NEURAL SUCCESSION: A MESOSCOPIC THEORY OF INVASION, COEXISTENCE, AND STABILIZATION IN CONTINUAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual learning is usually studied through mechanisms that preserve old knowledge. Here we study what happens when a new task enters a representation that already supports earlier tasks. We develop *Successional Learning Theory* (SLT), a mesoscopic account in which the current representation is a resident community, the incoming task is an invader, forgetting is resident displacement, joint retention is coexistence, replay is resident reinforcement, and training moves from establishment toward stabilization. Its empirical coordinate is directional pre-invasion compatibility, measured on the resident model before the incoming task is learned. Across eight experiments, compatibility orders later forgetting on the 20 directed Split-CIFAR-10 transitions (three-repeat $r = -0.789$, incoming-task cluster 95% CI $[-0.90, -0.72]$; every repeat alone $r \leq -0.67$), forecasts held-out forgetting with 24% lower error than a no-information baseline, and reproduces under controlled MNIST permutations and CIFAR-10 rotations ($r = -0.804$, -0.718). The same coordinate also resolves coexistence. The three natural transitions that coexist are exactly the three most compatible (AUC 1.00). On an 84-transition suite, compatibility separates coexistence from exclusion at every retention threshold (AUC 0.93–0.97). Replay repairs every transition with at most 325 stored examples and is most efficient where displacement is largest. The predictive structure appears at the transition scale. Compatibility reaches $|r| = 0.720$, while activation, representation, Jacobian, and fixed-coefficient Lotka–Volterra specializations do not. Plasticity and feature turnover fall reliably from early to late training (15/15 and 14/15 runs). We formalize a minimum habitat-modification bound, a displacement floor, a sufficient coexistence condition, an identifiability law with a range-restriction corollary, successional stabilization, and local reinforcement. The identifiability law also predicts where the coordinate loses leverage, and the prediction matches three CIFAR-100 partitions and five optimizer regimes. SLT is a pre-adaptation diagnostic that complements replay, regularization, and projection methods.

1 INTRODUCTION

Continual learning modifies a model that already supports earlier tasks (McCloskey & Cohen, 1989; French, 1999; Goodfellow et al., 2013). When task B arrives after task A, it is not learned in an empty parameter space. It enters a representation shaped by the resident task. Some incoming tasks reuse what is already present and cause little damage. Others require broader reorganization and displace previous behavior. Most continual-learning work starts after this transition has begun. It asks how to regularize parameters (Kirkpatrick et al., 2017; Zenke et al., 2017; Aljundi et al., 2018; Li & Hoiem, 2017), constrain or project gradients (Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2019a; Farajtabar et al., 2020; Saha et al., 2021; Zeng et al., 2019; Wang et al., 2021), replay old examples (Rebuffi et al., 2017; Rolnick et al., 2019; Chaudhry et al., 2019b; Buzzega et al., 2020), or isolate capacity (Rusu et al., 2016; Mallya & Lazebnik, 2018); see Parisi et al. (2019); De Lange et al. (2022); Wang et al. (2024) for surveys. We instead study the *transition itself*.

We call this view *Successional Learning Theory* (SLT). It borrows its questions from ecological succession, the ordered replacement of communities in which early occupants modify the habitat

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.

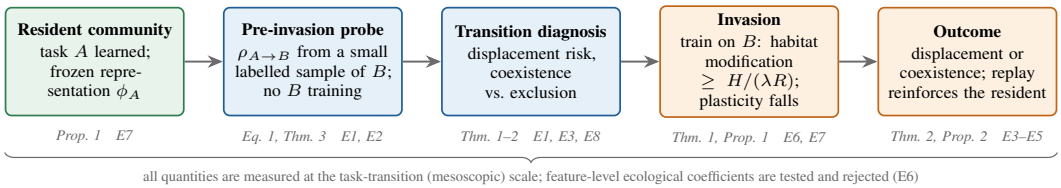


Figure 1: **The SLT pipeline.** A resident representation is probed *before* the incoming task is learned. The resulting coordinate diagnoses the transition, which then unfolds as establishment followed by stabilization and ends in displacement or coexistence. Grey labels map each stage to the result that formalizes it and the experiment that tests it.

that later arrivals must use (Clements, 1916; Odum, 1969; Connell & Slatyer, 1977; Odling-Smee et al., 2013). In this view, a learned representation is the resident habitat and an incoming task is an invader. The representation change required to solve it is establishment demand. Damage to resident-sensitive directions is displacement, joint retention is coexistence, and replay reintroduces resident signal. Within-task dynamics move from high-plasticity establishment toward stabilization. The ecology supplies a linked set of questions. Every formal object is defined directly through neural risk, representation change, or optimization.

The distinction from task-similarity work is scope. Task relation is known to matter for transfer, forgetting, and task ordering (Nguyen et al., 2019; Ramasesh et al., 2021; Lee et al., 2021; Doan et al., 2021; Hiratani, 2024; Li & Hiratani, 2025; Wang et al., 2025; 2026). SLT does not claim that a transfer score is new. It asks whether one pre-invasion coordinate can organize outcomes that are usually analyzed separately, including resident displacement, coexistence versus exclusion, response to resident reinforcement, and the move from establishment to stabilization. We also test the neural scale at which this ecological interpretation is predictive.

Across these tests, pre-invasion compatibility orders forgetting on natural transitions and under controlled manipulations. It also forecasts held-out forgetting and separates coexistence from exclusion. It is more predictive than activation covariance, representation similarity, raw Jacobian overlap, and stationary microscopic ecological specializations. Within an invasion, plasticity and feature turnover fall from early to late training. The theory predicts where a prospective coordinate should lose leverage, and the data follow this prediction. Our contributions are:

- **A mesoscopic theory of neural succession** (§3): a minimum habitat-modification bound, a displacement floor, a sufficient coexistence condition, an identifiability law with a range-restriction corollary, successional stabilization, and a local reinforcement result.
- **A prospective compatibility principle** (§5.1): a directional score measured before learning B orders later forgetting, supports held-out prediction, and reproduces under controlled task manipulations.
- **Phase, scale, and stage results** (§5.2–5.4): the coordinate separates coexistence from exclusion, dominates the tested microscopic specializations, and adaptation shows a reproducible establishment-to-stabilization profile.
- **Robustness and a predictive scope law** (§5.6): the ordering holds across estimator settings, label budgets, optimizer regimes, and a pretrained backbone, and the identifiability law predicts the regimes in which it loses leverage.

2 Related Work and Positioning

Invasion and displacement. Theory links forgetting to task geometry through NTK overlap (Jacot et al., 2018; Doan et al., 2021), teacher–student similarity (Lee et al., 2021; Hiratani, 2024), linear-regression analyses (Evron et al., 2022; Lin et al., 2023), and representation-level anatomy (Ramasesh et al., 2021). Computing optimal continual learners is intractable in general (Knoblauch et al., 2020). Transferability scores estimate how well frozen features support a new task (Tran et al., 2019; Nguyen et al., 2020; You et al., 2021), and fine-tuning can distort useful features (Kumar et al., 2022). Task relations have also been used to predict forgetting risk (Nguyen et al., 2019; Wang et al., 2025), to

Table 1: Positioning by the question each family answers. "Pre": computable before task-B adaptation.

Family	Main object	Pre	Coexistence	Scale test	Stages
Transferability / SPOT / NTK	task relation, kernel	✓			
EWC / SI / MAS	parameter importance				
GEM / OGD / GPM	gradient subspaces		implicit		
Replay / DER++	stored examples, logits				
Task ordering	sequence relation	✓			
Plasticity studies	training dynamics				✓
SLT (ours)	**resident–invader transition**	✓	✓	✓	✓

regularize dissimilar tasks (Wang et al., 2026), to route similar and dissimilar tasks (Ke et al., 2020; Wang et al., 2022; Adel, 2024), and to order tasks (Bell & Lawrence, 2022; Li & Hiratani, 2025). Our score follows the normalized-transfer construction of Li & Hiratani (2025). SLT adds a resident-side consequence, a lower bound on the habitat modification that establishment demands (Theorem 1).

Coexistence. In ecology, coexistence requires sufficient niche differentiation (Gause, 1934; Hardin, 1960; MacArthur & Levins, 1967; Tilman, 1982; Chesson, 2000). Meszéna et al. (2006) separate a species' impact on regulating factors from its sensitivity to them. The ML analogue is protected-subspace learning, which steers new updates away from directions that matter for old tasks (Zeng et al., 2019; Farajtabar et al., 2020; Saha et al., 2021; Wang et al., 2021). Theorem 2 turns this shared intuition into a sufficient condition phrased in resident-sensitive displacement.

Reinforcement. Invasion biology identifies propagule pressure as a main determinant of establishment (Lockwood et al., 2005), and resident communities resist invasion (Levine et al., 2004). In continual learning, replay and gradient constraints protect the resident (Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2019a; Riemer et al., 2019; Aljundi et al., 2019; Buzzega et al., 2020; Seo et al., 2025). Proposition 2 states the local first-order condition under which reintroduced resident signal counters displacement.

Successional stages. Deep networks change quickly early in training and stabilize later (Achille et al., 2019; Frankle et al., 2020; Gur-Ari et al., 2018), and continual training erodes plasticity (Ash & Adams, 2020; Lyle et al., 2023; Dohare et al., 2024). Training regime modulates forgetting (Mirzadeh et al., 2020; 2021). We measure feature-level plasticity and turnover within a transition and give a sufficient condition for finite total feature variation.

Ecology as a model of learning. Some learning systems admit exact ecological dynamics (Howell et al., 2020), and Lotka–Volterra competition (Lotka, 1925; Volterra, 1926) is a natural candidate for feature interaction. We test that candidate directly (E6) and find that the predictive ecological structure lives at the transition scale, consistent with ecological evidence that coarse community-level variables can be more predictive than pairwise coefficients (Mayfield & Levine, 2010).

3 Successional Learning Theory

3.1 Pre-invasion compatibility and establishment

Following Li & Hiratani (2025), the directional pre-invasion score is

$$\rho_{A\to B} = 1 - \sqrt{\frac{e_B(\theta_A)}{e_{B,\mathrm{sf}}(\theta_A)}}, \qquad \Omega_{A\to B} = 1 - \rho_{A\to B}, \tag{1}$$

where $e_B(\theta_A)$ is the error of a constrained incoming-task probe on the frozen resident representation and $e_{B,\mathrm{sf}}(\theta_A)$ is a shuffled-label reference. Larger ρ means B is already better supported. We treat Ω as an empirical coordinate and do *not* identify it with the theoretical demand below.

Hypothesis 1 (Successional compatibility principle)**.** *Within a fixed task ecology and training regime, expected resident forgetting $\mathbb{E}[F_{A\to B} \mid \Omega_{A\to B}]$ is increasing in $\Omega_{A\to B}$.*

Let ϕ_A be the resident representation and

$$\mathcal{R}_B^{(R_B)}(\phi) = \inf_{\|w\| \le R_B} \mathbb{E}_B\, \ell(w^\top \phi(x), y), \qquad H_{A\to B} = \left[\mathcal{R}_B^{(R_B)}(\phi_A) - \ell_B^\star\right]_+. \tag{2}$$

Let $\mathcal{E}_B$ be the reachable representations whose constrained probe risk is at most $\ell_B^\star$, and $D_B(\phi) = \left(\mathbb{E}_B \|\phi(x) - \phi_A(x)\|^2\right)^{1/2}$.

Theorem 1 (Minimum habitat modification)**.** *If ℓ is λ_B-Lipschitz in the logits, every $\phi \in \mathcal{E}_B$ satisfies $D_B(\phi) \ge H_{A\to B}/(\lambda_B R_B)$.*

An unsupported invader must change the representation before it can establish. The required change grows with its establishment deficit. Let $S_A(\phi)$ measure resident-sensitive displacement.

Assumption 1 (Resident sensitivity band)**.** *On the reachable region there are $0 < m_A \le M_A < \infty$ with $\frac{m_A}{2} S_A(\phi)^2 \le [L_A(\phi) - L_A(\phi_A)]_+ \le \frac{M_A}{2} S_A(\phi)^2$.*

Define $\sigma_{A\to B} = \inf_{\phi \in \mathcal{E}_B, D_B(\phi) > 0} S_A(\phi)/D_B(\phi)$ and $\Pi_{A\to B} = \sigma_{A\to B} H_{A\to B}/(\lambda_B R_B)$.

Corollary 1 (Displacement floor)**.** *Every establishing representation satisfies $S_A(\phi) \ge \Pi_{A\to B}$, hence $[L_A(\phi) - L_A(\phi_A)]_+ \ge \frac{m_A}{2}\Pi_{A\to B}^2$. If $\Pi_{A\to B} > \sqrt{2\delta_A/m_A}$, no B-establishing representation keeps A within loss tolerance δ_A.*

Testable consequence. The floor increases with $H_{A\to B}$; if Ω tracks H, forgetting should increase with Ω (Hypothesis 1). E1, E2, and E8 test the ordering, and a bridge analysis tests whether Ω tracks H.

3.2 Coexistence

Let $s_{A\to B}^\star = \inf_{\phi \in \mathcal{E}_B} S_A(\phi)$ be the least resident-sensitive displacement compatible with establishment.

Theorem 2 (Niche-separable coexistence)**.** *Under Assumption 1, if $s_{A\to B}^\star \le \sqrt{2\delta_A/M_A}$, then for every $\epsilon > 0$ some reachable representation establishes B while increasing resident loss by at most $\delta_A + \epsilon$ (at most δ_A if the infimum is attained).*

Corollary 1 and Theorem 2 define 2 sides of the phase picture. Small resident-sensitive establishment cost permits coexistence. Large unavoidable displacement forces exclusion. σ and $s^\star$ are latent, so the prediction is tested behaviorally (E3).

3.3 When a prospective coordinate can work

Theorem 3 (Identifiability)**.** *Within a fixed ecology, let $F = \alpha + \beta\Omega + \varepsilon$ with $\mathbb{E}[\varepsilon \mid \Omega] = 0$ and $\mathrm{Var}(\varepsilon) = \sigma_\varepsilon^2$. Then $R^2(F \mid \Omega) = \beta^2\sigma_\Omega^2/(\beta^2\sigma_\Omega^2 + \sigma_\varepsilon^2)$. If $\widetilde{\Omega} = \Omega + \eta$ with independent noise of variance σ_η^2, the squared correlation is multiplied by $\sigma_\Omega^2/(\sigma_\Omega^2 + \sigma_\eta^2)$.*

Corollary 2 (Range restriction)**.** *Suppose two ecologies share β and σ_ε but differ in the spread of the coordinate, $\sigma'_\Omega = s\,\sigma_\Omega$. If r is the correlation in the first, the correlation in the second is $|r'| = |r|\, s/\sqrt{1 - r^2 + r^2 s^2}$.*

Corollary 2 is the classical range-restriction relation (Thorndike, 1949). It gives a quantitative prediction for when the score should weaken. If transitions in an ecology are nearly equally compatible, the ordering must weaken even when the underlying mechanism is unchanged (§5.6).

3.4 Stabilization and reinforcement

Proposition 1 (Finite feature variation under convergent training)**.** *Suppose single-task training follows gradient flow on an L_s-smooth objective satisfying a μ-Polyak–Łojasiewicz inequality (Karimi et al., 2016), and active feature statistics $x_i(\theta) > 0$ satisfy $\|\nabla_\theta \log x_i(\theta)\| \le C_i$. Then $\int_0^\infty |\frac{d}{dt} \log x_i(\theta_t)|\, dt < \infty$.*

Total feature change is finite, so most change must occur early. Establishment precedes stabilization (E7). This is a sufficient idealized result, not a claim that ResNet training is globally PL.

Table 2: From theory to evidence. Each result is stated as an observable prediction and tested by a named experiment.

Result	Observable prediction	Tests	Evidence	Status
Thm. 1, Cor. 1	forgetting rises with Ω; Ω tracks demand	E1, E2, E8	$r = -0.789$; $-0.804/-0.718$; Ω– probe loss 0.767	supported
Thm. 2	Ω separates coexistence from exclusion	E3	AUC 1.00 (natural); 0.93–0.97 (84-suite)	supported
Thm. 3, Cor. 2	$\lvert r\rvert$ falls with coordinate and outcome spread	scope	CIFAR-100: $0.38/0.18/-0.05$ vs. predicted $0.32/0.30/0.13$	supported
Prop. 1	early $>$ late plasticity and turnover	E7	4.78 $[4.26, 5.53]$; 0.37 $[0.11, 0.48]$	supported
Prop. 2	replay repairs displacement; efficiency grows with displacement	E4	20/20 repaired ($\leq$ 325 ex.); $r(\eta, F_0) = 0.70$	supported
Scale claim	transition-level $\gg$ feature-level	E6	$\lvert r\rvert$: 0.72 vs. ≤ 0.45	supported

Proposition 2 (Local resident reinforcement)**.** *Consider one gradient step of size η on $L_q = (1-q)L_B + qL_A$, $q \in [0,1]$, with $g_A = \nabla L_A$, $g_B = \nabla L_B$, $a = \|g_A\|^2$, $b = \langle g_A, g_B\rangle$. The first-order change of the resident loss is $\delta(q) = -\eta[(1-q)b + qa]$. (i) δ is affine in q with slope $-\eta(a-b)$, so reinforcement lowers the first-order resident change whenever $b < a$. (ii) The marginal benefit at $q = 0$ is $\eta(a-b)$, larger for antagonistic invaders ($b < 0$) and for displaced residents (larger a). (iii) For $a > 0$, $\delta(q) \leq 0$ if and only if $q \geq [-b]_+/(a + [-b]_+)$. If L_A is M_A-smooth, the exact change is at most $\delta(q) + \frac{M_A\eta^2}{2}\|(1-q)g_B + qg_A\|^2$.*

Part (ii) predicts that replay is most efficient on displacement-prone transitions (E4). The result is local. It does not determine a global replay budget, and we do not claim one.

Table 2 maps each result to its observable prediction and test. Proofs are in Appendix A.

4 Experimental design

The 8 experiments test the predictions of the theory (Table 3). Full protocols are given in Appendix B.

Tasks and models. Natural transitions use the five binary Split-CIFAR-10 tasks $\{0,1\}, \ldots, \{8,9\}$ (Krizhevsky, 2009) and all 20 ordered pairs, with a multi-head (task-incremental) ResNet-18 (He et al., 2016; van de Ven & Tolias, 2019; van de Ven et al., 2022). Controlled transitions manipulate task relation directly with pixel permutations of MNIST (LeCun et al., 1998) and rotations of CIFAR-10. Coexistence is tested on the 20 natural transitions, on 60 pure-domain CIFAR-10 transitions (vehicle and animal binary tasks), and on an 84-transition MNIST suite (4 class pairs $\times$ 7 transforms $\times$ 3 seeds). Scope studies use CIFAR-100, a class-incremental construction, and an ImageNet-pretrained ResNet-50 (Deng et al., 2009).

Protocol. The default optimizer is Adam (Kingma & Ba, 2015) with learning rate 10^{-3} and batch size 128; E7 and the scope study vary it. $\rho_{A\to B}$ is computed on the frozen resident checkpoint *before* any task-B update, from a class-balanced incoming-task sample and shuffled-label references (256 examples and 20 shuffles in the estimator study). Forgetting is $F_{A\to B} = R_{AA} - R_{BA}$, the drop in task-$A$ test accuracy after learning B. A transition coexists when forgetting stays below a fixed threshold ($F < 0.20$ on natural transitions, $F < 0.10$ on the pure-domain construction) or, on the 84-transition suite, when both tasks keep a fixed fraction (0.80–0.99) of their reference accuracy.

Statistics. Transitions that share an incoming task are not independent. We therefore report pair means over repeats, incoming-task cluster bootstrap intervals (Efron & Tibshirani, 1993; Cameron et al., 2008), exact task-label permutation tests over all $5!$ relabelings (Mantel, 1967), leave-one-pair-out (LOPO) prediction with a no-information baseline, and paired-seed bootstrap intervals for interventions. Scale and mechanism tests used pass criteria fixed before the runs.

Table 3: The eight SLT experiments. Units are directed transitions unless noted.

ID	Question	Setting	Units	Primary statistic
E1	Does ρ order displacement?	Split-CIFAR-10, ResNet-18	20×3 repeats	$r(\rho, F)$ on pair means
E2	Is the order prospective?	same; ρ from resident only	20, LOPO	held-out r, MAE
E3	Is coexistence a phase?	natural; pure-domain; MNIST suite	20; 60; 84	rates; ROC AUC
E4	Does replay reinforce?	buffers 0–1000	20	efficiency, threshold vs. ρ
E5	Does ρ help an intervention?	CIFAR-10/100, 6 methods	3 seeds	ACC, BWT
E6	At which scale?	5 predictors; probe/LV tests	20	$\lvert r \rvert$, pre-registered gates
E7	Are there stages?	per-epoch feature statistics	15 runs	early−late contrasts
E8	Controlled relation?	MNIST perm.; CIFAR-10 rot.	96; 48	r, group-CV R^2

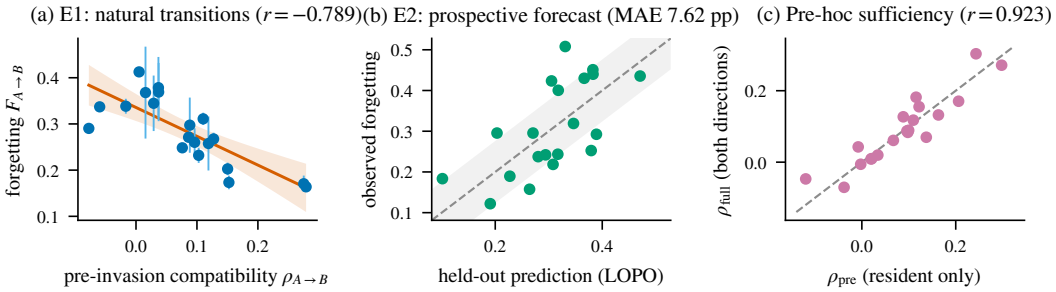


Figure 2: **Compatibility is prospective.** (a) Pair means over three independent repeats (bars: s.d.; band: OLS 95%). (b) Leave-one-pair-out forecasts (dashed: identity; band: ±MAE). (c) The resident-only score agrees with a score that also uses the reverse direction.

5 Results

5.1 Pre-invasion state orders later displacement (E1, E2, E8)

Across the three-repeat reproduction, compatibility strongly anticorrelates with later forgetting (Figure 2a): $r = -0.789$, Spearman -0.789, $R^2 = 0.622$. Resampling incoming tasks gives a 95% interval of $[-0.90, -0.72]$, and the association is negative within four of five incoming tasks. It holds in every repeat on its own ($r = -0.706, -0.720, -0.674$), across all 60 individual runs ($r = -0.695$), and in an independent single-run reproduction ($r = -0.710$, $p = 4.6 \times 10^{-4}$). The coordinate is highly repeatable (mean between-repeat correlation 0.96 for ρ and 0.77 for forgetting), and a separate revalidation with a task-label permutation test gives $r = -0.723$, $p = 0.033$. Forgetting spans 0.16–0.41 (mean 0.285), so the ordering is not produced by a few extreme pairs.

The same ordering is available before adaptation. Leave-one-pair-out forecasts reach $r = 0.648$ with MAE 7.62 percentage points, against 10.03 pp for a no-information predictor that uses the mean of the other pairs, a 24% error reduction (Figure 2b). A stricter split that holds out every transition into the same incoming task gives $r = 0.669$ and MAE 7.44 pp. A resident-only score agrees with a score that also uses the reverse direction ($r = 0.923$; Figure 2c) and loses little predictive power ($r = -0.753$ vs. -0.771 with forgetting).

Controlled manipulations recover the same direction (Figure 3, left). MNIST permutations give $r = -0.804$ over 96 transitions (group cross-validated $R^2 = 0.626$) and CIFAR-10 rotations give $r = -0.718$ over 48 (group-CV $R^2 = 0.484$). On MNIST, a quadratic term raises group-CV R^2 to 0.764, with a minimum near $\rho \approx 0.81$. Forgetting reaches a floor once the invader is almost fully supported. On CIFAR-10 the quadratic term adds nothing (0.465).

A bridge analysis relates the empirical coordinate to the demand in Theorem 1 without identifying the two. Within incoming task, Ω correlates 0.767 with the unclipped incoming probe loss (cluster 95% CI $[0.28, 0.95]$) and 0.692 with the clipped deficit H (CI $[-0.14, 0.94]$ with five clusters; Appendix C.2). The unclipped link is the robust one.

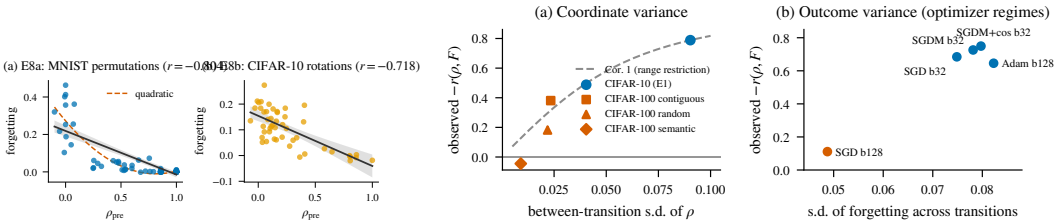


Figure 3: **Left:** controlled manipulations of task relation (E8); black line and band: linear fit with 95% interval; red dashed: quadratic fit on MNIST. **Right:** identifiability (Theorem 3). (a) Predictive strength falls with the between-transition spread of ρ along the range-restriction curve of Corollary 2, anchored at E1. (b) Across optimizer regimes, the one regime with collapsed outcome variance is the one where ordering vanishes.

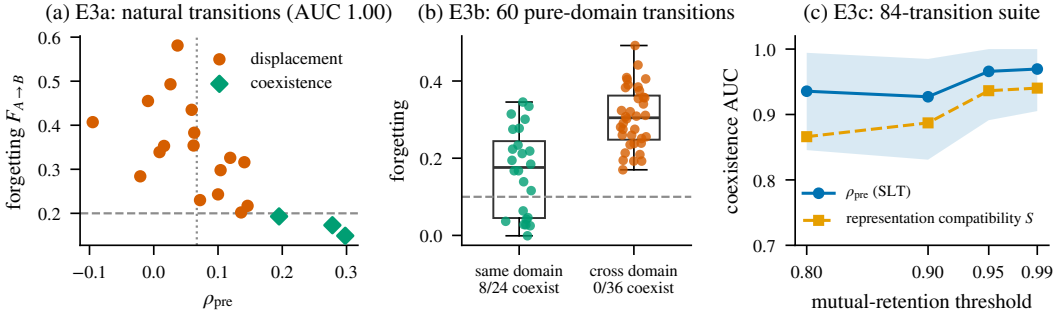


Figure 4: **Coexistence is a phase of the transition.** (a) Natural transitions: the three that coexist (green, $F < 0.20$) are the three most compatible; dotted line: compatibility split. (b) Pure-domain construction: only same-domain transitions reach near-zero forgetting. (c) 84-transition suite: coexistence AUC of ρ and of a representation-compatibility score S across mutual-retention thresholds (band: 95% cluster bootstrap over 28 class-pair×transform clusters).

5.2 Coexistence is a phase of the same transition (E3)

Corollary 1 and Theorem 2 predict a phase boundary. Compatible invaders can establish without displacing the resident, while incompatible ones cannot. The reproduced natural transitions follow this pattern (Figure 4a). Three of the 20 transitions coexist ($F < 0.20$), and they are the three with the highest compatibility (AUC 1.00). Splitting at $\rho = 0.067$, compatible transitions forget 0.235 on average and coexist in 30% of cases, whereas incompatible ones forget 0.408 and never coexist (Mann–Whitney $p = 2.2 \times 10^{-4}$). A stricter construction within semantic domains gives the same pattern. 8 of 24 same-domain transitions reach near-zero forgetting against 0 of 36 cross-domain transitions, with mean forgetting 0.166 versus 0.306 ($p = 1.25 \times 10^{-5}$; Figure 4b). On the 84-transition suite, where compatibility spans a wide range, ρ separates mutual coexistence from exclusion at every retention threshold, with AUC 0.936, 0.927, 0.966, and 0.969 at 0.80, 0.90, 0.95, and 0.99 (Figure 4c; at 0.90, 95% CI $[0.83, 0.98]$). It exceeds the representation-compatibility score S at all four thresholds, most clearly at 0.80 ($+0.070$, CI $[0.001, 0.169]$), keeps AUC 0.903 after removing the 12 identity transitions, and correlates -0.82 with forgetting in this suite (Appendix C.1).

5.3 The useful ecological scale is mesoscopic (E6)

We next test whether a literal microscopic ecological model is needed. On the same 20 natural transitions, compatibility has $|r| = 0.720$ with forgetting, compared with 0.447 for one-step gradient alignment, 0.143 for activation covariance, 0.109 for raw Jacobian overlap, and 0.005 for representation similarity (Figure 5a). Feature-level competition coefficients fail pre-registered tests in a regime with substantial forgetting (s.d. 0.064): task-level $r = 0.177$ (exact permutation $p = 0.54$); increasing Jacobian probes from 64 to 1024 leaves the median partial correlation at or below 0.105 (Figure 5b); fixed-coefficient Lotka–Volterra fits beat simpler baselines in 0 of 60 transitions (median held-out $R^2 = -29.2$); none of 330 local windows has stable coefficients; and neural R^* resources

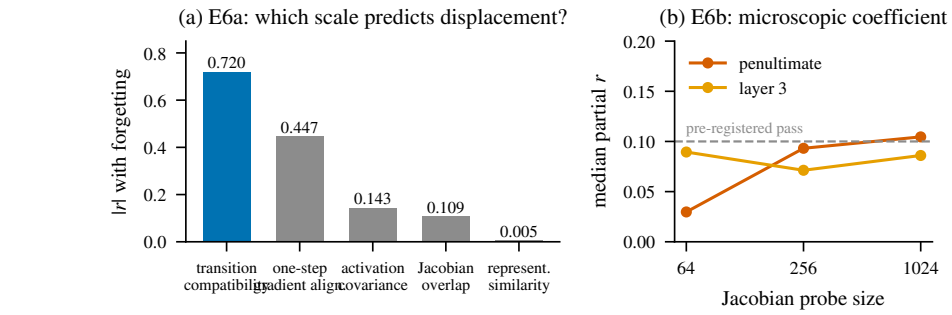


Figure 5: **Scale selection.** (a) On the same 20 transitions, the transition-level coordinate predicts displacement far better than feature- or step-level quantities. (b) Increasing Jacobian probes 16-fold at two layers does not produce a predictive microscopic coefficient.

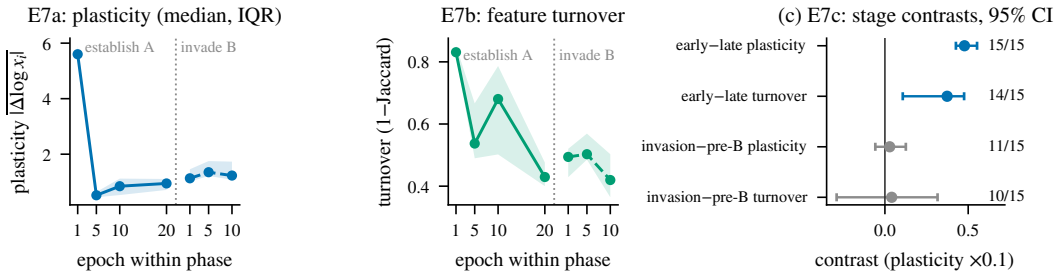


Figure 6: **Establishment precedes stabilization.** (a, b) Median and interquartile range over 15 runs of per-epoch plasticity and turnover of the top-quartile resident features, while A is established and then while B invades. (c) Pre-specified contrasts with 95% intervals; grey: interval includes zero.

are identifiable for only 1.9% of candidate directions (Appendix C.6). These results support *scale selection*. The ecological structure is predictive after coarse-graining to a resident–invader transition, while the tested feature-level coefficients do not behave as stationary population parameters.

5.4 Learning exhibits successional stages (E7)

The succession account also makes a temporal prediction. Across 15 runs, early-minus-late plasticity is 4.776 (95% CI $[4.262, 5.532]$, 15/15 positive) and early-minus-late feature turnover is 0.374 (CI $[0.107, 0.475]$, 14/15 positive; Figure 6). Large early reorganization gives way to a more stable regime, consistent with Proposition 1. The invasion does not reliably reopen a pioneer phase. invasion-minus-pre-B contrasts are positive in 11/15 and 10/15 runs but their intervals include zero, so we do not claim renewed establishment after invasion.

5.5 Reinforcement and interventions (E4, E5)

Replay reinforces the resident in every transition. Without replay, mean forgetting is 0.289. All 20 transitions are brought within 10% of their resident accuracy using at most 325 stored examples (median 100). Proposition 2(ii) predicts that the per-example benefit grows with the displacement being repaired: initial replay efficiency correlates 0.70 with forgetting at zero replay and -0.42 with compatibility (Spearman -0.46), so the most incompatible transitions gain most from their first replayed examples (Appendix C.3). The exact buffer at which a transition crosses the 10% line is not ordered by ρ ($r = 0.01$), consistent with the local scope of Proposition 2. The coordinate also carries over to interventions (E5; Appendix C.4). Combined with replay, compatibility-modulated regularization outperforms fine-tuning, EWC, and A-GEM on both Split-CIFAR-10 and Split-CIFAR-100 (ACC 0.815 and 0.531), and gating EWC-DR by Ω raises mean ACC from 0.800 to 0.824 while cutting forgetting by 40% (three seeds). DER++ remains the strongest single method, so SLT is best used as a diagnostic layer that tells such methods where protection is needed.

Table 4: Main results with uncertainty. CI: 95% interval (cluster bootstrap unless stated).

Claim	Result	Uncertainty / comparison
E1 natural ordering	$r = -0.789$, $R^2 = 0.622$	CI $[-0.90, -0.72]$; each repeat ≤ -0.67; single run -0.710
E2 held-out forecast	MAE 7.62 pp, $r = 0.648$	no-information MAE 10.03 pp; LOTO $r = 0.669$
E8 controlled relation	$r = -0.804$ / -0.718	group-CV R^2 0.626 / 0.484
E3 coexistence phase	AUC 1.00 / 0.927	natural / 84-suite (CI $[0.83, 0.98]$); 8/24 vs. 0/36
E4 reinforcement	20/20 repaired	≤ 325 examples (median 100); $r(\eta, F_0) = 0.70$
E5 intervention	ACC 0.815 / 0.531	above fine-tune, EWC, A-GEM on both benchmarks
E6 transition vs. micro scale	$\lvert r \rvert = 0.720$	gradient 0.447; others ≤ 0.143; LV 0/60
E7 successional stages	$+4.78$ plasticity, $+0.37$ turnover	CI $[4.26, 5.53]$, $[0.11, 0.48]$

5.6 Robustness and scope

The ordering persists across estimator settings. all 24 probe-size and shuffle combinations give r between -0.833 and -0.662, whereas CKA (Kornblith et al., 2019) gives $+0.113$ and a plain linear-probe accuracy -0.464 (Table 4). The score needs few labels, but the tested label-free variants do not recover the signal. 10 labels per class correlate 0.816 with the full estimate, while two label-free variants carry no signal. On an ImageNet-pretrained ResNet-50 the direction holds ($r = -0.426$, 10/10 negative leave-one-task-out slopes) with a wide cluster interval. Details are in Appendix C.5.

Theorem 3 predicts where a prospective coordinate should lose leverage. The observed pattern follows this prediction (Figure 3, right). On three CIFAR-100 partitions of ten 10-class tasks, the spread of ρ falls from 0.090 to 0.023, 0.022, and 0.009, and $-r$ falls from 0.789 to 0.381, 0.182, and -0.045. Anchored at E1, Corollary 2 predicts 0.32, 0.30, and 0.13, matching both the ordering and the size of the contraction. Across five optimizer regimes, four give r between -0.646 and -0.749; the fifth (plain SGD, batch 128) is the one whose forgetting spread collapses (0.049 vs. ≥ 0.075), and there the ordering vanishes ($r = -0.111$), again as the theorem requires.

6 Discussion

The evidence supports a task-scale claim rather than a literal ecological simulator of neural features. Continual-learning transitions show reproducible *successional structure* at this scale. Before adaptation, directional compatibility orders how much the resident will be displaced and resolves a coexistence/exclusion phase; replay repairs displacement where it is largest; during adaptation, plasticity and turnover fall from establishment toward stabilization. The scale result distinguishes SLT from an ecological relabeling of task similarity. The predictive abstraction is a transition state, not a static pairwise coefficient or a neuron-level resource law. The identifiability law specifies when that state should be measurable. Accuracy-oriented methods determine how to preserve knowledge. SLT instead diagnoses the upcoming transition from a frozen model and a small labelled sample, so it can be used alongside replay, regularization, projection, and isolation methods.

Limitations. The evidence comes from vision task-incremental settings with explicit task boundaries and a small labelled sample of the incoming task; wide tasks (CIFAR-100) and single-head class-incremental learning compress the between-transition signal, as Theorem 3 anticipates. $\sigma_{A\to B}$ and $s^\star_{A\to B}$ are latent, so E3 validates the phase rather than the quantities, and within one semantic construction domain membership carries information beyond ρ (Appendix C.1). The coordinate characterizes transitions that differ in task relation. Replay is described by a local principle rather than a global budget, and Proposition 1 is a sufficient idealized result. SLT is a diagnostic framework and is not proposed as a new accuracy-leading continual-learning algorithm.

7 Conclusion

SLT treats continual learning as a sequence of invasions into already-occupied representations. The transition can be diagnosed before adaptation, ends in coexistence or displacement, responds to resident reinforcement, and follows an establishment-to-stabilization path; the identifiability law predicts where this mesoscopic description stops working.

AI USE STATEMENT

Generative AI tools were used for literature exploration, code assistance, statistical cross-checking, figure preparation, and language editing. The authors verified mathematical statements, experimental procedures, numerical claims, citations, code, and final prose against executed analyses and primary sources. The authors are responsible for the submission.

ETHICS STATEMENT

This work uses standard public image-classification benchmarks and does not involve human subjects, private data, or deployment decisions.

REPRODUCIBILITY STATEMENT

The supplementary material contains the training code for all experiments, the raw result records and reproduction logs, and one analysis script, `scripts/make_figures_and_stats.py`, which regenerates every figure and every derived statistic (bootstrap intervals, permutation tests, held-out baselines, range-restriction predictions) into `data/derived_revision_stats.json`. Appendix B gives per-experiment protocols.

REFERENCES

Alessandro Achille, Matteo Rovere, and Stefano Soatto. Critical learning periods in deep networks. In *International Conference on Learning Representations*, 2019.

Tameem Adel. Similarity-based adaptation for task-aware and task-free continual learning. *Journal of Artificial Intelligence Research*, 80:377–417, 2024.

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision*, pp. 139–154, 2018.

Rahaf Aljundi, Eugene Belilovsky, Tinne Tuytelaars, Laurent Charlin, Massimo Caccia, Min Lin, and Lucas Page-Caccia. Online continual learning with maximal interfered retrieval. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jordan T. Ash and Ryan P. Adams. On warm-starting neural network training. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Samuel J. Bell and Neil D. Lawrence. The effect of task ordering in continual learning. *arXiv preprint arXiv:2205.13323*, 2022.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. In *Advances in Neural Information Processing Systems*, volume 33, pp. 15920–15930, 2020.

A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller. Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427, 2008.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with A-GEM. In *International Conference on Learning Representations*, 2019a.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K. Dokania, Philip H. S. Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019b.

Peter Chesson. Mechanisms of maintenance of species diversity. *Annual Review of Ecology and Systematics*, 31:343–366, 2000.

Frederic E. Clements. *Plant Succession: An Analysis of the Development of Vegetation*. Publication 242. Carnegie Institution of Washington, 1916.

Joseph H. Connell and Ralph O. Slatyer. Mechanisms of succession in natural communities and their role in community stability and organization. *The American Naturalist*, 111(982):1119–1144, 1977.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Thang Doan, Mehdi Abbana Bennani, Bogdan Mazoure, Guillaume Rabusseau, and Pierre Alquier. A theoretical analysis of catastrophic forgetting through the NTK overlap matrix. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1072–1080. PMLR, 2021.

Shibhansh Dohare, J. Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A. Rupam Mahmood, and Richard S. Sutton. Loss of plasticity in deep continual learning. *Nature*, 632: 768–774, 2024.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

Itay Evron, Edward Moroshko, Rachel Ward, Nathan Srebro, and Daniel Soudry. How catastrophic can catastrophic forgetting be in linear regression? In *Proceedings of the 35th Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 4028–4079. PMLR, 2022.

Mehrdad Farajtabar, Navid Azizan, Alexander Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pp. 3762–3773, 2020.

Jonathan Frankle, David J. Schwab, and Ari S. Morcos. The early phase of neural network training. In *International Conference on Learning Representations*, 2020.

Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3 (4):128–135, 1999.

Georgii F. Gause. *The Struggle for Existence*. Williams & Wilkins, 1934.

Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013.

Guy Gur-Ari, Daniel A. Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018.

Garrett Hardin. The competitive exclusion principle. *Science*, 131(3409):1292–1297, 1960.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Naoki Hiratani. Disentangling and mitigating the impact of task similarity for continual learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Owen Howell, Wenping Cui, Robert Marsland III, and Pankaj Mehta. Machine learning as ecology. *Journal of Physics A: Mathematical and Theoretical*, 53(33):334001, 2020.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 795–811, 2016.

Zixuan Ke, Bing Liu, and Xingchang Huang. Continual learning of a mixed sequence of similar and dissimilar tasks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

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

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114 (13):3521–3526, 2017.

Jeremias Knoblauch, Hisham Husain, and Tom Diethe. Optimal continual learning has perfect memory and is NP-hard. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5327–5337. PMLR, 2020.

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

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Sebastian Lee, Sebastian Goldt, and Andrew Saxe. Continual learning in the teacher-student setup: Impact of task similarity. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6109–6119. PMLR, 2021.

Jonathan M. Levine, Peter B. Adler, and Stephanie G. Yelenik. A meta-analysis of biotic resistance to exotic plant invasions. *Ecology Letters*, 7(10):975–989, 2004.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017.

Ziyan Li and Naoki Hiratani. Optimal task order for continual learning of multiple tasks. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 34578–34603. PMLR, 2025.

Sen Lin, Peizhong Ju, Yingbin Liang, and Ness Shroff. Theory on forgetting and generalization of continual learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21078–21100. PMLR, 2023.

Julie L. Lockwood, Phillip Cassey, and Tim Blackburn. The role of propagule pressure in explaining species invasions. *Trends in Ecology & Evolution*, 20(5):223–228, 2005.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Alfred J. Lotka. *Elements of Physical Biology*. Williams & Wilkins, 1925.

Clare Lyle, Zeyu Zheng, Evgenii Nikishin, Bernardo Avila Pires, Razvan Pascanu, and Will Dabney. Understanding plasticity in neural networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23190–23211. PMLR, 2023.

Robert MacArthur and Richard Levins. The limiting similarity, convergence, and divergence of coexisting species. *The American Naturalist*, 101(921):377–385, 1967.

Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7765–7773, 2018.

Nathan Mantel. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2):209–220, 1967.

Margaret M. Mayfield and Jonathan M. Levine. Opposing effects of competitive exclusion on the phylogenetic structure of communities. *Ecology Letters*, 13(9):1085–1093, 2010.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pp. 109–165. Academic Press, 1989.

Géza Meszéna, Mats Gyllenberg, Liz Pásztor, and Johan A. J. Metz. Competitive exclusion and limiting similarity: A unified theory. *Theoretical Population Biology*, 69(1):68–87, 2006.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Razvan Pascanu, and Hassan Ghasemzadeh. Understanding the role of training regimes in continual learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Dilan Gorur, Razvan Pascanu, and Hassan Ghasemzadeh. Linear mode connectivity in multitask and continual learning. In *International Conference on Learning Representations*, 2021.

Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. LEEP: A new measure to evaluate transferability of learned representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7294–7305. PMLR, 2020.

Cuong V. Nguyen, Alessandro Achille, Michael Lam, Tal Hassner, Vijay Mahadevan, and Stefano Soatto. Toward understanding catastrophic forgetting in continual learning. *arXiv preprint arXiv:1908.01091*, 2019.

John Odling-Smee, Douglas H. Erwin, Eric P. Palkovacs, Marcus W. Feldman, and Kevin N. Laland. Niche construction theory: A practical guide for ecologists. *The Quarterly Review of Biology*, 88 (1):3–28, 2013.

Eugene P. Odum. The strategy of ecosystem development. *Science*, 164(3877):262–270, 1969.

German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

Vinay V. Ramasesh, Ethan Dyer, and Maithra Raghu. Anatomy of catastrophic forgetting: Hidden representations and task semantics. In *International Conference on Learning Representations*, 2021.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *International Conference on Learning Representations*, 2019.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *International Conference on Learning Representations*, 2021.

Minhyuk Seo, Hyunseo Koh, and Jonghyun Choi. Budgeted online continual learning by adaptive layer freezing and frequency-based sampling. In *International Conference on Learning Representations*, 2025.

Robert L. Thorndike. *Personnel Selection: Test and Measurement Techniques*. Wiley, New York, 1949.

David Tilman. *Resource Competition and Community Structure*, volume 17 of *Monographs in Population Biology*. Princeton University Press, 1982.

Anh T. Tran, Cuong V. Nguyen, and Tal Hassner. Transferability and hardness of supervised classification tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1395–1405, 2019.

Gido M. van de Ven and Andreas S. Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019.

Gido M. van de Ven, Tinne Tuytelaars, and Andreas S. Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4:1185–1197, 2022.

Vito Volterra. Fluctuations in the abundance of a species considered mathematically. *Nature*, 118: 558–560, 1926.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5362–5383, 2024.

Shipeng Wang, Xiaorong Li, Jian Sun, and Zongben Xu. Training networks in null space of feature covariance for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 184–193, 2021.

Sijia Wang, Yoojin Choi, Junya Chen, Mostafa El-Khamy, and Ricardo Henao. Toward sustainable continual learning: Detection and knowledge repurposing of similar tasks. *arXiv preprint arXiv:2210.05751*, 2022.

Xulong Wang, Yu Zhang, Tong Liu, Zhipeng Yuan, Kang Liu, Vitaveska Lanfranchi, and Po Yang. SPOT: An efficient training-free task similarity quantification method for continual learning. *Pattern Recognition Letters*, 197:168–174, 2025.

Xulong Wang, Tong Liu, Menghui Zhou, Yu Zhang, Zhipeng Yuan, Kang Liu, and Po Yang. Adapting to dissimilar tasks for continual learning via gradient norm regularisation. *Neurocomputing*, 2026.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. LogME: Practical assessment of pre-trained models for transfer learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12133–12143. PMLR, 2021.

Guanxiong Zeng, Yang Chen, Bo Cui, and Shan Yu. Continual learning of context-dependent processing in neural networks. *Nature Machine Intelligence*, 1:364–372, 2019.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3987–3995. PMLR, 2017.

A PROOFS

Throughout, $\|\cdot\|$ is the Euclidean norm on features and logits and the operator norm on linear read-outs; $\mathbb{E}_B$ and $\mathbb{E}_A$ denote expectations over the incoming and resident task distributions.

A.1 PROOF OF THEOREM 1

Setting. A read-out w maps features to logits, $z = w\,\phi(x)$, and the loss satisfies $|\ell(z,y) - \ell(z',y)| \le \lambda_B\|z - z'\|$ for all z, z', y. For a representation ϕ write $\mathcal{R}(\phi) = \mathcal{R}_B^{(R_B)}(\phi) = \inf_{\|w\|\le R_B} \mathbb{E}_B\,\ell(w\phi(x), y)$.

Step 1 (near-optimal probe for the target). Fix $\phi \in \mathcal{E}_B$, so $\mathcal{R}(\phi) \le \ell_B^\star$. For any $\epsilon > 0$ choose w_ϵ with $\|w_\epsilon\| \le R_B$ and $\mathbb{E}_B\ell(w_\epsilon\phi(x), y) \le \mathcal{R}(\phi) + \epsilon \le \ell_B^\star + \epsilon$.

Step 2 (transport the probe to the resident). Because w_ϵ is feasible for the resident representation,

$$\mathcal{R}(\phi_A) \le \mathbb{E}_B\ell(w_\epsilon\phi_A(x), y) \le \mathbb{E}_B\ell(w_\epsilon\phi(x), y) + \lambda_B\,\mathbb{E}_B\|w_\epsilon(\phi_A(x) - \phi(x))\| \le \ell_B^\star + \epsilon + \lambda_B R_B\,\mathbb{E}_B\|\phi_A(x) - \phi(x)\|,$$

using Lipschitzness in the logits and $\|w_\epsilon v\| \le \|w_\epsilon\|\|v\|$.

Step 3 (conclude). Jensen's inequality gives $\mathbb{E}_B\|\phi_A - \phi\| \le (\mathbb{E}_B\|\phi_A - \phi\|^2)^{1/2} = D_B(\phi)$. Hence $\mathcal{R}(\phi_A) - \ell_B^\star \le \epsilon + \lambda_B R_B D_B(\phi)$ for every $\epsilon > 0$, so $\mathcal{R}(\phi_A) - \ell_B^\star \le \lambda_B R_B D_B(\phi)$. Since $D_B(\phi) \ge 0$, the same bound holds for the positive part, $H_{A\to B} \le \lambda_B R_B D_B(\phi)$, which is the claim. If $H_{A\to B} = 0$ the statement is trivial. □

A.2 PROOF OF COROLLARY 1

If $H_{A\to B} = 0$ then $\Pi_{A\to B} = 0$ and both claims are immediate. Otherwise Theorem 1 gives $D_B(\phi) \ge H_{A\to B}/(\lambda_B R_B) > 0$ for every $\phi \in \mathcal{E}_B$, so every establishing representation lies in the set over which $\sigma_{A\to B}$ is defined and satisfies $S_A(\phi) \ge \sigma_{A\to B} D_B(\phi)$. Combining the two inequalities, $S_A(\phi) \ge \sigma_{A\to B} H_{A\to B}/(\lambda_B R_B) = \Pi_{A\to B}$. The lower side of Assumption 1 then gives $[L_A(\phi) - L_A(\phi_A)]_+ \ge \frac{m_A}{2} S_A(\phi)^2 \ge \frac{m_A}{2}\Pi_{A\to B}^2$. Finally, if $\Pi_{A\to B} > \sqrt{2\delta_A/m_A}$ then $\frac{m_A}{2}\Pi_{A\to B}^2 > \delta_A$, so every B-establishing representation raises the resident loss by more than δ_A. □

A.3 PROOF OF THEOREM 2

Let $g(t) = \frac{M_A}{2}(s_{A\to B}^\star + t)^2$ for $t \ge 0$. By hypothesis $g(0) \le \delta_A$, and g is continuous and increasing, so for any $\epsilon > 0$ there is $t_\epsilon > 0$ with $g(t_\epsilon) \le \delta_A + \epsilon$. By definition of the infimum there is $\phi_\epsilon \in \mathcal{E}_B$ with $S_A(\phi_\epsilon) \le s_{A\to B}^\star + t_\epsilon$. This ϕ_ϵ establishes B (it lies in $\mathcal{E}_B$), and the upper side of Assumption 1 gives $[L_A(\phi_\epsilon) - L_A(\phi_A)]_+ \le \frac{M_A}{2} S_A(\phi_\epsilon)^2 \le g(t_\epsilon) \le \delta_A + \epsilon$. If the infimum is attained by some $\phi^\star \in \mathcal{E}_B$, the same argument with $t = 0$ gives an increase of at most $g(0) \le \delta_A$. □

A.4 PROOF OF THEOREM 3

Write $\sigma_\Omega^2 = \mathrm{Var}(\Omega)$. Because $\mathbb{E}[\varepsilon \mid \Omega] = 0$, $\mathrm{Cov}(\Omega, \varepsilon) = \mathbb{E}[(\Omega - \mathbb{E}\Omega)\,\mathbb{E}[\varepsilon \mid \Omega]] = 0$. Hence

$$\mathrm{Cov}(F, \Omega) = \beta\sigma_\Omega^2, \qquad \mathrm{Var}(F) = \beta^2\sigma_\Omega^2 + \sigma_\varepsilon^2, \qquad R^2(F \mid \Omega) = \frac{\mathrm{Cov}(F,\Omega)^2}{\mathrm{Var}(F)\,\mathrm{Var}(\Omega)} = \frac{\beta^2\sigma_\Omega^2}{\beta^2\sigma_\Omega^2 + \sigma_\varepsilon^2}.$$

For the measured coordinate $\widetilde{\Omega} = \Omega + \eta$ with η independent of (Ω, ε) and variance σ_η^2, $\mathrm{Cov}(F, \widetilde{\Omega}) = \mathrm{Cov}(F, \Omega) = \beta\sigma_\Omega^2$ and $\mathrm{Var}(\widetilde{\Omega}) = \sigma_\Omega^2 + \sigma_\eta^2$. Therefore

$$\mathrm{Corr}(F, \widetilde{\Omega})^2 = \frac{\beta^2\sigma_\Omega^4}{(\beta^2\sigma_\Omega^2 + \sigma_\varepsilon^2)(\sigma_\Omega^2 + \sigma_\eta^2)} = R^2(F \mid \Omega) \cdot \frac{\sigma_\Omega^2}{\sigma_\Omega^2 + \sigma_\eta^2},$$

which is the stated attenuation. Both factors lie in $[0, 1]$, so predictive power is limited by the between-transition spread of the coordinate relative to outcome noise and to measurement noise. □

A.5 Proof of Corollary 2

In the first ecology, Theorem 3 gives $r^2 = \beta^2\sigma_\Omega^2/(\beta^2\sigma_\Omega^2 + \sigma_\varepsilon^2)$; if $r \neq 0$ this rearranges to $\sigma_\varepsilon^2 = \beta^2\sigma_\Omega^2(1-r^2)/r^2$. In the second ecology the slope and noise are unchanged and the spread is $s\sigma_\Omega$, so

$$r'^2 = \frac{\beta^2 s^2 \sigma_\Omega^2}{\beta^2 s^2 \sigma_\Omega^2 + \beta^2\sigma_\Omega^2(1-r^2)/r^2} = \frac{s^2 r^2}{s^2 r^2 + 1 - r^2}.$$

Taking square roots, $|r'| = |r|s/\sqrt{1 - r^2 + r^2 s^2}$, and r' has the sign of β. The map $s \mapsto |r'|$ is increasing, equals $|r|$ at $s = 1$, and tends to 0 as $s \to 0$. □

A.6 Proof of Proposition 1

Let $\Delta(t) = L(\theta_t) - L^* \geq 0$. Along gradient flow, $\dot\theta_t = -\nabla L(\theta_t)$ and $\dot\Delta = -\|\nabla L(\theta_t)\|^2$. The PL inequality $\frac{1}{2}\|\nabla L(\theta)\|^2 \geq \mu(L(\theta) - L^*)$ gives $\dot\Delta \leq -2\mu\Delta$, hence $\Delta(t) \leq e^{-2\mu t}\Delta(0)$ by Grönwall's inequality.

Finite path length. Whenever $\Delta(t) > 0$,

$$\frac{d}{dt}\sqrt{\Delta(t)} = -\frac{\|\nabla L(\theta_t)\|^2}{2\sqrt{\Delta(t)}} \leq -\sqrt{\frac{\mu}{2}}\,\|\nabla L(\theta_t)\|,$$

using $\|\nabla L\| \geq \sqrt{2\mu\Delta}$. Integrating from t to T gives $\int_t^T \|\dot\theta_s\|\,ds \leq \sqrt{2/\mu}\left(\sqrt{\Delta(t)} - \sqrt{\Delta(T)}\right) \leq \sqrt{2\Delta(t)/\mu}$. If Δ reaches 0 at a finite time, the gradient vanishes there and the trajectory stops, so the bound still holds. Letting $T \to \infty$, $\int_t^\infty \|\dot\theta_s\|\,ds \leq \sqrt{2\Delta(0)/\mu}\,e^{-\mu t} < \infty$.

Feature variation. By the chain rule and the gradient bound on active features, $|\frac{d}{dt}\log x_i(\theta_t)| = |\langle \nabla_\theta \log x_i(\theta_t), \dot\theta_t\rangle| \leq C_i\|\dot\theta_t\|$. Therefore

$$\int_t^\infty \left|\frac{d}{dt}\log x_i(\theta_s)\right| ds \leq C_i\sqrt{\frac{2\Delta(0)}{\mu}}\,e^{-\mu t},$$

which is finite for $t = 0$ and decays exponentially in t: all but an exponentially small part of each feature's total change occurs early, which is the establishment-then-stabilization profile tested in E7. Smoothness is not needed for this argument; it only guarantees that gradient flow is well defined. □

A.7 Proof of Proposition 2

The step on $L_q = (1-q)L_B + qL_A$ is $\Delta\theta = -\eta[(1-q)g_B + qg_A]$. The first-order change of the resident loss is $\langle g_A, \Delta\theta\rangle = -\eta[(1-q)\langle g_A, g_B\rangle + q\|g_A\|^2] = -\eta[(1-q)b + qa] = \delta(q)$. (i) δ is affine in q with $\delta'(q) = -\eta(a-b)$, which is negative exactly when $b < a$; increasing q then lowers the first-order change. (ii) The marginal benefit of reinforcement is $-\delta'(q) = \eta(a-b)$. It increases with $a = \|g_A\|^2$, which is large when the resident has been displaced from its optimum, and decreases with b, so it is largest for antagonistic invaders with $b < 0$. (iii) $\delta(q) \leq 0$ iff $q(a-b) \geq -b$. If $b \geq 0$ this holds for every $q \in [0,1]$. If $b < 0$, then $a - b > 0$ and the condition is $q \geq -b/(a-b) = [-b]_+/(a + [-b]_+)$, which lies in $(0,1]$. For the exact change, if ∇L_A is M_A-Lipschitz then $L_A(\theta+\Delta\theta) = L_A(\theta) + \int_0^1 \langle \nabla L_A(\theta + t\Delta\theta), \Delta\theta\rangle dt \leq L_A(\theta) + \langle g_A, \Delta\theta\rangle + \frac{M_A}{2}\|\Delta\theta\|^2$ (the descent lemma), which gives the stated bound with $\|\Delta\theta\|^2 = \eta^2\|(1-q)g_B + qg_A\|^2$. □

B Experimental protocols

E1 natural transitions. Five binary Split-CIFAR-10 tasks; all 20 ordered pairs; task-incremental ResNet-18 with one head per task. For each pair, task A is learned, $\rho_{A\to B}$ is computed on the frozen checkpoint, task B is learned, and $F_{A\to B}$ is recorded. The whole pipeline is repeated three times and analyzed on pair means. The task-label permutation test permutes the identities of the five tasks in the ρ matrix, keeping the forgetting matrix fixed, and recomputes the correlation over the 20 off-diagonal entries for all 120 relabelings. On the three-repeat pair means this test gives one-sided $p = 0.067$ (8 of 120 relabelings at least as extreme; with five tasks the smallest attainable value is $1/120$). The separate revalidation used the permutation-invariant ρ with 256 probe examples and 20 shuffles. The single-run reproduction repeats the whole pipeline once with a new seed.

E2 prospective forecast. For each of the 20 pairs, a linear model of forgetting on ρ is fit to the other 19 and used to predict the held-out pair. The no-information baseline predicts the mean forgetting of the other 19. E1.5 compares the resident-only score with a score that also uses the reverse direction.

E3 coexistence. (a) The 20 natural E1 transitions; coexistence when $F < 0.20$; transitions are split at $\rho = 0.067$ into 10 compatible and 10 incompatible. (b) Binary tasks built within a single semantic domain of CIFAR-10 (vehicles: airplane, automobile, ship, truck; animals: the remaining six classes); 24 same-domain and 36 cross-domain directed transitions; coexistence when $F < 0.10$. (c) MNIST suite: class pairs $\{0/1, 2/7, 3/5, 4/9\}$ under identity, rotations of $15°, 45°, 90°$, and pixel permutations of 30%, 70%, 100%, three seeds, 84 transitions; mutual coexistence at threshold τ when the resident keeps at least a fraction τ of its accuracy and the invader reaches the same fraction of its reference accuracy, $\tau \in \{0.80, 0.90, 0.95, 0.99\}$. Intervals resample the 28 class-pair×transform clusters.

E4 replay. Same 20 transitions; replay buffers from 0 to 1000 resident examples in steps of 25, mixed into B training. M^* is the smallest buffer reducing forgetting below $0.1\ R_{AA}$; efficiency η is the forgetting reduction per stored example between buffers 0 and 25.

E5 interventions. Six methods on Split-CIFAR-10 and Split-CIFAR-100 (task-incremental, three seeds): fine-tuning, EWC, A-GEM, DER++, compatibility-weighted EWC, and compatibility-weighted EWC with replay. The gated test compares an EWC regularization baseline (EWC-DR) against the same method with its regularization strength multiplied by the mean of Ω from the resident tasks to the incoming task, and a gradient-gated replay against uniform replay at the matched mean weight. The pre-specified pass rule was $+1$ pp ACC with BWT not worse.

E6 scale. On the 20 E1 transitions: activation covariance, centered representation similarity, raw Jacobian overlap, and one-step gradient alignment, each measured before B training. Mechanism tests (identifiable regime: plain SGD, lr 0.03, batch 32) evaluate Jacobian-derived competition coefficients at the feature and task levels, fixed-coefficient and locally stationary Lotka–Volterra fits against constant-rate and shuffled-coefficient baselines, a 64/256/1024-probe by two-layer sweep, and R^* resource identification.

E7 stages. Fifteen runs (five transitions, three seeds). Feature statistics are recorded at epochs $\{0, 1, 5, 10, 20\}$ of task A and $\{0, 1, 5, 10\}$ of task B. Active features are the top quartile by resident strength; plasticity is the mean absolute change of log-strength between checkpoints; turnover is one minus the Jaccard overlap of active sets. Early/late contrasts compare the first and last checkpoint intervals of task-A training; invasion contrasts compare the first interval of B with the last interval of A.

E8 controlled. MNIST: fixed pixel permutations at graded fractions (96 transitions). CIFAR-10: rotations at graded angles (48 transitions). Group cross-validation holds out whole manipulation groups.

Robustness and scope. Estimator grid: $\{32, 64, 128, 256, 512, 1000\}$ probe examples $\times$ $\{1, 5, 20, 100\}$ shuffles. Label budgets: 2–100 labels per class, plus two label-free variants. Optimizers: Adam (lr 10^{-3}, batch 128), SGD (lr 0.03, batch 32 or 128), SGD with momentum (lr 0.03, batch 32), and SGD with momentum and cosine schedule (lr 0.1, batch 32). Pretrained: ImageNet ResNet-50. CIFAR-100: ten 10-class tasks under contiguous, random, and superclass-based partitions (90 directed transitions each).

C ADDITIONAL RESULTS

C.1 COEXISTENCE DETAILS

Table 5: E3c: coexistence AUC on the 84-transition suite (95% cluster-bootstrap intervals).

Threshold τ	Coexistence rate	ρ_{pre} (SLT)	Representation compatibility S	Difference
0.80	0.667	0.936 $[0.85, 0.99]$	0.866 $[0.74, 0.96]$	$+0.070$ $[0.001, 0.169]$
0.90	0.571	0.927 $[0.83, 0.98]$	0.887 $[0.76, 0.97]$	$+0.040$ $[-0.014, 0.113]$
0.95	0.476	0.966 $[0.89, 1.00]$	0.936 $[0.81, 1.00]$	$+0.030$ $[-0.005, 0.106]$
0.99	0.310	0.969 $[0.91, 1.00]$	0.940 $[0.85, 0.99]$	$+0.029$ $[-0.026, 0.112]$

Excluding the 12 identity transitions, AUC at $\tau = 0.90$ is 0.903 for ρ and 0.850 for S. In the pure-domain construction (E3b) the phase separation follows domain membership; within that construction ρ is only weakly related to forgetting ($r = -0.19$), so task relation beyond ρ contributes to coexistence there.

C.2 ESTABLISHMENT BRIDGE

Theorem 1 is stated for the establishment deficit $H_{A\to B}$, while the experiments measure Ω. The bridge analysis tests whether the two move together. For each Split-CIFAR-10 transition we take Ω from the estimator study (256 probe examples, 20 shuffles) and the incoming probe loss and clipped deficit from probes fitted on the frozen resident; both are averaged over three seeds. Because incoming tasks differ in difficulty, we remove each incoming task's mean before correlating, so the comparison is between residents facing the same invader. Within incoming task, Ω tracks the probe loss closely and the clipped deficit more loosely, because clipping at zero removes variation among well-supported transitions. On semantic CIFAR-100 the coordinate has almost no spread and the link disappears, which is the same contraction that Theorem 3 predicts for the behavioral ordering.

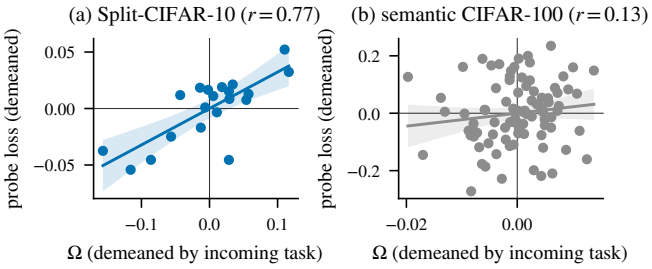


Figure 7: Establishment bridge. Within incoming task, Ω tracks the incoming probe loss on Split-CIFAR-10 ($r = 0.767$, cluster 95% CI $[0.28, 0.95]$; clipped deficit H: 0.692, $[-0.14, 0.94]$ with five clusters) but not on semantic CIFAR-100 ($r = 0.131$), where the coordinate has little spread.

C.3 REPLAY

Each of the 20 transitions is retrained with resident buffers from 0 to 1000 examples in steps of 25. Forgetting falls with buffer size in every transition, and all 20 reach the 10% criterion within 325 examples (median 100). The initial efficiency η, the drop in forgetting per stored example over the first 25 examples, is largest where displacement without replay is largest ($r = 0.70$) and on incompatible transitions ($r = -0.42$, Spearman -0.46). Proposition 2(ii) predicts both relations, since the marginal value of reinforcement grows with the resident gradient that displacement creates. The buffer at which the criterion is first met is not ordered by ρ, which fits the local character of the proposition: it describes where each additional replayed example helps most, not a universal budget.

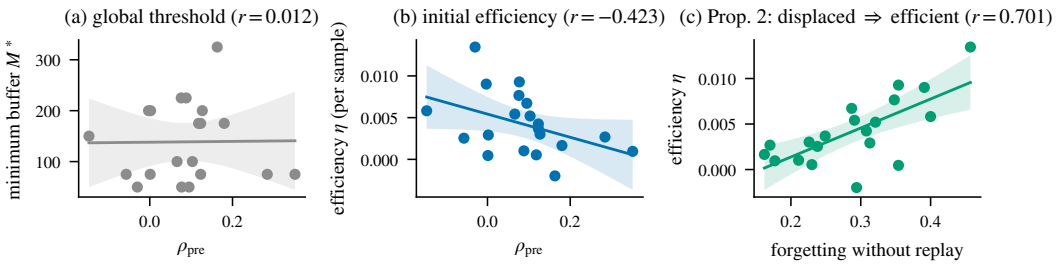


Figure 8: Replay (E4). (a) The global minimum buffer is not ordered by ρ. (b) Initial per-sample efficiency decreases with ρ. (c) Efficiency rises with forgetting at zero replay, as Proposition 2(ii) predicts; part of this relation is expected from headroom.

C.4 Interventions

The benchmark places compatibility-aware variants among standard task-incremental methods trained with the same backbone and schedule. Adding replay to compatibility-modulated regularization lifts it above fine-tuning, EWC, and A-GEM on both datasets; modulation alone helps on CIFAR-10 (ACC 0.569 vs. 0.526 for EWC) but not on CIFAR-100, and DER++ remains the strongest single method. In a separate protection study on the 20 Split-CIFAR-10 transitions (three seeds), every transition gains from protection, and the gain rises with incompatibility: across pair means the reduction in forgetting correlates 0.69 with Ω for DER++ (mean gain 0.162) and 0.55 for EWC-DR (mean gain 0.216). This is the pattern Proposition 2 predicts, since incompatible transitions displace the resident most and so leave the most to recover. The gated sign tests below compare each gate with its ungated counterpart at matched budget.

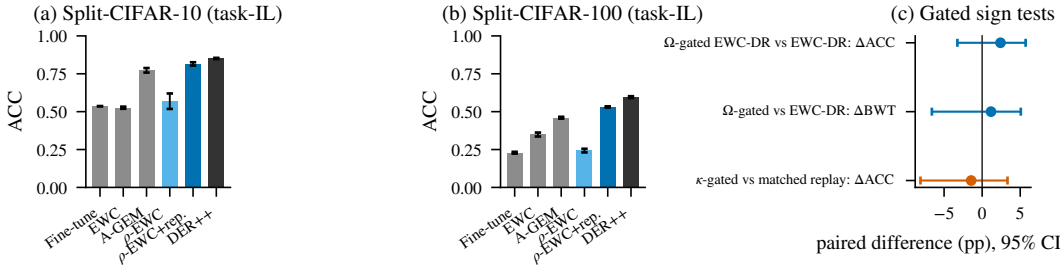


Figure 9: Interventions (E5). (a, b) Task-incremental accuracy (mean $\pm$ s.d., three seeds). (c) Paired differences for the gated sign tests.

Table 6: E5 benchmark (task-incremental, three seeds): ACC / BWT.

	Fine-tune	EWC	A-GEM	ρ-EWC	ρ-EWC+replay	DER++
Split-CIFAR-10	0.536 / −0.480	0.526 / −0.492	0.773 / −0.184	0.569 / −0.436	0.815 / −0.111	0.851 / −0.068
Split-CIFAR-100	0.229 / −0.522	0.349 / −0.330	0.460 / −0.259	0.243 / −0.501	0.531 / −0.117	0.596 / −0.071

Gated tests (three seeds): Ω-gated EWC-DR vs. EWC-DR ΔACC $+2.41$ pp $[-3.25, 5.71]$, ΔBWT $+1.16$ pp $[-6.60, 5.07]$; gradient-gated replay vs. matched-mean replay ΔACC -1.44 pp $[-8.10, 3.33]$. DER++ in the same run reaches ACC 0.862. A custom class-incremental Seq-CIFAR-100 projection test moves ACC from 0.107 to 0.121 and BWT from -0.856 to -0.841 over three seeds at poor absolute accuracy; we report it only as a sign.

C.5 Robustness

The robustness studies vary 1 part of the pipeline at a time and recompute the compatibility–forgetting correlation on the 20 Split-CIFAR-10 transitions. The estimator grid varies the number of probe

examples and shuffled references and leaves the correlation between -0.83 and -0.66 in all 24 settings; generic representation similarity (CKA) and plain probe accuracy do much worse, indicating that the normalized, directional construction carries the relevant signal. The label study replaces the full labelled sample with a few labels per class; agreement with the full score rises quickly with the label budget, while label-free variants lose the signal. The optimizer and pretrained-backbone studies retrain the residents under different regimes and initializations.

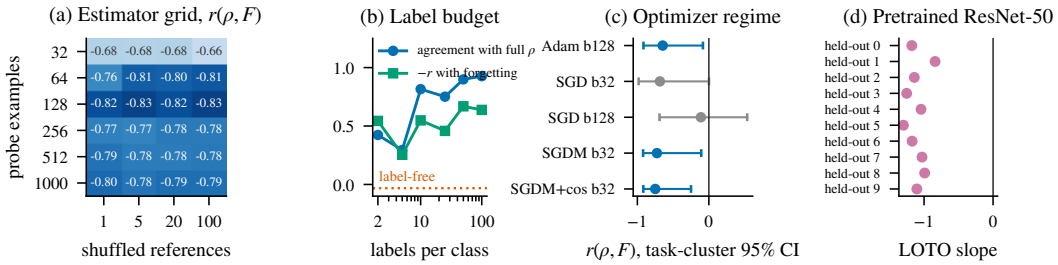


Figure 10: Robustness. (a) All 24 estimator settings. (b) Label budget: agreement with the full score and correlation with forgetting; red dotted: label-free k-means variant. (c) Optimizer regimes with task-cluster intervals (grey: interval includes zero). (d) Leave-one-task-out slopes on a pretrained ResNet-50.

Optimizer regimes (Pearson r, task-cluster 95% CI, negative LOTO slopes): Adam -0.646 $[-0.91, -0.08]$, 5/5; SGD b32 -0.685, 5/5; SGD b128 -0.111, 4/5; SGDM -0.726, 5/5; SGDM+cosine -0.749, 5/5. Label budgets (agreement with full score / r with forgetting): 2: $0.42/-0.54$; 5: $0.29/-0.25$; 10: $0.82/-0.55$; 25: $0.75/-0.46$; 50: $0.90/-0.67$; 100: $0.93/-0.64$; label-free k-means $0.05/+0.03$; label-free CKA $-0.44/+0.03$. A single-head class-incremental construction gives $r = -0.201$ ($p = 0.396$); the results are therefore stated for the task-incremental setting.

C.6 Scale tests

The scale tests examine whether a literal feature-level ecological model can replace the transition-level coordinate. All run in a regime with substantial forgetting, so a failure cannot be attributed to a lack of signal. Jacobian-derived competition coefficients are tested at the task level and per feature, with probe sets of 64, 256, and 1024 examples at two layers. Lotka–Volterra dynamics are fitted with coefficients fixed from the Jacobians and, separately, within short windows where the coefficients might be locally stable, and both are compared with constant-rate and shuffled-coefficient baselines on held-out trajectory segments. Resource-competition (R^*) constructions and a niche/fitness decomposition are tested as coexistence mechanisms. None reaches its pass criterion. The transition-level coordinate orders forgetting on the same transitions, which motivates the mesoscopic interpretation in §5.3.

Table 7: Pre-registered microscopic tests (E6). All run in a regime with forgetting s.d. 0.064.

Test	Criterion	Result
Jacobian competition, task level	$r \geq 0.5$, perm. $p \leq 0.05$	$r = 0.177$, $p = 0.54$
Jacobian competition, feature level	median partial $r > 0.10$	≤ 0.105 at 1024 probes; $\leq 6.7\%$ significant
Fixed-coefficient LV	beats baselines (Wilson > 0.5)	0/60; median held-out $R^2 = -29.2$
Locally stationary LV	stable windows exist	0/330 stable; unstable $R^2 = -1.08$
Neural R^*	resources identifiable	1.9%; precision 0.0 vs. base 0.58
Niche/fitness (ND–FD) decomposition	explains forgetting	$R^2 = 0.21$ (held-out 0.11), 45 pairs