Beyond Spread: Distinct Organizational Pathways in Predictive JEPA

Anonymous authors
Paper under double-blind review

Abstract

Existing analyses of geometric regularization in predictive JEPA often emphasize endpoint properties such as feature spread, isotropy, or collapse avoidance. We argue that these endpoint summaries are too coarse to explain how different regularizers shape predictive representations during training. Motivated by observations from a shell-based neighborhood regularization strategy, we compare shell-based neighborhood regularization (`shell_context_knn`), whitening decorrelation (`whitening_decorrelation`) (Ermolov et al., 2021), VICReg-style regularization (`vicreg_like`) (Bardes et al., 2022), and a repaired SIGReg-like Gaussian matching configuration (`sigreg_like`) (Balestriero & LeCun, 2025) in a controlled PatchJEPA setting (Assran et al., 2023). We observe consistent differences in linear probe accuracy, effective rank, and trajectory dynamics across methods. In particular, shell and whitening reach similar endpoint probe performance through visibly different training trajectories, while SIGReg exhibits stronger context-prediction coupling but weaker structural expansion and lower final probe. We further show that these pathway differences remain visible in a joint rank-coupling state space, persist under low-cost hyperparameter perturbations, and respond sharply to mid-training intervention. When the regularizer is switched at step 100, context-prediction coupling rapidly follows the new regularizer style, whereas final probe remains only partially plastic. These patterns motivate an organizational view of regularization in which fast coupling variables and slower predictive-structural variables separate in time. In our controlled setting, these pathways provide a diagnostic observational framework that complements spread-based summaries for understanding downstream behavior.

1 Introduction

Regularization in predictive JEPA is often described in terms of endpoint geometry: spread, isotropy, whitening, variance preservation, or collapse avoidance (Assran et al., 2023; Ermolov et al., 2021; Bardes et al., 2022). This language is useful, but it leaves an important question unresolved: why can methods with similar endpoint statistics still arise through very different training histories?

Motivated by observations from a shell-based neighborhood regularization strategy, we study this question in a controlled PatchJEPA setting. Shell and whitening can end with very similar downstream probe performance while following clearly different training trajectories. This is the entry point of the paper. The point is not that shell is uniquely important, but that it exposed a broader phenomenon: comparable endpoint statistics can arise from qualitatively different formation processes.

Existing spread-based analyses mainly describe final representation states (Ansuini et al., 2019; Papyan et al., 2020). They do not describe the trajectory through which representations are formed, and therefore cannot capture whether predictive utility emerges through fast early coupling, gradual structural expansion, or a mismatch between the two. In that sense, spread is not wrong; it is incomplete. Here, organization does not refer to a single geometric property, but to the evolving coordination between representation structure and predictive utility.

We therefore shift the object of analysis from spread alone to *organizational pathways*. We use this term to describe the joint evolution of three coupled observables during training:

1. downstream predictive outcome, measured by linear probe accuracy;
2. structural expansion, measured by effective rank;
3. diagnostic coupling dynamics, measured by context-prediction alignment statistics.

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.

This paper focuses on formation-level organizational dynamics, not on identifying a unique underlying quantity.

Our central empirical finding is the following:

> Different regularizers do not merely differ in performance, but induce qualitatively distinct organizational pathways toward predictive regimes.

This claim is intentionally narrower than a full mechanistic theory. We do not claim a unique optimal regularizer, a unique measure of organization, or a universal account of predictive learning. Instead, we show that spread-centric summaries are insufficient in our controlled setting, and that a pathway-level description is more faithful to the observed training dynamics (Achille et al., 2019). As the paper develops, we also show that parts of these pathways are directly intervention-sensitive: coupling can be redirected quickly by changing the regularizer mid-training, while final predictive quality remains only partially plastic. Shell is therefore positioned in this paper as a representative context-conditioned family, and as the motivating entry point of the investigation, rather than as the sole object of explanation.

The rest of the paper is organized as follows. Section 2 positions the paper relative to JEPA, spread-centric regularization, and organization-oriented analysis. Section 3 defines the setup and canonical method naming. Section 4 defines the predictive metrics, structural proxies, and diagnostic observables used throughout the paper. Section 5 presents the main results in five layers: predictive outcomes, structural geometry, formation trajectories, pathway geometry, and hyperparameter continuity. Section 6 discusses interpretation, scope, and limitations.

2 Related Work and Positioning

Spread-centric regularization. Many regularizers in self-supervised learning are motivated through dispersion, decorrelation, or anti-collapse language. Whitening-based objectives (Ermolov et al., 2021) and VICReg-style penalties (Bardes et al., 2022) are standard examples, while Gaussian matching variants (Balestriero & LeCun, 2025) can also be interpreted through this lens. These methods are often compared through endpoint representation quality or downstream performance. Our paper does not primarily ask which family is globally best. Instead, it asks how different regularization families organize predictive representations over training.

JEPA-style predictive learning. JEPA-style methods are organized around latent prediction rather than direct discrimination (Assran et al., 2023). They are therefore often interpreted through representation quality, prediction loss, or collapse avoidance. Our perspective is different: we treat predictive JEPA as a setting in which training dynamics themselves become an object of study. Rather than focusing only on whether a representation is well behaved at the endpoint, we analyze how different regularizers induce different formation paths.

Pathways rather than winners. The contribution of this paper is not a benchmark-style statement that one regularizer strictly dominates the others. The stronger contribution is a dynamics decomposition of predictive JEPA. We compare regularization families through the pathways they induce, and use endpoint probe, effective rank, and trajectory evidence as the empirical anchor for that comparison.

Gap summary. Existing studies have provided valuable explanations from distributional, geometric, and predictive perspectives (Ansuini et al., 2019; Papyan et al., 2020). However, these analyses primarily characterize what representations become, rather than how they form. Our work complements these perspectives by studying regularization-induced formation trajectories in predictive JEPA.

3 Setup and Canonical Methods

3.1 Controlled PatchJEPA Setting

We work in a controlled PatchJEPA setting (Assran et al., 2023) in which a context encoder processes visible patches, a target encoder supplies the prediction target, and a predictor maps context-derived features to target-space predictions. Regularization is applied on the predictive branch so that different methods can be compared under a shared predictive objective.

The experiments reported here are drawn from two closely related stages. First, an exploratory regularizer comparison establishes the broader landscape across shell, whitening, VICReg-like, and SIGReg-like variants. Second, a focused pathway-comparison experiment studies three representative lines in detail: shell, whitening, and SIGReg.

3.2 Method Canonical Naming

Table 1 fixes the canonical naming used in this paper. This table is included to avoid ambiguity between implementation names, regularization families, and paper-level terminology.

Table 1: Canonical naming used in this paper. The pathway analysis focuses on representative families rather than treating the work as a broad regularizer survey.

Paper name	Code identifier	Role in this paper
Shell-based neighborhood regularization	`shell_context_knn`	Context-conditioned family; main pathway line
Whitening decorrelation	`whitening_decorrelation`	Decorrelation-based family; strong baseline and main pathway line
VICReg-style regularization	`vicreg_like`	Decorrelation-related comparison method; supports family-level grouping
SIGReg-like Gaussian matching	`sigreg_like`	Dispersion / Gaussian-matching family; main pathway line

The paper-level concept figure operates at the level of *families* rather than individual methods. This is deliberate: the paper is about organizational pathways in predictive learning, not about maximizing the number of baselines shown in one figure.

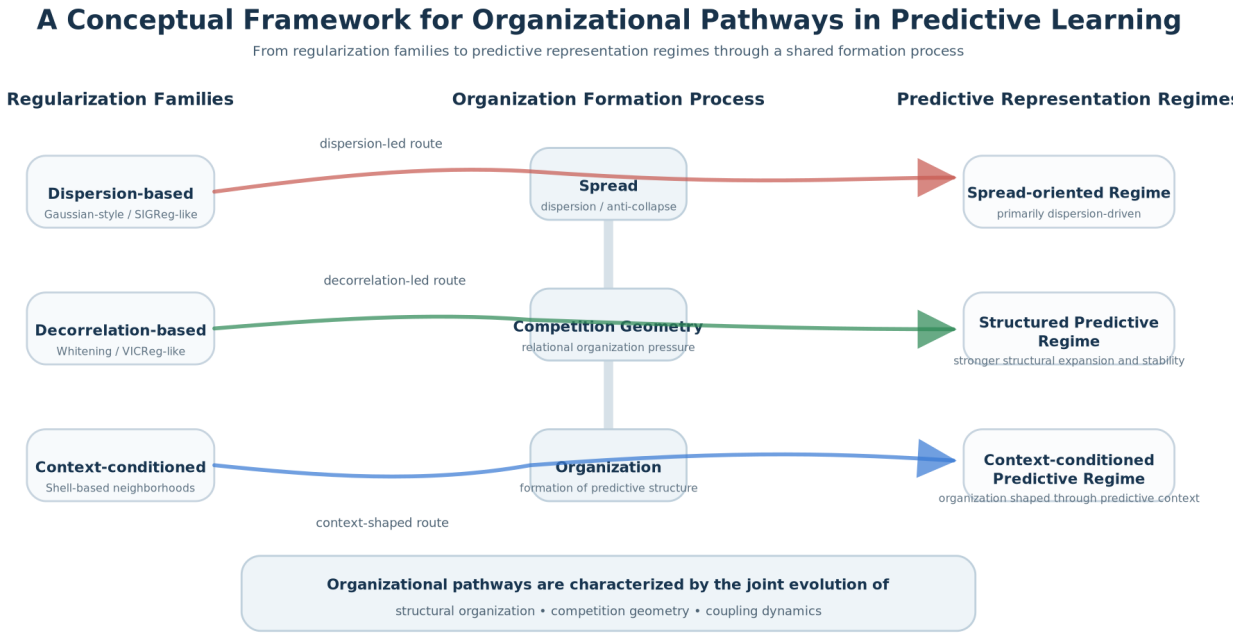


Figure 1: A conceptual framework for organizational pathways in predictive learning. Rather than relating regularizers directly to endpoint performance, we view different regularization families as inducing distinct organizational pathways that progressively shape predictive representations into different structural regimes. The shared formation process is illustrated as a progression from spread to competition geometry and then to organization, while different regularization families traverse this process differently.

4 Metrics and Organizational Proxies

4.1 Prediction Metrics

We use linear probe accuracy as the primary downstream predictive outcome and predictive loss as the training-side performance summary. Probe is the main endpoint metric in the current paper because it tracks downstream readout quality while remaining compatible with our controlled predictive setting.

4.2 Geometry Metrics

We use effective rank as the primary structural proxy. Effective rank measures whether representation dimensions are being used in a concentrated or expanded way, and therefore serves as a compact summary of structural expansion beyond a single spread statistic (Ansuini et al., 2019).

4.3 Coupling Diagnostics

We use `ctx_pred_paircorr` as a diagnostic measure of context-prediction coupling, and trajectory views of this quantity to compare how different methods build predictive structure over training. We treat coupling as a diagnostic signal rather than as a direct causal quantity. In particular, strong coupling is not assumed to be sufficient for strong downstream behavior (Assran et al., 2023).

4.4 Operational Definition of Organizational Pathways

Throughout the paper, we use "organizational pathways" to refer to the joint evolution of:

1. probe accuracy,
2. effective-rank structure,
3. and context-prediction coupling dynamics.

We treat these metrics as proxies rather than direct measures of underlying mechanisms. We emphasize that our goal is not to define a unique measure of organization, but to compare how different training regimes affect these coupled observables. In practice, we regard two pathways as meaningfully distinct when the early coupling trajectories differ by roughly 0.2 or more over the first 30% of training, which corresponds to a substantial shift relative to the full correlation range $[-1, 1]$, and when effective-rank trajectories remain separated by about 1.0 or more over a sustained part of training, which is a substantial portion of the observed rank scale in our focused setting. These are preliminary empirical criteria rather than formal thresholds.

4.5 A Geometric Rationale for Pathway Differences

We do not claim a theorem for the full nonlinear PatchJEPA dynamics, but the observed pathway differences admit a useful geometric rationale. Let $z \in \mathbb{R}^d$ denote predictive features with covariance $\Sigma = \mathbb{E}[(z-\mu)(z-\mu)^\top]$. A whitening-style regularizer can be idealized as

$$\mathcal{L}_{\text{white}} = \|\Sigma - I\|_F^2.$$

Its gradient is global: it directly suppresses covariance anisotropy and pushes the representation toward a decorrelated, isotropic state. This type of pressure naturally acts on dominant correlated directions early, which is consistent with the rapid rise of coupling observed under whitening and with the immediate coupling jump after the shell-to-whitening intervention.

By contrast, a shell-style regularizer is local rather than global. At a stylized level, it acts through neighborhood relations,

$$\mathcal{L}_{\text{shell}} \approx \sum_i \phi\big(\|z_i - z_{\pi_k(i)}\|_2\big),$$

where $\pi_k(i)$ denotes a neighborhood-defined partner index and ϕ is a distance-shaping potential. The resulting gradient depends on local geometry instead of the full covariance matrix. This makes the pressure more incremental: structure expands through repeated local rearrangement rather than immediate global decorrelation. Such a mechanism is consistent with slower coupling rise but sustained rank growth.

This contrast provides a compact rationale for the qualitative regimes observed in the experiments. Whitening behaves like a "fast align, then relax" regularizer because the covariance constraint acts globally and immediately. Shell behaves more like "slow structure, steady organize" because its neighborhood pressure accumulates through local geometric updates. The switch intervention is informative precisely because it changes the active gradient geometry. Coupling responds rapidly because it is sensitive to the currently active alignment pressure, whereas mature rank and final probe reflect slower accumulation of structural

change over many updates. This rationale is geometric rather than formal, but it supplies a theoretical spine for the empirical fast-variable / slow-variable interpretation.

5 Results

The results are organized from endpoint to process. We first show that the considered methods differ in predictive outcome. We then show that representation geometry alone is insufficient to summarize those differences. Finally, we show that the strongest evidence lies in the trajectories themselves: different regularizers follow distinct organizational pathways during training (Achille et al., 2019).

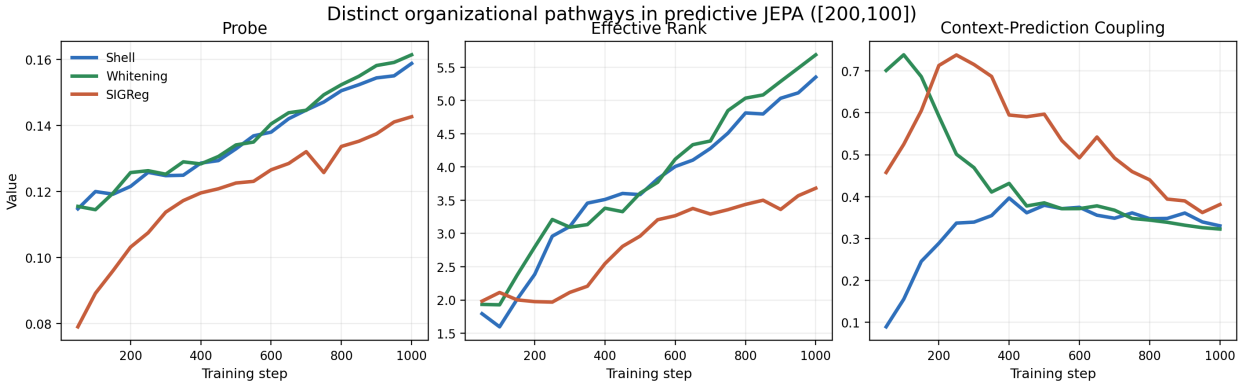


Figure 2: Different regularizers can achieve similar endpoint performance while following distinct formation trajectories. We compare representative shell, whitening, and SIGReg lines at regime $[200, 100]$ using probe, effective rank, and context-prediction coupling. Shell and whitening finish with similar endpoint probe performance but follow visibly different paths, while SIGReg reaches a different regime characterized by stronger early coupling and weaker structural expansion.

5.1 Different Regularizers Achieve Distinct Predictive Outcomes

At the focused pathway-comparison setting $[200, 100]$, whitening reaches the strongest final probe (0.1614), shell follows closely (0.1588), and SIGReg remains lower (0.1427). This already separates the three representative lines into distinct predictive outcomes, but the gap is not large enough to support a purely winner-take-all interpretation. Shell and whitening are endpoint-near yet not identical, while SIGReg is clearly not a trivial failure baseline.

This point is important for interpretation. If SIGReg had immediately collapsed, the comparison would have remained ambiguous. If it had simply matched whitening and shell on every axis, there would be little reason to ask about pathway differences. The actual pattern is more informative: SIGReg is competitive enough to matter, yet still settles into a different predictive regime.

Table 2: Endpoint comparison in the focused pathway setting $[200, 100]$. Shell and whitening are probe-near, while SIGReg remains lower but non-trivial.

Method	Probe	Effective rank	`ctx_pred_paircorr`
Shell-based neighborhood regularization	0.1588	5.3509	0.3292
Whitening decorrelation	0.1614	5.6865	0.3307
SIGReg-like Gaussian matching	0.1427	3.6810	0.3731

5.2 Endpoint Geometry Alone Does Not Capture These Differences

Effective rank clearly differentiates the methods, but it does not by itself explain the whole picture. Whitening and shell both end in a high-rank regime, with whitening slightly higher (5.6865 vs. 5.3509), while SIGReg remains structurally weaker (3.6810). This supports the claim that structural expansion matters. However, endpoint geometry alone still misses an important mismatch.

The most useful failure case is SIGReg. At the endpoint, SIGReg shows the strongest context-prediction pair correlation among the three representative methods (0.3731), yet its probe and rank remain lower. This

means that coupling cannot be reified as a single causal explanation. A method can display strong coupling diagnostics without reaching the same structural regime or predictive utility as shell or whitening.

This is exactly why the paper does not claim that one observable, such as coupling, is "the mechanism." Probe is the downstream outcome, effective rank is the structural proxy, and coupling is a diagnostic signal. The contribution of the paper is the organization of these signals into a pathway-level account, not the elevation of any single one into a universal quantity.

5.3 Training Trajectories Reveal Distinct Organizational Pathways

The strongest evidence comes from training trajectories, summarized in Figure 2. Although shell and whitening finish with similar endpoint probe and similar endpoint context-prediction correlation, they do not get there in the same way.

Whitening begins with very strong context-prediction coupling, with `ctx_pred_paircorr` near 0.70 at early steps, and then progressively relaxes while effective rank expands from roughly 1.93 to 5.69. Shell shows the opposite broad pattern: it starts with much weaker coupling, near 0.09, and gradually builds toward a moderate organized regime while rank expands from roughly 1.79 to 5.35. The endpoint similarity therefore hides a large trajectory difference. Existing endpoint summaries would largely miss this distinction.

Fairness-repaired SIGReg follows a third path. Its coupling signal rises rapidly and peaks strongly early in training, but structural expansion lags behind the whitening and shell lines for most of training, and the final probe remains lower. In other words, SIGReg does not simply fail, nor does it match the other methods through the same route. It follows a distinct trajectory characterized by strong coupling but weaker structural expansion.

These patterns support the central paper-level statement: different regularizers do not merely alter spread, but induce different organizational pathways toward predictive regimes. The pathway comparison is therefore not a visual supplement to the endpoint table. It is the main empirical reason to prefer a pathway-centered interpretation over a spread-centered one. Put differently, Figure 2 is not just a trajectory visualization; it is the empirical anchor for the paper's claim that formation process carries explanatory information beyond endpoint spread summaries.

The robustness picture is also consistent across multiple random seeds. The seed-banded trajectory views show that the main pathway differences are not driven by a single lucky run: whitening still exhibits a high-coupling early phase, shell still organizes more gradually from low initial coupling, and SIGReg still remains structurally compressed relative to the other two lines. This matters because it upgrades the pathway claim from a single-trajectory observation to a family-level pattern.

5.4 Pathway Geometry and Trajectory Separation

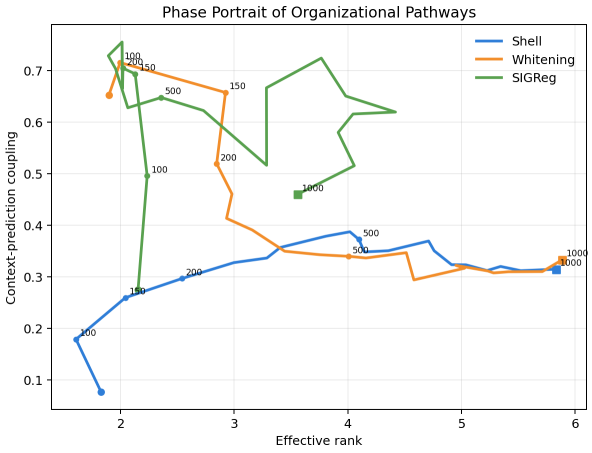
Trajectory differences become even clearer when viewed in a joint state space rather than only as metric-versus-time curves. Figure 3 plots the three focused pathways in the plane spanned by effective rank and `ctx_pred_paircorr`. This view highlights the qualitative difference between a whitening-style "fast align, then relax" regime, a shell-style "slow structure, steady organize" regime, and a SIGReg-style "early saturate, late lag" regime. These labels are descriptive rather than causal, but they make the geometric differences easier to retain.

The phase portrait shows that shell, whitening, and SIGReg do not merely end at different points; they occupy different regions of pathway space while training unfolds. Whitening moves rapidly into a high-coupling region and then relaxes while rank expands. Shell remains much lower in coupling at the start and climbs gradually as rank increases. SIGReg reaches strong coupling early but remains compressed in rank relative to the other two methods.

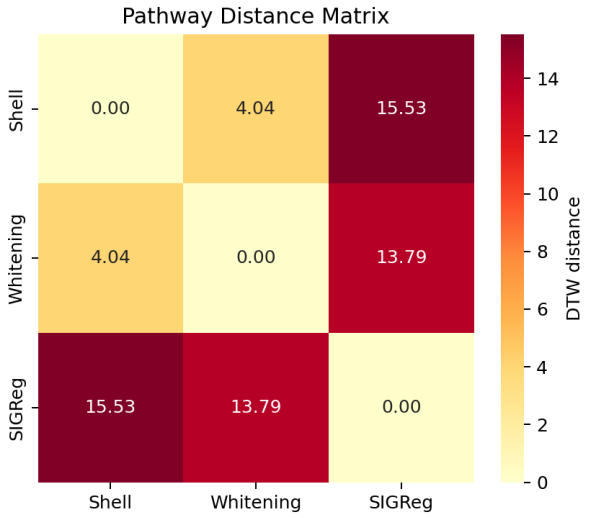
To quantify this separation, we compute dynamic time warping (DTW) distances over the 2D pathway $(\text{effective-rank}_t, \texttt{ctx_pred_paircorr}_t)$. The resulting distances are approximately 4.04 for shell versus whitening, 15.53 for shell versus SIGReg, and 13.79 for whitening versus SIGReg. These values support the visual interpretation that SIGReg occupies a substantially more separated pathway region, while shell and whitening are closer in pathway space but still clearly non-identical.

5.5 Intervening on the Pathway Reveals Fast and Slow Variables

The observational pathway analysis shows that different regularizers follow different routes, but it does not yet tell us how plastic those routes are. We therefore add a direct intervention experiment at the focused



(a) Phase portrait of the focused pathways in the joint state space of effective rank and `ctx_pred_paircorr`. This view makes the organizational differences more geometric: whitening enters a high-coupling region early and later relaxes, shell rises more gradually while expanding in rank, and SIGReg saturates coupling earlier while remaining structurally compressed.



(b) DTW distance matrix over the 2D pathway (effective-rank$_t$, `ctx_pred_paircorr`$_t$). The matrix turns qualitative trajectory differences into a simple quantitative separation result: shell and whitening are pathway-near but not identical, whereas SIGReg is much farther from both.

Figure 3: Pathway geometry and trajectory separation. (a) Phase portrait; (b) DTW distance matrix.

regime [200, 100]: at step 100, we either keep the regularizer unchanged or switch between shell and whitening. The resulting trajectories are shown in Figure 4.

The main finding is asymmetric in timescale rather than in direction. Context-prediction coupling reacts extremely quickly to intervention, as seen in Figure 4, panel A. When we switch from shell to whitening, mean `ctx_pred_paircorr` rises from about 0.123 at step 100 to about 0.679 at step 150, a jump of roughly +0.556. Conversely, when we switch from whitening to shell, mean coupling falls from about 0.704 to about 0.184, a drop of roughly −0.519. The matching control runs do not show anything comparably abrupt: shell-to-shell rises only from about 0.123 to 0.201, and whitening-to-whitening remains high over the same window.

This sharp redirection indicates that coupling is not simply a frozen imprint of the early regime. It behaves as a fast variable that can rapidly adopt the local style imposed by the active regularizer. The shaded bands in Figure 4, panel A show that this plasticity is consistent across random seeds rather than being a single lucky trajectory. At the same time, endpoint probe shows a much weaker degree of plasticity, summarized in panel B. The final probe values remain ordered within a narrow range: about 0.158 for shell-to-shell, 0.161 for whitening-to-whitening, 0.155 for shell-to-whitening, and 0.155 for whitening-to-shell. Thus, changing the regularizer mid-training strongly alters the coupling pathway, but only partially alters final predictive quality.

This is a stronger result than either complete stickiness or complete plasticity. It suggests a separation between a fast diagnostic variable and slower predictive-structural variables. Early organization does not rigidly lock later coupling dynamics, but it still leaves a residual imprint on the final predictive regime. Panel C visualizes the same result in state-space form across the full trajectory: after intervention, the pathways redirect quickly in coupling while still moving toward different final effective-rank regions. This separation gives the pathway view a more intervention-sensitive interpretation: pathways are not merely descriptive traces, but partially controllable training objects whose components adapt on different timescales.

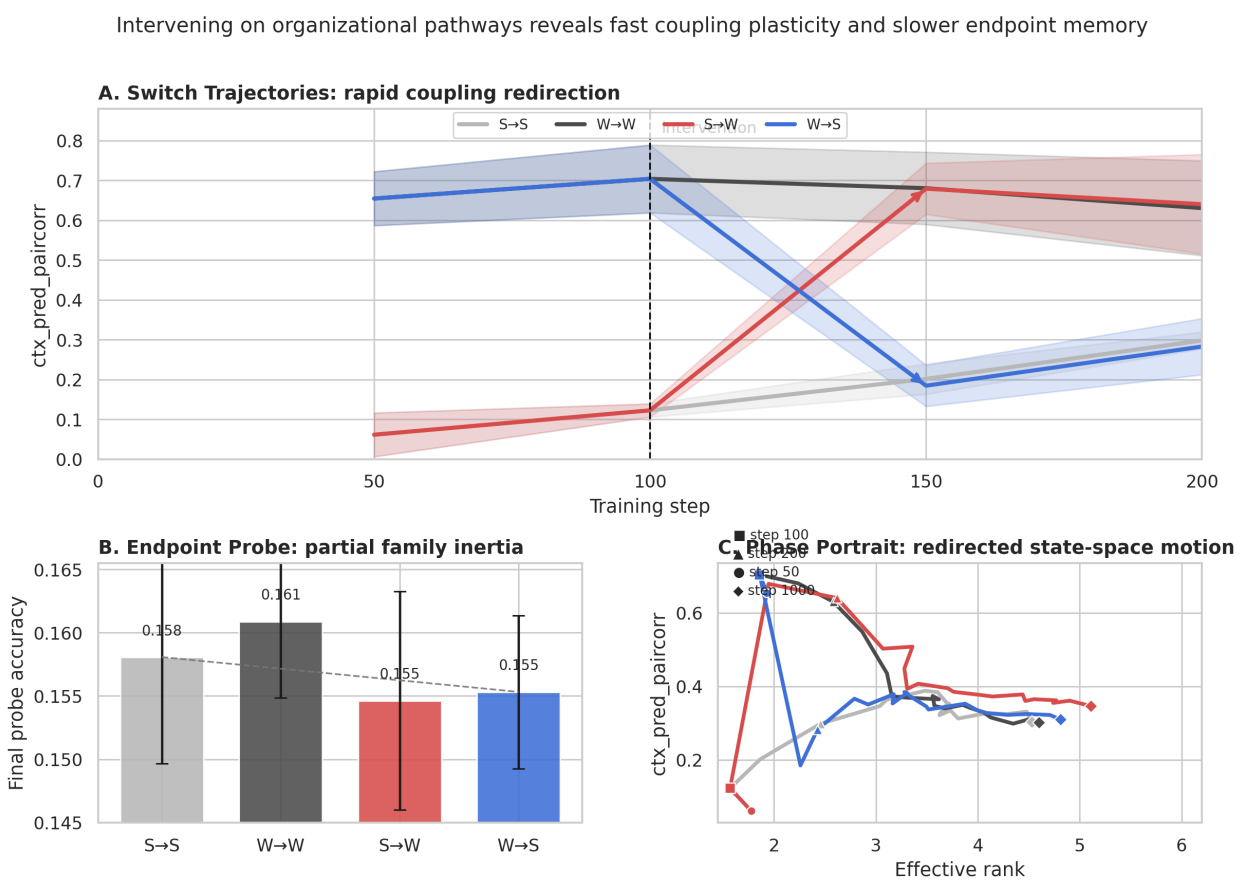


Figure 4: Intervention response under mid-training regularizer switching at step 100. Panel A shows the rapid redirection of `ctx_pred_paircorr`: shell-to-whitening sharply jumps upward, while whitening-to-shell sharply drops, with shaded seed bands indicating that the effect is consistent across multiple random seeds. Panel B shows that final probe accuracy retains family inertia: the switched runs do not fully recover the natural endpoint of the destination family. Panel C visualizes the redirected motion in the joint state space of effective rank and coupling over the full trajectory, including the final region above effective rank 5 when reached. Together, the panels show that coupling is highly plastic under intervention, whereas final predictive quality is only partially plastic.

5.6 Hyperparameter Continuity Adds Robustness Rather Than Noise

If organizational pathways were only a fragile artifact of one exact setting, small hyperparameter changes would be expected to destroy the family-level picture. We therefore add three low-cost checks at the focused regime $[200, 100]$: a regularization-strength scan, a shell neighborhood-size scan, and a whitening schedule check. The results (detailed in Appendix A) confirm that the family-level distinction remains intact under moderate perturbations.

6 Discussion

The paper's main claim should be read at the formation level. We do not argue that any one proxy uniquely identifies mechanism, and we do not claim that one regularizer is universally optimal. Instead, the evidence supports a narrower but more stable statement: in our controlled setting, organizational pathways provide a more informative description of downstream behavior than spread-based summaries alone.

The newly added pathway-geometry, multi-seed, hyperparameter-continuity, and switching-intervention checks sharpen this interpretation. They show that the pathway claim is not merely a matter of eyeballing one figure. Different methods separate in a joint state space, remain measurably different under trajectory-distance analysis, retain family-level differences under modest hyperparameter changes, and re-

spond systematically to direct intervention. This makes the pathway view closer to a diagnostic framework than to an anecdotal visualization.

The switching result is especially informative because it reveals a timescale separation. Coupling behaves as a fast variable: it can be redirected sharply within a short window after the regularizer changes. Final predictive quality behaves more like a slower variable: it moves, but does not immediately forget the earlier pathway. This pattern is more subtle than a binary "sticky versus plastic" story and therefore more valuable. It suggests that organizational pathways should be interpreted as multi-timescale objects rather than as monolithic curves with a single notion of inertia.

This interpretation also clarifies the role of each method family. Shell is not presented as the sole effective route. Whitening shows that strong predictive outcomes can emerge from a decorrelation-based route. SIGReg shows that spread-oriented or Gaussian-matching pressure is not simply invalid, but can organize prediction differently and more weakly in the present setting. VICReg-like results, while not central to the focused pathway figure, support the broader grouping of whitening-style and decorrelation-oriented methods into a related family.

More broadly, the paper is best understood as a dynamics decomposition of predictive JEPA rather than as an architecture paper or a regularizer leaderboard. The question is not primarily which method wins, but how predictive representations become organized under different regularization regimes, and which parts of that organization are rapidly steerable versus more persistent. While this account is empirical rather than theorem-driven, it serves a methodological purpose: the observed pathways, phase portraits, intervention responses, and rank-coupling summaries can be used diagnostically, much like loss curves are used to diagnose optimization behavior without constituting a formal causal proof (Fort & Jastrzebski, 2019).

7 Limitations

The present evidence is intentionally controlled and therefore limited in scope. First, the strongest pathway evidence currently comes from a compact PatchJEPA setting (Assran et al., 2023) rather than a large-scale image regime. Second, the pathway comparison focuses on a small number of representative methods, which is appropriate for a concept paper but does not exhaust the design space of predictive regularization. Third, our metrics are organizational proxies rather than direct measurements of an underlying "true" mechanism (Ansuini et al., 2019). Fourth, the schedule check reported here should be read as a cheap falsification-oriented experiment, not as a definitive statement about optimal whitening schedules.

These limitations are compatible with the paper's intended contribution. The goal here is not to prove a universal theory of predictive organization, but to show that spread alone is too coarse and that pathway-sensitive analysis reveals meaningful differences missed by endpoint summaries.

8 Conclusion

We presented a pathway-centered view of regularization in predictive JEPA. Across shell-based, whitening-based, VICReg-style, and SIGReg-style regularization, the main empirical pattern is that similar endpoint behavior can arise through different formation trajectories. This matters because it suggests that spread-based endpoint summaries alone are too coarse to characterize how predictive representations form. By extending the analysis from time-series trajectories to pathway geometry, low-cost hyperparameter continuity checks, seed-banded robustness views, and direct switching interventions, we also show that these pathways function as a diagnostic observational framework rather than as a single anecdotal plot. The switching experiment further suggests a useful refinement of this view: coupling behaves as a fast pathway variable, whereas final predictive quality is only partially plastic and retains some pathway memory. At the same time, we do not claim a unique mechanism or a universally optimal regularizer; our contribution is a controlled formation-level analysis.

AI use statement

This work was prepared with the assistance of generative AI tools (e.g., large language models) for the purposes of grammar refinement, sentence rephrasing, and LaTeX formatting assistance. The research ideas, experimental design, code implementation, data analysis, and core scientific arguments were entirely developed by the authors. All AI-assisted text and content have been thoroughly reviewed, validated, and revised by the authors. The authors take full responsibility for the accuracy, originality, and integrity of all content presented in this manuscript, including any text or claims generated with the aid of AI tools.

Reproducibility statement

We provide comprehensive descriptions of the proposed methods and experimental settings (Section 3). Detailed hyperparameter configurations, including regularization weights, batch sizes, and training steps, are specified in the main text and the Appendix. The exact training logs and checkpoints used to generate the figures (including the seed-banded trajectories and intervention results) are retained in the current workspace. The source code and data required to reproduce the main empirical findings will be made publicly available upon the paper's acceptance, in accordance with ICLR's reproducibility guidelines.

Ethics statement

This work does not involve human subjects, personal data, or sensitive applications. The experiments are conducted on standard synthetic/controlled patch-level benchmarks using publicly available practices, and there are no foreseeable harmful societal impacts or ethical conflicts associated with this methodology. The authors adhere to the ICLR Code of Ethics and have no conflicts of interest to declare.

References

A. Achille, M. Rovere, and S. Soatto. Critical learning periods in deep networks. In *ICLR*, 2019.

A. Ansuini, A. Laio, J. H. Macke, and D. Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *NeurIPS*, 2019.

M. Assran et al. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, 2023.

R. Balestriero and Y. LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

A. Bardes, J. Ponce, and Y. LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR*, 2022.

A. Ermolov, A. Siarohin, E. Sangineto, and N. Sebe. Whitening for self-supervised representation learning. In *ICML*, 2021.

S. Fort and S. Jastrzebski. Understanding generalization via learning dynamics. *arXiv preprint arXiv:1906.08326*, 2019.

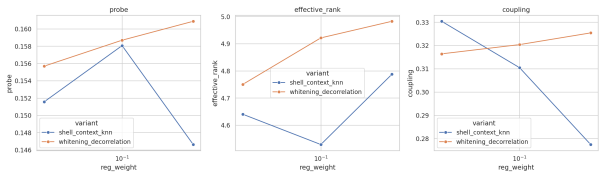
V. Papyan, X. Y. Han, and D. L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *PNAS*, 2020.

A Hyperparameter Robustness Details

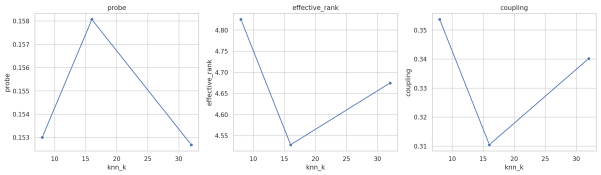
The regularization-strength scan is the most informative of the three. For shell, increasing regularization weight from 0.03 to 0.30 lowers final coupling from about 0.330 to 0.277, while increasing the very-early rank level from 1.63 to 2.39 at step 50. For whitening, increasing regularization slightly improves both final rank and final probe, from roughly 4.75 to 4.98 in rank and from 0.156 to 0.161 in probe. More importantly, the broad family-level distinction remains intact: whitening retains its early high-coupling peak, while shell continues to organize more gradually. This is consistent with a continuity view of pathways, where hyperparameters move a method within a pathway family without erasing the family-level identity.

The shell neighborhood-size scan shows a similar robustness pattern. Varying the neighborhood size over $k \in \{8, 16, 32\}$ changes the exact tradeoff between final probe, rank, and coupling, but all three settings remain recognizably shell-like. In particular, $k = 16$ gives the strongest final probe in the tested range, while $k = 8$ yields the largest final rank. The relevant point for the present paper is not which k is globally best, but that moderate neighborhood changes do not convert shell into a whitening-style trajectory.

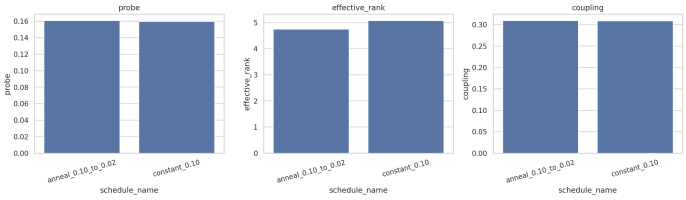
Finally, we tested a simple whitening schedule hypothesis motivated by the trajectory view: if whitening's early high-coupling phase is useful but its later relaxation is excessive, annealing the whitening strength after the first 30% of training might preserve expansion while reducing over-relaxation. In this cheap check, the annealed schedule does not support the strong version of that hypothesis. Final coupling is nearly unchanged between the constant and annealed schedules, while final rank becomes slightly lower under annealing. We include this result because it shows that the pathway framework generates falsifiable schedule ideas rather than only post hoc descriptions.



(a) Regularization-strength scan: strength changes move shell and whitening within their respective pathway families, but do not collapse the family-level distinction.



(b) Neighborhood-size scan for shell-based regularization: moderate changes in k alter the exact tradeoff, but do not overturn the broader shell-style pathway identity.



(c) Cheap schedule check for whitening: annealing does not preserve higher rank while reducing late coupling relaxation in this setting. The pathway view supplies testable hypotheses, not guaranteed improvements.

Figure A1: Low-cost hyperparameter checks confirm that organizational pathways behave as structured regions rather than isolated points. (a) Regularization-strength scan; (b) Neighborhood-size scan; (c) Whitening anneal check.

B Additional Experimental Details

B.1 Metric Definitions and Implementation Notes

We summarize here the operational definitions used throughout the main text.

Context-prediction coupling. The quantity `ctx_pred_paircorr` is computed as the correlation between the context-conditioned summary features passed to the predictor and the predictor outputs. In the paper, this observable is treated as a diagnostic measure of coupling rather than as a direct causal variable.

Effective rank. Effective rank is computed from the representation covariance spectrum as a compact proxy for dimensional utilization. We use it as a structural observable rather than as a sufficient summary of representation quality.

Trajectory distance. The DTW distances in the main text are computed over the 2D pathway

$$\big(\text{effective-rank}_t,\ \texttt{ctx_pred_paircorr}_t\big),$$

using snapshot values recorded every 50 training steps.

Switch intervention. In the switch experiment, the model parameters are not reinitialized at the intervention step. Only the active regularizer term is changed. This is important: the experiment probes pathway plasticity under intervention, not restart behavior under a new objective.

C Seed-Level Robustness

Figure A2 reports the focused pathway trajectories with seed-banded summaries at regime $[200, 100]$. The shell, whitening, and SIGReg families remain separated under seed variation, which supports the family-level interpretation adopted in the main text.

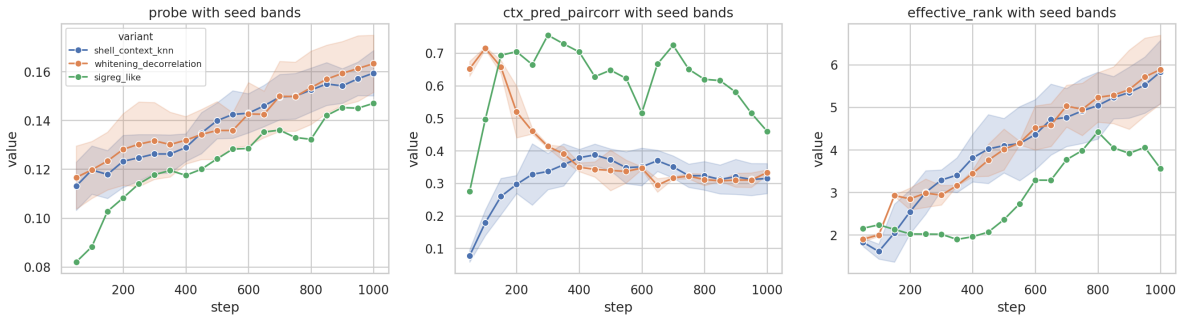


Figure A2: Focused pathways with mean trajectories and seed bands at regime $[200, 100]$. The qualitative separation between shell, whitening, and SIGReg remains visible under seed variation.

D Switch Intervention Details

Table A1 reports the main summary statistics for the regularizer-switch intervention. The most informative column is the coupling change from step 100 to step 150, which isolates the immediate post-intervention response.

Table A1: Summary of the regularizer-switch intervention at regime $[200, 100]$. Means and standard deviations are reported across 4 seeds.

Switch	c_{100}	c_{150}	$\Delta c_{100\to150}$	Final probe	Final rank
$S \to S$	0.123 ± 0.017	0.201 ± 0.038	0.079 ± 0.028	0.158 ± 0.008	4.529 ± 0.840
$S \to W$	0.123 ± 0.017	0.679 ± 0.064	0.556 ± 0.054	0.155 ± 0.009	5.108 ± 1.050
$W \to S$	0.704 ± 0.085	0.184 ± 0.052	-0.519 ± 0.077	0.155 ± 0.006	4.813 ± 0.825
$W \to W$	0.704 ± 0.085	0.680 ± 0.091	-0.024 ± 0.071	0.161 ± 0.006	4.599 ± 1.190

For convenience, Figure A3 reproduces the composite intervention figure from the main text. In the appendix it can be referenced independently when reviewers want to inspect the three panels together.

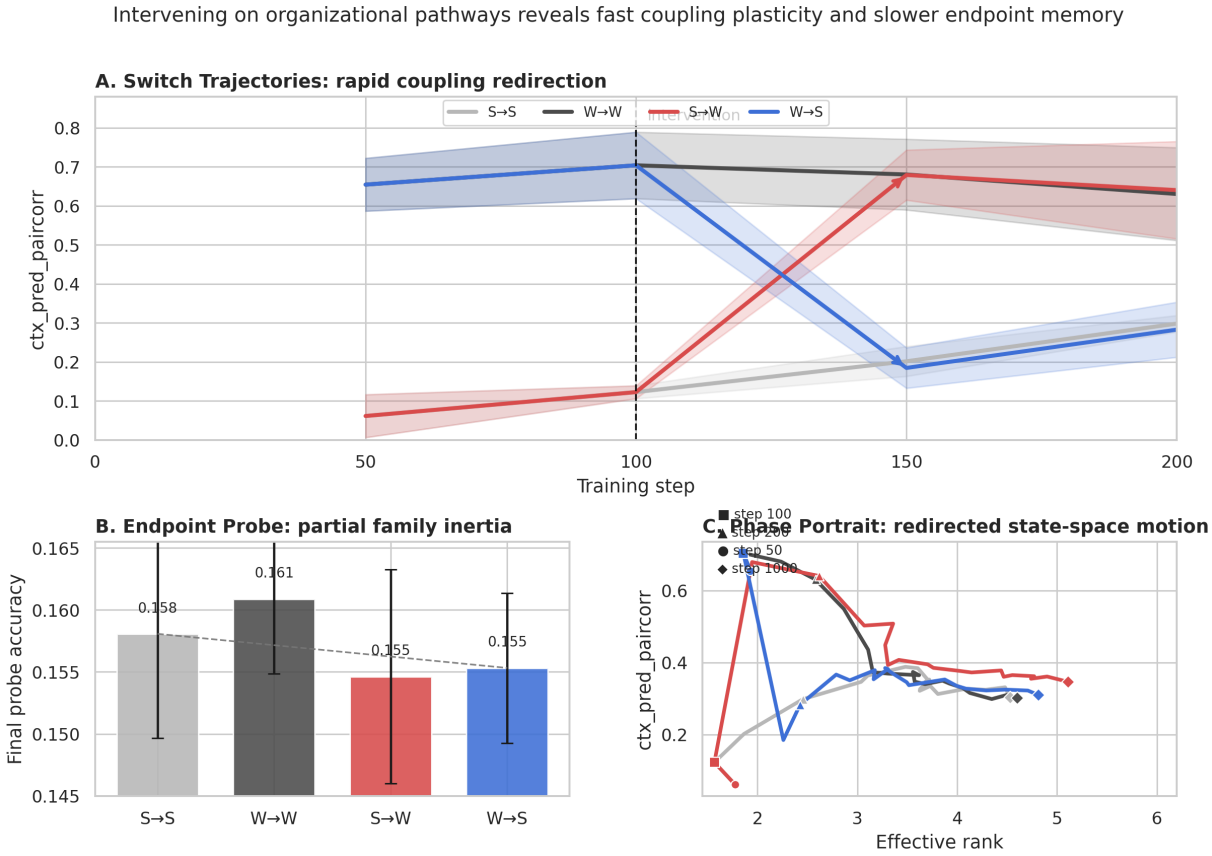


Figure A3: Composite view of the switch intervention. Panel A highlights the rapid coupling redirection after intervention, panel B shows partial inertia in final probe, and panel C shows redirected motion in the joint rank-coupling state space.

E Hyperparameter Continuity Tables

The following compact tables collect the summary values used in the hyperparameter-continuity discussion.

E.1 Regularization-Strength Scan

Table A2: Regularization-strength scan at regime $[200, 100]$.

Variant	Weight	Final probe	Final rank	Final coupling
Shell	0.03	0.152	4.640	0.330
Shell	0.10	0.158	4.529	0.310
Shell	0.30	0.147	4.787	0.277
Whitening	0.03	0.156	4.750	0.316
Whitening	0.10	0.159	4.921	0.320
Whitening	0.30	0.161	4.983	0.325

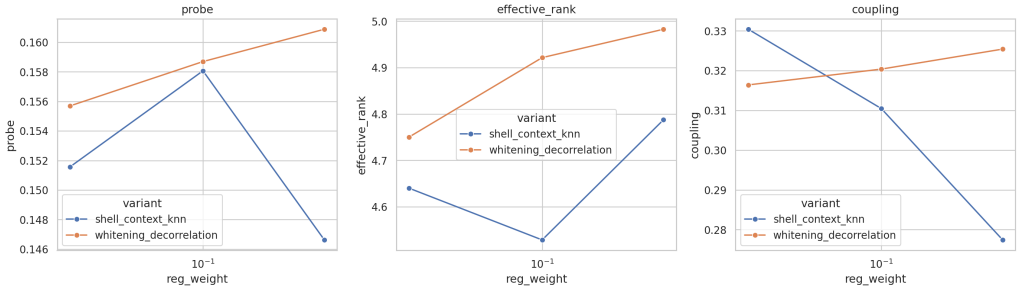


Figure A4: Summary visualization for the regularization-strength scan.

E.2 Shell Neighborhood-Size Scan

Table A3: Neighborhood-size scan for shell-based regularization at regime $[200, 100]$.

k	Final probe	Final rank	Final coupling
8	0.153	4.825	0.354
16	0.158	4.529	0.310
32	0.153	4.675	0.340

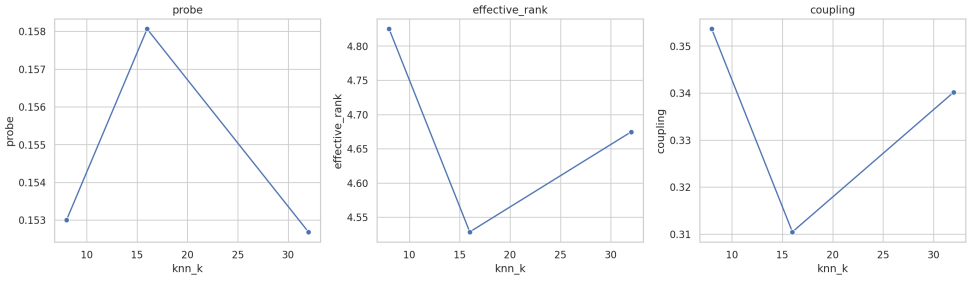


Figure A5: Summary visualization for the shell neighborhood-size scan.

E.3 Whitening Anneal Check

Table A4: Whitening schedule check at regime $[200, 100]$.

Schedule	Final probe	Final rank	Final coupling
Constant 0.10	0.159	5.074	0.309
Anneal $0.10 \rightarrow 0.02$	0.161	4.745	0.309

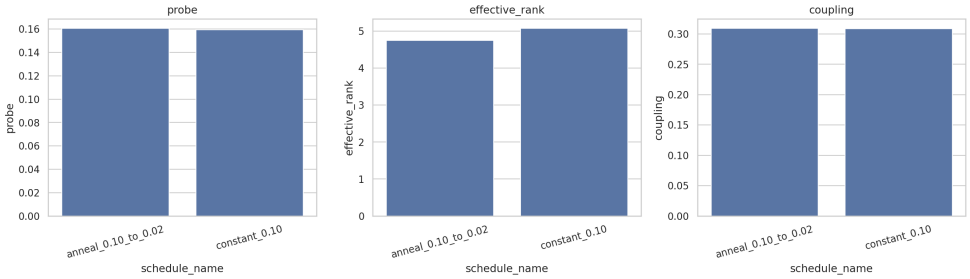


Figure A6: Summary visualization for the whitening schedule check.

F Appendix Discussion of the Negative Schedule Check

The whitening anneal experiment is best interpreted as a small falsification-oriented check rather than as a failed main result. Its value is methodological: the pathway view generates concrete intervention hypotheses, but those hypotheses still require empirical validation. In the present setting, annealing whitening strength after the early phase does not preserve the intended rank advantage while reducing late coupling relaxation. This makes it a suitable appendix result rather than a central claim.