A PRE-REGISTERED INVARIANCE AUDIT OF POST-HOC OUT-OF-DISTRIBUTION DETECTORS UNDER EXACT NETWORK SYMMETRIES

Anonymous authors
Paper under double-blind review

ABSTRACT

We present, to our knowledge, the first pre-registered invariance test for post hoc OOD detectors: is detector behaviour a property of the network function or of its coordinates? Two exact symmetries rescale or permute the penultimate representation without changing predictions. Before training, we registered predictions for 28 constants across 21 detectors; across six backbones, 13 of 16 testable rows matched, and the three misses are reported as falsifications. Whether rescaling reaches the penultimate feature depends on the classifier head, not the backbone. Rebuilt references gave $\Delta\mathrm{AUROC} \approx 10^{-8}$; frozen ones shifted AUROC by up to 0.64, and frozen unit-indexed masks by up to 0.63 under permutation. Algebra fixes whether a detector moves and how far its score shifts, with closed forms matched for two detectors, and approximated for a third, across backbones and a second seed; AUROC loss and its direction depend on checkpoint and OOD set. A post hoc ImageNet-1K replication, on two architectures and five OOD sets, reproduces the refit–frozen split. ViM moved in float32 but was ranking-invariant in float64. As practical guidance, we recommend logit-only scores, or references rebuilt on the deployed checkpoint, and avoiding constants fixed in absolute units, as in NCI and GMM, which move regardless. Predictions, manifests, results and a claim-checking script are released; registrations are verifiable from git ancestry.

1 INTRODUCTION

Post-hoc out-of-distribution alerts are often treated as properties of a deployed classifier, yet many depend on an arbitrary scaling of the final hidden layer — the degree of freedom Dinh et al. (2017) used to make minima arbitrarily sharp without changing the network function. Two gauge-equivalent parameterisations whose logits match to roughly 10^{-6} can therefore raise different alarms if a percentile threshold or feature mask is calibrated on one gauge and reused on another. The practical rule is simple: use logit-only detectors, or refit feature references on the exact checkpoint being deployed; a constant fixed in absolute units, as in NCI and GMM, moves either way. Gradient-based scores are covered by neither branch: they read the feature, so they are not logit-only, and they hold no reference to refit. GradNorm's score scales by $|c|$ for this reason.

OOD detectors are typically evaluated as properties of trained networks: this checkpoint, this detector, this AUROC (Yang et al., 2022; Zhang et al., 2024), which assumes the detector output is a function of the network. But networks are represented in redundant coordinates: multiplying one ReLU layer by $c > 0$ and dividing the next by c leaves all outputs unchanged by positive homogeneity, and for non-increasing-width architectures such rescalings and permutations exhaust the invariances (Phuong and Lampert, 2020). Any statistic that changes when only these coordinates change is not a property of the function the network computes — though statistics of the parameterisation itself, such as Hessian eigenvalues or weight norms, are meant to vary, as Dinh et al. (2017) exploit.

That yields a clean test. We apply an exact reparameterisation that leaves the classifier unchanged and every predicted label identical, while scaling the penultimate feature by a known c: one internal quantity changes in a controlled way, the decision rule does not, and we ask which detectors respond. Nothing is approximated and no baseline is improved. Predictions for 28 detector constants were

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.

committed and tagged in version control before training — logit-only scores invariant, feature-based scores invariant when references are recomputed on the evaluated model, frozen references changed, single-linear heads exposing no reachable gauge — and three were wrong and are reported as such.

Contributions. The rule tested here — a detector holding a constant in absolute units moves under the symmetry, one holding it relative to the current activation scale does not — was observed post hoc in earlier work. Existing OOD benchmarks compare performance across datasets, architectures or detectors; this test instead holds the classifier function and the evaluation data fixed and changes only arbitrary internal coordinates, which makes five questions measurable.

- **Does the rule survive prospective testing?** Largely: 13 of 16 testable rows match, and the three that do not are reported as falsifications rather than revised.
- **When does a symmetry reach the detector's representation?** Not always. A reachability gate rejects runs where the transformation leaves the inspected feature unmoved, separating a detector that passed from one never tested; two further gates reject non-symmetries and nondeterministic detectors.
- **Is apparent backbone robustness actually head dependence?** Yes. BatchNorm (Ioffe and Szegedy, 2015) residual networks with single-linear heads possess rescaling gauges, and the one we register does not reach the penultimate feature; a two-layer head on the same trunk restores the symmetry, and a trunk-level gauge that does makes frozen references fail there too.
- **Which frozen references fail, and under which symmetry?** All under rescaling, only the coordinate-indexed ones under permutation. ReAct and DICE with recomputed versus frozen references, differing only in when the percentile or mask is built, differ by eight orders of magnitude in $\Delta\mathrm{AUROC}$ across four backbones.
- **Do algebraic predictions survive finite precision?** Not exactly. The transformed network reproduces ch to 2×10^{-7} relative, bit-identically only at powers of two; ViM is non-invariant in float32 and its ranking exactly invariant in float64.

Predictions, checkpoint manifests, and every void run with its rejection reason are released together.

2 Related work

Parameter-space symmetry. S1 and S2 are group actions on parameters that leave the network function unchanged on every input (Zhao et al., 2026): S1 is positive scaling of a homogeneous activation, S2 permutation of a pointwise one. Godfrey et al. (2022) argue via the intertwiner group that hidden representations are meaningful only up to such symmetries, so representation-similarity measures should be invariant to them; our claim is narrower and operational, testing whether a given implementation is invariant and locating failures in code. The logic follows Dinh et al. (2017), who used this same rescaling to build functionally equivalent networks with arbitrarily sharper minima and concluded that flatness measures cannot explain generalisation: apply an exact symmetry, observe which statistic moves. Related audits cover transformer sharpness (da Silva et al., 2025), attribution (Crabbé and van der Schaar, 2023; Sundararajan et al., 2017), approximate Bayesian inference (Roy et al., 2024), input perturbations of OOD detectors (Lorenz et al., 2024), and representation-coordinate statistics (Karasan, 2026), where column-permutation parallel analysis selects different component counts under function-preserving changes of basis. The principle — that a proposed invariant be checked against the equivalence relation its own theory implies — predates machine learning; in ergodic theory Dooley and Mortiss (2009) extended the critical dimension, an invariant of metric isomorphism, to all product odometers. These redundant coordinates form a gauge (Saul, 2023; Hashimoto et al., 2024).

Relation to concurrent work by the same authors. A companion submission, under review elsewhere and unnamed for anonymity, applies the same two symmetries to nine post-hoc detectors on one four-layer MLP within a broader training-stability study. That audit produced the hypothesis tested here — detectors whose constants are fixed in absolute units move under symmetry, those expressed relative to the current activation scale do not — and identified implementation pairs differing only in the treatment of those constants, a design generalised here. This paper is distinct in

registering predictions before training, registering 21 detectors and measuring 15 across six trained backbones rather than nine on one network, introducing and testing the reachability condition and the single-linear-head result, and giving symmetry effects in closed form for two detectors and approximately for a third, rather than empirically.

Completeness. We focus on permutation and positive scaling because for architectures of non-increasing width Phuong and Lampert (2020) show these are the only function-preserving weight transformations. The result is not universal: Grigsby et al. (2023) show deeper networks can carry additional hidden symmetries, more prevalent with depth. Passing our tests therefore establishes invariance under the two symmetries applied, not under the network's full symmetry group.

3 SETUP

We audit post-hoc OOD detectors from OpenOOD (Yang et al., 2022; Zhang et al., 2024) (commit `3c35632`) under two exact reparameterisation symmetries. The evaluated detectors are MSP (Hendrycks and Gimpel, 2017), Mahalanobis (Lee et al., 2018), KNN (Sun et al., 2022), SHE (Zhang et al., 2023), ReAct (Sun et al., 2021), DICE (Sun and Li, 2022), ViM (Wang et al., 2022), Energy (Liu et al., 2020), MaxLogit (Hendrycks et al., 2022), GEN (Liu et al., 2023), NCI (Liu and Qin, 2025), GradNorm (Huang et al., 2021), Gram (Sastry and Oore, 2020) and the Gaussian mixture postprocessor. ASH (Djurisic et al., 2023), VRA (Xu et al., 2023) and SCALE (Xu et al., 2024) are excluded for the reasons discussed in Section 4.1. **S1** scales the final hidden linear layer by c and divides the output layer by c; **S2** permutes penultimate-feature coordinates, drawn uniformly from the symmetric group with `torch.randperm` under three seeded generators. Both preserve the network function exactly, so any score movement is attributable to the detector.

Metrics and gates. We report $\Delta\mathrm{AUROC} = |\mathrm{AUROC}' - \mathrm{AUROC}|$ and

$$\Delta\mathrm{score} \;=\; \max_{x} \frac{|s'(x) - s(x)|}{\max(|s(x)|,\, 10^{-12})},$$

taking the larger over the in- and out-of-distribution sets. ΔAUROC records rank damage; Δscore records worst relative drift, scale-free and comparable across detectors, but as a maximum it can be dominated by one near-zero score even when the distribution is unchanged. We read the two jointly.

Labels near the 10^{-6} verdict tolerance are not read literally: detectors crossed it on different backbones without behaving differently. MSP reported INVARIANT at $\Delta\mathrm{score} = 6.6 \times 10^{-7}$ on one backbone and SCORE-ONLY at 1.5×10^{-6} on another; ReAct gave $\Delta\mathrm{AUROC} = 2.9 \times 10^{-6}$ on one while its Δscore stayed at 2.07×10^{-7}, unchanged to three significant figures across all c. Conclusions rest on the numerical values, not the label, and the registered 10^{-6} is never relaxed.

Before any detector is scored, *validity* requires $\|\Delta z\|_\infty = \max_{x,i} |z_i(x) - z_i'(x)| < 10^{-5}$ on the logits, so the transformed network computes the same function, and *reachability* requires $\|\Delta h\|_\infty = \max_{x,j} |h_j(x) - h_j'(x)| > 10^{-5}$ on the penultimate feature, so the transformation perturbs what the detector reads; observed margins exceed it by four to six orders. A health gate, added after the audit, voids runs whose untransformed scores are non-finite or constant (Section 5). Runs failing any gate are discarded: passing an invisible symmetry reveals architecture, not detector invariance.

Experimental design. Predictions for 28 detector constants were pre-registered before training (tags `prereg-v1`, `prereg-v2`). Six backbones were trained: VGG-16-BN (Simonyan and Zisserman, 2015), ResNet-18 (He et al., 2016) with a two-layer head, and stock ResNet-18, each on CIFAR-10 and CIFAR-100 (Krizhevsky, 2009). CIFAR-10 models are evaluated on SVHN (Netzer et al., 2011), Textures (Cimpoi et al., 2014), CIFAR-100 and Imagenette (Howard, 2019); CIFAR-100 models on SVHN, Textures, CIFAR-10 and Imagenette. For S1, $c \in \{0.01, 0.5, 3.0, 5.0\}$, chosen to span a factor of 500 around unity while including one exact binary value (Section D.4). For S2, three random permutation seeds.

From 28 predictions to the measured set. The 28 pre-registered rows are constants of 21 detectors, several of which register more than one; each row predicts both symmetries. Fifteen were measured, with two unregistered refit control arms, giving 24 (detector, symmetry) units: 16 scored

rows, 3 refit controls, ReAct under S2 (excluded below), and 4 whose constants are sheltered from both symmetries. Nine rows were not measured: ASH, VRA and SCALE are excluded below; ODIN needs an input-space gradient the wrapper's `no_grad` forward cannot supply, though lifting it confirms the registered prediction; NNGuide is present but unreachable through the loader wrapper; and Residual's registered `dim` equals the feature width, so its residual subspace is empty and the score constant.

Why stock ResNet-18 has no reachable S1. The issue is not an absence of gauge freedom. Positive homogeneity always permits rescaling across ReLU-separated layers (Dinh et al., 2017; Phuong and Lampert, 2020). The question is whether the symmetry reaches the penultimate feature, which is what feature-based detectors consume. A single-linear classification head does not permit this: there is no hidden layer between the penultimate feature and the logits to absorb the inverse scale, and the transformation cannot be propagated backward. At evaluation BatchNorm is affine, $y \mapsto (\gamma/\sigma)y + (\beta - \gamma\mu/\sigma)$, so scaling its input scales only the linear part, and a residual connection adds a scaled branch to an unscaled identity path. We report this as a result rather than a limitation: **such a network has rescaling gauges, and the one registered here does not reach its penultimate feature.** A gauge acting on BatchNorm's parameters does reach it, and under it frozen references fail as they do elsewhere (Appendix E); we claim no exhaustiveness over all rescaling gauges. The two-layer-head ablation makes S1 reachable — the feature drift clears the reachability gate at every c while the transform remains an exact symmetry, $\|\Delta z\|_\infty \leq 5.7 \times 10^{-6}$ against a validity gate of 10^{-5} — showing that S1's reachability depends on the classification head, not on residual architectures.

4 RESULTS

4.1 THE PRE-REGISTERED PREDICTIONS, SCORED

Predictions were registered and scored at the *(detector, symmetry)* level, where they were written; scoring per c would quadruple the count without adding independent evidence. Every outcome is assessed against its original registered prediction, none reclassified afterwards.

Table 1: Registered predictions against measured outcomes, 16 scored pairs. Left: SCORE-ONLY counted as movement, since the statistic did move. Right: folded into INVARIANT, which is what a benchmark reporting AUROC alone would conclude.

	strict		AUROC only	
	meas. INV	meas. MOVES	meas. INV	meas. MOVES
predicted INVARIANT	5	3	7	1
predicted MOVES	0	8	2	6
correct	13/16		13/16	

The panels agree on the count and disagree on which rows are wrong, which is the more useful comparison. Under the strict reading no detector predicted to move failed to move, and all three errors run one way — predicted invariant, measured otherwise — and are diagnosed in Section 4.4, where all three prove invariant in exact arithmetic and move only in float32. Under the AUROC-only reading one error of that kind remains and two are new: SHE and GradNorm were correctly registered as score-only, their scores moving by $|c^2 - 1|$ and $|c - 1|$ while their rankings hold, so a benchmark seeing only AUROC records both as invariant and both predictions as wrong. AUROC-only evaluation costs not a lower score but a different set of errors, falling on exactly the detectors whose statistic the symmetry moves.

One prediction was right in outcome and incomplete in scope. GMM was registered as moving only at $c \leq 0.01$, yet moved at every c. The mechanism is confirmed for small c, where the covariance contracts until an absolute floor dominates; residual movement at larger c arises from a different, weaker effect. It is scored correct at the registered outcome level, though not in its full explanation.

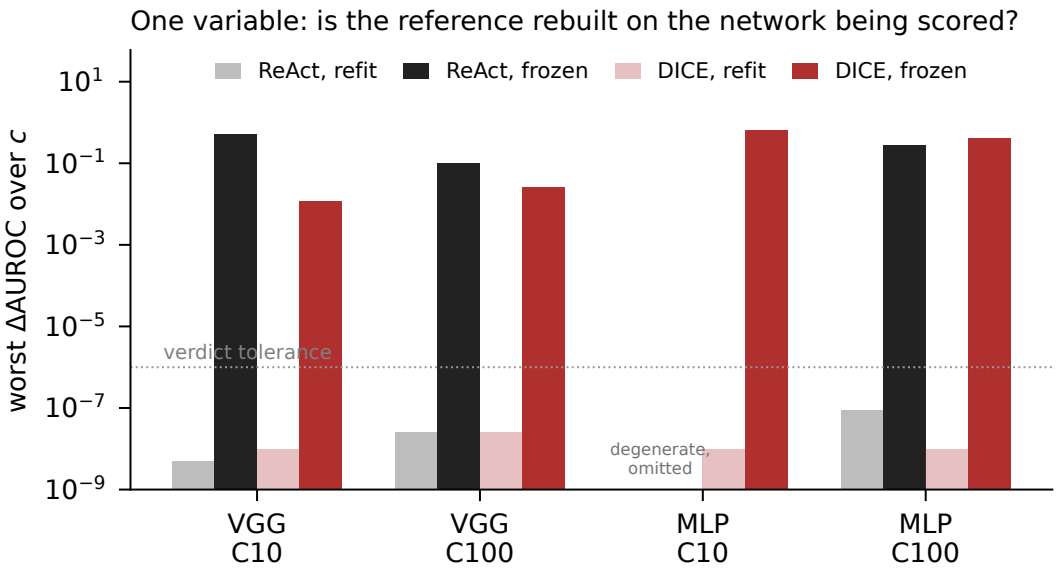

Figure 1: The ablation across all four backbones and both detectors. Bars are the worst ΔAUROC over $c \in \{0.01, 0.5, 3, 5\}$, on a log scale: the two arms differ by up to eight orders of magnitude. The refit arms sit below the 10^{-6} verdict tolerance in every case; the frozen arms never do. The only difference between a light bar and its dark neighbour is whether the detector's internal reference was rebuilt on the network being scored. A paired bootstrap separates the two arms by about $5{,}200\times$ (Appendix C).

Exclusions. Four rows are excluded rather than scored, since scoring them would convert untested predictions into artificial passes or failures. *ASH*: cached and refit arms are bit-identical, because the percentile is computed per sample and no reference is stored, so the ablation cannot operate. *ReAct under S2*: a percentile over flattened activations is permutation-invariant, making cached and refit thresholds identical. *SCALE* (Xu et al., 2024): present in the pinned commit but not in its postprocessor factory, and inert once reached — it passes its percentile number where a threshold value is expected, so at 85 against a feature maximum of 11.9 the clip never fires and its arms are bit-identical. *VRA*: reduces to ReAct on post-ReLU features, where the lower clipping threshold is zero, so it provides no independent evidence. LINe (Ahn et al., 2023) was not tested; the rule predicts its absolute clip moves under S1 even when rebuilt.

Controls. Refit arms are reported as controls rather than scored predictions. DICE and Gram were pre-registered only in their frozen forms; their refit results appear throughout as the complementary half of each comparison but are not counted in Table 1.

4.2 Refit versus frozen: one variable, eight orders of magnitude

The central experiment changes exactly one implementation detail: whether the detector's internal reference is rebuilt on the evaluated network or reused from an equivalent parameterisation (Figure 1). For DICE (Sun and Li, 2022), the sparsity mask is derived from mean activations; rebuilt on the current network it remains invariant, reused from a gauge-equivalent network it does not.

Table 2: DICE under S1, refit against frozen reference. Worst ΔAUROC over $c \in \{0.01, 0.5, 3, 5\}$. The detector, the data, the symmetry and the OOD set are identical between columns; only the reference differs.

backbone	refit	frozen	ratio
VGG-16-BN, CIFAR-10	1.0×10^{-8}	1.18×10^{-2}	10^{6}
ResNet-18-MLP, CIFAR-10	1.0×10^{-8}	$\mathbf{6.40 \times 10^{-1}}$	10^{8}
VGG-16-BN, CIFAR-100	2.5×10^{-8}	2.59×10^{-2}	10^{6}

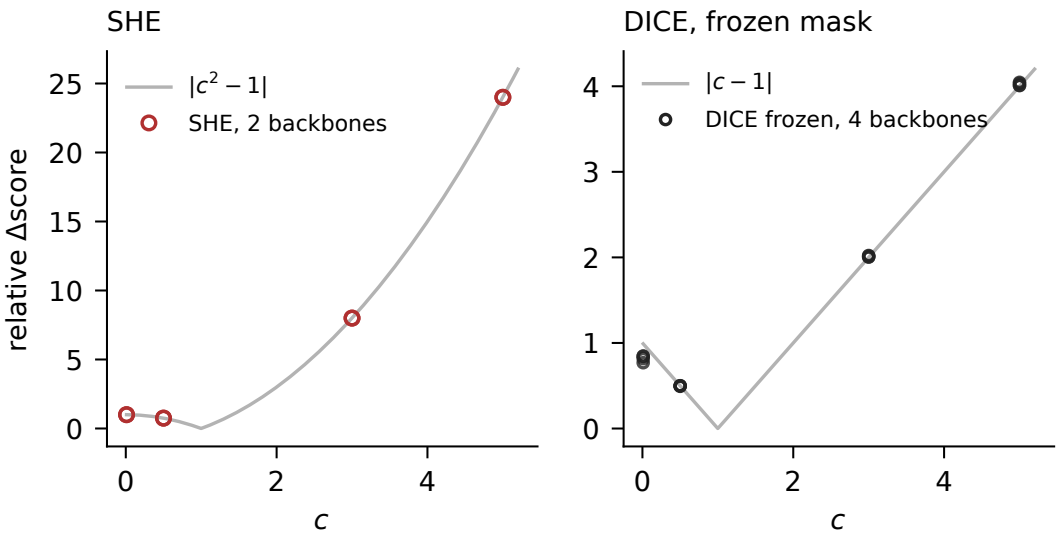

Figure 2: Measured score movement against closed form. Left: SHE's score scales by c^2, so the relative movement is $|c^2 - 1|$; measured values agree to five decimal places. Right: frozen DICE follows $|c - 1|$ on four backbones for $c \geq 0.5$, and sits 14–22% below it at $c = 0.01$, where the closed form is only approximate (Appendix B). Points are plotted unfilled and overlap.

Across both datasets and architectures the refit arm stays invariant to roughly 10^{-8} to 10^{-9}. The split holds post hoc on ImageNet-1K (Deng et al., 2009): across 50 detector–architecture–dataset combinations, MSP and the rebuilt references stay below 5×10^{-8} while the frozen ones move by up to 0.405 (Appendix F). ReAct shows the same pattern on VGG-16-BN/CIFAR-10 with a frozen percentile in place of a frozen mask: 5×10^{-9} refit against 5.29×10^{-1} frozen. Both arms use a validation split held out from training and disjoint from the evaluation set; the package also ships an earlier ReAct run whose validation split pointed at the test set, marked superseded and retained rather than deleted.

4.3 Two closed forms and one approximation

For three detectors the score movement reproduces a closed form, measured across backbones (Figure 2).

SHE scores by $\sum_i h_i \bar{h}_i^{(k)}$ against class-mean activations. Both factors scale with c, so the score scales by c^2 and the relative movement is $|c^2 - 1|$. Measured at $c = 0.01, 0.5, 3, 5$:

predicted $\lvert c^2 - 1 \rvert$	0.9999	0.75	8	24
VGG-16-BN/CIFAR-10	0.999900	0.750000	8.000002	24.000005
ResNet-18-MLP/CIFAR-10	0.999900	0.750000	8.000005	24.000014

Agreement to six significant figures on two architectures. ΔAUROC never exceeds 2×10^{-8}, so the ranking is untouched.

DICE with a frozen mask nearly scales the score by c, giving $|c - 1|$. Across four backbones it measures 2.004–2.022 at $c = 3$ and 4.008–4.046 at $c = 5$, against 2 and 4. The agreement is approximate and degrades at small c (Appendix B).

GradNorm reaches the same form by an unrelated route, and more exactly. Its score is $\|\nabla_W \mathcal{L}\|_1$, the L_1 norm of the gradient of the KL divergence to a uniform posterior taken with respect to the final classifier weights, so the gradient is $\delta \otimes h$; S1 leaves the logits unchanged, so δ is unchanged, and the gradient becomes $\delta \otimes (ch)$. An L_1 norm of that is exactly c times the original, so the relative movement is again $|c - 1|$. Measured on VGG-16-BN/CIFAR-10 at $c = 0.01, 0.5, 3, 5$: 0.9900000, 0.5000000, 2.0000010, 4.0000017, against 0.99, 0.5, 2, 4 — agreement to seven significant figures, closer than DICE because nothing is masked. ΔAUROC never exceeds 2.5×10^{-8}, so the ranking is preserved to eight decimal places — and yet a calibrated threshold is destroyed. At $c = 0.01$

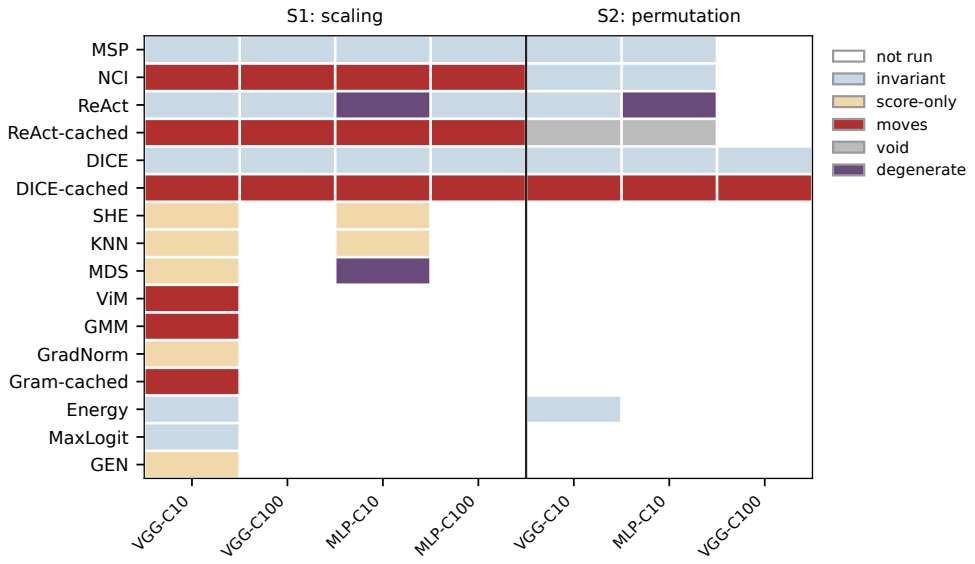

Figure 3: Every (detector, symmetry, backbone) cell we ran except the seeded Gram refit control (Appendix C), and what it returned. White is not run. The three purple cells are degenerate scores, omitted rather than counted: a constant score is invariant under anything, so an INVARIANT verdict there would be unearned. Grey is void — the ablation could not act. Energy, MaxLogit and GEN read only logits or normalised probabilities, so they are sheltered rather than scored. Labels read rather than thresholded, GEN's included, are explained in Appendix H.

and $c = 0.5$ the true positive rate falls to 0.0 and 0.033, so the detector fires on almost nothing; at $c = 3$ and $c = 5$ both the true and false positive rates reach 1.0 and it fires on everything. The symmetry changed no prediction and no ranking, yet took a deployed detector from flagging nothing to everything.

Algebra fixes displacement, not ranking damage. DICE's Δscore is nearly identical across all four backbones while its ΔAUROC varies by two orders of magnitude; changing only the OOD set moves the AUROC damage by a factor of 3.9 to 15.6 for DICE and 0.40 to 0.57 for ReAct-cached, in opposite directions. What was frozen also determines how it fails: DICE's mask degrades worst at $c = 0.01$ on all four audited checkpoints, ReAct's percentile worst at $c = 5$ on both VGG models but at $c = 0.01$ on the healthy ResNet-18-MLP. The direction is not a stable property: retraining VGG-16-BN/CIFAR-10 under a second seed, registered as `prereg-v3` before the checkpoint existed, reverses DICE's direction — worst at $c = 5$ rather than $c = 0.01$, differing by a factor of 14 there — while its Δscore still tracks $|c - 1|$ to 2% and ReAct-cached replicates at every c. The score law belongs to the detector; the AUROC damage belongs to the checkpoint. A frozen percentile over flattened activations survives permutation because the multiset is preserved; a frozen unit-indexed mask does not, and the pinned implementation caches $W \odot M$ rather than M, which is why the refit arm cancels exactly while the frozen arm moves as $|c - 1|$. Halving is exact in binary floating point, so every detector whose ranking we predicted would hold returned exactly zero ΔAUROC at $c = 0.5$ while every mechanical failure moved as at any other c: clean at $c = 0.5$ and dirty elsewhere is the signature of a numerical cause (Appendix D).

4.4 THREE FAILURES, NONE OF THEM A FAILURE OF THE RULE

Three detectors moved where invariance was predicted: KNN, MDS and ViM. All three satisfy the free-constant rule and are invariant in exact arithmetic; all three move only in float32, through L2 normalisation, matrix inversion and a residual eigendecomposition respectively. All three are falsified as registered — each said invariant, each moved — so the diagnoses change what the failures mean, not the count. Each was right about the algebra and wrong about the arithmetic, a distinction the predictions did not draw because it did not occur to us to. GMM, whose absolute floor the rule would have caught, is not among them: it was registered as moving and moved.

Table 3: The three misses, diagnosed, with GMM's scope correction alongside. Only GMM's cause lies inside the rule: its registered scope was too narrow.

detector	cause	in the rule?	response
MDS	float32 error in matrix inversion	no	Δscore 2.2×10^{-3}; falls two orders under float64
GMM ($c \leq 0.01$)	`reg_covar`$= 10^{-6}$ is absolute	**yes**	pass `reg_covar` $= 10^{-6}c^2$
ViM	float32 in the residual eigendecomposition	no	float64: ΔAUROC exactly 0
KNN	float32 in the L2 normalisation	no	float64: $2.0 \times 10^{-3} \rightarrow 8.4 \times 10^{-7}$

ViM has no absolute constant, and still moves. ViM (Wang et al., 2022) subtracts a virtual logit, built from the feature residual, from an energy score (Liu et al., 2020). Every constant in its fit is relative: the offset $u = -W^{+}b$ uses a singular-value cutoff relative to the largest singular value, and the residual basis comes from an eigendecomposition whose spectrum scales uniformly with c^2. With no absolute threshold anywhere ViM cannot fail as GMM does, yet it moves. On VGG-16-BN/CIFAR-10 against SVHN, ΔAUROC is 8.4×10^{-4}, 1.1×10^{-3} and 1.2×10^{-3} at $c = 0.01$, 3 and 5, and exactly zero at $c = 0.5$. Δscore reaches 5.4×10^{2} to 7.8×10^{2}, the largest of any detector predicted invariant and the clearest case of the caveat in Section 3: the score has mean -10.5 and standard deviation 30.7, so the distribution straddles zero and the maximum is carried by near-zero samples. Re-running this cell reproduces the verdicts, ΔAUROC and the fitted constants exactly, but not the Δscore magnitudes, which moved by up to a factor of two in both directions; the statistic is a maximum over a ratio whose smallest denominator here is 8.7×10^{-4}, so which sample attains it is not stable. We report the reproduced values. Threshold transfer is undamaged, the false positive rate staying within one percentage point across all four c, and the ranking is essentially intact.

ViM is invariant in exact arithmetic; the movement is float32. We refit ViM twice in float64, once on the network as trained and once with the transform applied analytically, so no float32 step intervenes. The fit is exactly scale-covariant: α tracks $1/c$ and $\|u\|$ tracks c to ten significant figures, the residual subspace rotates by 2.1×10^{-6} degrees, the energy term drifts by 1.9×10^{-15}, and ΔAUROC is identically zero. So the movement the audit records is absent in exact arithmetic, which is consistent with the $c = 0.5$ diagnostic of Section D.4: ViM is clean at that value and dirty elsewhere, the signature that rule assigns to a numerical rather than a mechanical cause.

We do not claim to have isolated the cause: the float64 refit uses a symmetric eigensolver and applies the transform analytically, where the audit uses the pinned implementation's general solver and a forward pass. The claim is the measurement: a detector can satisfy the symmetry exactly in the algebra and still move at the precision it ships in, and refitting does not address that.

A detector can have every constant in its fit correctly relative and still fail, if its score is a difference of two large nearly-equal terms. The issue is how the score is assembled, not what the constants are.

GMM confirms the registered mechanism at small c. Its covariance scales as c^2, so at $c = 0.01$ it shrinks by 10^{-4} and its entries become comparable to `reg_covar`, which then dominates the fit: the precision matrix and the mixture weights both come out wrong, and ΔAUROC reaches 2.7×10^{-3}, an order of magnitude above its value at any larger c. At $c = 3$ and $c = 5$ the covariance *grows*, the floor is relatively $200\times$ smaller, and the precision is correct to 0.3%.

5 Limitations

This is an audit, not a pre-registered theory test. Three patterns were noticed only after scoring: failure direction, a scalar-versus-vector score split, and the dependence of AUROC damage on the OOD set. None was pre-registered. Two proposed explanations for failure direction, a $\max/p_{90}$ rule and then a dead-fraction rule, were both falsified by subsequent runs and are retained only as failed hypotheses. No mechanism claimed here should be treated as established.

Coverage is uneven (Figure 3). Under S1, MSP, NCI and DICE were tested on four backbones; ReAct on three; SHE and KNN on two; MDS, ViM, GMM, GradNorm, Gram, Energy, MaxLogit and GEN on one, counting non-degenerate cells. Under S2, both DICE arms ran on three backbones; MSP and NCI on two; ReAct on one, its frozen arm on two (both void), and Energy on one; the other nine detectors were not run. Gradient-based detectors are represented by GradNorm alone. Three MLP-C10 cells passed for reasons unrelated to invariance and are excluded: ReAct's score is constant at $\log 10$ under both symmetries (Figure 5), and a constant is invariant under anything; MDS produced non-finite scores, which fail every $\geq$ comparison and fall through to INVARIANT. A health gate, added after these runs, now voids constant or non-finite runs (Appendix A). S2 covers 39% of S1's design cells, one on CIFAR-100 (Appendix C). Results should not be extrapolated to unmeasured backbone–method cells.

Comparability is limited in two ways. Hyperparameters use OpenOOD defaults with automatic tuning disabled, matching the pre-registration but not comparable to published OpenOOD AUROCs. Imagenette replaces TinyImageNet, which has no stable versioned source, and Places365 is omitted. The registered experiments are CIFAR-scale, with near-OOD evidence from a single dataset. A post hoc ImageNet-1K replication (Appendix F) adds NINCO (Bitterwolf et al., 2023) and SSB-hard (Vaze et al., 2022) and reproduces the refit–frozen split, but it was not registered and covers seven detectors.

Finally, passing S1 and S2 demonstrates invariance to those symmetries only (Grigsby et al., 2023).

6 DISCUSSION

What the algebra fixes, and what it does not. The verdict under rescaling — rebuilt references invariant, frozen ones moving — held on every backbone, dataset and seed we tested, including a post hoc ImageNet-1K replication (Appendix F). The score law is nearly as stable: frozen DICE's Δscore tracks $|c - 1|$ on all four backbones, with equal values at $c = 3$ and 5 on two seeds. The ranking damage is not: frozen AUROC loss spans two orders of magnitude, and its direction reversed between those seeds. The score law depends only on how a detector's formula transforms as the feature scales; the damage depends on where particular samples sit in one network's feature space. The rule is about units, not detector families: a constant relative to the current activation scale travels with the representation, one fixed in absolute units does not.

The question applies to our own statistics. Re-running ViM reproduces its verdicts exactly but not its Δscore, a maximum that one near-zero sample decides. A statistic can fail to be a property of the thing it names without any symmetry applied at all.

What this asks of OOD benchmarks. The conclusions hold at every tolerance from 10^{-6} to 10^{-4}, and a paired bootstrap separates refit from frozen arms by about $5{,}200\times$ (Appendix C). What AUROC cannot show is threshold transfer, or whether a reference travels with the representation. A benchmark reporting both, under one exact symmetry per detector, would catch every failure in Table 1 at the cost of one transformed network.

7 CONCLUSION

Two exact function-preserving symmetries, predictions registered before training, and six trained backbones test whether a post hoc OOD detector is a property of the function a network computes or of the coordinates representing it. Thirteen of sixteen testable rows matched their registered predictions; the other three are retained as falsifications, diagnosed as finite-precision effects rather than failures of the rule. Verdicts and score laws follow from each detector's algebra; the size and direction of AUROC damage follow from the checkpoint. In practice: use logit-only scores, or rebuild feature references on the exact checkpoint deployed, and avoid constants fixed in absolute units, which move regardless. Three directions follow: a registered ImageNet-scale test extending Appendix F, symmetry groups beyond rescaling and permutation, and detectors with references covariant by construction.

8 REPRODUCIBILITY STATEMENT

The supplementary package includes the pre-registration, the prediction file with its git tags, every result JSON, and the scripts that reproduce each table and figure. Three properties are enforced by the code rather than asserted here.

Predictions precede measurements. `predictions.csv` was committed and tagged before any network was trained, and the audit driver reads it without writing to it. The ordering is verifiable from an anonymous git history: every manifest records the commit its checkpoint was trained under, and ancestry is established by `git merge-base --is-ancestor prereg-v1 95dba870`. Unlike a timestamp, ancestry cannot be set by hand. Two further tags post-date the original audit and precede only their own measurements: `prereg-v3` registers the seed replication of Section D.1, and `prereg-v4` the trunk-gauge audit of Appendix E. No prediction in `predictions.csv` was altered when either was written.

Provenance travels with the artefact. Each manifest stores that commit, the checkpoint's `SHA-256`, library versions, device, seed and determinism flags; the checkpoint itself carries the RNG state of Python, NumPy, Torch and CUDA, so a resumed run is the same experiment as an uninterrupted one. The seven manifests are released in `checkpoints/`, and the post hoc ImageNet replication ships its per-image scores with a script that recomputes Table 5. Deterministic cuBLAS is set for training the CIFAR checkpoints and for the later ImageNet runs, but not for the CIFAR audit's scoring or the first ImageNet runs. Results reproduce from the shipped files, but a fresh scoring run is not guaranteed to reproduce them, as ViM's Δscore shows (Section 4.4).

Invalid results are refused. Runs failing the gates of Section 3 return VOID with a stated reason rather than a number. The shipped results contain four such rows, all Gram failing the determinism gate, alongside four arms excluded before measurement and three degenerate cells excluded in the analysis.

Trained weights are not distributed, since six checkpoints exceed the package size limit. `train_backbones.py` regenerates them from the seeds recorded in each manifest, at roughly 45–90 minutes per backbone on one accelerator. Expected accuracies are listed in the package so a regenerated checkpoint can be checked against ours.

AI USE STATEMENT

Generative AI tools were used to support literature discovery, code development, manuscript preparation, and review of mathematical arguments such as the SCALE invariance argument, trunk-gauge construction, ViM cancellation bound, and GradNorm closed-form derivation. AI assistance was used to draft initial pre-registration predictions based on the study's key assumptions. These predictions were subsequently subjected to independent review, validation, and revision by the authors prior to registration. AI did not generate datasets or experimental results; all reported results were produced using author-executed code, and all AI-assisted outputs were reviewed and validated by the authors.

REFERENCES

Y. H. Ahn, G.-M. Park, and S. T. Kim. LINe: out-of-distribution detection by leveraging important neurons. *CVPR*, 2023.

J. Bitterwolf, M. Müller, and M. Hein. In or out? Fixing ImageNet out-of-distribution detection evaluation. *ICML*, 2023.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. *CVPR*, 2014.

J. Crabbé and M. van der Schaar. Evaluating the robustness of interpretability methods through explanation invariance and equivariance. *NeurIPS*, 2023.

M. F. da Silva, F. Dangel, and S. Oore. Hide & seek: transformer symmetries obscure sharpness & Riemannian geometry finds it. *ICML*, 2025.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: a large-scale hierarchical image database. *CVPR*, 2009.

L. Dinh, R. Pascanu, S. Bengio, and Y. Bengio. Sharp minima can generalize for deep nets. *ICML*, 2017.

A. Djurisic, N. Bozanic, A. Ashok, and R. Liu. Extremely simple activation shaping for out-of-distribution detection. *ICLR*, 2023.

A. H. Dooley and G. Mortiss. On the critical dimensions of product odometers. *Ergodic Theory and Dynamical Systems*, 29(2):475–485, 2009.

C. Godfrey, D. Brown, T. Emerson, and H. Kvinge. On the symmetries of deep learning models and their internal representations. *NeurIPS*, 2022.

J. E. Grigsby, K. Lindsey, and D. Rolnick. Hidden symmetries of ReLU networks. *ICML*, 2023.

K. Hashimoto, Y. Hirono, and A. Sannai. Unification of symmetries inside neural networks: transformer, feedforward and neural ODE. *Machine Learning: Science and Technology*, 5(2), 2024.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CVPR*, 2016.

D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *ICLR*, 2017.

D. Hendrycks, S. Basart, M. Mazeika, A. Zou, J. Kwon, M. Mostajabi, J. Steinhardt, and D. Song. Scaling out-of-distribution detection for real-world settings. *ICML*, 2022.

J. Howard. Imagenette. GitHub repository `fastai/imagenette`, 2019.

R. Huang and Y. Li. MOS: towards scaling out-of-distribution detection for large semantic space. *CVPR*, 2021.

R. Huang, A. Geng, and Y. Li. On the importance of gradients for detecting distributional shifts in the wild. *NeurIPS*, 2021.

S. Ioffe and C. Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. *ICML*, 2015.

A. Karasan. Representation measurements under function-preserving reparameterizations. *arXiv preprint arXiv:2608.27020*, 2026.

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

K. Lee, K. Lee, H. Lee, and J. Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *NeurIPS*, 2018.

L. Liu and Y. Qin. Detecting out-of-distribution through the lens of neural collapse. *CVPR*, 2025.

W. Liu, X. Wang, J. D. Owens, and Y. Li. Energy-based out-of-distribution detection. *NeurIPS*, 2020.

X. Liu, Y. Lochman, and C. Zach. GEN: pushing the limits of softmax-based out-of-distribution detection. *CVPR*, 2023.

P. Lorenz, M. R. Fernandez, J. Müller, and U. Köthe. Deciphering the definition of adversarial robustness for post-hoc OOD detectors. *ICML Workshop on the Next Generation of AI Safety*, 2024.

Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

M. Phuong and C. H. Lampert. Functional vs. parametric equivalence of ReLU networks. *ICLR*, 2020.

H. Roy, M. Miani, C. H. Ek, P. Hennig, M. Pförtner, L. Tatzel, and S. Hauberg. Reparameterization invariance in approximate Bayesian inference. *NeurIPS*, 2024.

C. S. Sastry and S. Oore. Detecting out-of-distribution examples with Gram matrices. *ICML*, 2020.

L. K. Saul. Weight-balancing fixes and flows for deep learning. *Transactions on Machine Learning Research*, 2023.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.

Y. Sun and Y. Li. DICE: leveraging sparsification for out-of-distribution detection. *ECCV*, 2022.

Y. Sun, C. Guo, and Y. Li. ReAct: out-of-distribution detection with rectified activations. *NeurIPS*, 2021.

Y. Sun, Y. Ming, X. Zhu, and Y. Li. Out-of-distribution detection with deep nearest neighbors. *ICML*, 2022.

M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. *ICML*, 2017.

G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie. The iNaturalist species classification and detection dataset. *CVPR*, 2018.

S. Vaze, K. Han, A. Vedaldi, and A. Zisserman. Open-set recognition: a good closed-set classifier is all you need. *ICLR*, 2022.

H. Wang, Z. Li, L. Feng, and W. Zhang. ViM: out-of-distribution with virtual-logit matching. *CVPR*, 2022.

M. Xu, Z. Lian, B. Liu, and J. Tao. VRA: variational rectified activation for out-of-distribution detection. *NeurIPS*, 2023.

K. Xu, R. Chen, G. Franchi, and A. Yao. Scaling for training time and post-hoc out-of-distribution detection enhancement. *ICLR*, 2024.

J. Yang, P. Wang, D. Zou, et al. OpenOOD: benchmarking generalized out-of-distribution detection. *NeurIPS Datasets and Benchmarks*, 2022.

J. Zhang, Q. Fu, X. Chen, L. Du, Z. Li, G. Wang, X. Liu, S. Han, and D. Zhang. Out-of-distribution detection based on in-distribution data patterns memorization with modern Hopfield energy. *ICLR*, 2023.

J. Zhang, J. Yang, P. Wang, et al. OpenOOD v1.5: enhanced benchmark for out-of-distribution detection. *Journal of Data-centric Machine Learning Research*, 2024.

B. Zhao, R. Walters, and R. Yu. Symmetry in neural network parameter spaces. *Transactions on Machine Learning Research*, 2026.

A HARNESS CHANGES MADE AFTER THE AUDIT

Two changes were made to the harness after measurement began, both recorded here and in `REPRODUCIBILITY.md`.

The *health gate* voids a run whose untransformed scores are non-finite or have standard deviation below `TOL_SCORE`, before any delta is computed. A `NaN` fails every $\geq$ comparison in the verdict logic and falls through to INVARIANT; a constant score is invariant under anything. It postdates the audit runs: the three cells it would void, ReAct under both symmetries and MDS under S1 on ResNet-18-MLP/CIFAR-10, are excluded in the analysis instead.

The *reference-free exemption* concerns the freshness guard, which asserts that a postprocessor's `setup` leaves `setup_flag` set. GradNorm has no reference to build — its `setup` is `pass` in the pinned implementation — so the assertion fires on a detector it was never meant to catch. The harness exempts `GradNorm` by name rather than inferring which detectors are reference-free, and the guard remains in force for every other detector, including both cached arms.

B WHY DICE'S CLOSED FORM IS APPROXIMATE

SHE and GradNorm read quantities that are positively homogeneous in the penultimate feature, so their score movement under S1 is exact. DICE is not: it scores by a log-sum-exp over masked logits, and $\mathrm{logsumexp}(ca + b) \neq c\,\mathrm{logsumexp}(a + b)$. The $|c - 1|$ form is therefore a first-order approximation, accurate where the scaled term dominates and poor where it does not.

Measured Δscore for DICE-cached under S1, against $|c - 1|$, on all four backbones:

	$c = 0.01$	$c = 0.5$	$c = 3$	$c = 5$
predicted $\lvert c - 1\rvert$	0.99	0.50	2.00	4.00
ResNet-18-MLP/CIFAR-100	0.848	0.500	2.004	4.008
ResNet-18-MLP/CIFAR-10	0.843	0.500	2.005	4.011
VGG-16-BN/CIFAR-100	0.769	0.493	2.017	4.035
VGG-16-BN/CIFAR-10	0.817	0.498	2.022	4.046

Agreement is within 1.2% at $c = 0.5, 3, 5$ and between 14% and 22% low at $c = 0.01$. The deviation is systematic rather than stochastic: every cell with $c > 1$ lies above the prediction and every cell with $c < 1$ below it. The registered prediction for DICE-cached is MOVES, which all sixteen cells satisfy; the $|c - 1|$ form was derived afterwards and is reported here with its range of validity.

C TOLERANCE SENSITIVITY AND SYMMETRY COVERAGE

Three counts appear in this paper and they are not interchangeable. Table 1 scores the 16 registered (detector, symmetry) rows at the registered 10^{-6}; that is the confirmatory result. The first sweep below rescores those same 16 rows at other tolerances. The second counts individual result rows, of which 106 carry an INVARIANT prediction, and is a different denominator entirely. Only Table 1 at 10^{-6} is confirmatory; everything in this appendix is post hoc sensitivity and revises no registered prediction. Both tables are recomputed from the shipped result files by `results/tolerance_and_coverage.json`, which records the rule used.

How the headline count moves with the tolerance. A row is scored by the worst cell in its designed backbone set: it MOVES if any cell's ΔAUROC reaches the tolerance, is SCORE-ONLY if only Δscore does, and is INVARIANT otherwise, with SCORE-ONLY counted as movement.

tolerance	10^{-8}	10^{-7}	10^{-6}	2×10^{-6}	3×10^{-6}	10^{-5}	10^{-4}
rows matching prediction	8	8	11	12	13	13	13

The count is flat at 13 from 3×10^{-6} upward, so the headline result does not depend on where the tolerance is placed above that point. It is not 13 at the registered 10^{-6} read strictly: the two rows in the gap are MSP under each symmetry, whose Δscore sits at 1.0–2.3×10^{-6}. Those are the cells discussed in Section 3 and marked in Figure 3; scored on the threshold alone rather than on the reading given there, the count is 11 of 16.

Below 10^{-7} the count falls to 8 of 16 and every row registers movement. This is consistent with accumulated float32 error rather than with any detector property, though we have not isolated rounding as the cause and do not claim to have located a precision floor.

At the cell rather than the row level the picture is sharper. Of the 121 live rows whose registered prediction is INVARIANT, 110 satisfy $|\Delta\mathrm{AUROC}| \leq 10^{-6}$, rising to 118 at 3×10^{-6} and flat thereafter. Every one of the eight rows that enters between 10^{-6} and 3×10^{-6} comes from a ResNet-18-MLP/CIFAR-10 cell already excluded as degenerate (Figure 3): six ReAct, whose score on that cell is pinned at $\log 10$, and two MDS, whose scores there are non-finite. Restricting the sweep to the cells the design includes leaves 102 rows, of which 99 pass at 10^{-6} — and 99 at 2×10^{-6}, 3×10^{-6}, 10^{-5} and 10^{-4} alike. Loosening the tolerance by two orders of magnitude admits no further row. There is no boundary band in the measured cells; the apparent one is an artefact of counting cells the audit excludes. The three rows that fail at every tolerance are ViM, a registered miss diagnosed in Table 3 as a float32 effect in the residual eigendecomposition. The registered 10^{-6} is the only confirmatory threshold; this sweep revises no prediction.

Uncertainty on the refit/frozen contrast. The result files record point estimates, so Section 4.2's contrast carries no sampling uncertainty. A paired stratified bootstrap supplies it on VGG-16-BN/CIFAR-10 against far/SVHN: ID and OOD indices resampled separately, the same indices applied to baseline and transformed scores, both AUROCs recomputed, 10,000 replicates at a recorded seed. Across the twelve refit cells the largest observed $\Delta\mathrm{AUROC}$ is 3.5×10^{-8} and the largest interval upper bound 1.05×10^{-7}, with no replicate anywhere reaching 10^{-6}. Across the eight frozen cells the smallest lower bound is 5.49×10^{-4} and every replicate exceeds 10^{-6}. The two sets of intervals are separated by a factor of about $5{,}200$ and neither approaches the tolerance from the wrong side, so the eight-order contrast is not an artefact of two point estimates that happen to differ. At $c = 0.5$ the refit interval is exactly $[0, 0]$ for all three detectors while ReAct-frozen still moves by 4.60×10^{-3}: a power-of-two scale removes the numerical component and leaves the stale reference untouched. Full intervals are in `results/bootstrap_intervals.json`; this is post-hoc sensitivity and revises no registered verdict.

How thin the permutation evidence is. Counting (detector, backbone) pairs actually measured:

	detectors	(detector, backbone) pairs	pairs on CIFAR-100
S1 (rescaling)	17	38	12
S2 (permutation)	7	15	2

Three of those pairs are the degenerate ResNet-18-MLP/CIFAR-10 cells noted above (two under S1, one under S2); counting only cells the design includes gives 36 and 14. On those design cells S2 reaches 39% of S1's pairs, and **only one CIFAR-100 backbone was measured under S2**: the DICE refit/frozen pair on VGG-16-BN. On that cell the registered prediction holds and the contrast is sharp. With the mask rebuilt on the permuted network, $\Delta\mathrm{AUROC}$ is 5×10^{-9}, 1×10^{-8} and 1×10^{-8} across the three seeds and the operating point transfers exactly ($\mathrm{TPR} = 0.95$); with the mask frozen, $\Delta\mathrm{AUROC}$ is 0.043, 0.056 and 0.022 and TPR transfer falls to 0.68–0.69. Both arms see the same transformation: logit drift is 3.81×10^{-6} and reach drift 4.27–4.60 in each. Every other permutation result here still rests on CIFAR-10, and the refit arm carries no registered prediction because `predictions.csv` registers only the frozen arm; it is recorded as NONE and reported as a control, not scored.

D ALGEBRA, DIRECTION AND THE $c = 0.5$ DIAGNOSTIC

D.1 SCORE MOVEMENT IS FIXED; RANKING MOVEMENT IS NOT

The algebra determines the score displacement; its ranking consequence depends on the network and on the OOD set.

For DICE, Δscore is nearly identical across all four backbones — at $c = 0.01, 0.5, 3, 5$ it measures $0.82/0.50/2.02/4.05$, $0.77/0.49/2.02/4.04$, $0.84/0.50/2.01/4.01$ and $0.85/0.50/2.00/4.01$ — while ΔAUROC varies by roughly two orders of magnitude.

Changing only the OOD set isolates the effect. On VGG-16-BN/CIFAR-10, DICE's Δscore against near-OOD (CIFAR-100) matches its far-OOD (SVHN) value to three decimals: 0.817, 0.498, 2.022,

4.046 in both. The displacement belongs to the symmetry, not the data. The AUROC change does not: near-OOD damage is 3.9 to 15.6 times larger, rising with c. ReAct-cached runs the other way, its near-OOD damage at 0.40 to 0.57 of the far-OOD value at every c. The same displacement is therefore amplified by one benchmark and damped by the other, so a single far-OOD number understates one detector's sensitivity and overstates the other's. This is one backbone and one near set; we do not generalise the ratio.

D.2 Failure depends on what was frozen

The direction of failure depends on the frozen reference. DICE's frozen mask degrades worst at $c = 0.01$ on every backbone. ReAct's frozen percentile degrades worst at $c = 5$ on both VGG models, but at $c = 0.01$ on the single ResNet-18-MLP backbone where its score is healthy (Figure 7). Because that reversal rests on one ResNet-18-MLP instance, we report the contrast as an observation rather than a rule.

D.3 Scalar and vector references fail under different symmetries

ReAct freezes a scalar percentile; DICE freezes a matrix, and which matrix matters. The pinned `calculate_mask` caches `masked_w` $= W \odot M$: the *weights* after masking, not M. A boolean M would be invariant under S1, since rebuilding gives $(ch)^\top (W/c \odot M) = h^\top (W \odot M)$ and the factor cancels — what the refit arm measures at 10^{-8}. The stale $W \odot M$ instead leaves the vote scaled by c, so the logits shift by $(c-1)\, h^\top (W \odot M)$ and the score moves as $|c-1|$: the registered closed form, derived from the implementation rather than fitted.

Table 4: Frozen references under both symmetries, VGG-16-BN/CIFAR-10, ΔAUROC.

detector	frozen object	S1 (rescaling)	S2 (permutation)
ReAct	scalar percentile	5.29×10^{-1}	*void: identical to refit*
DICE	masked weights $W \odot M$	1.18×10^{-2}	1.06×10^{-1}

A frozen scalar carries magnitude only: `np.percentile` over flattened activations is order-independent, so permutation leaves the threshold correct and the ablation has nothing to act on. A frozen $W \odot M$ carries magnitude *and* coordinate identity, so both symmetries invalidate it, and DICE moves harder under permutation than rescaling. On `resnet18_mlp`/CIFAR-10 the pair is wider still: ΔAUROC of 5.0×10^{-9} to 3.5×10^{-8} rebuilt against 0.300, 0.614 and 0.630 frozen across three permutation seeds. That spread is the split of Section D.1 under permutation: score displacement is fixed at 0.84 whichever permutation is drawn, while ranking damage depends on which units swap.

D.4 A diagnostic: $c = 0.5$

Halving and doubling are exact in binary floating point, so a transform at $c = 0.5$ leaves every intermediate bit-identical. Every detector whose ranking we predicted would hold returned exactly zero ΔAUROC there — MSP, MDS, ViM, KNN, the rankings of SHE and GradNorm, and the refit arms of both ReAct and DICE — while every mechanical failure (ReAct's and DICE's frozen references, NCI's absolute α, GMM's `reg_covar` floor) moved at $c = 0.5$ as at every other value.

If a detector is clean at $c = 0.5$ and dirty elsewhere, its movement is numerical rather than mechanical. GMM is the case that could have broken this and does not. Its `reg_covar` floor is absolute, so one might expect a covariance scaled by c^2 to clear it at $c = 0.5$; in fact GMM moves there too, by 1.8×10^{-4} against 2.7×10^{-3} at $c = 0.01$ — smaller, because the floor binds less, but not zero. A threshold never approached at $c = 0.5$ would still pass this test, so the diagnostic is evidence rather than proof.

It works because the transform is not exact in float32. Measured directly against ch on two backbones, the transformed network's features deviate by 1.6 to 2.2×10^{-7} relative at every c except the powers of two, where they are *bit-identical*: scaling by 2^k shifts the exponent and leaves the mantissas untouched. In float64 the deviation falls to 4×10^{-16}. At $c = 0.5$ a numerically driven detector therefore has nothing to respond to, while a stale reference or absolute constant fails either

way (`results/forward_pass_fidelity_S1.json`). The separation is exact rather than approximate: of the 55 rows measured at $c = 0.5$ in the shipped results, 29 are identically zero and 26 are not, and no detector appears on both sides. The one borderline case is Gram's seeded refit arm at 10^{-8}, below tolerance but not identically zero. We recommend including $c = 0.5$ in any such sweep.

E A REACHABLE GAUGE ON STOCK RESNET-18

Section 3 reports that the registered S1 does not reach the penultimate feature of a stock ResNet-18. That leaves open whether the network is immune or the gauge merely fails to reach it. We test this with a second gauge, registered as `prereg-v4` after the original audit and after the gauge's gates were measured, but before any detector was scored under it.

Construction. Scale every BatchNorm affine pair (γ, β) by c, divide every convolution weight except the stem by c, and divide the linear head's weight by c, leaving its bias. Each convolution then sees its original input, so frozen running statistics stay valid; each BatchNorm output scales by c; ReLU is positively homogeneous; and a residual sum of two c-scaled terms stays c-scaled. The penultimate feature scales by exactly c and the logits are unchanged.

Gates. On `resnet18_cifar10_seed0`, $\|\Delta z\|_\infty = 0$ exactly at $c = 0.5$ and $c = 2.0$, with the feature ratio exactly c and $\|\Delta h\|_\infty$ of 1.002 and 2.005. At $c = 3.0$, $\|\Delta z\|_\infty = 4.15 \times 10^{-5}$ fails the 10^{-5} validity gate: the construction divides nineteen convolutions, and each non-power-of-two division rounds. Only $c \in \{0.5, 2.0\}$ are scored.

Result. Every arm returns the outcome the rule predicts, 26 cells in all. Frozen references move: ReAct-cached by $\Delta\mathrm{AUROC}$ 7.1×10^{-4} and 3.8×10^{-3}, DICE-cached by 0.120 and 0.018. Their refit arms return exactly zero. MSP, Energy, MaxLogit, GEN, MDS and KNN return exactly zero at both scales. NCI, whose α is absolute, moves by 1.2×10^{-4} and 4.9×10^{-4}. SHE and GradNorm are score-only with $\Delta\mathrm{AUROC}$ exactly zero and Δscore exactly $|c^2 - 1|$ and $|c - 1|$ — the closed forms of Section 4.3, holding under a construction that shares no parameter with S1. The laws belong to how the feature scales, not to how the scaling is produced.

What it settles. A stock ResNet-18 is not immune to frozen-reference failure. Under the registered S1 it looked immune because that gauge cannot reach its feature; under a gauge that can, DICE-cached loses 0.12 AUROC. Head dependence is a property of the transformation searched, not of the architecture.

How it was registered. `prereg-v4` sorted detectors by reference handling, predicting rebuilt references invariant and frozen ones moving. That grouping omits the rule's other axis, units: it would have called SHE and GradNorm, which are score-only by construction, and NCI, whose constant is absolute, invariant. Scored against that literal text the result is 20 of 26; scored against the specific `predictions.csv` registrations, which the reclassification followed and which was announced before each batch ran, it is 26 of 26. Both are reported. Mahalanobis-family float32 effects cannot appear here, since only powers of two are valid and they introduce no rounding. ViM, GMM and Gram were not run; Gram reads intermediate activations that the penultimate-feature path does not supply.

F IMAGENET-1K REPLICATION

This appendix is post hoc. No prediction was registered for it, it changes no row of Table 1, and it tests whether the refit–frozen split found at CIFAR scale survives at ImageNet scale and on near-OOD data.

Models and symmetries. We use torchvision's pretrained VGG-16-BN and ResNet-50, whose top-1 accuracy on a 1,562-image check was 0.739 and 0.765 against published values of 0.734 and 0.761. VGG-16-BN has two hidden linear layers, so the registered S1 reaches its penultimate feature;

we apply it at $c \in \{0.01, 0.5, 3, 5\}$, with logit drift at most 5.7×10^{-6}, the same float32 floor as at CIFAR scale. ResNet-50 has a single linear head, so S1 cannot reach it; we apply the trunk gauge of Appendix E, which is exact at $c \in \{0.5, 2\}$ and fails the validity gate at $c = 3$ with drift 5.2×10^{-5}.

Data and references. Images are streamed from the webdataset release of ImageNet-1K. Following the OpenOOD source at the pinned commit, ReAct fits its 90th-percentile threshold on validation data and DICE its mean activation on training data: here, 6,250 validation images and 10,009 training images. The ID test set is the remaining 43,750 validation images, disjoint from ReAct's fitting set. OOD sets are NINCO (5,879; Bitterwolf et al. 2023), SSB-hard (49,000; Vaze et al. 2022), iNaturalist (10,000; Van Horn et al. 2018, as curated by Huang and Li 2021), OpenImage-O (15,869; Wang et al. 2022) and Textures (5,160; Cimpoi et al. 2014), from the OpenOOD v1.5 test lists (Zhang et al., 2024). The rebuilt references behave as the algebra requires: ReAct's refit threshold divided by c agrees to six digits at every scale, and DICE's refit mask is identical to the baseline mask entry for entry.

Table 5: Worst ΔAUROC over c, rebuilt against frozen reference. MSP is logit-only. ResNet-50 values are exactly zero because the trunk gauge is exact at powers of two.

		MSP	ReAct refit	ReAct frozen	DICE refit	DICE frozen
VGG-16-BN	NINCO	3.7×10^{-8}	3.3×10^{-8}	0.289	7.8×10^{-9}	0.083
	SSB-hard	2.0×10^{-8}	2.8×10^{-9}	0.109	4.0×10^{-9}	0.053
	iNaturalist	4.0×10^{-8}	6.9×10^{-9}	0.376	3.4×10^{-9}	0.163
	OpenImage-O	2.3×10^{-8}	2.9×10^{-9}	0.321	9.4×10^{-9}	0.091
	Textures	4.4×10^{-8}	2.0×10^{-8}	0.301	1.6×10^{-8}	0.044
ResNet-50	NINCO	0	0	0.217	0	0.046
	SSB-hard	0	0	0.027	0	0.024
	iNaturalist	0	0	0.198	0	0.020
	OpenImage-O	0	0	0.294	0	0.041
	Textures	0	0	0.405	0	0.006

Result. All 50 detector–architecture–dataset combinations return the outcome the registered rule predicts: MSP and both rebuilt references stay below 5×10^{-8}, and both frozen references move on every combination, by up to 0.405. The magnitude of the damage varies by more than an order of magnitude with architecture and OOD set, as Section D.1 would lead one to expect. Frozen ReAct was worst at the largest c on all ten architecture–dataset pairs. Frozen DICE was worst at the smallest c on nine, but on ResNet-50 with Textures it was worst at $c = 2$: together with the seed replication, a second case of direction reversing, which is why we do not generalise it.

Permutation, SCALE and GradNorm. Three further post hoc runs use NINCO. Under S2 on VGG-16-BN, with three permutation seeds and the full ID test, frozen DICE moves by between 0.064 and 0.075, while MSP and both rebuilt references stay below 5×10^{-8}. Frozen ReAct's threshold is a percentile over pooled activations, which permutation leaves unchanged, so its two arms coincide on every image and the row is VOID, as at CIFAR scale. SCALE, run as its authors implement it rather than through the pinned OpenOOD path, which clips instead of scaling, has no fitted reference: it multiplies the feature by $\exp(s_1/s_2)$, a ratio of two sums of that feature. On a 6,250-image ID set its score changed by at most 8.4×10^{-7} relative on VGG-16-BN under S1 and S2, and not at all on ResNet-50 under the trunk gauge, with ΔAUROC below 5×10^{-8}. The registered SCALE-cached prediction, that it moves, therefore has nothing to ablate: the rule, applied to what SCALE computes, predicts invariance, and the registration misclassified its constant as absolute. GradNorm, computed as $\|Kp-1\|_1\|h\|_1$ for K classes, softmax p and feature h, and checked in the run against OpenOOD's per-image loop to 2.3×10^{-7}, moves its score by $|c-1|$ to within 2×10^{-6} per image under S1 and by less than 5×10^{-7} under S2, with ΔAUROC below 7×10^{-8}: score-only under S1 and invariant under S2, as registered.

Deviations and limits. The ID split is ours, not OpenOOD's split file, since the streamed shards do not carry OpenOOD's filenames. ReAct and DICE were reimplemented from the OpenOOD source rather than run through its harness; the percentile, masking rule and energy score match that source. Seven detectors were run — the five of Table 5, and SCALE and GradNorm on NINCO

— against fifteen at CIFAR scale. Textures was scored in a later session, against the saved ID-test scores with every reference refitted; re-scoring 781 ID images in that session reproduced the saved scores to a relative difference of at most 3.9×10^{-6}, and the refitted thresholds matched the originals to every printed digit. The ID test and the first four OOD sets were scored, as in the CIFAR audit, without deterministic cuBLAS — the deterministic flag in the CIFAR manifests describes checkpoint training, and these models are torchvision's pretrained weights — whereas the later Textures, permutation, SCALE and GradNorm runs set it. Every comparison between a baseline and its transformed network was computed on the same batch, side by side, where no drift appears — the ResNet-50 refit arms return exactly zero — but the scores are not guaranteed to reproduce bit for bit in a new session. Per-image scores, the extraction code and a script that recomputes this table from them are in `results/imagenet/` of the supplementary package.

G THE AUDIT IN ONE DIAGRAM

Figure 4 sets out the structure tested in this paper: the two exact symmetries, the two symmetry gates every run must pass, and the two branches that differ only in whether the detector's internal reference is rebuilt on the transformed checkpoint or carried over from the original. Branch A is the invariant path and Branch B is where every failure in Section D.1 occurs. The measured "unchanged" of Branch A is unchanged to within the verdict tolerance, $\Delta\mathrm{AUROC} \approx 10^{-8}$, not identically zero; the panel on finite precision is the reason, and Section D.4 gives the measurements.

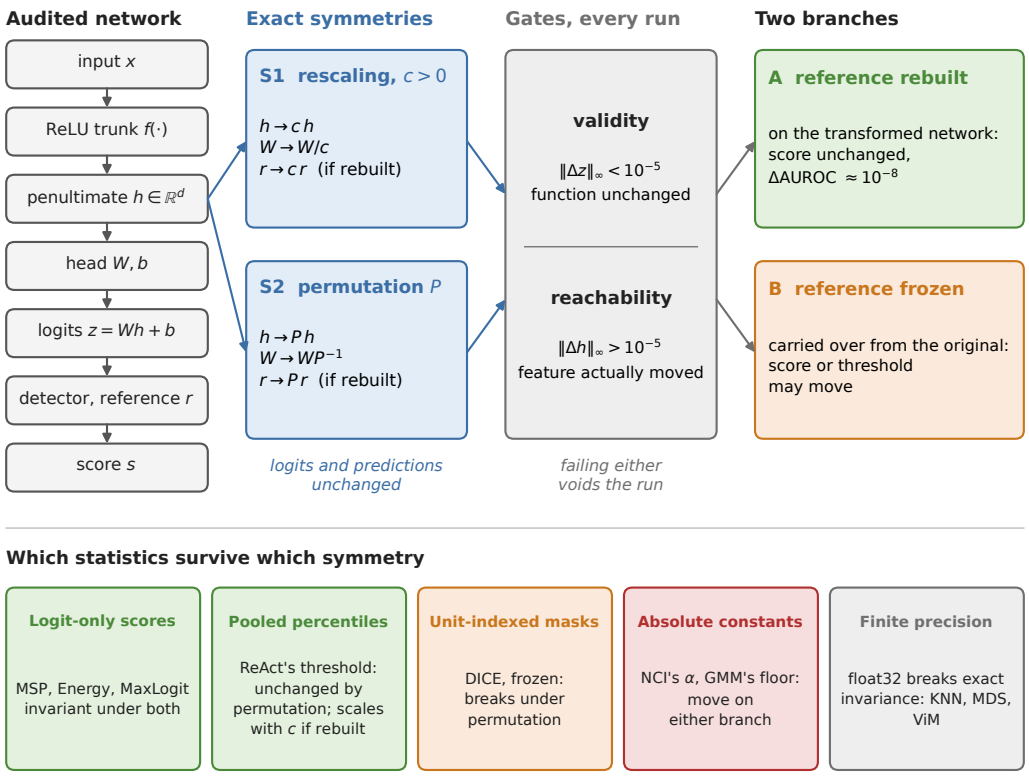

Figure 4: The audited object and the two branches. Blue: the exact symmetries, which leave logits and predictions unchanged. Grey: the gates every run must pass. Green: the reference is rebuilt on the network being scored, and the score is unchanged. Orange: the reference is frozen, and the score or the threshold may move. The lower panels summarise which kinds of statistic survive which symmetry; red marks constants fixed in absolute units, which move on either branch.

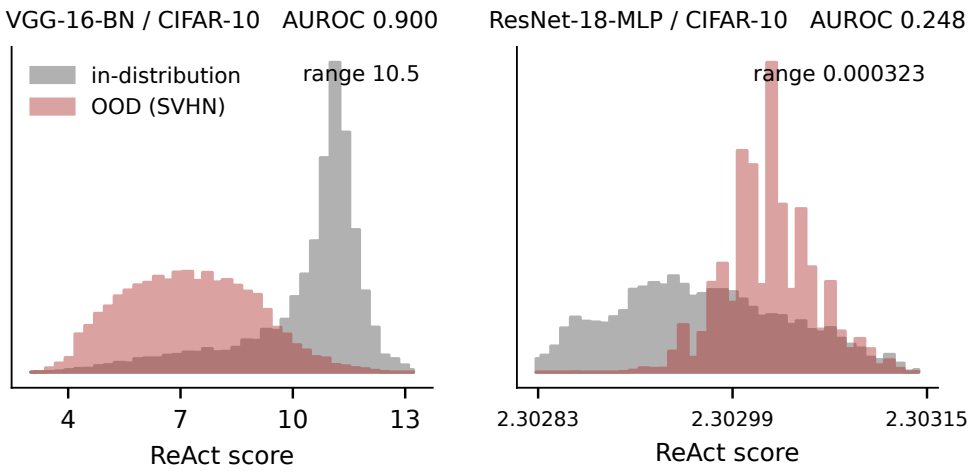

Figure 5: ReAct's in-distribution and OOD score distributions, same detector and same OOD set, on two backbones. Left: separation over a range of 10.5. Right: the entire range is 3.2×10^{-4} and the OOD mass sits above the in-distribution mass, which is why the AUROC is 0.248 rather than 0.5 — the ranking is of float32 noise. The harness returns the registered invariance here for the wrong reason, so the cell is excluded rather than scored; the health gate, added after these runs, now catches such scores (Appendix A).

H SUPPORTING FIGURES

Figure 6 shows the two laws of Section D.1 separately, for NCI rather than DICE: the same pattern holds for a detector with a different score, so the collapse is a property of the algebra and not of one implementation. Figure 7 plots ΔAUROC against c for the two frozen references. Figure 3 records every (detector, symmetry, backbone) cell we ran and what it returned. Its VGG-C100 column under S2 was measured after the original audit; its frozen arm carries a registered prediction, its refit arm is a control and carries none. GEN is shaded SCORE-ONLY there because that is what the gate returned, but its median relative score movement across 20,000 samples is exactly zero and only 0.06% of samples exceed the tolerance; its threshold transfer is unchanged at $\mathrm{TPR} = 0.95$. It is the clearest case in the audit of a maximum-based statistic reporting a tail as though it were the distribution. Energy, MaxLogit and GEN are shown for completeness and are not among the scored rows: each reads only logits or normalised probabilities, which neither symmetry moves, so their registered constants are never exercised and a pass is sheltered rather than evidence for the rule. Two further groups of cells carry a label that is read rather than thresholded, as Section 3 describes: MSP's four SCORE-ONLY cells sit at Δscore 1.0–2.3×10^{-6} against a 10^{-6} tolerance and are shown INVARIANT; ReAct-cached under S2 is grey because the ablation cannot act — its Δscore is identical across seeds — not because a gate fired. Figure 5 shows the score distributions behind ReAct's degenerate cells.

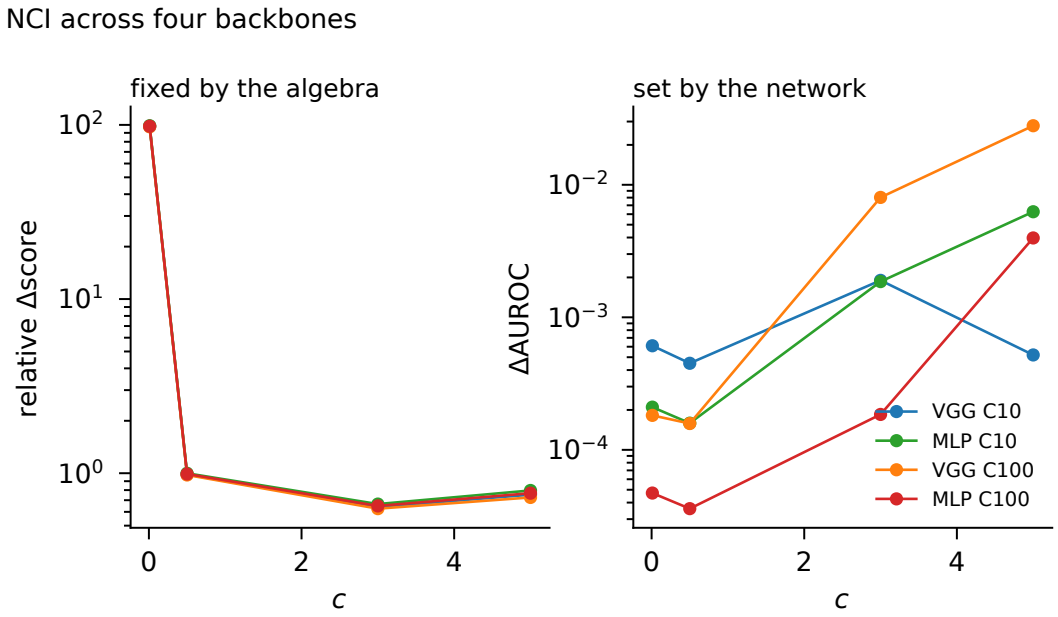

Figure 6: NCI on four backbones. Left: relative Δscore, four curves collapsing onto one — the movement is fixed by the score's algebraic form and stays within 10% across two architectures and two datasets. Right: ΔAUROC for the same runs, spanning three orders of magnitude.

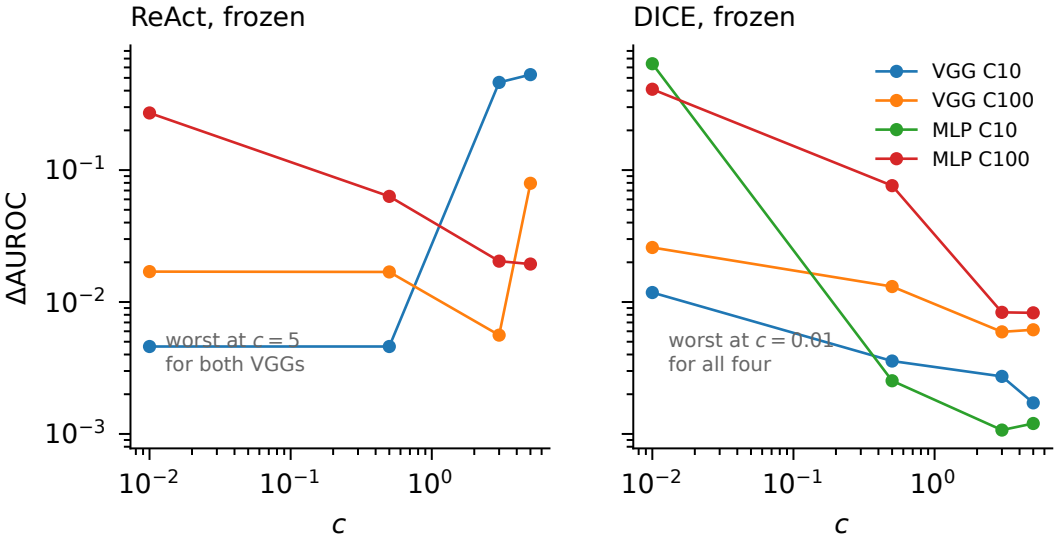

Figure 7: ΔAUROC against c for the two frozen references on clean validation splits: DICE on all four backbones, ReAct on three (MLP-C10 is omitted, its ReAct score being degenerate — Section 5). Each backbone keeps one colour across this figure and Figure 6. DICE is worst at $c = 0.01$ on all four of these checkpoints, though a second training seed reverses that (Section 4.3). ReAct is worst at $c = 5$ on both VGGs but at $c = 0.01$ on the ResNet-MLP. Plotted per detector, the two curves differ in shape.