A MODEL ZOO IS NOT A LIST: CLONE- AND REPARAMETERIZATION-INVARIANT PREDICTIVE AGGREGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A model zoo is usually treated as a list whose rows receive prior mass. Saving a checkpoint twice, sampling a regularization path more densely, or renaming its coordinate changes the aggregate, although the encoded predictive family is unchanged. We remove this *registry representation bias* with representation-invariant measure ensembling (RIME), an output-only few-shot adapter that replaces row counts with quotient mass over exact predictive atoms and Hellinger arc mass along declared scalar paths. Under the declared family, path-order, embedding, and target-slice conventions, RIME is provably invariant to exact clone splitting and smooth monotone reparameterization; finite grid refinement is only approximately stable, with a bounded residual. We evaluate a 41-row registry on four benchmarks, two of them under pre-specified protocols. Duplicating one PACS branch sixteen times moves flat-row aggregation by up to 0.583 total variation, whereas RIME is invariant up to floating-point precision and cuts near-clone drift 440-fold. A NICO++ confirmation, planned before its download, reproduces both effects. Invariance is not paid for in accuracy: RIME has the lowest mean Brier of fifteen output-only methods. Against a matched family-balanced control it ties at 4 labels/class and gains 0.00069 at 16, 0.7% of the adaptation gain and unconfirmed under target-level clustering. The evidence supports representation stability at no measured predictive cost, not general baseline superiority: a zoo's bookkeeping can stop being a modeling decision. Code is available at `https://anonymous.4open.science/r/rime-B03C`.

1 INTRODUCTION

Modern deployment increasingly begins with a collection of trained models rather than source data. A client may receive checkpoints from several domains, snapshots along a fine-tuning trajectory, regularization paths, or independently trained seeds, and then acquire only a small labeled support set at a new target. Model selection, stacking, mixtures of experts, and generalized Bayesian aggregation all offer natural ways to use such a zoo (Wolpert, 1992; Lacoste et al., 2014; Lakshminarayanan et al., 2017; Bissiri et al., 2016). A common finite implementation, however, quietly identifies the registry with the hypothesis space: a candidate row is given a unit of prior or search mass.

That identification creates a nuisance degree of freedom. Suppose one checkpoint is saved twice. A flat prior doubles its prior mass, even though the predictive family is unchanged. Equal mass per named family fixes cross-family duplication but not the density of rows within a regularization path. Declaring a hyperparameter coordinate uniform is also not intrinsic: the same path can be called uniform in λ, $\log \lambda$, or any monotone transform. Clone/relabel adds no predictor; refinement may add predictors but keeps the intended continuous family fixed. Under row counting, however, all three alter the prior measure, the predictions, and the release decisions.

We argue that the correct object is a measure on the *realized predictive family*. Let $p_h(\cdot \mid x)$ denote the class probabilities of candidate h. Exact equality of deployed prediction arrays defines quotient atoms. A smooth scalar family becomes a curve after the square-root embedding $p \mapsto \sqrt{p}$ in an empirical L_2 space over unlabeled target inputs, and arc measure on this curve does not depend on

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.

how the curve is parameterized. Labels should then determine which predictive regions perform well, not how much arbitrary registry mass those regions received.

This viewpoint leads to RIME. Exactly identical deployed predictors are merged into quotient atoms, while declared scalar families receive empirical Hellinger arc mass. Standard cross-fitted Brier evidence reweights this base measure, and the output is one deterministic posterior-mean predictor. The resulting claim is deliberately modest but, we argue, more useful than a leaderboard gain: representation invariance can be obtained for free.

Our contributions are:

- We diagnose *registry representation bias* and define two precise nuisance transformations: exact clone splitting and smooth monotone relabeling of a declared scalar predictive path.
- We propose quotient predictive mass plus empirical Hellinger arc mass. We prove exact invariance for the two transformations, prove that no order-only rule can achieve it, and give approximation bounds, rather than an exact claim, for finite grid refinement.
- We show that this invariance has no measurable predictive cost. On PACS, RIME has the lowest mean Brier loss of the fifteen output-only methods evaluated; on a pre-specified NICO++ confirmation it reproduces the invariance while matching a family-balanced control. We report the small PACS advantage together with its most conservative clustered analysis, under which it is no longer significant.

2 RELATED WORK

Prediction aggregation and model zoos. Stacking and Bayesian predictive stacking learn combinations from held-out predictions (Wolpert, 1992; Raftery et al., 1997; Yao et al., 2018). Exponential weighting and PAC-Bayesian aggregation trade empirical loss against a prior-relative complexity (McAllester, 1999; Catoni, 2007; Lacoste et al., 2014; Dalalyan & Tsybakov, 2008). Deep ensembles, trajectory ensembles, weight averaging, and model merging expose multiple trained predictors or checkpoints for robustness or accuracy (Lakshminarayanan et al., 2017; Huang et al., 2017; Garipov et al., 2018; Izmailov et al., 2018; Wortsman et al., 2022; Matena & Raffel, 2022; Ilharco et al., 2023). These methods require a finite prior or candidate list, but generally do not require the aggregate to be unchanged when one predictive path is cloned or resampled. Our question is how to define that finite base measure when the list is a discretization.

Dilution, objective priors, and information geometry. Bayesian dilution priors address redundant model spaces (George, 2010), while Jeffreys-type constructions and information geometry provide reparameterization-invariant volume elements (Jeffreys, 1946; Amari, 2016). Dilution corrects redundant discrete model spaces, whereas a Jeffreys volume requires a specified statistical parameterization. Neither directly supplies a finite rule that both pushes mass through exact predictive equivalence classes and approximates length along a target-slice prediction curve. RIME instantiates that rule; its Hellinger choice and scalar-path boundary are stated below.

Active model evaluation and selection. MODEL SELECTOR and CODA actively identify a strong pretrained model from a candidate pool (Okanovic et al., 2025; Kay et al., 2025), establishing that limited-label model-zoo decisions are important. Exact prediction-hash deduplication removes repeated atoms but leaves grid density among distinct path points unresolved. RIME addresses this remaining measure choice for fixed stratified labels and a deterministic mixture; active query design is orthogonal.

Transfer without source data. Our setting is few-shot and output-only, which places it between two literatures. Hypothesis transfer and multi-source adaptation combine fixed source predictors, with or without unlabeled target data and with varying levels of model access (Kuzborskij & Orabona, 2013; Tommasi et al., 2014; Mansour et al., 2008; Guo et al., 2018; Ahmed et al., 2021; Liang et al., 2020); training-time domain generalization instead shapes a single model to be stable across environments (Arjovsky et al., 2019; Krueger et al., 2021; Cha et al., 2021). We borrow only their benchmarks and shift structure (Li et al., 2017; Gulrajani & Lopez-Paz, 2021; Koh et al., 2021;

Zhang et al., 2023; Zhou et al., 2023): they motivate the multi-domain stress tests, not an equivalence of learning settings. A target-only feature-head diagnostic is kept in a separate access track because weight or feature access changes the comparison.

3 PROBLEM: THE REGISTRY IS A NUISANCE REPRESENTATION

A registry is a discretization of a predictive family, and the choice of discretization should not change the scientific conclusion. Let $\mathcal{F}$ index named model families and let $\mathcal{H}_f = (h_{f1}, \ldots, h_{fJ_f})$ be a finite registry. Candidate h produces $p_h(x) \in \Delta^{C-1}$. The target data enter through four roles:

- an *input slice* $X_{1:N}$, unlabeled, on which all predictions are observed and the predictive geometry is built;
- an *adaptation subset* A of labeled examples, which supplies all evidence used to reweight candidates;
- a *confirmation subset* V, labeled and disjoint from A, used only by the optional release gate;
- a *test subset*, disjoint from both, materialized only for final evaluation.

A family-balanced prior assigns an externally fixed mass π_f to family f and π_f/J_f to each row. This insulates one family's mass from the row count in another, but not from refining only one region of $\mathcal{H}_f$. A flat row prior is sensitive to both. Our scientific specification fixes the family partition, order within every scalar path, square-root embedding, target input slice, and immutable branch keys before outcomes. Changing that specification—for example by splitting a family, retracing a path, truncating endpoints, or rewriting its order—can change the result and is not an invariance claim. Conditional on this specification, we require two exact invariances.

Clone splitting. Every declared candidate descriptor carries an immutable base-branch key. Suppose one branch is replaced by exact copies—adjacent on a scalar path or unordered within a family—that inherit its key and evidence. These copies should split the branch's mass rather than recreate it, so the aggregate must not change.

Coordinate relabeling. If a scalar path is relabeled by a smooth monotone diffeomorphism, its measure and aggregate must be unchanged. Our finite rule on ordered grids satisfies a stronger property: it is invariant under any strictly monotone relabeling, because it never uses coordinate spacing.

We do *not* require a coarse grid and a refined grid to produce exactly the same answer: the refined points may be genuinely new predictors. Instead, the discrete rule should converge to a coordinate-free path measure as predictive and evidence mesh widths vanish.

The two propositions below show that this nuisance is real and that no cheap fix removes it. The first shows that a registry can silently encode an arbitrary prior; the second shows that no rule depending only on row order and row count can be made neutral to refinement.

Proposition 1 (A registry can encode an arbitrary prior)**.** *Under a flat row prior, let one branch have positive Gibbs factor q and let all other rows have finite total factor C. Replacing that branch by k exact copies gives its predictive atom posterior mass $kq/(kq + C) \to 1$ as $k \to \infty$. Moreover, for any positive continuous probability density ρ on a scalar path $t \in [0, 1]$, there is a monotone coordinate u such that the uniform row measure on grids equally spaced in u converges weakly to $\rho(t)dt$. If prediction and evidence are continuous, the corresponding row-Gibbs laws converge to the law with base density ρ. Thus multiplicity and coordinate choice can implement a prior without changing the continuous predictive path or the observed labels.*

Proposition 2 (Order alone cannot make refinement neutral)**.** *Let R_n assign strictly positive probability weights to every ordered n-row registry using only n and row rank. Suppose that whenever one row is inserted between two adjacent rows as an exact predictive clone of either neighbor, quotienting the inserted row with that neighbor recovers the original registry weights. Then no such family $\{R_n\}_{n\geq 2}$ exists.*

4 REPRESENTATION-INVARIANT MEASURE ENSEMBLING

4.1 PREDICTIVE QUOTIENT AND FAMILY MASS

Exact predictive equality, not registry position, defines the atoms that receive mass. Canonicalize each prediction array as contiguous float64 and set $h \sim h'$ exactly when the full shapes and bytes of $\{p_h(X_i)\}_{i=1}^N$ agree; write $a = [h]$. We preassign fallback mass π_0 and equal total mass to every non-fallback family. Alternative family masses are possible, but they are part of the scientific specification and must not depend on target outcomes.

Let $b(h)$ denote the immutable base-branch key in candidate h's descriptor, and let $e(h)$ extend $b(h)$ with the identity of the complete out-of-fold (OOF) evidence used by that fitted branch. Within one predictive node, geometric mass is divided equally among unique $e(h)$ values, so saved copies sharing $b(h)$ and evidence divide that branch's mass rather than recreate it. A copy thus cannot take mass from another evidence branch sharing the deployed predictor. Distinct descriptors remain distinct evidence branches when their deployed arrays coincide but their OOF evidence differs.

For an unordered family, divide its mass equally over unique quotient atoms. For a singleton, assign the full family mass. Scalar families use the construction below. After within-family assignment, quotient exact atoms globally by summing their base masses. Equality is exact, not tolerance based: near-clones are genuinely distinct predictors and belong to the approximation analysis rather than the exact theorem.

4.2 HELLINGER PREDICTIVE ARC MEASURE

Once the predictive embedding is fixed, arc length is the *only* local path measure that survives every monotone recoordinatization; the remaining design choice is the embedding itself. We choose the square-root embedding because its chords are bounded and symmetric under class permutation, stay finite when probabilities reach zero, and control ℓ_1 prediction error and hence Brier loss. These criteria motivate Hellinger geometry but do not make it unique among all possible embeddings; the theorem establishes uniqueness of arc mass only *after* the embedding is fixed. Appendix Q measures what changes under three alternatives. Embed a scalar predictive path $t \mapsto p_t$ as

$$\Psi(t)(x, c) = \sqrt{p_t(c \mid x)} \quad \text{in} \quad L_2(Q \times [C]), \tag{1}$$

where Q is the empirical uniform distribution over the complete unlabeled target slice. Its normalized arc measure is

$$d\mu_f(t) = \frac{\|\dot{\Psi}(t)\|_{L_2(Q)} dt}{\int \|\dot{\Psi}(s)\|_{L_2(Q)} ds}. \tag{2}$$

On a fixed ordered grid, use the chord

$$d(j, k) = \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} (\sqrt{p_j(c \mid X_i)} - \sqrt{p_k(c \mid X_i)})^2 \right)^{1/2}. \tag{3}$$

To discretize, merge consecutive exactly identical nodes, give each remaining node half of each incident chord, and normalize. This is trapezoidal quadrature for square-root arc measure, computed with Euclidean Hellinger chords rather than finite Fisher–Rao geodesics.

Theorem 1 (Unique local path mass for a fixed embedding)**.** *Fix the predictive embedding* (1) *and a nonconstant C^1 path. Suppose a nonzero unnormalized path measure is local and has density $g(v(t))dt$, where $v(t) = \|\dot{\Psi}(t)\|$, $g : [0, \infty) \to [0, \infty)$ is measurable, and the rule is invariant under every smooth monotone diffeomorphic reparameterization. Then $g(v) = \kappa v$ for some $\kappa > 0$. Thus every such path has the normalized arc measure in* (2)*. The discrete rule is exactly unchanged when grid values are monotonically relabeled without changing their order or predictions.*

4.3 CROSS-FITTED EVIDENCE BRANCHES

Labels enter only through cross-fitted evidence, which reweights the base measure without reintroducing row counts. Let $\Delta_i(j)$ be the out-of-fold multiclass Brier loss of candidate branch j on adaptation example i, minus the loss of the pooled fallback on that example; the Brier score is a

Algorithm 1 RIME adaptation and exact release

1: Evaluate the declared model registry on the complete target input slice.
2: Assign fallback/family mass; quotient exact predictors, compute path chords, and split node mass over immutable base-branch keys.
3: Cross-fit every candidate on the current adaptation labels to obtain $\Delta_i(j)$.
4: Refit candidate branches on all adaptation labels; form the branch-marginalized Gibbs law and posterior mean.
5: Fix $\bar{p}$ before materializing any test label.
6: Optionally, on disjoint confirmation labels, release $\bar{p}$ only if the e-value exceeds its threshold; otherwise return the exact fallback (Appendix B.3).

quadratic strictly proper scoring rule (Brier, 1950; Gneiting & Raftery, 2007). If atom a contains branches $j \in a$ with base masses μ_j, define

$$W_a = \sum_{j \in a} \mu_j \exp\left\{-\tau \sum_{i \in A} \Delta_i(j)\right\}, \qquad w_a = \frac{W_a}{\sum_b W_b}. \tag{4}$$

Distinct evidence branches of one deployed predictor stay separate in W_a. The adapted predictor is the deterministic posterior mean

$$\bar{p}(x) = \sum_a w_a p_a(x). \tag{5}$$

Algorithm 1 states the complete procedure; every step before the release gate is a deterministic function of the declared registry and the adaptation labels.

Theorem 2 (Exact clone-splitting invariance)**.** *Suppose a scalar-path branch is replaced by adjacent exact prediction-and-evidence copies, or an unordered-family branch by exact copies, and every copy inherits the original immutable base-branch key. The two-level quotient then splits the original base mass. Every quotient mass W_a, the Gibbs normalizer, family-posterior mass, and predictor* (5) *are unchanged. The same holds for global exact atoms shared across families, including when their evidence branches differ, provided all branches are marginalized as in* (4)*.*

Proposition 3 (Native-registry transport)**.** *If every row of a native registry has a distinct immutable base-branch key, the two-level clone-complete allocation equals the original equal-row allocation inside each predictive node. Hence predictions, evidence, row and atom priors, posterior means, confirmation decisions, and test metrics are identical before any clone is inserted.*

5 APPROXIMATION AND STATISTICAL SCOPE

The two exact invariances do not extend to grid refinement, so we bound the residual rather than claim it away. For probability vectors p, q, $\|p - q\|_1 \le 2\|\sqrt{p} - \sqrt{q}\|_2$ by Cauchy–Schwarz, and multiclass Brier loss is 2-Lipschitz in ℓ_1; the chord (3) is therefore tied to the operational loss.

Proposition 4 (Grid approximation)**.** *Let μ be normalized predictive arc measure and let a grid map every path point to a representative within mean predictive Hellinger distance δ_p, with error at most δ_L in the* total *evidence statistic used in the Gibbs exponent. Let $\widetilde{P}_G$ be the posterior obtained by pushing μ through that map and then using representative evidence, let P_G be the implemented half-chord posterior, and write $\epsilon_G = \mathrm{TV}(\widetilde{P}_G, P_G)$ for the posterior quadrature error in total variation (TV). At inverse temperature τ, the continuous and implemented grid posterior means satisfy*

$$\mathbb{E}_Q\|\bar{p}_\mu(X) - \bar{p}_G(X)\|_1 \le 2\delta_p + 2\tanh(\tau\delta_L) + 2\epsilon_G, \tag{6}$$

and their Brier risks differ by at most $4\delta_p + 4\tanh(\tau\delta_L) + 4\epsilon_G$.

Proof sketch. Transport each predictor to its representative for the first term. Evidence perturbations bound the likelihood ratio of normalized Gibbs laws, whose total variation is at most $\tanh(\tau\delta_L)$. A probability mixture changes in ℓ_1 by at most twice that variation. Replacing transported arc mass by the implemented quadrature contributes $2\epsilon_G$. Apply Brier Lipschitzness. □

Two scope restrictions deserve explicit mention. First, the chords themselves are estimated from the finite target slice; Appendix B.2 gives the uniform concentration rate, which covers target-input sampling only and not model-training randomness, feature refitting, label sampling, or adaptive family discovery. Second, an optional e-value gate on the disjoint confirmation subset can return the pooled fallback exactly when evidence is insufficient (Appendix B.3). That gate is an evaluation protocol, not the source of representation stability, and every invariance result below holds without it.

6 EXPERIMENTS

6.1 EVALUATION AXES AND MATCHED-ACCESS PROTOCOL

We separate two questions that a single benchmark number conflates: whether the aggregate is stable under the declared nuisance transformations, and what predictive performance that stability costs. The first uses exact-clone, near-clone, coordinate, and six-family refinement stresses. The second compares native-registry test metrics with output-only baselines and the pooled fallback.

On PACS, leave-one-domain-out banks use ResNet-50, ViT-B/16, and CLIP ViT-B/32. Each target split contains nested adaptation sets, a disjoint confirmation set, and a held-out test set; Appendix C gives quotas, seeds, optimization, calibration, and cross-fitting details. Throughout, an *architecture–target unit* is one backbone paired with one held-out PACS domain, after averaging over its splits and training seeds. Because all three architectures are evaluated on the same four target domains, the 12 units resolve only four independent targets. We report Brier loss, NLL, and expected calibration error (ECE), with lower values favorable.

The matched-access control is family-balanced exponential weighting (which we denote FABLE): it uses the same candidate library and cross-fitted Brier Gibbs update, but assigns equal row mass within each family. Two PACS campaigns with pre-specified protocols are reported. The *single-seed* campaign fixes one model-bank seed; the *five-seed same-cap* campaign fixes both methods at 20,000 solver iterations, averages five new training seeds per unit, and is the source of every performance number in §6.3. A same-fit, same-solver development ablation further isolates the measure change (Appendix P). Solver-cap and convergence checks are in Appendix F; measure construction itself costs about 0.5% of the shared candidate-fitting time at the native 41-row registry (Appendix U). Stronger-access methods are kept out of the output-only ranking.

Appendix roadmap. The appendix is organized so that each question a reader is likely to ask has one address: registry and protocol (Appendices C and E); the complete invariance stresses (Appendices S and T); design sensitivity to the embedding and to the two fixed constants (Appendices Q and R); which component of RIME moves native predictions (Appendix P); the complete comparator tables (Appendices M, O, and I); clustering and multiplicity robustness (Appendices K and L); release-gate behavior (Appendices B.3 and N); computational cost (Appendix U); and external replications including a failure analysis (Appendices W and X).

6.2 ROW PRIORS DRIFT; QUOTIENT ARC MASS REMOVES THE TARGETED DRIFT

Registry multiplicity has an operational effect, not merely a prior-level one. Figure 1a shows the drift accumulating as a single strong row is duplicated: flat-row and equal-family aggregates move steadily, while the quotient stays at floating-point zero. Over the complete 60-cell stress, multiplying one branch by 16 moves flat-row aggregation by up to 0.583 TV, enough to flip a release decision, whereas exact quotienting and RIME are invariant up to floating-point precision (maximum TV 5.2×10^{-16}; Appendix S).

Table 2 separates the two axes that a single benchmark number would confound. Quotienting barely changes native performance (-0.0615 versus -0.0637 mean ΔBrier), yet it is the only component that removes the 0.583 clone drift; arc mass supplies near-clone continuity, cutting mean drift from 3.65×10^{-3} to 4.49×10^{-5}. The stepwise decomposition in Appendix P localizes the native prediction change to exactly one step: arc mass reallocating weight inside scalar paths. Exact and global quotienting change nothing natively—the native registry has few cross-family coincidences—which is why their necessity has to be demonstrated under stress rather than on the leaderboard.

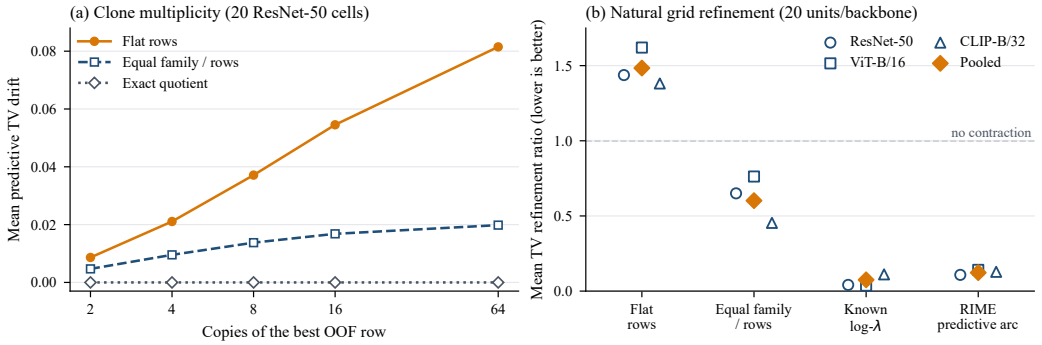


Figure 1: Row counting drifts under clone insertion and refinement, whereas quotient arc mass removes the targeted drift on the development seed. **(a)** Mean predictive TV after duplicating the best OOF row, over four targets and five splits at 4 labels/class, for the 20 ResNet-50 cells; the quotient stays at floating-point zero (maximum 5.2×10^{-16} over the complete 60-cell grid). **(b)** Half-to-quarter TV divided by coarse-to-half TV for a fixed signed-ridge path; values below one contract. Points are backbone means and diamonds pool 60 architecture–target–split cells. "Known $\log \lambda$" is an oracle that presumes the intrinsic coordinate; RIME does not.

Table 1: Every declared scalar family contracts under refinement, and arc mass sharply reduces near-clone sensitivity. Values are mean predictive TV over 60 development-seed cells. The first two columns are RIME aggregates compared with the common quarter grid; the last two compare equal-row and RIME mass under 16 inserted near-clones.

Family	RIME: coarse TV	RIME: half TV	Near clone: equal rows	Near clone: RIME
Entropy	6.27×10^{-4}	9.51×10^{-5}	6.90×10^{-4}	9.50×10^{-8}
Multinomial	2.31×10^{-3}	3.10×10^{-4}	6.44×10^{-3}	4.75×10^{-6}
Nonnegative ridge	9.90×10^{-4}	1.20×10^{-4}	1.30×10^{-3}	2.99×10^{-7}
Shrinkage path	1.13×10^{-4}	2.83×10^{-5}	2.01×10^{-4}	1.19×10^{-5}
Signed ridge	9.29×10^{-4}	1.10×10^{-4}	1.44×10^{-3}	4.70×10^{-7}
Target-logit head	9.04×10^{-4}	6.03×10^{-5}	6.45×10^{-3}	7.48×10^{-6}

Refinement behaves the same way. Figure 1b divides the drift caused by a half-to-quarter grid refinement by the drift caused by a coarse-to-half one, so values below one indicate convergence. Flat rows do not contract at all, equal-family mass contracts only weakly, and RIME contracts to a pooled ratio of 0.12. The oracle group is the informative one: a practitioner who already knows that the intrinsic coordinate of this path is $\log \lambda$ can match RIME by hand (0.07); RIME's claim is not that it beats that oracle but that no such knowledge is needed, and that the behavior holds for all six scalar families (Table 1), not only the one whose intrinsic coordinate happens to be known. Monotone coordinate relabeling leaves every RIME aggregate unchanged; changing path order or family membership remains outside the claim.

Table 2: Native utility and counterfactual correctness are distinct ablation axes, and only the second separates the variants. ΔBrier is method minus fallback on the native registry, averaged over 60 development-seed cells; the TV columns report predictive drift under inserted-copy stresses. Quotienting is silent natively but necessary for exact clones; arc mass adds near-clone continuity. All three variants have one Brier-harmful architecture–target unit, so the difference between them is entirely counterfactual.

Variant	Native ΔBrier	Exact-clone max TV	Near-clone mean TV
Flat row prior	-0.0637	0.583	1.97×10^{-2}
Equal family + exact quotient	-0.0615	5.01×10^{-16}	3.65×10^{-3}
RIME-Adapt	-0.0620	5.18×10^{-16}	4.49×10^{-5}

6.3 INVARIANCE IS FREE: A FIVE-SEED SAME-CAP COMPARISON

The advantage over the matched control is present but budget dependent, and it is so small that its sign is more informative than its size. Table 3 reports the pre-specified five-seed same-cap campaign, which fixes both methods at the same solver cap, the same 41-row registry, five target splits, all five budgets, and the analysis plan before any of its training-seed artifacts existed. At 16 labels/class all five training seeds agree in sign and the 12-unit bootstrap interval excludes zero; at 4 labels/class the effect disappears, with two thirds of the units favoring the control. At both budgets the NLL intervals still cross zero, so NLL does not separate the two methods.

We report the conservative reading alongside the favorable one. Because the 12 units resolve only four PACS targets, we recomputed every interval after first collapsing the three architectures within each shared target. Under that clustering the 16-label Brier interval is $[-0.00132, +0.00007]$ and no longer excludes zero (Appendix L also reports two-budget Bonferroni intervals). The five training-seed means span 1.8×10^{-3}, which is larger than the 6.9×10^{-4} effect itself: what replicates across seeds is the sign, not the magnitude.

The scale of the effect settles its interpretation. Adaptation moves Brier from the pooled fallback's 0.2847 to 0.1852, so the entire RIME–FABLE gap is about 0.7% of what adaptation itself buys (Table 4). We therefore read this campaign as evidence that representation invariance is *free*, not that it is *profitable*.

RIME has the lowest mean Brier of all fifteen output-only methods at both primary budgets, ahead of multinomial stacking and CAP (Appendix M). The stronger-access target-only feature head is better (0.2141 and 0.1632) and is kept outside this ranking. The ordering is not universal: in the single-seed campaign multinomial stacking is better by 0.0020 at 16 labels/class (Appendix O), which is why we claim compatibility with strong baselines rather than dominance over them.

6.4 NICO++ REPRODUCES THE MECHANISM, NOT THE ADVANTAGE

NICO++ separates the two claims cleanly: the mechanism transfers, but the advantage does not. Its corpus rules were fixed before the public release was downloaded. On this benchmark, exact-clone invariance survives at floating-point precision (5.62×10^{-16} maximum TV) while the family-balanced control drifts by 0.149, and mean half-to-quarter refinement drift falls 15.5-fold, from 6.18×10^{-4} to 3.98×10^{-5}. These are the same two effects observed on PACS. Predictive performance tells the opposite story: the two methods are indistinguishable ($+1.34\times10^{-6}$ Brier, 95% CI $[-1.21\times10^{-4}, 1.24\times10^{-4}]$ over six target contexts, with three contexts favoring each), a difference two orders of magnitude smaller than the PACS effect it fails to reproduce. Both gates abstain in all 300 cells, so this benchmark contributes evidence about representation, not about deployment. Corpus construction, cell counts, and the complete numerical disclosure are in Appendix V.

7 LIMITATIONS

What would change the conclusions. The small PACS advantage is fragile. Its 16-label interval excludes zero over 12 architecture–target units but crosses zero once those units are collapsed into their four shared targets, and the two budgets are not multiplicity adjusted. The single-seed campaign rests on one model-training seed and on 12 correlated task summaries over only four target clusters,

Table 3: Five-seed same-cap PACS comparison. Differences are RIME-Adapt minus FABLE-Adapt; negative is favorable. Unit intervals use 20,000 bootstrap resamples of the 12 architecture–target units after averaging five splits and five training seeds; target intervals first collapse the three architectures within each of the four shared PACS targets and are the more conservative reading. "Units fav. FABLE" counts units favoring the control; "Fav. seeds" counts training-seed means favoring RIME.

Labels	Metric	Mean	95% unit CI	95% target CI	Units fav. FABLE	Fav. seeds
4	Brier	-0.000454	$[-0.001655, +0.000601]$	$[-0.001739, +0.000832]$	8/12	3/5
	NLL	$+0.000692$	$[-0.003850, +0.004719]$	$[-0.004292, +0.005677]$	8/12	3/5
16	Brier	-0.000690	$[-0.001166, -0.000195]$	$[-0.001315, +0.000070]$	4/12	5/5
	NLL	-0.000083	$[-0.001912, +0.001905]$	$[-0.002167, +0.003395]$	5/12	4/5

Table 4: Three representative comparators of the fifteen output-only methods evaluated in the five-seed same-cap campaign, with the pooled fallback for scale; means are over all 300 split–training-seed cells per budget. CAP is a one-standard-error cross-validated Brier selector. Lower is better and the best displayed value is bold; the complete fifteen-method table is Table 18 in Appendix M, where RIME is also lowest.

Method	4 labels/class	16 labels/class
Pooled fallback	0.2847	0.2847
CAP-Adapt	0.2431	0.2009
Multinomial stacking	0.2408	0.1866
FABLE-Adapt	0.2215	0.1859
RIME-Adapt	**0.2211**	**0.1852**

and its target-cluster range crosses zero. The five-seed campaign removes these seed and solver-cap limitations, but its 12 units still share four PACS targets. The reported effect is therefore reproducible finite-benchmark evidence, not a population-level significance or superiority claim; in particular, we do not claim superiority over the matched control. The NICO++ parity result is likewise not formal equivalence, because no equivalence margin was fixed before evaluation.

What limits extrapolation. RIME is conditional on a declared family partition, path order, square-root embedding, target-input slice, and calibrated probabilities. Family splitting or merging, path retracing, endpoint truncation, and order rewrites are modeling changes and can alter the base measure. Exact quotienting covers clones, but unordered near-clones remain distinct, and scalar arcs do not solve singular multidimensional families. Empirically, six selected NICO++ contexts and four PACS targets are too few and too specific for population-wide inference. Office-Home and ACSIncome use one bank seed; ACSIncome harms the fallback and all 80 external cells abstain, which supports mechanism transfer rather than utility, and Appendix X implicates ranking/calibration mismatch without identifying a unique cause.

Technical caveats. The release gate's guarantee is a per-cell finite-population Brier statement, not a population, deployment-shift, benchmark-wide familywise, or NLL guarantee; on NICO++ and the external settings it abstains everywhere and therefore provides no deployment-utility evidence. The post-outcome target-slice diagnostic covers one ResNet seed and split.

8 CONCLUSION

Counting model-zoo rows lets registry bookkeeping choose a prior. RIME instead measures exact predictive atoms and length along declared predictive paths, which makes the aggregate exactly stable to clone splitting and path-coordinate relabeling and approximately stable under refinement. Two benchmarks with pre-specified protocols show that this stability is available at no measured predictive cost: on PACS RIME has the lowest mean Brier among fifteen output-only methods, and on NICO++ it matches a family-balanced control while reproducing the invariance. The claim is narrow but useful: when a zoo discretizes predictive paths, measure those paths rather than count their saved rows.

AI USE STATEMENT

Large language models (LLMs) were used as assistive tools in two roles. First, they helped polish the writing, improving grammar, wording, and readability of author-written text. Second, they helped check code, for example by reviewing implementation and analysis scripts for bugs and for consistency between reported numbers and the underlying experimental artifacts. LLMs were not used to generate research ideas, the method, theoretical results, or experimental data. The authors reviewed and verified every LLM-assisted change and take full responsibility for the content of this paper.

REPRODUCIBILITY STATEMENT

We have taken several steps to support reproduction of our results. The method is fully specified in Section 4 and Algorithm 1, including the quotient construction, the Hellinger chord quadrature, and the cross-fitted Gibbs update; all proofs appear in Appendices A and B. Every fixed constant (fallback mass, Gibbs temperature, solver cap, release margin and threshold, budgets, and bootstrap settings) is listed in Table 8, and the complete 41-row candidate registry with executable family definitions and hyperparameter grids is given in Appendix E. Table 5 reports data splits, optimizers, augmentation, learning-rate grids, split seeds, and software versions for the PACS model banks. All four benchmarks (PACS, NICO++, Office-Home, and ACSIncome) are publicly available; the corpus-construction and deduplication rules for NICO++, Office-Home, and ACSIncome are described in Appendices V and W. The pre-specified evaluation protocols and label-access rules are summarized in Appendix D. Anonymized source code, with a single entry point that runs every experiment and regenerates every table and figure, is available at `https://anonymous.4open.science/r/rime-B03C`.

REFERENCES

Sk Miraj Ahmed, Dripta S. Raychaudhuri, Sujoy Paul, Samet Oymak, and Amit K. Roy-Chowdhury. Unsupervised multi-source domain adaptation without access to source data. In *CVPR*, pp. 10103–10112, 2021.

Shun-ichi Amari. *Information Geometry and Its Applications*. Springer, 2016. doi: 10.1007/978-4-431-55978-8.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. URL `https://arxiv.org/abs/1907.02893`.

Pier Giovanni Bissiri, Chris C. Holmes, and Stephen G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B*, 78(5):1103–1130, 2016.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2.

Olivier Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. Institute of Mathematical Statistics, 2007.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. SWAD: Domain generalization by seeking flat minima. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22405–22418, 2021. URL `https://proceedings.neurips.cc/paper/2021/hash/bcb41ccdc4363c6848a1d760f26c28a0-Abstract.html`.

Arnak S. Dalalyan and Alexandre B. Tsybakov. Aggregation by exponential weighting, sharp PAC-bayesian bounds and sparsity. *Machine Learning*, 72(1–2):39–61, 2008. doi: 10.1007/s10994-008-5051-0.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL `https://proceedings.neurips.cc/paper/2021/hash/32e54441e6382a7fbacbbbaf3c450059-Abstract.html`.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P. Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of DNNs. In *Advances in Neural Information Processing Systems*, volume 31, pp. 8803–8812, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/be3087e74e9100d4bc4c6268cdbe8456-Abstract.html.

Edward I. George. Dilution priors: Compensating for model space redundancy. In *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*, pp. 158–165. Institute of Mathematical Statistics, 2010.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

Jiang Guo, Darsh Shah, and Regina Barzilay. Multi-source domain adaptation with mixture of experts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4694–4703, 2018. doi: 10.18653/v1/D18-1498. URL https://aclanthology.org/D18-1498/.

Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E. Hopcroft, and Kilian Q. Weinberger. Snapshot ensembles: Train 1, get m for free. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=BJYwwY9ll.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6t0Kwf8-jrj.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, pp. 876–885, 2018. URL https://www.auai.org/uai2018/proceedings/papers/313.pdf.

Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A*, 186(1007):453–461, 1946.

Justin Kay, Grant Van Horn, Subhransu Maji, Daniel Sheldon, and Sara Beery. Consensus-driven active model selection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4594–4604, 2025. doi: 10.1109/ICCV51701.2025.00437.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 5637–5664, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (REx). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 5815–5826, 2021. URL https://proceedings.mlr.press/v139/krueger21a.html.

Ilja Kuzborskij and Francesco Orabona. Stability and hypothesis transfer learning. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 942–950, 2013.

Alexandre Lacoste, Mario Marchand, François Laviolette, and Hugo Larochelle. Agnostic bayesian learning of ensembles. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pp. 611–619, 2014. URL https://proceedings.mlr.press/v32/lacoste14.html.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5542–5550, 2017. doi: 10.1109/ICCV.2017.591.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 6028–6039, 2020. URL https://proceedings.mlr.press/v119/liang20a.html.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. *Advances in Neural Information Processing Systems*, 21, 2008. URL https://proceedings.neurips.cc/paper/2008/hash/0e65972dce68dad4d52d063967f0a705-Abstract.html.

Michael S. Matena and Colin A. Raffel. Merging models with fisher-weighted averaging. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17703–17716, 2022. doi: 10.52202/068431-1287.

David A. McAllester. Pac-bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pp. 164–170, 1999.

Patrik Okanovic, Andreas Kirsch, Jannes Kasper, Torsten Hoefler, Andreas Krause, and Nezihe Merve Gürel. All models are wrong, some are useful: Model selection with limited labels. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258, pp. 2035–2043, 2025. URL https://proceedings.mlr.press/v258/okanovic25a.html.

Adrian E. Raftery, David Madigan, and Jennifer A. Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191, 1997. doi: 10.1080/01621459.1997.10473615.

Tatiana Tommasi, Francesco Orabona, and Barbara Caputo. Learning categories from few examples with multi model knowledge transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(5):928–941, 2014.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5018–5027, 2017.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 23965–23998, 2022. URL https://proceedings.mlr.press/v162/wortsman22a.html.

Ziyu Xu and Aaditya Ramdas. Online multiple testing with e-values. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238, pp. 3997–4005, 2024. URL https://proceedings.mlr.press/v238/xu24a.html.

Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman. Using stacking to average bayesian predictive distributions (with discussion). *Bayesian Analysis*, 13(3):917–1007, 2018. doi: 10.1214/17-BA1091.

Xingxuan Zhang, Yue He, Renzhe Xu, Han Yu, Zheyan Shen, and Peng Cui. NICO++: Towards better benchmarking for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16036–16047, 2023. doi: 10.1109/CVPR52729.2023.01539. URL https://openaccess.thecvf.com/content/CVPR2023/html/Zhang_NICO_Towards_Better_Benchmarking_for_Domain_Generalization_CVPR_2023_paper.html.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2023. doi: 10.1109/TPAMI.2022.3195549.

A PROOFS FOR THE EXACT REPRESENTATION RESULTS

Proof of Proposition 1. The clone formula follows by grouping identical unnormalized factors. Set $u = F(t) = \int_0^t \rho(s)ds$. The inverse images under F of a uniform u-grid have empirical measures converging to the distribution with density ρ. Continuity and positivity of the Gibbs factor give convergence after reweighting and normalization. □

Proof of Proposition 2. Write $R_2 = (a, 1-a)$ and $R_3 = (b, c, d)$. If the middle row clones the left endpoint, quotient consistency requires $(b + c, d) = (a, 1 - a)$. If it clones the right endpoint, it requires $(b, c + d) = (a, 1 - a)$. Rank-only dependence makes the same R_3 serve both, so $c = 0$, contradicting positivity. At rank $i+1$ the same comparison gives $b_i + b_{i+1} = a_i$ and $b_i = a_i$, forcing $b_{i+1} = 0$. Thus every neutral insertion is ignored, so no rank-only rule can be neutral to refinement; RIME instead uses predictive equality and distances. □

Proof of Theorem 1. At any point choose a local coordinate change with derivative $a > 0$. The new speed is v/a and $du = a\,dt$, so invariance requires $ag(v/a) = g(v)$ for every $a, v > 0$. Setting $a = v$ gives $g(v) = vg(1)$. At $v = 0$, $ag(0) = g(0)$ for every a forces $g(0) = 0$. The nonzero-measure assumption forces $\kappa = g(1) > 0$. Conversely, $g(v) = \kappa v$ obeys the change-of-variables identity. The discrete rule uses only declared order and predictive chords, never coordinate magnitudes. A zero-length singleton receives its family's atomic mass. For an orientation-reversing diffeomorphism, replace the derivative by its absolute value; path orientation reverses but the induced arc measure is unchanged. □

Proof of Theorem 2. Within the affected atom, identical prediction-and-evidence copies contribute

$$\sum_k \mu_{jk} \exp(-\tau \textstyle\sum_i \Delta_i(j)) = \mu_j \exp(-\tau \textstyle\sum_i \Delta_i(j)).$$

Other atom masses are unchanged. Distinct evidence branches already enter through the same finite sum, so global quotienting does not discard evidence. Normalization and Equation (5) follow. □

Proof of Proposition 3. For a node with r native rows, distinct keys create r singleton base branches, so both rules assign one rth of the fixed geometric node mass to each row. All later operations are the same deterministic functions of those masses and the unchanged prediction and evidence arrays. □

B DETAILED APPROXIMATION ARGUMENTS

B.1 PREDICTION TRANSPORT AND GIBBS PERTURBATION

We give the details behind Proposition 4. Couple every continuous-path branch h with its grid representative $r(h)$. By the square-root inequality,

$$\mathbb{E}_Q \|p_h(X) - p_{r(h)}(X)\|_1 \le 2\delta_p. \tag{7}$$

Let P be the continuous Gibbs law and let $\widetilde{P}_G$ use the representative evidence while retaining the transported base mass. If the total evidence statistic changes by at most δ_L, the density ratio before normalization lies in $[e^{-\tau\delta_L}, e^{\tau\delta_L}]$. After normalization, the likelihood ratio between P and $\widetilde{P}_G$ lies in $[e^{-2\tau\delta_L}, e^{2\tau\delta_L}]$. The sharp bounded-ratio inequality gives

$$\mathrm{TV}(P, \widetilde{P}_G) \le \tanh(\tau\delta_L). \tag{8}$$

Transport changes the mean predictor by at most (7); changing the mixing law changes a probability-valued mean by at most twice (8). Finally, the implemented chord posterior P_G differs from $\widetilde{P}_G$ by ϵ_G in total variation and hence by at most $2\epsilon_G$ in posterior-mean prediction. Summing proves (6). Since the gradient of $L(p, y) = \|p - e_y\|_2^2$ has ℓ_∞ norm at most two on the simplex, Brier loss is 2-Lipschitz in ℓ_1, yielding the stated risk bound.

The ϵ_G term is necessary at finite resolution: coordinate invariance does not make half-chord quadrature equal to transported continuous arc mass. For a regular finite-length path with uniformly continuous prediction and evidence integrands, both chord arc sums and the associated endpoint quadrature converge under vanishing arc mesh, so $\epsilon_G \to 0$. We measure finite-grid contraction directly in Figure 1 rather than assuming this asymptotic regime has already been reached.

An explicit chord-quadrature rate. Let $0 = t_0 < \cdots < t_J = 1$, $h = \max_j(t_j - t_{j-1})$, and $\ell = \int_0^1 \|\dot\Psi(t)\|dt > 0$. If $\dot\Psi$ is K-Lipschitz, then the polygonal length $\ell_G = \sum_j \|\Psi(t_j) - \Psi(t_{j-1})\|$ satisfies

$$0 \le \ell - \ell_G \le Kh/3. \tag{9}$$

Let $\widetilde{\mu}_G$ give each node half of each adjacent *true* arc segment and let μ_G use the implemented half-chords. Then $\mathrm{TV}(\widetilde{\mu}_G, \mu_G) \le Kh/(3\ell)$. If the common representative Gibbs factors obey $0 < g_- \le g_j \le g_+$, the posterior quadrature term in Proposition 4 obeys

$$\epsilon_G \le \min\left\{1, \frac{2g_+}{g_-}\frac{Kh}{3\ell}\right\}. \tag{10}$$

Indeed, on an interval of width Δ, with $\bar v = \Delta^{-1}\int \dot\Psi(t)dt$,

$$\int \|\dot\Psi(t)\|dt - \Delta\|\bar v\| \le \int \|\dot\Psi(t) - \bar v\|dt \le K\Delta^2/3.$$

Summing proves (9). The true half-arc node masses dominate the half-chord masses componentwise and differ in total by $\ell - \ell_G$; normalization therefore gives the base-measure TV bound. Reweighting two laws by the same factor in $[g_-, g_+]$ expands TV by at most $2g_+/g_-$, proving (10).

B.2 Unlabeled target-input concentration

For fixed candidate pairs, each squared chord summand lies in $[0, 2]$. Hoeffding and a union bound over M fixed chords give, with probability $1 - \delta$,

$$\max_{m \le M} |\widehat{d}_m^2 - d_m^2| \le \sqrt{\frac{2\log(2M/\delta)}{N}}, \qquad \max_m |\widehat{d}_m - d_m| \le \left(\frac{2\log(2M/\delta)}{N}\right)^{1/4}. \tag{11}$$

The fourth-root distance bound is conservative near zero. This result covers target input sampling only, not model-training randomness, feature refitting, label sampling, or adaptive family discovery.

For a fixed pair (j, k) define $Z_i = \sum_c (\sqrt{p_{jc}(X_i)} - \sqrt{p_{kc}(X_i)})^2 \in [0, 2]$. Hoeffding gives

$$\Pr(|N^{-1}\textstyle\sum_i Z_i - \mathbb{E}Z_i| > \epsilon) \le 2\exp(-N\epsilon^2/2). \tag{12}$$

A union bound over M pairs proves the squared-distance part of (11). For nonnegative a, b, $|\sqrt{a} - \sqrt{b}| \le \sqrt{|a - b|}$, which gives the conservative fourth-root distance bound. The implementation uses the complete finite target slice and does not reinterpret this statement as control over model or label randomness.

B.3 Optional disjoint confirmation release

RIME-Deploy is an optional evaluation protocol layered on top of the adapted predictor; none of the invariance results depend on it. Let $D = L(\bar p, Y) - L(p_0, Y)$ for Brier loss and fixed margin $m = 0.002$. Under the fixed stratified without-replacement confirmation design, the betting product below is an e-value for the finite eligible-pool null $\bar D + m \ge 0$ (Xu & Ramdas, 2024). RIME-Deploy releases $\bar p$ only when the e-value exceeds the fixed threshold 160, giving a per-cell false-release bound of $1/160$ under that null, and otherwise returns p_0 exactly. This guarantee concerns Brier loss in the finite eligible pool; it is not a population, deployment-shift, benchmark-wide familywise, or negative log-likelihood (NLL) guarantee. The gate is deliberately selective: it releases 8/60 and 15/60 cells on the development seed, 11/60 and 13/60 in the single-seed campaign, and 47/300 and 81/300 in the five-seed same-cap campaign at the two primary budgets, while abstaining in all 300 NICO++ cells and all 80 external cells. Its realized utility is analyzed in Appendix N.

Condition on the fixed maximum-adaptation reservation, the current-budget adaptation data, all prediction arrays, and the eligible finite population in each prespecified stratum. The probability below is over the uniform stratified confirmation assignment before its seed-keyed stable-hash permutation is realized, not over labels after a realized split is fixed. In stratum s, let $a_{sj} = 1 - \lambda(D_{sj} + m) > 0$ for the N_s remaining rows, and draw n_s rows uniformly without replacement. Maclaurin's inequality for elementary symmetric means gives

$$\mathbb{E}\left[\prod_{j \in V_s} a_{sj}\right] = \frac{e_{n_s}(a_{s1}, \ldots, a_{sN_s})}{\binom{N_s}{n_s}} \le \left(1 - \lambda(\bar D_s + m)\right)^{n_s}. \tag{13}$$

The strata are sampled independently. Writing $n = \sum_s n_s$ and $\bar{D} = \sum_s (n_s/n)\bar{D}_s$, weighted AM–GM therefore yields

$$\mathbb{E}\left[\prod_s \prod_{j \in V_s} a_{sj}\right] \le \prod_s \big(1 - \lambda(\bar{D}_s + m)\big)^{n_s} \le \big(1 - \lambda(\bar{D} + m)\big)^n \le 1 \tag{14}$$

under the macro-null $\bar{D} + m \ge 0$. The betting grid has $\lambda \in \{0.1, \dots, 0.9\}/(2+m)$, so positivity follows from $D_{sj} \in [-2, 2]$. A fixed mixture of the nine products remains an e-value; Markov's inequality gives the stated per-cell $1/160$ false-release bound. Equal class quotas make $\bar{D}$ the macro-class finite-population contrast. No claim is made after conditioning on the confirmation outcomes or uniformly over the 60 evaluated cells.

B.4 Variational characterization and quotient push-forward

Let Π be the branch-level base measure after family and path mass assignment, and let $\widehat{L}(h) = \sum_{i \in A} \Delta_i(h)$. The Gibbs law

$$dQ^*(h) = \frac{e^{-\tau \widehat{L}(h)} d\Pi(h)}{\int e^{-\tau \widehat{L}(g)} d\Pi(g)} \tag{15}$$

uniquely minimizes, up to Π-null sets,

$$\mathbb{E}_Q \widehat{L}(h) + \tau^{-1} \mathrm{KL}(Q \| \Pi). \tag{16}$$

This is the standard Gibbs variational identity, not a new PAC-Bayes theorem. Pushing Q^* through $h \mapsto [p_h]$ produces exactly the atom weights in Equation (4). This explains why distinct evidence branches inside one predictive atom must be summed rather than arbitrarily retaining one branch.

C Reproducibility specification

Table 5 collects the PACS protocol; the released code implements exactly these choices. No language model is part of the method or the experimental pipeline.

Table 5: PACS training, calibration, and evaluation specification.

Component	Specification
Source partition	Per-domain, class-stratified content-hash split: 70% train, 10% learning-rate selection, 10% early stopping, 10% temperature calibration
Optimization	AdamW, cosine decay, inverse-square-root class weighting, mixed precision, label smoothing 0.05; at most 12 epochs, patience 3 on validation NLL
Augmentation	Random resized crop to 224 (scale 0.7–1.0), horizontal flip, color jitter; validation uses resize 256 and center crop 224
Backbones	ResNet-50: layer4+classifier, batch 12, weight decay 10^{-4}; ViT-B/16: last two blocks+head, batch 6, weight decay 10^{-2}; CLIP ViT-B/32: last two blocks+projection+head, batch 6, weight decay 10^{-2}
Learning rates	Source-only grids: ResNet $\{3\times10^{-5}, 10^{-4}, 3\times10^{-4}, 10^{-3}\}$; ViT $\{10^{-5}, 3\times10^{-5}, 10^{-4}, 3\times10^{-4}\}$; CLIP $\{5\times10^{-6}, 10^{-5}, 3\times10^{-5}, 10^{-4}\}$; head multipliers 10, 10, and 100
Target splits	Split seeds 20260821–20260825; nested budgets 1, 2, 4, 8, 16 labels/class; 24 confirmation labels/class; disjoint test remainder
Cross-fitting	Stratified leave-one-out at one label/class; otherwise $k = \min(5, b)$ class-stratified folds from the stable split/target/budget seed
Evaluation software	Python 3.10.13, NumPy 2.2.6, pandas 2.3.3, SciPy 1.15.3, scikit-learn 1.7.2; exact package versions are pinned in the released code

D Evaluation protocol and label access

Two PACS campaigns with pre-specified protocols are reported. The *single-seed* campaign uses model-bank seed 20261014; its seed, constants (Table 8), analysis unit, endpoints, and reporting

conditions (Appendix H) were fixed before any model for that seed was trained. The *five-seed same-cap* campaign is described in Appendix D.1.

The single-seed campaign was first evaluated with equal row mass inside each predictive node. Proposition 3 shows that the clone-complete branch allocation coincides with this rule on every native registry. Recomputing all 300 architecture–target–split–budget cells, including 41 cells with multiple evidence branches at one predictive atom, gives a maximum row-prior difference of exactly zero and identical outputs. The reported posterior means, release decisions, and single-seed metrics are therefore those of the clone-complete method; only counterfactual registries containing inserted copies are affected by the branch split.

D.1 FIVE-SEED SAME-CAP CAMPAIGN PROTOCOL

The five-seed same-cap protocol was fixed before any artifact for its five deterministically derived training seeds existed. It fixes five seeds, three architectures, four PACS targets, five split seeds, budgets 1/2/4/8/16, equal 20,000-iteration solver caps, and 20,000 cluster-bootstrap resamples with seed 2026492702. The independent unit is architecture–target after averaging both split and training seeds. The aggregate contains 3,000 RIME cells, 3,000 FABLE cells, and 24,000 full-baseline cells.

Table 6: Pre-specified primary comparison for the five-seed same-cap campaign. "95% cluster boot." is the bootstrap over the 12 architecture–target units; the target-collapsed intervals of Table 3 are computed post-outcome from the same units.

Labels/class	ΔBrier	95% cluster boot.	Positive clusters	Favorable seeds
4	-0.000454	$[-0.001655, +0.000601]$	8/12	3/5
16	-0.000690	$[-0.001166, -0.000195]$	4/12	5/5

Table 7: Five-seed same-cap training-seed means of RIME-Adapt minus FABLE-Adapt Brier over 12 architecture–target units. Negative is favorable. The five means span 1.8×10^{-3}, which exceeds the pooled effect: the sign replicates, the magnitude does not.

Training seed	4/class	16/class
13807578	$+0.000519$	-0.000524
80452493	-0.000698	-0.000677
70145267	-0.000810	-0.000766
22616572	-0.001327	-0.000765
12496800	$+0.000048$	-0.000717

One numerical rule applies to every signed-ridge and SCLA fit in all experiments: when L-BFGS-B ends with an abnormal line search or at its function-evaluation limit, the fit is re-solved with SLSQP from the identical objective, initialization, and bounds. On PACS this occurs in one full-baseline unit (ViT-B/16, photo, seed 22616572); it affects descriptive baseline reporting only, and the primary RIME–FABLE comparison is unchanged.

The evaluator reads only calibrated logits and canonical row identifiers from the prediction bank and never deserializes checkpoints; the hidden features stored for the separately labeled feature-access baseline are used only by that baseline. For each budget, candidate fitting, cross-fitting, and posterior construction use labels of the current adaptation rows only, the release decision uses the confirmation rows only, and test labels are used only to score the returned predictor. The confirmation labels affect the binary release but never candidate fitting or posterior weights.

For H cached predictors, N inputs, and C classes, measure construction is $O(HNC)$ and avoids an H^2 similarity graph; candidate fitting and cross-fitting, shared with the matched adapters, remain the dominant cost.

Table 8: Fixed RIME constants and analysis units for the single-seed PACS campaign. The five-seed same-cap campaign reuses every constant and replaces the bootstrap seed with 2026492702 (Appendix D.1). The exact-invariance tolerance is the pre-specified acceptance threshold; measured exact-clone drift is four orders of magnitude smaller.

Quantity	Value
Fallback mass / Gibbs temperature	0.10 / 1.0
Logistic solver / maximum iterations	L-BFGS / 20,000
Confirmation Brier margin / e-threshold	0.002 / 160
Budgets per class	1, 2, 4, 8, 16
Primary budgets	4, 16
Prespecified analysis unit	architecture–target after split averaging
Bootstrap	20,000 unit resamples; seed 20261015
Exact-invariance tolerance	2×10^{-12}

E COMPLETE CANDIDATE REGISTRY

The registry is shared with FABLE, and its order was fixed before either confirmation campaign. It contains 41 rows: one pooled fallback; uniform sources and uniform all; three fixed sources; probability and convex-logit stacking; signed and nonnegative ridge at $\{10, 1, 0.1, 0.01, 0.001, 0\}$; Source-Contrastive Logit Adaptation (SCLA) at $\{1, 0.1, 0.01, 0.001, 0\}$; entropy gates at $\{0, 0.5, 1, 2, 4, 8\}$; and target-logit and multinomial-stacking heads at $C \in \{0.01, 0.1, 1, 10, 100\}$. These last six are scalar families, and their fixed registry order defines path order. Fixed source is unordered; every remaining family is a singleton. Each distinct non-fallback family receives equal total mass after the fallback's 0.10 mass is removed.

Table 9: Executable definitions for the fitted output-only candidate families. Every objective uses adaptation labels only; hyperparameters and fitted predictions enter the same cross-fitting protocol. Probability stacking follows the classical stacked generalization template (Wolpert, 1992); nonnegative transfer variants follow hypothesis-transfer regularization (Kuzborskij & Orabona, 2013; Tommasi et al., 2014).

Family	Fitted objective / constraints	Path or solver grid
Probability stacking	Target-label NLL of a simplex mixture of source probabilities	SLSQP simplex optimization
Convex-logit stacking	Target-label NLL of a simplex source-logit mixture with temperature	Temperature $[0.25, 4]$
Signed ridge	Fallback-centered source-logit directions; NLL plus $\lambda\|w\|_2^2$	Coefficients bounded from -3 to 3; the six-value λ grid above
Nonnegative ridge	Same objective with nonnegative equivalent source weights	Same λ grid
SCLA	Signed-ridge objective plus a penalty on negative equivalent mass	$\gamma \in \{1, 0.1, 0.01, 0.001, 0\}$
Entropy mixture	Source-only softmax gate using negative predictive entropy	$\eta \in \{0, .5, 1, 2, 4, 8\}$
Target-logit head	L2 multinomial logistic regression on fallback logit contrasts	$C \in \{0.01, 0.1, 1, 10, 100\}$; L-BFGS
Multinomial stacking	L2 multinomial logistic regression on concatenated source-logit contrasts	Same C grid and solver

The CAP baseline selects from its fixed adapter portfolio by cross-validated Brier loss with a one-standard-error rule. FABLE assigns equal family and row mass before the same cross-fitted Brier Gibbs update; RIME replaces only that base measure and adds the clone-complete branch construction described in Algorithm 1.

F RIME AND FABLE REMAIN CLOSE ON THE DEVELOPMENT SEED

At 4 labels per class, RIME-Adapt minus FABLE-Adapt has mean Brier difference -0.000439 and mean NLL difference $+0.001757$ across the 12 architecture–target units; RIME wins Brier on 6/12 units. At 16 labels per class the corresponding values are -0.000408, $+0.000800$, and 7/12. Adapted predictors have 1 Brier-harmful and 3 NLL-harmful units at budget 4, and zero at budget

16. The disjoint gate reduces both Brier and NLL harmful deployment-unit counts to zero at both primary budgets.

The logistic convergence analysis, run on additional model-bank seed 20261003, spans 240 sensitivity cells and 16,500 fits. Raising the iteration cap to 20,000 removes all final and CV warnings. Relative to a 3,000-iteration cap, no release decision changes, maximum posterior ℓ_1 change is 0.0151, maximum aggregate probability change is 0.0277, and maximum single-cell Brier-delta change is 0.00110. These diagnostics support numerical convergence but are not confirmation evidence.

The single-seed campaign ran RIME and FABLE at logistic iteration caps of 20,000 and 3,000 respectively. A post-outcome same-cap reconstruction of all 120 FABLE cells changed no release decision; the maximum change in a split-cell Brier difference is 3.89×10^{-16}. This rules out cap sensitivity in those cells, although the check is retrospective. The five-seed campaign fixes both methods at 20,000 iterations in advance, so its comparison needs no such check.

Table 10: Development-seed RIME minus FABLE mean Brier. Negative is favorable; intervals resample 12 architecture–target summaries.

Budget	Adapt diff.	95% unit boot.	Deploy diff.	Releases
4/class	-0.0004	$[-0.0013,\ 0.0004]$	-0.0001	8/60
16/class	-0.0004	$[-0.0009,\ 0.0000]$	-0.0000	15/60

G ADDITIONAL MODEL-BANK SEEDS BOUND TRAINING-SEED SENSITIVITY

The pre-specified claim remains attached only to seed 20261014. As a post-outcome sensitivity analysis, we evaluated RIME on three additional PACS model-bank seeds (the development seed and two post-hoc seeds) and paired each cell with the FABLE evaluation of the same cell. Table 11 reports split-averaged RIME–FABLE Brier differences and labels the development, post-hoc, and single-seed rows explicitly; seed means are descriptive because the four seeds were not sampled as a training-randomness study.

Table 11: Training-seed sensitivity over the same 12 architecture–target units per seed. Negative values favor RIME. "Single seed" is the pre-specified single-seed campaign; the remaining rows are post-outcome evidence.

Model-bank seed	4 labels/class	16 labels/class
20260820 (development)	-0.0004	-0.0004
20260922 (post-hoc)	-0.0009	-0.0012
20261003 (post-hoc)	-0.0007	-0.0004
20261014 (single seed)	-0.0006	-0.0016
Mean $\pm$ SD across seeds	-0.0006 $\pm$ 0.0002	-0.0009 $\pm$ 0.0006

H PRE-SPECIFIED PERFORMANCE REPORTING CONDITIONS

The single-seed aggregation pairs RIME, FABLE, and baseline cells over all 12 architecture–target units, five splits, and five budgets. It outputs paired cells and split-averaged architecture–target units, 20,000-resample intervals, and two reporting conditions:

1. The performance-compatibility condition requires nonpositive mean RIME–FABLE Brier at both primary budgets, a strict improvement for at least one, and zero Brier/NLL harmful RIME-Deploy units versus fallback.
2. The interval-reporting condition additionally requires the upper 95% unit-bootstrap Brier endpoint to be below zero for at least one primary budget.

Both conditions passed, but neither evaluates representation invariance and neither is a population or multiplicity-adjusted hypothesis test.

I STRONG BASELINES EXPOSE BUDGET-DEPENDENT ORDERING

Table 12: Single-seed output-only descriptive comparison. Values average the 12 architecture–target units after split averaging; lower is better. No row is selected for inferential emphasis.

Method	Brier (4)	NLL (4)	Brier (16)	NLL (16)
Pooled fallback	0.309294	0.724783	0.309294	0.724783
Uniform sources	0.326948	0.725604	0.326948	0.725604
Uniform all	0.285240	0.650511	0.285240	0.650511
Best source	0.338545	0.815495	0.331783	0.815185
Probability stacking	0.251644	0.567603	0.241258	0.539883
Convex-logit stacking	0.234332	0.616636	0.222944	0.494253
Entropy MoE	0.282302	0.625832	0.283149	0.628299
Nonnegative ridge HTL	0.232180	0.574389	0.222314	0.488134
Signed ridge HTL	0.234782	0.587926	0.223758	0.492911
SCLA-CV	0.236856	0.630491	0.224694	0.498410
SCLA-BrierCV-1SE	0.253396	0.619558	0.237893	0.537614
CAP-Adapt	0.240292	0.559439	0.201800	0.442780
Target-only logit head	0.257136	0.624755	0.215056	0.510468
Multinomial stacking	0.226660	0.530541	0.185026	0.407038
FABLE-Adapt	0.217573	0.472167	0.188601	0.398595
RIME-Adapt	0.216949	0.472103	0.187036	0.396562

Table 13: Single-seed confirmation-release track. These methods use the disjoint confirmation labels for a release decision and are not ranked with the adaptation-only track.

Method	Brier (4)	NLL (4)	Brier (16)	NLL (16)
Pooled fallback	0.309294	0.724783	0.309294	0.724783
CAP-Deploy	0.246571	0.560514	0.222590	0.497106
FABLE-Deploy	0.236767	0.534168	0.218137	0.481849
RIME-Deploy	0.236705	0.533936	0.218056	0.481685

Table 14: Single-seed white-box feature-access diagnostic, reported separately from output-only methods.

Method	Brier (4)	NLL (4)	Brier (16)	NLL (16)
Target-only feature head	0.216547	0.491556	0.167363	0.366479

J TARGET CLUSTERING WEAKENS THE SINGLE-SEED INTERVAL CLAIM

The table reports leave-one-target/architecture-out (LOTO) sensitivity in addition to cluster ranges. The dependence-robust Brier check passed; the release-cluster safety check passed.

Table 15: Single-seed dependence sensitivity across shared target and architecture clusters. A/D denote Adapt/Deploy; negative is favorable. Ranges contain all cluster means; worst LOTO is the largest leave-one-cluster-out mean. These diagnostics are descriptive.

Budget	Comparison	Metric	Task mean	Cluster range	Worst LOTO
Target clusters (three architectures per target)					
4	RIME-A − FABLE-A	BRIER	-0.000624	[-0.002408, 0.000798]	-0.000029
4	RIME-A − FABLE-A	NLL	-0.000064	[-0.007704, 0.007808]	0.002482
4	RIME-D − fallback	BRIER	-0.072589	[-0.280850, 0.000000]	-0.003169
4	RIME-D − fallback	NLL	-0.190848	[-0.728492, 0.000000]	-0.011633
16	RIME-A − FABLE-A	BRIER	-0.001564	[-0.004055, 0.000516]	-0.000734
16	RIME-A − FABLE-A	NLL	-0.002033	[-0.009946, 0.005877]	0.000604
16	RIME-D − fallback	BRIER	-0.091238	[-0.324948, 0.000000]	-0.013335
16	RIME-D − fallback	NLL	-0.243098	[-0.838324, 0.000000]	-0.044689
Architecture clusters (four targets per architecture)					
4	RIME-A − FABLE-A	BRIER	-0.000624	[-0.001059, -0.000380]	-0.000406
4	RIME-A − FABLE-A	NLL	-0.000064	[-0.000496, 0.000480]	0.000151
4	RIME-D − fallback	BRIER	-0.072589	[-0.110775, 0.000000]	-0.053496
4	RIME-D − fallback	NLL	-0.190848	[-0.296997, 0.000000]	-0.137773
16	RIME-A − FABLE-A	BRIER	-0.001564	[-0.002932, -0.000386]	-0.000881
16	RIME-A − FABLE-A	NLL	-0.002033	[-0.004820, 0.000831]	-0.000640
16	RIME-D − fallback	BRIER	-0.091238	[-0.151876, 0.000000]	-0.060919
16	RIME-D − fallback	NLL	-0.243098	[-0.418347, 0.000000]	-0.155473

K CLUSTER AND MULTIPLICITY SENSITIVITY REMOVES THE INFERENTIAL HEADLINE (SINGLE-SEED CAMPAIGN)

The pointwise unit interval at 16 labels/class excludes zero, as does the two-budget Bonferroni unit interval by a narrow margin. Once the three architectures are first collapsed within their four shared targets, however, the multiplicity-adjusted interval crosses zero. With only four target clusters, this bootstrap is low-resolution and is not a population test; this is why we treat the single-seed result as a small matched-control effect rather than an inferential superiority claim.

Table 16: Post-outcome dependence and multiplicity sensitivity for single-seed RIME-Adapt minus FABLE-Adapt Brier. Bonferroni intervals use 97.5% marginal coverage for the two primary budgets. Unit intervals resample 12 architecture–target means; cluster intervals resample four target means after averaging the three architectures.

Labels/class	Mean	Task Bonferroni CI	Target-cluster Bonferroni CI
4	-0.0006	[-0.0018,+0.0003]	[-0.0019,+0.0005]
16	-0.0016	[-0.0035,-0.0001]	[-0.0036,+0.0003]

L CLUSTERING AND MULTIPLICITY IN THE FIVE-SEED SAME-CAP CAMPAIGN

The 12 architecture–target units of the five-seed campaign resolve only four PACS targets, so the pre-specified unit interval is the optimistic reading of its dependence structure and the two primary budgets are not multiplicity adjusted. Table 17 therefore recomputes every interval under the two conservative alternatives. At 16 labels/class the unit interval survives a two-budget Bonferroni adjustment, but the target-cluster interval crosses zero at either coverage; at 4 labels/class no interval excludes zero. With four clusters the bootstrap is low-resolution and is not a population test. This sensitivity is why we claim "invariance is free" rather than "invariance is better".

Table 17: Post-outcome clustering and multiplicity sensitivity for five-seed same-cap RIME-Adapt minus FABLE-Adapt Brier. Unit intervals resample the 12 architecture–target means; target-cluster intervals resample the four target means obtained after averaging the three architectures. Bonferroni intervals use 97.5% marginal coverage for the two primary budgets.

Labels/class	Mean	Unit 95% CI	Unit Bonferroni CI	Target-cluster Bonferroni CI
4	-0.0005	[-0.0017,+0.0006]	[-0.0018,+0.0007]	[-0.0019,+0.0010]
16	-0.0007	[-0.0012,-0.0002]	[-0.0012,-0.0001]	[-0.0014,+0.0002]

M COMPLETE FIVE-SEED SAME-CAP COMPARATOR TABLE

Table 4 shows three representative comparators for reasons of space. Table 18 gives all methods evaluated in the same campaign, each averaged over the identical 300 split–training-seed cells per budget, and groups them by access track. RIME-Adapt has the lowest mean Brier of the fifteen output-only methods at both primary budgets, and RIME-Deploy is lowest among the gated variants; the supervised feature head is stronger but uses feature access and is therefore reported outside the ranking. Ordering among the weaker comparators is budget dependent, which is why we make a compatibility claim rather than a dominance claim.

Table 18: All methods evaluated in the five-seed same-cap campaign, sorted within access track by mean Brier loss at 16 labels/class over 300 split–training-seed cells per budget. Bold marks the best output-only value in each column. Gated rows abstain to the pooled fallback in most cells, so their means are not comparable with the ungated rows.

Access track	Method	4 labels/class	16 labels/class
Reference	Pooled fallback	0.2847	0.2847
Output-only	RIME-Adapt	**0.2211**	**0.1852**
	FABLE-Adapt	0.2215	0.1859
	Multinomial stacking	0.2408	0.1866
	CAP-Adapt	0.2431	0.2009
	Target-logit head	0.2515	0.2076
	Nonnegative ridge	0.2314	0.2196
	Signed ridge	0.2336	0.2197
	SCLA (CV)	0.2368	0.2209
	Convex-logit stacking	0.2346	0.2210
	Probability stacking	0.2439	0.2346
	SCLA (1-SE)	0.2568	0.2397
	Uniform all	0.2805	0.2805
	Entropy gate	0.2956	0.2916
	Best single source	0.3291	0.3182
	Uniform sources	0.3271	0.3271
Output-only, gated	RIME-Deploy	0.2438	0.2153
	FABLE-Deploy	0.2439	0.2167
	CAP-Deploy	0.2506	0.2245
Feature access	Target-only feature head	0.2141	0.1632

N THE PACS GATE IS SELECTIVE BUT NON-VACUOUS

This section is a post-outcome descriptive analysis of the single-seed cells; it does not alter the pre-specified threshold, margin, endpoints, or gates. At the fixed $E = 160$ rule, RIME releases 11/60 cells at 4 labels/class and 13/60 at 16 labels/class. Every released cell improves test Brier, NLL, accuracy, and ECE over its exact fallback, and split-averaged deployment remains non-harmful on all 12 architecture–target units. The conditional mean Brier differences among released cells are -0.395940 and -0.421100, respectively; zero gated harm therefore does not come from vacuous release, although the rule is deliberately selective.

Table 19: Single-seed campaign release operating point. Differences are deployed minus fallback, so negative Brier is favorable. "Harmful rel." counts released split cells; "Harmful tasks" counts architecture–target units after split averaging. This table is post-outcome descriptive.

Budget	Releases	Rate	ΔBrier	Harmful rel.	Harmful tasks
4	11/60	0.183	-0.072589	0/11	0/12
16	13/60	0.217	-0.091238	0/13	0/12

O RIME MATCHES FIVE OF SIX STRONG OUTPUT-ONLY COMPARATORS

Table 20 pairs RIME-Adapt with prespecified strong output-only methods on the same five split seeds and then averages to the 12 architecture–target units. Pointwise unit-bootstrap intervals are descriptive and are not multiplicity adjusted. At 4 labels/class, RIME has favorable mean Brier against all six comparators, with intervals below zero for CAP, convex-logit, nonnegative-ridge, and probability stacking. At 16 labels/class it remains favorable against five comparators; multinomial stacking has lower mean Brier by 0.002010, with an interval crossing zero. This supports a stability–performance compatibility claim, not universal output-only dominance.

Table 20: Post-outcome paired output-only Brier comparisons. Differences are RIME-Adapt minus comparator; negative is favorable. Intervals resample the 12 architecture–target units with 20,000 replicates. "Worst" is the largest task difference.

Budget	Comparator	ΔBrier	95% interval	Wins	Worst
4	CAP_Adapt	-0.023343	[-0.037006, -0.012256]	11/12	+0.000877
4	FABLE_Adapt	-0.000624	[-0.001617, +0.000232]	6/12	+0.001359
4	convex_logit	-0.017383	[-0.029297, -0.006208]	8/12	+0.005058
4	multinomial_stacking	-0.009711	[-0.024416, +0.003899]	6/12	+0.034350
4	nonnegative_ridge_htl	-0.015231	[-0.026693, -0.005280]	8/12	+0.003629
4	probability_stacking	-0.034695	[-0.056029, -0.015677]	12/12	-0.000823
16	CAP_Adapt	-0.014764	[-0.020473, -0.009241]	12/12	-0.001025
16	FABLE_Adapt	-0.001564	[-0.003271, -0.000248]	9/12	+0.001086
16	convex_logit	-0.035908	[-0.062325, -0.014594]	12/12	-0.000996
16	multinomial_stacking	+0.002010	[-0.002983, +0.007474]	5/12	+0.022053
16	nonnegative_ridge_htl	-0.035278	[-0.061663, -0.013648]	11/12	+0.000657
16	probability_stacking	-0.054222	[-0.089677, -0.022355]	12/12	-0.003090

P ARC MASS CHANGES NATIVE PREDICTIONS; QUOTIENTING MATTERS UNDER STRESS

This post-outcome mechanism study uses only development seed 20260820 and all 60 backbone–target–split cells at 4 labels/class. Table 21 adds family balancing, exact quotienting, Hellinger arc mass, global quotienting, evidence-branch marginalization, and the disjoint release gate in sequence. The native registry contains few exact cross-family coincidences, so exact and global quotient steps are numerically silent there; their necessity is isolated by the stress tests below. Evidence-branch marginalization is small but not silent: it changes Brier in 13/60 split cells, with mean absolute change 4.68×10^{-6} and maximum 1.11×10^{-4}. Arc mass changes the native aggregate slightly while retaining the mean improvements. Adaptation has one Brier-harmful architecture–target unit, whereas the gate releases 8/60 cells and removes Brier and NLL harm on all 12 architecture–target units.

Table 21: Development-seed component ablation. Differences are method minus the target's pooled fallback; negative Brier/NLL/ECE and positive accuracy are favorable. "Harm" is the number of Brier-harmful architecture–target units after averaging five splits; "Rel." is released split cells.

Variant	ΔBrier	ΔNLL	ΔAcc.	ΔECE	Harm	Rel.
Flat row prior	-0.0637	-0.1654	+0.0281	-0.0512	1/12	60/60
Equal family / row	-0.0615	-0.1607	+0.0270	-0.0487	1/12	60/60
+ exact quotient	-0.0615	-0.1607	+0.0270	-0.0487	1/12	60/60
+ arc, no global quotient	-0.0620	-0.1590	+0.0274	-0.0442	1/12	60/60
+ quotient, no branch marg.	-0.0620	-0.1590	+0.0274	-0.0442	1/12	60/60
Full RIME–Adapt	-0.0620	-0.1590	+0.0274	-0.0442	1/12	60/60
Full RIME–Deploy	-0.0323	-0.0766	+0.0153	-0.0301	0/12	8/60
RIME–Adapt, no gate	-0.0620	-0.1590	+0.0274	-0.0442	1/12	60/60

Table 22: Increment from branch-marginalized Gibbs evidence, comparing the otherwise identical fifth and sixth component variants over 60 development-seed split cells.

Metric	Nonzero cells	Mean absolute change	Maximum absolute change
BRIER	13/60	4.68e-06	1.11e-04
NLL	13/60	8.89e-06	1.63e-04
ECE	13/60	1.00e-05	3.35e-04

Q Alternative predictive geometries preserve exact clones

Hellinger geometry is a declared design choice, not a uniqueness claim. We therefore recomputed the development-seed base measure and posterior under centered log probability, probability-Euclidean, and root Jensen–Shannon embeddings while holding the registry, evidence, and Gibbs update fixed. Every embedding preserves exact-clone invariance to below 7×10^{-16}. Root Jensen–Shannon is closest to Hellinger; centered log probability moves the posterior mean by up to 0.134 TV, although its mean Brier change versus fallback remains close. The invariance theorem therefore holds for a fixed embedding, whereas finite-grid mass and predictions still depend on the choice of geometry.

Table 23: Post-outcome geometry sensitivity over 60 development-seed cells. Prior ℓ_1 and predictive TV are relative to Hellinger; ΔBrier is the method minus the pooled fallback.

Geometry	Prior L1	Mean TV	Max TV	ΔBrier
Centered log probability	0.182	9.42e-03	1.34e-01	-0.0623
Hellinger (paper)	0.000	0.00e+00	0.00e+00	-0.0620
Probability Euclidean	0.043	1.41e-03	2.40e-02	-0.0617
Root Jensen–Shannon	0.003	9.63e-05	2.38e-03	-0.0620

R Mean utility is stable across a small mass–temperature grid

The fallback mass and Gibbs temperature τ are fixed design parameters. Across a 3×3 post-outcome grid, mean Brier improvement versus fallback varies only from -0.0581 to -0.0628. Individual predictions are less stable: moving τ away from 1.0 can produce maximum cell-level TV above 0.36. Thus the average-utility conclusion is insensitive to these values on this grid, but the chosen values remain part of the method definition rather than nuisances to which invariance is claimed. Likewise, splitting or merging declared families is a modeling change; the explicit registry stress in Table 25 moves predictions by up to 0.111 TV.

Table 24: Post-outcome fallback-mass and Gibbs-temperature sensitivity over 60 development-seed cells. TV is relative to the default $(0.10, 1.0)$ setting, marked by an asterisk; ΔBrier is the method minus pooled fallback.

Fallback mass	τ	Mean TV	Max TV	ΔBrier
0.05	0.5	2.82e-02	3.56e-01	-0.0598
0.05	1.0	5.15e-03	9.02e-02	-0.0626
0.05	2.0	2.45e-02	3.87e-01	-0.0628
0.10	0.5	2.50e-02	3.66e-01	-0.0594
0.10	1.0*	0.00e+00	0.00e+00	-0.0620
0.10	2.0	2.45e-02	3.64e-01	-0.0621
0.20	0.5	2.43e-02	3.92e-01	-0.0581
0.20	1.0	8.99e-03	1.49e-01	-0.0604
0.20	2.0	2.69e-02	3.63e-01	-0.0607

S ARC MASS IS EXACT AT CLONES AND CONTINUOUS NEARBY

The complete stress grid contains all three backbones, four targets, and five splits. Exact multiplication of one base branch by 16 moves flat-row aggregation by mean TV 0.0464 (maximum 0.583), while exact quotienting and RIME are invariant up to floating-point precision. The effect is operational rather than merely prior-level: in an initial ResNet-only stress, cloning the best OOF candidate 64 times changed the flat-row aggregate by up to 0.1351 Brier and 0.7218 sample-level TV, including one release flip. Exact quotienting alone is not a continuity construction: under 16 distinct near-clones spanning the first 10% of a path chord its mean TV is 3.65×10^{-3}, compared with 4.49×10^{-5} for predictive arc mass. The three transformations covered by the invariance claim—row permutation preserving within-path order, path reversal, and monotone coordinate relabeling—have maximum TV 5.75×10^{-16}. Retracing, endpoint truncation, and splitting one family are modeling changes rather than claimed invariances and can move predictions, with maximum TV 0.111.

Table 25: Predictive total-variation drift under exact and near duplication and registry transformations, over 60 development-seed cells.

Method / change	Exact clone $\times 16$		Near clone $K = 16$, $\alpha = .1$	
	Mean TV	Max TV	Mean TV	Max TV
Flat rows	4.64e-02	5.83e-01	1.97e-02	3.39e-01
Exact quotient	2.19e-17	5.01e-16	3.65e-03	8.94e-02
Tolerance 10^{-3}	2.30e-17	6.13e-16	2.58e-03	8.89e-02
RIME arc	3.48e-17	5.18e-16	4.49e-05	4.94e-03
Guaranteed registry transforms	–	5.75e-16	–	–
Modeling-choice transforms	–	1.11e-01	–	–

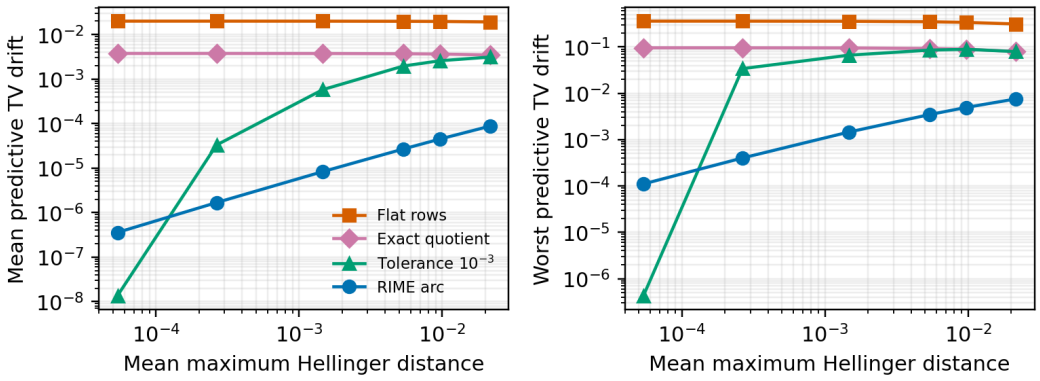


Figure 2: Near-clone continuity as the inserted atoms move away from their anchor. Panels show mean and worst predictive TV; exact quotienting removes the discontinuity at zero but Hellinger arc mass is substantially less sensitive nearby.

T ALL SIX SCALAR FAMILIES CONTRACT UNDER REFINEMENT

We refine each of the six scalar families from its natural coarse grid to half-step and quarter-step grids while leaving all other families at the paper registry. The union quarter grid is fit once per cell, so every comparison reuses identical candidate fits, OOF folds, and test rows. For every family, mean RIME TV to the quarter grid contracts from coarse to half resolution (Table 26). The same suite clones each complete within-family predictive equivalence class; if a fitted path collapses to one atom, the nearest distinct nonfallback atom defines the reported local direction. Every near candidate interpolates both deployed and complete OOF predictions and recomputes its OOF Brier evidence. The mean near-clone TV of RIME is 9.5×10^{-8} to 1.19×10^{-5} across families, versus 2.01×10^{-4} to 6.45×10^{-3} for equal-family/equal-row mass. Exact clone TV is at most 5.60×10^{-16} and monotone coordinate-name relabeling is exactly zero. Coordinate-uniform native, log, and nonlinear quadratures do differ, as shown by their per-cell Brier ranges; this is the nuisance that predictive arc length removes.

Table 26: Six-family refinement and near-clone study over 60 development-seed cells. H_q is the named family's quarter-grid row count. TV columns average target-example prediction drift; coordinate columns summarize the range across native/log/nonlinear coordinate quadratures.

Family	H_q	Grid TV to quarter		Near-clone mean TV		Coord. ΔBrier range	
		Coarse	Half	Equal-row	RIME	Mean	Max
Entropy	17	6.27e-04	9.51e-05	6.90e-04	9.50e-08	2.33e-04	2.40e-03
Multinomial	17	2.31e-03	3.10e-04	6.44e-03	4.75e-06	3.96e-03	4.64e-02
Nonnegative ridge	18	9.90e-04	1.20e-04	1.30e-03	2.99e-07	9.99e-04	3.64e-03
SCLA	14	1.13e-04	2.83e-05	2.01e-04	1.19e-05	1.84e-04	1.74e-03
Signed ridge	18	9.29e-04	1.10e-04	1.44e-03	4.70e-07	9.57e-04	3.29e-03
Target-logit head	17	9.04e-04	6.03e-05	6.45e-03	7.48e-06	1.67e-03	1.59e-02

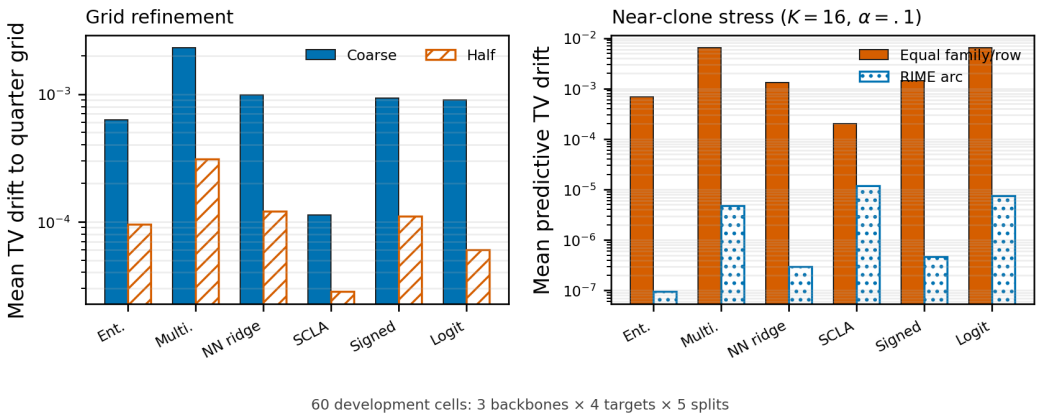


Figure 3: All-family grid contraction (left) and near-clone sensitivity (right). Near candidates consistently interpolate deployment and OOF predictions before recomputing evidence. Bars aggregate 3 backbones, 4 targets, and 5 fixed splits; axes are logarithmic.

U MEASURE CONSTRUCTION SCALES LINEARLY BUT IS NOT FASTEST AT CURRENT SIZE

The measure core was benchmarked with $C = 7$, three repeats, $H = 32$–2048, and $N = 64$–2048 on the stated 12-core CPU. At fixed $N = 128$, its fitted log–log slope in H is 1.04, while an explicitly materialized all-pairs similarity graph has slope 1.89. The dense vectorized reference is faster at these moderate sizes, but RIME does not incur its quadratic memory and runtime growth: measure construction stores $O(HNC)$ prediction/evidence input and no $H \times H$ graph. Candidate fitting and cross-fitting remain shared upstream costs and are not hidden in this timing.

Table 27: CPU wall time at $N = 128$, $C = 7$. Slopes use the high-H log–log tail; absolute timings should not be compared as if the dense reference implemented the same statistical operation.

Stage	$H = 32$	$H = 256$	$H = 2048$
RIME measure	0.0057s	0.0417s	0.3654s
Dense H^2 graph	0.0002s	0.0018s	0.0881s
Log–log slope in H	RIME 1.04; dense graph 1.89		

A separate real-cell decomposition uses the first split for every one of the 12 PACS architecture–target units and repeats each stage three times. Complete candidate fitting and cross-fitting dominate at median 6.71s; checked prediction loading takes 0.055s, the RIME measure plus posterior 0.033s, and confirmation 0.0011s. Thus the new measure core is about 0.5% of the shared fitting time at the native 41-row registry, rather than a hidden end-to-end bottleneck.

Table 28: Real PACS stage decomposition, 36 observations per row. "Peak" is the within-call process/tracemalloc increment and excludes already resident arrays.

Stage	Median	P90	Max	Median peak
Prediction loading + checks	0.0547s	0.1047s	0.1915s	23.3 MiB
Candidate fitting + cross-fitting	6.7077s	8.5668s	9.3475s	7.6 MiB
RIME measure + posterior	0.0327s	0.0455s	0.0490s	5.0 MiB
Confirmation e-value	0.0011s	0.0014s	0.0017s	0.3 MiB

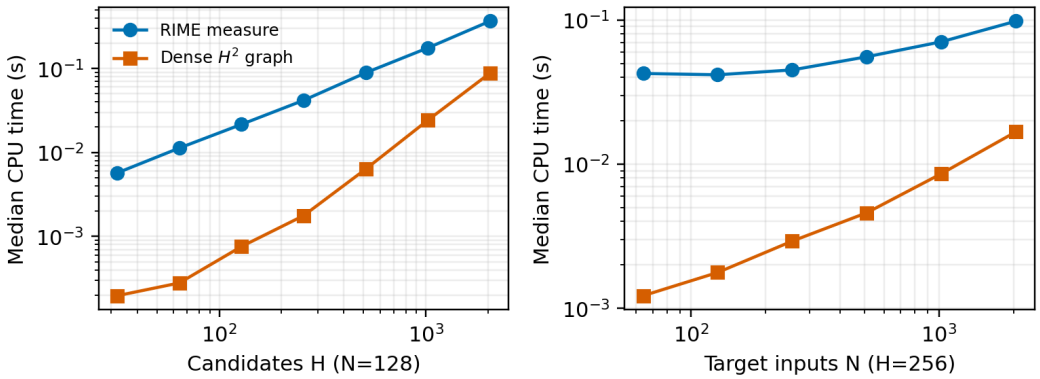


Figure 4: CPU runtime scaling in registry size H and target-slice size N. Points are medians over three repeats; dashed guides show the stated linear and quadratic growth contracts. Peak memory is reported in Table 27, not in this figure.

V NICO++ CONFIRMATION PROVENANCE AND NUMERICAL DISCLOSURE

Table 29: Pre-specified NICO++ confirmation over 300 target–split–seed– architecture cells. Adaptation differences are method minus fallback; R−F is RIME-Adapt minus FABLE-Adapt. The R−F interval is a t interval over six target contexts after averaging architectures, training seeds, and splits.

Quantity	RIME	FABLE	Paired interpretation
Mean adapted Brier	0.242493	0.242492	R−F $= +1.34\times10^{-6}$
ΔBrier vs fallback	-0.004072	-0.004073	R−F 95% CI $[-1.21\times10^{-4}, 1.24\times10^{-4}]$
Exact-clone maximum TV	5.62×10^{-16}	0.149	invariant only for RIME
Mean half-to-quarter TV	3.98×10^{-5}	6.18×10^{-4}	15.5× smaller
Confirmation releases	0/300	0/300	total abstention

NICO++ defines category labels jointly with common and category-specific contexts (Zhang et al., 2023). Our pre-specified corpus rule uses the aligned contexts available in the public challenge release, removes content-hash conflicts and duplicate digests globally, and retains only classes meeting the fixed per-context coverage rule. This yields six contexts, 58 classes, and 85,026 readable unique images. The design, corpus rule, and evaluation settings were fixed before any evaluation outcome was computed. Validation of the outputs accounts for 1,200 utility rows, 600 exact-clone stress rows, and 1,800 refinement rows with no duplicate analysis-grain rows.

The L-BFGS-B signed-path fit reached its function/gradient-evaluation limit in 24 of 34,500 fits (0.070%), all at the zero-penalty endpoint. These fits were re-solved with SLSQP from the identical objective, initialization, and bounds, the same numerical rule used on PACS (Appendix D), which was fixed before any complete evaluation outcome existed; every re-solve is recorded with its training rows. This fallback changes numerical execution, not candidate definitions, labels, folds, aggregation, metrics, or the release rule.

A separate post-outcome diagnostic tests how much the result depends on the chosen target-input slice, using the actual, unevenly saved source-checkpoint trajectories. For each of six targets, 20% of test images are held out from geometry and evaluated inductively; the remaining pool supplies five deterministic subsets at each of 64, 128, 256, 512, 1024, and 2048 unlabeled images. Relative to full-pool geometry, prior-predictive mean TV decreases from 1.70×10^{-3} at 64 images to 3.20×10^{-4} at 2048, worst-example TV from 1.04×10^{-2} to 1.43×10^{-3}, and mean absolute Brier change from 6.93×10^{-4} to 1.36×10^{-4}. All six targets contract between the endpoints. Fixed adaptation evidence concentrates the posterior in this bank, so posterior-predictive mean TV is at most 5.84×10^{-12}. The base-measure metrics were added after inspecting this concentrated-posterior behavior; the analysis uses one architecture, bank seed, and target split and is therefore explicitly diagnostic.

W REPRESENTATION STABILITY TRANSFERS; ADAPTATION UTILITY REVERSES

Office-Home contains four visual domains and 65 categories (Venkateswara et al., 2017). We extract frozen ImageNet ResNet-50 and ViT-B/16 features, train one calibrated linear head trajectory per domain at uneven $5/4/3/4$ checkpoint schedules, and train a leave-one-domain-out pooled fallback for each target. The resulting target-excluded registries contain 11–13 real saved source checkpoints. Before fitting any reported head, a global SHA-256 content check removes all 87 digest groups with conflicting labels and retains one canonical row for every remaining digest. The resulting 15,089-image corpus is globally content-unique, so source/target and adaptation/confirmation/test partitions cannot share an image. Source files are checksum-verified and row retention is deterministic. This replication is retrospective because an Office-Home subset had already been used for learning-rate selection.

ACSIncome is a binary income task derived from American Community Survey records; the source data span states and years and were introduced specifically to support geographic and temporal shift studies (Ding et al., 2021). Before evaluating outcomes we fixed CA/TX/NY/FL as domains and selected the lowest original row indices within each state and class, yielding a deterministic balanced sample of 96,000 rows. Linear and two-layer MLP trajectories use the same uneven checkpoint schedule and source-only per-model standardization. These data were not used earlier in this project, although the study still uses one training seed per model family.

The external studies separate representation stability from adaptation utility. Exact clone invariance transfers to every evaluated cell, and signed-ridge refinement drift remains small (Table 30). Adaptation improves on Office-Home but harms the strong fallback on ACSIncome; RIME and FABLE remain descriptively close within each setting. The fixed gate abstains in all 80 cells, so the deployed predictor equals the fallback everywhere. These results support the representation mechanism, not external deployment utility.

Table 30: External replications. Differences are method minus the setting's exact leave-one-domain-out fallback. R−F is RIME-Adapt minus FABLE-Adapt after averaging five splits within target; harmful counts are target-domain units.

Dataset	Backbone	Adapt ΔBrier			Deploy	Rel.	Harmful tasks		Exact TV
		RIME	FABLE	R−F	RIME		Adapt	Deploy	
Office-Home	ResNet-50	-0.0292	-0.0294	+0.0002	+0.0000	0/20	0/4	0/4	7.7e-16
	ViT-B/16	-0.0121	-0.0120	-0.0001	+0.0000	0/20	0/4	0/4	1.8e-16
ACSIncome	Linear	+0.0125	+0.0127	-0.0001	+0.0000	0/20	4/4	0/4	3.4e-16
	MLP	+0.0046	+0.0046	+0.0001	+0.0000	0/20	4/4	0/4	4.6e-16

X ACSINCOME FAILURE CASES IMPLICATE EVIDENCE AND CALIBRATION MISMATCH

We selected the three largest ACSIncome test Brier harms over all 40 RIME-Adapt split cells, with lexical tie-breaking fixed in the diagnostic script; all three use the linear model trajectory. The TX case is the clearest reversal: its selected aggregate improves confirmation Brier by 0.0358 yet harms test Brier by 0.0803 and ECE by 0.1262; candidate OOF and test rankings have Spearman correlation -0.222, while only 0.023 posterior mass remains on the fallback. NY has positive confirmation and test harm, whereas CA is near zero on confirmation but harmful on test. All three have $\log E \ll \log 160 \simeq 5.08$, so the release rule correctly abstains. These cases locate the utility failure in noisy or non-transferable few-shot evidence and calibration, rather than in exact-clone geometry, but they do not identify a single causal mechanism.

Table 31: Deterministically selected worst three ACSIncome test-Brier harms. Changes are against the pooled leave-one-domain-out fallback; confirmation rows are disjoint from adaptation and test rows.

Target/split	Top family / mass	Test ΔBrier	Confirm ΔBrier	ΔECE	$\log E$
TX/20261112	convex-logit / 0.445	$+0.0803$	-0.0358	$+0.1262$	0.00
NY/20261113	uniform-sources / 0.166	$+0.0203$	$+0.1111$	$+0.0363$	-0.44
CA/20261113	SCLA / 0.180	$+0.0190$	$+0.0032$	$+0.0245$	-0.03

Y BROADER IMPACT

Replacing row count by predictive mass removes one avenue for accidental or strategic registry manipulation, but family declarations and branch metadata remain trusted inputs and could themselves be manipulated. Miscalibrated probabilities or an unrepresentative target slice can distort both geometry and release evidence. The fallback gate limits some harms under its finite-population assumptions, yet conservative abstention has an opportunity cost and is not a guarantee under deployment shift.