11%	133
P69905	$\sigma=1$	0.039	0.0458	−21%	116
P69905	$\sigma=2$	0.079	0.0401	−31%	80
P69905	$\sigma=4$	0.157	0.0274	−53%	47
P69905	$\sigma=8$	0.315	0.0227	−61%	26
P62241	frozen ($\gamma=0$)	—	0.0355	—	0
P62241	$\sigma=0.5$	0.030	0.0321	−9%	96
P62241	$\sigma=1$	0.059	0.0294	−17%	56
P62241	$\sigma=2$	0.118	0.0287	−19%	40
P62241	$\sigma=4$	0.236	0.0247	−30%	22
P62241	$\sigma=8$	0.473	0.0208	−41%	15

which is the oracle anchor used in Table 2. A synthetic study confirms the mechanism and locates its feasibility boundary: with predictor/ground-truth correlation ρ , replay *hurts* for $\rho \lesssim 0.25$ (i.e. $\sigma_p > \sigma_d$, consistent with (15)) and helps by 15–27% for $\rho \geq 0.5$; at a saturated step amplitude (3 Å/block) the benefit collapses to 6%.

4.5.1 How to read the $\mathcal{D}_w$ rows

All reductions are measured against the frozen baseline ($\gamma = 0, B \equiv x^0$): $\text{reduction}(\sigma) = 1 - \mathcal{D}_w(\sigma)/\mathcal{D}_w(\text{frozen})$ is the fraction of the frozen-vs-PB weighted topology gap that is *closed*, so −41% means 41% closed and 59% left (the sign is a direction convention only). Likewise, “ $\mathcal{D}_w$ decreases” means that the refined edge set comes closer to PB in weighted mass ((4)), not that the graph becomes sparser.

The κ_{med} column converts the dimensionless σ into Å: since $\kappa_i = \sigma m_i/2$, an entry equals σ times half the margin median of the sample (0.0787 Å for P69905 and 0.118 Å for P62241, Table 3), so $\sigma = 1$ permits a median step of 0.039 Å and the five-point sweep runs from $\kappa_{\text{med}} = m_{\text{med}}/4$ ($\sigma = 1/2$) to $4m_{\text{med}}$ ($\sigma = 8$). The default $\sigma = 1/2$ therefore confines the median residue to a quarter of the margin that separates its k -th from its $(k+1)$ -th neighbour; read together with the $|\Omega|$ column, which counts the residues whose displacement reaches the trust radius, this is the quantitative content of “ $\sigma = 1/2$ is too tight” and of the margin-based calibration advocated above.

4.6 Discussion

(i) *Metric proxy*: $w_e = 1/(1 + \|d_e\|)$ approximates the attention mass; replacing it by the loss sensitivity $|L(G \cup e) - L(G)|$ would make $\mathcal{D}_w$ directly comparable to training metrics. (ii) *Sample degeneracy*: one widely used sample maps 24/103 residues, which makes margins identically zero and inflates all topology distances; conclusions about margins and absolute geometry are drawn only from the two well-mapped samples. (iii) *Complete-graph identity*: for $L \leq 65$, $k = \min(64, L - 1)$ makes the graph complete, so $J \equiv 1$ is an identity and such samples cannot test stability. (iv) *Trajectory coupling*: runtime $\mathcal{D}$ is measured against a PB reference of the same family, but each mode integrates its own trajectory, so $\mathcal{D}$ mixes topology error with trajectory divergence; the offline replay on recorded coordinates (Table 5) is the clean instrument. (v) *Accuracy gap unknown*: the existing F vs PB comparison has a seed spread (± 1.3 on a loss of ≈ 15.5) as large as the effect, so the benefit ceiling of the whole line of work is still unmeasured. (vi) *Sampling of crossings*: $x(t)$ is recorded only at the T block boundaries, so the crossings counted in (28) are a *boundary-sampled* statistic — a margin that is crossed and crossed back inside one block leaves no trace — and the measured churn is therefore a *lower bound* on the true flip rate (the near-tie model of §3.5 predicts 0.26–0.78 crossings per residue per boundary against a sampled ≈ 0.64 , consistent with this sampling bias).

5. CONCLUSION

We formulated topology-pass refinement as a damped fixed-point problem with a unique PB fixed point, proved exact non-flip certificates with two constants ($m_i/4$ two-sided, $m_i/2$ one-sided), derived the minimum-variance replay gain and the $\sigma_p = \sigma_d$ feasibility boundary, and quantified edge flapping through an analytic near-tie spacing law that matches measurement to 1%. The

implementation reproduces PB exactly when the trust region is removed and reduces to the frozen pipeline when compensation is disabled, giving a continuous and verifiable family of schedules.

The compensation ladder works as a *fidelity* mechanism (L_1 within 0.6%, L_3 interpolating, L_2 exact), and the recorded two-pass trajectory delivers a real $1.76\times$ epoch speed-up with no measurable loss gap. Because the frozen pipeline already matches 99.4% of the PB edge set, the whole compensation family can only recover the remaining 0.6%, so three follow-ups dominate further engineering: (1) the same-seed F vs PB comparison of loss, FAPE and the chi-angle error on well-mapped samples, which decides whether those edges matter at all; (2) a small-drift regime study (lower per-block cap, later epochs) that characterises how the risk set contracts during training; and (3) replacing the edge-weight proxy by a loss-sensitivity weight, which would put $\mathcal{D}_w$ on the same scale as the training objective.

REFERENCES

- [1] J. Jumper *et al.*, “Highly accurate protein structure prediction with AlphaFold,” *Nature*, vol. 596, pp. 583–589, 2021.
- [2] J. Abramson *et al.*, “Accurate structure prediction of biomolecular interactions with AlphaFold 3,” *Nature*, vol. 630, pp. 493–500, 2024.
- [3] R. Krishna *et al.*, “Generalized biomolecular modeling and design with RoseTTAFold All-Atom,” *Science*, vol. 384, no. 6693, p. eadl2528, 2024.
- [4] N. Thomas, T. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff, and P. Riley, “Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds,” *arXiv:1802.08219*, 2018.
- [5] F. Fuchs, D. Worrall, V. Fischer, and M. Welling, “SE(3)-Transformers: 3D roto-translation equivariant attention networks,” in *Proc. NeurIPS*, 2020.
- [6] E. Rossi, B. Chamberlain, F. Frasca, D. Eynard, F. Monti, and M. Bronstein, “Temporal graph networks for deep learning on dynamic graphs,” *arXiv:2006.10637*, 2020.
- [7] Y. A. Malkov and D. A. Yashunin, “Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 4, pp. 824–836, 2020.
- [8] D. A. Spielman and N. Srivastava, “Graph sparsification by effective resistances,” *SIAM J. Comput.*, vol. 40, no. 6, pp. 1913–1926, 2011.
- [9] J. Topping, F. Di Giovanni, B. P. Chamberlain, X. Dong, and M. M. Bronstein, “Understanding over-squashing and bottlenecks on graphs via curvature,” in *Proc. ICLR*, 2022.
- [10] Y. Bengio, N. Léonard, and A. Courville, “Estimating or propagating gradients through stochastic neurons for conditional computation,” *arXiv:1308.3432*, 2013.
- [11] Y. Xie, H. Dai, M. Chen, B. Dai, T. Zhao, H. Zha, W. Wei, and T. Pfister, “Differentiable top- k with optimal transport,” in *Proc. NeurIPS*, 2020.
- [12] V. A. Zorich, *Mathematical Analysis I*, 2nd ed. Berlin, Germany: Springer (Universitext), 2015, sec. 2.3.2, pp. 71–72; sec. 2.4.1, p. 74; sec. 4.2, p. 162; secs. 8.5.4 and 8.6.1, pp. 490–491, 498–499; app. B.2, pp. 569–573.
- [13] V. A. Zorich, *Mathematical Analysis II*, 2nd ed. Berlin, Germany: Springer (Universitext), 2016, sec. 9.7, pp. 34–39; sec. 10.7, pp. 97–105; sec. 11.1.2, pp. 112–113.
- [14] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. New York, NY, USA: Springer, 2006, ch. 2, pp. 10–29; ch. 4, pp. 66–100; ch. 10, pp. 262–266.
- [15] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. Englewood Cliffs, NJ, USA: Prentice Hall, 1993, ch. 6, sec. 6.3–6.5, pp. 134–139.
- [16] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD, USA: Johns Hopkins Univ. Press, 1996, sec. 7.2, p. 320.
- [17] G. Last and M. Penrose, *Lectures on the Poisson Process*. Cambridge, U.K.: Cambridge Univ. Press, 2018, sec. 3.1, p. 19; sec. 5.1, p. 38.
- [18] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, 2nd ed. Boca Raton, FL, USA: CRC Press, 2025, sec. 3.4, p. 143.
- [19] J.-M. Azaïs and M. Wschebor, *Level Sets and Extrema of Random Processes and Fields*. Hoboken, NJ, USA: Wiley, 2009, sec. 3.1, pp. 68–79; sec. 6.2, pp. 177–181.