SAME CLASSIFIER, DIFFERENT OOD DECISIONS: THE HIDDEN-BASIS DEPENDENCE OF ACTIVATION CLIPPING

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-hoc OOD detectors that shape the penultimate representation coordinate by coordinate—REACT, ASH, DICE—are popular because they are nearly free. We show their verdicts are not a function of the classifier alone: they depend on the network's *hidden basis*, and we ask which reparameterization symmetries a detector can and should respect, at three tiers. Under the architecture's own symmetry (positive channel rescaling), no global threshold reproduces REACT's clipping, however re-calibrated; demanding basis-agnosticism rules out every continuous coordinate-wise map except uniform rescaling and, among continuous per-sample shapers, forces the radial form $T(h) = \alpha(\|h\|)\, h$; within radial norm-based constructions, affine covariance leads to the Mahalanobis geometry of Maha-Radial. Empirically, across random bases of the *same* classifier function, REACT's AUROC swings by up to 15.8 points and an adversarially optimized basis inverts it entirely ($0.97 \to 0.003$), while the exactly invariant, training-free Radial-ReAct and Maha-Radial remove the dependence exactly—Radial-ReAct essentially matching raw energy at ImageNet scale, Maha-Radial exposing a calibration-rank tradeoff that covariance shrinkage mitigates. At scale the artifact dominates the benefit: on an untouched ImageNet ResNet-50 over six standard OOD suites, REACT beats raw energy at the trained basis on all six sets, yet on every one its worst of 20 random bases falls below the energy score—reparameterization alone erases the entire gain—and the swing explodes as the penultimate dimension shrinks (23 points at $d \leq 16$), exactly where compact deployed heads sit. Code reproducing every result: `https://anonymous.4open.science/r/react_invariance-A33F/`.

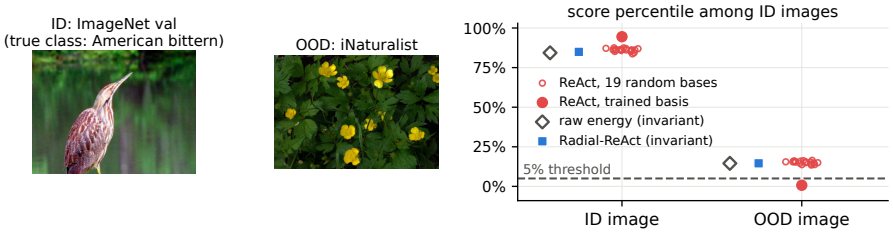


Figure 1: **One image, one classifier, two verdicts.** An untouched ImageNet ResNet-50 scores an ID bittern and an iNaturalist wildflower (no matching class); markers give each image's OOD score percentile among $10{,}000$ ID images (dashed: 95%-TPR threshold). At the trained basis REACT (filled red) flags the wildflower—its documented benefit—but under 19 random bases of the *same* function (open circles; logits identical to 3×10^{-14}) the score moves above the threshold, landing beside the basis-invariant radial shaper and raw energy (square, diamond). Only the basis changed.

1 INTRODUCTION

A large family of training-free out-of-distribution (OOD) detectors reads the network's penultimate representation $h(x) \in \mathbb{R}^d$ and transforms it before scoring. REACT (rectified activations) clips each

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.

activation at a percentile threshold (Sun et al., 2021), ASH (activation shaping) prunes and binarizes the largest activations per sample (Djurisic et al., 2023), and DICE (directed sparsification) masks head weights by an in-distribution (ID) contribution ranking (Sun & Li, 2022): these act *coordinate-wise* (per unit of the trained basis, in h or in W). A second group is *geometric*: ViM (virtual-logit matching) scores the residual of h to the ID principal subspace (Wang et al., 2022), and Mahalanobis-type distances score against an ID-estimated metric (Mahalanobis, 1936; Lee et al., 2018)—covariant with the basis by construction. The coordinate-wise family is popular because it is nearly free, and it quietly ties the detector to the network's *hidden basis*—the topic of this paper (Figure 1).

Coordinates of a penultimate layer are not intrinsic objects. ReLU networks possess an exact, architecture-internal symmetry: $h_D(x) = Dh(x)$, $W_D = WD^{-1}$ for any positive diagonal D (or permutation) yields the *same architecture* computing the *same function*; the construction realizes any orthogonal Q by a change-of-basis map before the head. No experiment on the classifier's function can reveal which parameterization is "the real one," yet a coordinate-wise OOD rule sees a different vector in each: a global percentile cannot track heterogeneous channel scales (Proposition 1), and the clipping map does not commute with rotations, $R_c(Qh) \neq QR_c(h)$.

This paper asks the question this raises, answers it theoretically, and closes it constructively: *Which reparameterization symmetries can a post-hoc OOD detector respect, what does respecting them force it to give up, and how much of its behavior is an artifact of the ones it violates?*

What we show. (1) Coordinate-wise shaping is provably non-invariant—under the architecture's own symmetry, and generically under rotations (§3.2, §3). Positive channel rescaling $D = \mathrm{diag}(e^{\sigma s})$ is a symmetry ReLU networks possess exactly, yet *no* global threshold, however re-calibrated, reproduces REACT's clipping operator: the conjugation $D^{-1}R_{c'}D$ clips channels at heterogeneous levels c'/d_j and equals no R_c unless D is scalar (Proposition 1). Per-channel thresholds restore exact equivariance (Proposition 2) but remain coordinate-wise—and no continuous pointwise coordinate-wise map is equivariant under rotations except a uniform rescaling (Proposition 3): *coordinate semantics and basis-agnosticism are structurally incompatible*. Under rotations the failure is generic (Theorem 1; Proposition 4).

(2) The symmetry demanded forces the shaper: a hierarchy of equivariant detectors (§3.3, §3.3). Equivariance under the natural group admits per-channel thresholds; orthogonal equivariance forces the radial form $T(h) = \alpha(\|h\|)\, h$ among continuous per-sample shapers (Theorem 2); affine covariance yields the Mahalanobis geometry of Maha-Radial. The tiers are realized by PC-REACT, Radial-ReAct, Maha-Radial—training-free, one line of code each; the first two are incomparable (a radial map ignores non-scalar rescalings; a per-channel map ignores rotations).

(3) The artifact is large, lawful, family-wide, and removable—and at scale it dominates the benefit (§4). Across five trained-from-scratch settings plus frozen ImageNet-pretrained backbones (3 seeds, 100 reparameterizations each, exact logit preservation verified for every transform):

- *Large and family-wide.* Under the natural rescaling, frozen REACT degrades FPR_{95} by up to $+20.1$ points at σ=1.5, re-calibration leaving monotone residuals of $+5$–10. Across random re-calibrated bases, REACT's AUROC (area under the ROC curve) swings by 2.0–9.4 points (Table 1); ASH-B and DICE swing still more (18–75 and 5–29), PC-REACT still moves—the hierarchy is strict—while the geometric control (ViM, re-fit PCA; principal component analysis) is exactly flat.
- *Systematic.* The swing follows a *dimension-dependent trend*, shared across all four task families on $d \in \{8, \ldots, 256\}$: it peaks at compact widths $d \leq 16$ (up to 23.3 AUROC / 53 FPR_{95} points) and decays with d roughly as a power law (log–log slopes -0.3 to -1.1 per family; Figure 3)—a practitioner choosing a compact head sits exactly where the artifact is largest. It also decreases monotonically in the clipping percentile (Appendix Figure 6).
- *Removable exactly—and the artifact dominates at scale.* Radial-ReAct/Maha-Radial are exactly invariant (zero variance over $1500+$ transforms), Maha-Radial matching or exceeding REACT's average AUROC in every setting. On the untouched ImageNet ResNet-50 (§4.4), REACT beats raw energy at the identity on all six OOD sets ($+1.7$ to $+7.9$), yet on all six its worst of 20 random bases falls below energy: reparameterization erases the entire gain.

The symmetry group is a design choice, and the tiers price it differently. The results establish the cost of leaving the choice implicit: *when two parameterizations realize the identical classifier func-*

tion, a post-hoc safety score that differs between them reports properties of the parameterization, not of the input—and, as the dimension trend shows, those properties can dominate the reported metric. Making the choice explicit, pricing each tier, and offering exactly invariant alternatives is the contribution.

2 PRELIMINARIES

Classifier and notation. Let $f(x) = Wh(x)+b$ with feature map $h : \mathcal{X} \to \mathbb{R}^d$, head $W \in \mathbb{R}^{K\times d}$, $b \in \mathbb{R}^K$. All experiments use the post-ReLU penultimate unless stated. Data matrices $h \in \mathbb{R}^{n\times d}$ carry one feature row per sample; $\mathcal{G}$ denotes the distribution from which reparameterizations are drawn.

Definition 1 (Function-preserving reparameterization)**.** *An invertible $A \in \mathbb{R}^{d\times d}$ reparameterizes f as $h_A(x) = Ah(x)$, $W_A = WA^{-1}$ (rows: $h_A = hA^\top$). Then $W_A h_A(x) + b = Wh(x) + b$ for all x: f is unchanged as a function. (The affine tier of equation 1 adds translations $h \mapsto hA^\top + a$ with $b \mapsto b - Wa$, preserving f identically.) We use (i)* rotations $A = Q \in O(d)$ *sampled Haar-uniformly (Mezzadri, 2007), and (ii)* positive channel rescalings $A = D = \mathrm{diag}(e^{\sigma s})$, $s_j \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, $\mathrm{cond}(D)$ *growing with σ.*

These families define equivariance *requirements* of increasing strength (arrows below; G_{nat} and $O(d)$ are incomparable subgroups of $GL(d)$, the affine tier contains both):

$$\underbrace{G_{\text{nat}} = \{PD :\ P \text{ permutation, } D \text{ positive diagonal}\}}_{\text{architecture's own symmetry}} \longrightarrow \underbrace{O(d)}_{\text{basis-agnosticism}} \longrightarrow \underbrace{\{h \mapsto Ah + a\}}_{\text{affine covariance}}. \tag{1}$$

The first needs no argument: G_{nat}-reparameterizations are realized *within* the same ReLU architecture, so the rescaled network is not an insertion but the same model class. Orthogonal reparameterization is a stronger, *design* requirement—basis-agnosticism: it corresponds to allowing an arbitrary linear change of basis between the last hidden layer and the head, and no experiment on the classifier's function can rule it out. Affine covariance subsumes both and is the natural geometry of Mahalanobis-style scores. A detector can be equivariant at one tier and not another, and the tiers G_{nat} and $O(d)$ are *incomparable* (neither group contains the other's non-scalar elements; illustrated in Figure 5, Appendix C.1); §3–§3.3 work out which shapers each tier admits.

Remark 1 (Normalization layers and the natural group)**.** *Which parts of G_{nat} survive depends on the normalization and on where the hook sits: plain ReLU and BatchNorm with frozen inference statistics keep all of it; LayerNorm keeps all of it for* post-LN *features (a positive diagonal rescaling is absorbed into LN's own affine parameters γ, β) but only permutations $\oplus$ global scaling for* pre-LN *features (full statement: Appendix B.2).*

REACT and the scores. REACT (Sun et al., 2021) applies $R_c(h)_j = \min(h_j, c)$ with a global threshold c set to the p-th percentile of *all* ID calibration activations (p=90 unless stated). Scored detectors we evaluate: energy $E(x) = -\log \sum_k e^{z_k(x)}$ (Liu et al., 2020), MSP (maximum softmax probability) (Hendrycks & Gimpel, 2017), and MaxLogit. The OOD metrics are AUROC (FPR_{95}: false-positive rate of OOD at 95% true-positive rate on ID; AUPR (area under the precision–recall curve) in the appendix), ID as the positive class.

Sensitivity metrics. For a metric M and transform distribution $\mathcal{G}$:

$$\mathrm{RS}(M) = \mathrm{Std}_{A\sim\mathcal{G}}\left[M(A)\right], \qquad \mathrm{WD}(M) = M(I) - \min_{A\sim\mathcal{G}} M(A), \tag{2}$$

representation sensitivity and worst-case degradation. These quantify the artifact; they are not proposed as a benchmark.

Protocol. Features are extracted once in float64, every threshold *re-fit* on the transformed calibration set, and the run aborts unless logits are preserved to 10^{-13} (full protocol: Appendix B).

3 FROM HIDDEN-BASIS DEPENDENCE TO INVARIANT SHAPING

3.1 REACT IS NOT INVARIANT, AND RE-CALIBRATION DOES NOT SAVE IT

Theorem 1 (Non-invariance). *For every $d \geq 2$ there exist $h \in \mathbb{R}^d$, $W \in \mathbb{R}^{K\times d}$, $b \in \mathbb{R}^K$, $c > 0$, and $Q \in O(d)$ such that $WQ^\top Qh + b = Wh + b$ yet $W R_c(h) + b \neq WQ^\top R_c(Qh) + b$.*

(Proof: Appendix A.) The failure is not an accident of a special basis: conditioning on clipping being active and non-degenerate, the set of rotations admitting *any* recovering threshold—re-calibrated or adversarially optimized—has Haar measure zero (Proposition 4).

3.2 POSITIVE RESCALING: A GENUINE NETWORK SYMMETRY THAT GLOBAL THRESHOLDS CANNOT TRACK

Rotations can be dismissed as an adversarial insertion; $D = \mathrm{diag}(d_1, \ldots, d_d)$, $d_j > 0$, cannot: positive homogeneity makes the rescaled network the *same ReLU network*—the strongest reparameterization family *within* the architecture. Here the failure is sharp and distribution-free.

Proposition 1 (No global threshold tracks the rescaled clipper). *Let $D = \mathrm{diag}(d_1, \ldots, d_d)$ with $d_j > 0$, not all equal, and let $c' \in \mathbb{R}$ be arbitrary. The global clipper, conjugated into the rescaled coordinates, acts as a* per-channel *clip at heterogeneous levels:*

$$\big[D^{-1} R_{c'}\big(D\,h\big)\big]_j = \min\big(h_j,\ c'/d_j\big) \qquad \textit{for all } h \in \mathbb{R}^d. \tag{3}$$

As a map on $\mathbb{R}^d$, $D^{-1}R_{c'}D$ equals R_c iff $c'/d_j = c$ for every j—impossible for any $c > 0$ when the d_j differ (percentile calibration yields $c > 0$ whenever any activation is positive); $c = 0$ is the trivial exception, tracked by $c' = 0$. This is a statement about the operators themselves: it requires no assumption on the input distribution (no independence, no identical channel laws).

In quantile language: the re-calibrated threshold clips channel j at the wrong level c'/d_j, so the effective per-channel quantiles $F_j(c'/d_j)$ are heterogeneous as soon as the d_j (or the channel laws F_j) are—the semantics of a common quantile is destroyed (Corollary 1); §4.3 quantifies the resulting degradation.

The symmetry also dictates the fix: nothing forces the threshold to be *global*, and the natural group asks only for *per-channel* calibration—

Proposition 2 (Per-channel thresholds are exactly equivariant for the natural group). *For each channel j, let $\hat{c}_j$ be any positively homogeneous statistic of the ID calibration activations ($\hat{c}_j(\lambda\, h_{\cdot,j}) = \lambda\, \hat{c}_j(h_{\cdot,j})$ for $\lambda > 0$; the per-channel p-percentile qualifies), and define $PC(h)_j = \min(h_j, \hat{c}_j)$. Then for every positive diagonal D and every permutation matrix P: calibrating $\hat{c}$ on the transformed ID data and applying PC commutes with the transform,*

$$PC_{\hat{c}(D\cdot)}(Dh) = D\ PC_{\hat{c}}(h) \qquad \textit{(and likewise for } P\textit{)}, \tag{4}$$

so the post-shaping logits are identical *in the two parameterizations. Per-channel thresholds remove the basis-dependence of* REACT for the symmetry group the architecture actually has.

Both proofs are one line each (Appendix A). What per-channel thresholds cannot do is deliver *basis-agnosticism*—and the next result shows this is a structural impossibility, not an accident of the construction.

3.3 EQUIVARIANT SHAPING, TIER BY TIER

Per-channel thresholds are exactly equivariant for the natural group but remain coordinate-wise; if one additionally asks for equivariance under *all* of $O(d)$—basis-agnosticism—the two desiderata collide:

Proposition 3 (Coordinate-wise semantics and basis-agnosticism are incompatible). *Let $d \geq 2$ and let T be a continuous* pointwise coordinate-wise *map, $T(h)_j = \varphi_j(h_j)$, that is equivariant, $T(Qh) = QT(h)$ for all $Q \in O(d)$. Then $T(h) = k\,h$ for a constant k: a uniform rescaling. No clipping, thresholding, or sparsifying operator of this form is basis-agnostic.*

A detector must choose: keep coordinate-wise semantics and accept dependence on the hidden basis (global or per-channel thresholds, REACT, DICE-style sparsification), or ask for basis-agnosticism and give up coordinates. What remains in the second branch?

Theorem 2 (Radial characterization)**.** *Let $d \geq 2$ and $T : \mathbb{R}^d \to \mathbb{R}^d$ be a continuous per-sample shaper (acting on one vector h at a time, independent of the rest of the batch). Then T satisfies $T(Qh) = QT(h)$ for all $Q \in O(d)$, $h \in \mathbb{R}^d$,* if and only if *$T(0) = 0$ and there exists $\alpha \in C\big((0,\infty), \mathbb{R}\big)$ with $\lim_{r\downarrow 0} r\,\alpha(r) = 0$ such that $T(h) = \alpha(\|h\|_2)\,h$ for all $h \neq 0$, with α acting on the* unsquared *radius $r = \|h\|_2$.*

(Full proofs of Theorem 2 and Proposition 3: Appendix A.) The statement is sharp—$\alpha(r) = r^{-1/2}$ is admissible—and it says *radial maps are all there is* for the per-sample equivariant class; invariance of the whole detector then requires the scalar field (and any affine center) to be estimated covariantly from ID data. The three tiers of equation 1 each admit a canonical training-free shaper—one line of code, replacing pooled percentile clipping of *coordinates* with calibrated structures that transform covariantly.

Tier 1 (natural group): PC-REACT. Calibrate each channel separately: $PC(h)_j = \min(h_j, \hat{c}_j)$, $\hat{c}_j$ = p-th percentile of channel j—exactly G_{nat}-equivariant (Proposition 2), still coordinate-wise, hence not basis-agnostic (Proposition 3).

Tier 2 (orthogonal, basis-agnostic): Radial-ReAct. With $\tau = Q_p\big(\|h(X_{\text{cal}})\|_2\big)$,

$$T_\tau(h) \;=\; h \cdot \min\Big(1, \tfrac{\tau}{\|h\|_2 + \varepsilon}\Big). \tag{5}$$

$T_\tau(Qh) = QT_\tau(h)$ exactly, so with $W_Q = WQ^\top$ the post-shaping logits are identical in every basis (Theorem 2 with $\alpha(r) = \min(1, \tau/(r+\varepsilon))$).

Tier 3 (affine-covariant): Maha-Radial. Estimate (μ, Σ) on ID calibration features, define the Mahalanobis radius $r(h) = \sqrt{(h-\mu)^\top \Sigma^{-1}(h-\mu)}$ and $\tau = Q_p(r(h))$; then

$$T_{\mu,\Sigma,\tau}(h) \;=\; \mu + \min\Big(1, \tfrac{\tau}{r(h)+\varepsilon}\Big)\,(h-\mu). \tag{6}$$

If $h' = Ah + a$, $\mu' = A\mu + a$, $\Sigma' = A\Sigma A^\top$ then $r'(h') = r(h)$ and $T_{\mu',\Sigma',\tau}(h') = A\,T_{\mu,\Sigma,\tau}(h) + a$: equivariance under all invertible *affine* reparameterizations, a strictly stronger invariance than orthogonal. With rank-deficient Σ (typical for post-ReLU features) we use a fixed relative cutoff $10^{-12}\lambda_{\max}$ eigendecomposition inverse, which stays orthogonally equivariant; exact affine covariance holds for full-rank Σ and fails for rank-deficient Σ *even when the eigenspace is preserved* (Proposition 5(iii)), so at $n_{\text{cal}} < d$ the affine guarantee degrades strictly to the orthogonal one (§4.4 measures the 12–13 FPR_{95}-point failure on ResNet-101; Appendix C.10 repairs it with the shrinkage dial).

Shrunk covariance: the $n_{\text{cal}} < d$ repair. A Ledoit-style shrinkage $\Sigma_\lambda = (1-\lambda)\hat{\Sigma} + \lambda I$ restores full rank, and Proposition 6 states what is bought and sold: orthogonal equivariance survives *exactly* for every λ, while the affine tier acquires a bias vanishing as $\lambda \to 0$. Shrinkage turns Maha-Radial's two-case guarantee ("affine if invertible, otherwise orthogonal") into a tunable choice; §4.4 measures the dial.

4 EXPERIMENTS

4.1 SETUP AND SANITY

Tasks and models. Five trained-from-scratch settings—Digits, MNIST, and CIFAR-10 in near- and far-OOD splits (Appendix B)—on an MLP (d=32), a CNN (d=64), and ResNet-18-lite (d=128), 3 seeds each, plus the ImageNet-scale benchmark of §4.4 (untouched pretrained ResNet-50). **Detectors.** REACT; PC-REACT; ASH-B; DICE; ViM (re-fit per transform) (Djurisic et al., 2023; Sun & Li, 2022; Wang et al., 2022); kNN (k-nearest neighbors) (Sun et al., 2022); Radial-ReAct; Maha-Radial; raw energy (exact rules: Appendix Table 4). **Defaults.** p=90 percentile, energy

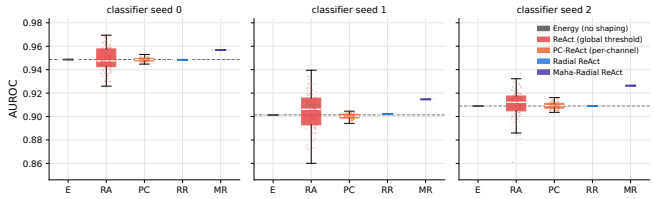


Figure 2: **Same classifier function,** 100 **hidden bases (MNIST near-OOD).** AUROC of each detector across Haar-random orthogonal reparameterizations (thresholds re-calibrated per basis; p=90; energy score; one panel per classifier seed). Raw energy, Radial-ReAct and Maha-Radial are exactly flat (std $< 10^{-15}$); REACT moves; PC-REACT moves less but is not flat. Remaining tasks and FPR_{95} panels: Figures 12–15.

score, 1000 ID calibration samples, 100 reparameterizations per classifier (m=0 identity, so WD is well-defined), thresholds always *re-calibrated* per basis unless labeled "frozen"; ASH-B pruning percentile 90, DICE sparsity 0.7 (the official ImageNet default), ViM D=1000 at d=2048 else $d/2$, all official (Appendix Table 4). **Sanity.** Across all 1500+ transforms the maximum logit deviation is 3.0×10^{-14}, raw-energy AUROC is constant to 10^{-16}, and the synthetic checks of Theorem 1 and Proposition 4 pass exactly.

4.2 ROTATION SENSITIVITY AT THE STANDARD OPERATING POINT

Table 1 (columns RS, range, WD) and Figure 2 give the central measurement. At p=90 with re-calibration, REACT's AUROC across bases has per-seed ranges 2.0–9.4 points (FPR_{95} up to 31.4), with worst-case degradation from the identity basis up to 6.3 points (per-seed panels: Figure 2). ASH-B swings by 18–75 points and DICE by 5–29—both far exceeding REACT's worst seed—while the geometric detectors (ViM with re-fit PCA, kNN (Sun et al., 2022)) are exactly flat ($< 10^{-15}$), confirming the taxonomy (Appendix Table 5). PC-REACT sits exactly in between, as Propositions 2–3 predict: rotation range 0.4–2.6 points—consistently nonzero, yet 2–10× smaller than global REACT's. The swing is not a calibration artifact: growing the ID calibration set 50× (100→5000) leaves the across-basis range unchanged in every task (Appendix Table 8). The swing also falls monotonically in the clipping percentile (below 2 points at p=99; Appendix Figure 6). The adversarial counterpart:

Adversarial basis search: the worst case is inversion. Optimizing Q directly (Cayley parameterization, Adam, train/test-half transfer; Appendix C.8), REACT falls to 0.003–0.297 across the four tasks—inversion—whereas the worst of 100 random bases never dropped below 0.87; raw energy and Maha-Radial are unchanged to four decimals. The instability arises under a symmetry the architecture realizes *natively*.

4.3 THE NATURAL SYMMETRY AND THE DIMENSION TREND

Appendix Figure 11(a,b,d,e) varies σ of $D = \mathrm{diag}(e^{\sigma s})$ (30 draws per σ, $\sigma \in \{0, \ldots, 1.5\}$; logits provably unchanged; full sweep: Appendix Table 10). The measurements separate the detectors exactly as Proposition 1 says they must:

(i) Global thresholds cannot be re-calibrated back. On MNIST vs Fashion-MNIST, *frozen* REACT thresholds degrade FPR_{95} by +20.1 points at σ=1.5, and *re-calibration* restores only the pooled clipping rate, leaving a monotone residual of +5.4; on CIFAR vs SVHN the residual is +9.6 (+16.2 frozen); bounding $\mathrm{cond}(D) \leq 32$ does not rescue it—already at κ=2 the re-calibrated degradation (FPR_{95} increase over the identity basis) is +0.5/+1.1 points, rising to +3.0/+4.6 at κ=32 (Appendix Table 9). ASH fares worst re-calibrated (+50.8 on MNIST far: its per-sample pruning set reshuffles under heterogeneous D).

(ii) Per-channel calibration is exact. PC-REACT's re-calibrated FPR_{95} residual at σ=1.5 is 0.0 on every task—Proposition 2 to machine precision. DICE is likewise exactly flat re-calibrated ($w_{cj}\mathbb{E}[h_j]$ is unchanged under the rescaling; Remark 3).

Table 1: **Main comparison** (mean over 3 seeds; 100 bases; p=90 re-calibrated; energy score). RS/range/WD in AUROC points ($\times$100); scale-std = AUROC std over re-calibrated rescalings at $\sigma \in \{0.5, 1.0\}$. Energy, Radial-ReAct, Maha-Radial: RS=WD=0 exactly; PC-REACT: scale-std 0 exactly, rotation columns nonzero (Props. 2, 3).

Task	Method	AUROC	FPR_{95}	AUPR	RS(100)	range	WD	scale-std
Digits (0–4 vs 5–9)	Energy	**0.96**	0.24	**0.89**	**0.0**	**0.0**	**0.0**	–
	ReAct	0.95	0.25	0.87	1.1	5.9	4.2	0.8
	PC-REACT	0.96	0.24	0.89	0.1	0.6	0.2	**0.0**
	Radial-ReAct	0.96	0.24	0.89	**0.0**	**0.0**	**0.0**	0.1
	Maha-Radial	0.96	**0.21**	0.88	**0.0**	**0.0**	**0.0**	**0.0**
MNIST (0–4 vs 5–9)	Energy	0.92	0.27	0.89	**0.0**	**0.0**	**0.0**	–
	ReAct	0.91	0.30	0.88	1.3	6.6	2.6	1.7
	PC-REACT	0.92	0.27	0.90	0.2	1.0	0.7	**0.0**
	Radial-ReAct	0.92	0.27	0.89	**0.0**	**0.0**	**0.0**	0.1
	Maha-Radial	**0.93**	**0.24**	**0.91**	**0.0**	**0.0**	**0.0**	**0.0**
MNIST vs Fashion	Energy	0.96	0.21	0.97	**0.0**	**0.0**	**0.0**	–
	ReAct	0.95	0.27	0.96	0.6	3.4	1.1	1.1
	PC-REACT	0.96	0.22	0.97	0.3	1.5	0.7	**0.0**
	Radial-ReAct	0.96	0.21	0.97	**0.0**	**0.0**	**0.0**	**0.0**
	Maha-Radial	**0.98**	**0.12**	**0.98**	**0.0**	**0.0**	**0.0**	**0.0**
CIFAR-10 vs SVHN	Energy	0.83	**0.82**	0.76	**0.0**	**0.0**	**0.0**	–
	ReAct	0.80	0.82	0.71	0.7	4.0	**0.0**	1.7
	PC-REACT	0.83	0.82	0.76	0.3	1.6	0.3	**0.0**
	Radial-ReAct	0.83	**0.82**	0.76	**0.0**	**0.0**	**0.0**	0.2
	Maha-Radial	**0.83**	0.83	**0.76**	**0.0**	**0.0**	**0.0**	**0.0**
CIFAR-10 (0–4 vs 5–9)	Energy	0.69	0.85	0.66	**0.0**	**0.0**	**0.0**	–
	ReAct	0.70	**0.85**	0.66	0.4	2.6	1.2	1.2
	PC-REACT	0.69	0.85	0.64	0.1	0.9	**0.0**	**0.0**
	Radial-ReAct	0.69	0.85	0.66	**0.0**	**0.0**	**0.0**	0.1
	Maha-Radial	**0.70**	0.85	**0.66**	**0.0**	**0.0**	**0.0**	**0.0**

(iii) The orthogonal tier does not. Radial-ReAct carries a small monotone residual (≤ 0.3 points) and ViM a large one ($+20.7$: a re-fit PCA basis does not transform with a non-orthogonal D); Maha-Radial is exactly flat on every task. The hierarchy of equation 1 is strict in both directions.

The dimension trend. Training all four task families across the full penultimate-dimension grid $d \in \{8, \ldots, 256\}$ (two seeds each), REACT's basis-sensitivity follows a *dimension-dependent trend shared across families* (Figure 3; full grid in Appendix Table 6): the across-basis AUROC range peaks at $d \leq 16$ in every family—on MNIST near-OOD, 23.3 points at d=8 vs. 2.7 at d=256—and decays with d approximately as a power law, with log–log slopes of -0.30 (CIFAR near), -0.46 (CIFAR/SVHN), -0.69 (MNIST near), and -1.09 (MNIST far) over the per-family means; the FPR_{95} swings are larger still (up to 53 points). The mechanism is dimensional concentration: few coordinates let each Haar rotation remap the pooled percentile onto heterogeneous per-channel quantiles. Compact penultimate layers—standard whenever a small projection head sits before the classifier (embedding look-up, efficiency-constrained or multi-head designs)—sit exactly in the regime where the artifact is largest. The trend is a within-regime statement: the clip rewrites only 11–30% of the logit signal on every from-scratch head (41 banks), but 52–56% on the pretrained backbones—and that high-impact regime re-opens large swings at any width (Appendix C.15).

4.4 SCALE: PRETRAINED BACKBONES AND THE IMAGENET REGIME

The scale objection—that small classifiers exaggerate the artifact—is addressed by freezing *ImageNet-pretrained* ResNet-18 (d=512) and ResNet-101 (d=2048) with a trained linear head on CIFAR-10 (three head seeds; Appendix Table 20). **(1) The effect replicates at scale**: against SVHN, REACT's across-basis AUROC range is 11–16 points—larger than on any small model—PC-REACT moves 7–11, DICE 1–5, official ASH-B 84–97; geometric detectors exactly flat. **(2) The strongest detector is the most fragile**: ASH-B is best at the identity (0.985 on ResNet-101 vs 0.91 energy) *and* most fragile (92-point swings); DICE matches on SVHN but collapses on CIFAR-100 (0.59–0.66)—task-dependent benefit, not basis coordination. **(3) The rank caveat is**

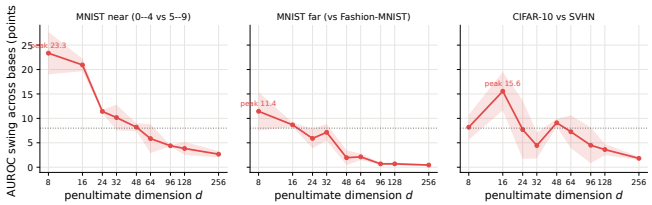


Figure 3: **The dimension trend.** REACT's across-basis AUROC range (points; 50 bases, p=90 re-calibrated; mean ± std over 2 seeds) vs. penultimate dimension d; dotted line = the 8-point strong-effect bar. The swing peaks at compact widths $d \leq 16$ in every family and decays with d (log–log slopes -0.3 to -1.1); the fourth family follows the same trend (Appendix Table 6).

visible: Maha-Radial is exactly flat over rotations, but at $n_{\text{cal}}<d$ its re-calibrated rescaling residual reaches 12–13 FPR_{95} points on ResNet-101 (exactly 0.0 at d=512; Proposition 5(iii)).

The standard large-scale regime: ImageNet OOD. REACT's documented benefit is largest with an ImageNet-1k classifier evaluated as trained: a torchvision ResNet-50 (He et al., 2016) (V1, as trained), ID = ImageNet-1k val (Russakovsky et al., 2015), against OpenOOD (Yang et al., 2022; Zhang et al., 2023) curated splits[1] of six standard OOD sets (iNaturalist, SSB-hard, NINCO, OpenImages-O, Textures (Cimpoi et al., 2014), Places365 (Zhou et al., 2018); 20 bases; Appendix Table 11).

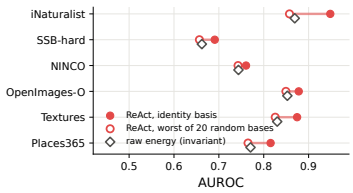


Figure 4: **Reparameterization erases REACT's ImageNet benefit.** REACT at the trained basis (filled) vs. its worst of 20 Haar bases of the *same* function (open), against invariant raw energy (diamond).

This is the regime where shaping genuinely helps—and random bases erase the help (Figure 4). On iNaturalist, REACT improves raw energy from 0.869 to 0.948 AUROC (+7.9 points), yet its worst of 20 reparameterizations of the *same* function falls to 0.857: *the entire advantage is erased*. This holds on all six sets: identity-basis gains of +1.7 to +7.9 (mean +4.0; REACT and ASH-B peak at the identity on all six, ASH-B reaching 0.954 on iNaturalist), worst-basis scores at or below the invariant raw energy. The 20-basis view is not a small-draw artifact: over 99 Haar bases the *median* REACT AUROC sits at the raw-energy level on every set, the identity basis is the best draw on all six, and $\Pr_Q[\text{REACT}(Q) < \text{energy}]$ averages 0.71 (1.00 on iNaturalist; Appendix Table 12, case studies in Appendix C.7). Radial-ReAct collapses to the energy score (mean 0.789 vs. 0.788): the gains live in coordinate structure a radial map must discard (Theorem 2); Maha-Radial is exactly flat but uneven at $n_{\text{cal}}<d$ (Appendix C.10); the rank caveat is also repairable by calibration alone: at n_{cal}=10,000 $\geq d$, exact affine covariance returns (rescaling residual 0.00, all six sets) *with higher* accuracy (0.814 vs. 0.725; Appendix C.15). In short (Appendix Table 11): the strongest identity detectors are precisely those whose verdicts move.

4.5 BENEFIT, FRAGILITY, AND THE FIXES

Does basis dependence appear only where shaping *fails*? No (Appendix Figure 10): the *downside* grows with the identity-basis gain (Spearman ρ=+0.34 across the coordinate-wise family, +0.97 for REACT alone; −0.04 for the raw swing, which counts *favorable* bases), positive-gain points carry a significantly larger worst case (p=0.007). Table 1 compares detectors in the *identity* basis and across bases. Maha-Radial matches or exceeds REACT's identity-basis AUROC in all five tasks (a tie on CIFAR near, 0.696 vs. 0.695) and raw energy in four of five; PC-REACT improves on REACT in four of five while remaining rotation-sensitive—per-channel calibration helps detection as well as equivariance. Against the geometric baselines (Appendix Table 19), Maha-Radial stays the only detector exactly flat under *both* transforms—normalized Mahalanobis (Müller & Hein, 2025) is stronger than raw but norm-based: rotation-flat with rescaling residuals up to +34 points. Even where

[1]Why not LSUN/iSUN: OpenOOD v1.5's curation removes the ID-class contamination Bitterwolf et al. (2023) documented in legacy OOD test sets; SVHN is our CIFAR-scale far-OOD set, and legacy LSUN/iSUN are omitted (scene content enters via Places365).

REACT sits *below* raw energy, an invariant variant removes the artifact entirely ($\mathrm{RS} = \mathrm{WD} = 0$). Ablations (appendix): calibration-set size, pre-ReLU hook, MSP/MaxLogit scoring, bootstrap CIs; ASH/DICE/ViM per task in Table 17; deployment guidance in Appendix Table 14.

5 RELATED WORK

Post-hoc OOD detection, two families. Logit- and confidence-based scores (Hendrycks & Gimpel, 2017; Liang et al., 2018; Liu et al., 2020; Guo et al., 2017) are basis-free by construction. Feature-based methods split into the *coordinate-wise* family—REACT (Sun et al., 2021), DICE (Sun & Li, 2022), ASH (Djurisic et al., 2023), and successors—SCALE (Xu et al., 2024) traces the benefit to scaling, DDCS (Yuan et al., 2024) selects channels by discriminability, DRL (Zhang et al., 2024) shapes features at training time, Zhao et al. (2024) give optimality theory, and ActSub (Zöngür et al., 2025) observes shaping prune different activations under function-preserving changes—a symptom of the dependence we measure, for which we supply cause (the symmetry hierarchy) and consequence (verdict flips). The *geometric* family—Mahalanobis distance (Lee et al., 2018), its feature-normalized variant (Müller & Hein, 2025), kNN (Sun et al., 2022), ViM's PCA residual (Wang et al., 2022), fDBD's decision-boundary margins (Liu & Qin, 2024)—is covariant under basis changes when its statistics are re-fit; DAVIS (Hassan et al., 2026) instead reads pre-pooling channel statistics and is coordinate-wise, so Theorem 1 applies to it. Measured on our banks (Appendix Table 19): fDBD is exactly rotation-flat; Mahalanobis++ is norm-based and pays Radial-ReAct's tier—rotation-flat, re-calibrated rescaling residual up to $+34\ \mathrm{FPR}_{95}$ points. Successors that act per-unit on h inherit Theorem 1 directly; official DICE masks W instead, and is exactly natural-symmetric yet not basis-agnostic (Remark 3). We add no new heuristic family to this race: our three formulations are existing paradigms made exactly equivariant, not new detectors (surveys: Yang et al., 2024). Beyond detection, Karasan (2026) shows parallel-analysis component counts can reflect the hidden coordinate choice rather than the model—the same function-preserving-reparameterization lens.

Reparameterization invariance and parameterization symmetries. Le Lan & Dinh (2021) prove perfect densities cannot guarantee anomaly detection, and Dauncey et al. (2024) build a representation-invariant OOD score for generative models. ReLU scaling symmetries are classical (Dinh et al., 2017); permutation symmetries complicate optimization and model merging (Simsek et al., 2021; Entezari et al., 2022; Ainsworth et al., 2023); SVCCA/CKA abstract over the basis freedom we exploit (Raghu et al., 2017; Kornblith et al., 2019). New here: the group reparameterizes a *discriminative classifier's hidden representation* while preserving the function, the measured object is *OOD decisions*, and the equivariant object is a post-hoc read-out, not an architecture (Cohen & Welling, 2016).

6 LIMITATIONS

(1) Scale: the effect replicates on frozen pretrained backbones, an untouched ResNet-50, and ViT-B/16 (Appendix C.17); full fine-tuning is untested. (2) Invariance is not sufficiency: an invariant detector can still be weak (Radial-ReAct discards coordinate-sparsity signal), equivariance costs compute (Maha-Radial's radius is $O(d^2)$ per sample against REACT's $O(d)$ clip), and the geometric tier conflates covariate shift with OOD (Maha-Radial flags nearly all corrupted ID inputs; Radial-ReAct inherits energy's robustness; Appendix C.15). (3) Under rank-deficient Σ, Maha-Radial is orthogonal-only; shrinkage (Proposition 6) turns the affine failure into a λ-bounded bias (12–13 points on ResNet-101 at $n_{\mathrm{cal}} < d$; exactly repaired at $n_{\mathrm{cal}} \geq d$, Appendix C.15). (4) The symmetry choice is a value judgment; the adversarial search is a white-box existence result.

7 CONCLUSION

Activation shaping pays for its simplicity with dependence on an object the classifier function does not define: the hidden basis. We made the dependence an explicit *choice of symmetry*, gave each tier an exactly invariant, training-free shaper, and showed the price: reparameterization alone erases REACT's entire ImageNet-scale advantage. OOD methods should state their symmetry position explicitly; open: whether natural training approaches the adversarial bases, and basis-agnostic recovery of Radial-ReAct's discarded signal.

AI USE STATEMENT

AI assistants were used for language polishing of parts of the manuscript and as a programming aid for parts of the experimental code. The research questions, theoretical results, experimental design, and analysis are the authors' own work, and all AI-assisted text and code were reviewed, tested, and approved by the authors, who take full responsibility for the content of this paper.

REFERENCES

Samuel K Ainsworth, Jonathan Hayase, and Siddhartha S Srinivasa. git re-basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations*, 2023.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Julian Bitterwolf, Maximilian Müller, and Matthias Hein. In or out? Fixing ImageNet out-of-distribution detection evaluation. In *International Conference on Machine Learning*, 2023.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning*, 2016.

Sam Dauncey, Christopher C. Holmes, Christopher K. I. Williams, and Fabian Falck. Approximations to the Fisher information metric of deep generative models for out-of-distribution detection. *Transactions on Machine Learning Research*, 2024.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, 2017.

Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. In *International Conference on Learning Representations*, 2023.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. In *International Conference on Learning Representations*, 2022.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.

Abid Hassan, Tuan Ngo, Saad Shafiq, and Nenad Medvidovic. DAVIS: OOD detection via dominant activations and variance for increased separation. *arXiv preprint arXiv:2601.22703*, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, 2015.

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

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, 2019.

Charline Le Lan and Laurent Dinh. Perfect density models cannot guarantee anomaly detection. *Entropy*, 23(12):1690, 2021.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, 2018.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

Litian Liu and Yao Qin. Fast decision boundary based out-of-distribution detector. In *International Conference on Machine Learning*, 2024.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 2020.

Prasanta Chandra Mahalanobis. On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2(1):49–55, 1936.

Francesco Mezzadri. How to generate random matrices from the classical compact groups. *Notices of the American Mathematical Society*, 54(5):592–604, 2007.

Michael Müller and Matthias Hein. Mahalanobis++: Improving OOD detection via feature normalization. In *International Conference on Machine Learning*, 2025.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, 2017.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Berfin Simsek, François Ged, Arthur Jacot, Francesco Spadaro, Clement Hongler, Wulfram Gerstner, and Johanni Brea. Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances. In *International Conference on Machine Learning*, pp. 9722–9732, 2021.

Yiyou Sun and Yixuan Li. DICE: Leveraging sparsification for out-of-distribution detection. In *European Conference on Computer Vision*, 2022.

Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems*, 2021.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, 2022.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. ViM: Out-of-distribution with virtual-logit matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4921–4930, 2022.

Kai Xu, Rongyu Chen, Gianni Franchi, and Angela Yao. Scaling for training time and post-hoc out-of-distribution detection enhancement. In *International Conference on Learning Representations*, 2024.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. OpenOOD: Benchmarking generalized out-of-distribution detection. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2022.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132:5635–5662, 2024.

Yue Yuan, Rundong He, Yicong Dong, Zhongyi Han, and Yilong Yin. Discriminability-driven channel selection for out-of-distribution detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. OpenOOD v1.5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023.

Yonggang Zhang, Jie Lu, Bo Peng, Zhen Fang, and Yiu-ming Cheung. Learning to shape in-distribution feature space for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 2024.

Qinyu Zhao, Ming Xu, Kartik Gupta, Akshay Asthana, Liang Zheng, and Stephen Gould. Towards optimal feature-shaping methods for out-of-distribution detection. In *International Conference on Learning Representations*, 2024.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018.

Barış Zöngür, Robin Hesse, and Stefan Roth. Activation subspaces for out-of-distribution detection. In *IEEE/CVF International Conference on Computer Vision*, 2025.

A PROOFS

Proof of Theorem 1. Let $a > 0$, c with $a/\sqrt{d} < c < a$, $h = ae_1$. Let $u = \frac{1}{\sqrt{d}}\mathbf{1}$ and let Q be any orthogonal map with $Qe_1 = u$ (e.g. the Householder reflection $Q = I - 2vv^\top/\|v\|^2$, $v = u - e_1$). Then $Qh = au$ has all entries $a/\sqrt{d} \in (0, c)$, so $R_c(Qh) = Qh$ and

$$Q^\top R_c(Qh) = Q^\top Qh = h, \qquad \text{while} \qquad R_c(h) = ce_1 \neq h \quad (\text{since } c < a). \tag{7}$$

Hence for any W with $Wh \neq W(ce_1)$ (e.g. generic W),

$$W\,R_c(h) = W(ce_1) \;\neq\; Wh = W\,Q^\top R_c(Qh). \tag{8}$$

The logits agree before clipping by construction. □

Proposition 4 (Genericity under re-calibration; restated from the main text)**.** *Let $d \geq 2$, $c > 0$, and fix $h \in \mathbb{R}^d$ such that clipping is active and non-degenerate: $R_c(h) \neq h$ and $R_c(h) \neq 0$. Then the set of rotations admitting* any *recovering threshold,*

$$\big\{Q \in O(d) :\ \exists\, c' \in \mathbb{R} \ \textit{with}\ \ Q^\top R_{c'}(Qh) = R_c(h)\big\}, \tag{9}$$

has Haar measure zero. In particular, for almost every Q no *threshold—the re-calibrated percentile, or any other choice, including an adversarially optimized one—recovers the original clipped features.*

The conditioning on active, non-degenerate clipping is the honest one: it holds on an open neighborhood of any h with coordinates on both sides of c (probability $1-p^d-(1-p)^d$ under iid coordinates, $\mathbb{P}(h_j \leq c) = p$). The proof below characterizes the bad set exactly—Q must map the clip pattern of h onto coordinates on which $Q\,R_c(h)$ is *constant*—and shows each pattern class is the zero set of a non-zero real-analytic function on $O(d)$.

Proof of Proposition 4. Write $g = R_c(h)$ and $v = h - g = (h - c\mathbf{1})_+$. The hypotheses give $v \neq 0$ (clipping active) and $g \neq 0$: indeed $g_j = \min(h_j, c) = 0$ forces $h_j = 0$ when $c > 0$ (if $h_j > 0$ then $\min(h_j, c) > 0$; if $h_j < 0$ then $g_j = h_j < 0$), so $g = 0$ would imply $h = 0$ and hence $v = 0$.

Suppose $Q^\top R_{c'}(Qh) = g$ for some $c' \in \mathbb{R}$, and set $u = Qh$, $w = Qg$, so $R_{c'}(u) = w$, i.e. coordinatewise $w_j = \min(u_j, c')$. Let $T := \{j : (Qv)_j \neq 0\} = \operatorname{supp}(u - w)$; $T \neq \emptyset$ because $Qv \neq 0$. For $j \in T$: $u_j \neq w_j$, so $w_j = c'$ exactly (the clipped value). For $j \notin T$: $u_j = w_j \leq c'$ (the minimum is attained by u_j). Hence *every* solution obeys, for some non-empty $T \subseteq [d]$ and some common value c',

$$(Qv)_{T^c} = 0 \qquad \text{and} \qquad (Qg)_T = c'\,\mathbf{1}_T. \tag{10}$$

(Conversely, equation 10 together with $(Qh)_{T^c} \le c'$ implies equality, but we only need the forward inclusion.) Fix $T \neq \emptyset$ and pick $j_0 \in T$; define the real-analytic map $\Phi_T : O(d) \to \mathbb{R}^{d-1}$ by

$$\Phi_T(Q) = \Big((Qv)_{T^c},\ \ (Qg)_j - (Qg)_{j_0} \text{ for } j \in T \setminus \{j_0\} \Big). \tag{11}$$

The bad set equation 9 is contained in $\bigcup_{T \neq \emptyset} \Phi_T^{-1}(0)$, a *finite* union—the auxiliary variable c' has been eliminated (it is the common value), so no uncountable union over thresholds is needed.

It remains to show each $\Phi_T^{-1}(0)$ is null. Since $\|\Phi_T\|^2$ is real-analytic on the compact analytic manifold $O(d)$, its zero set either is all of a connected component of $O(d)$ or has measure zero; we show $\Phi_T \not\equiv 0$. If $\Phi_T \equiv 0$ then $(Qg)_T \in \operatorname{span}(\mathbf{1}_T)$ for *every* $Q \in O(d)$. But $g \neq 0$ and $O(d)$ acts transitively on the sphere of radius $\|g\|$, so $\{Qg\}$ is that entire sphere; taking Q with $Qg = \|g\| e_j$ for $j \in T$ shows

$$(Qg)_T = \|g\|\, e_j \notin \operatorname{span}(\mathbf{1}_T) \qquad \text{whenever } |T| \ge 2, \tag{12}$$

and if $T = \{j_0\}$ then $\Phi_T(Q) = (Qv)_{T^c}$ alone (the g-component is empty and $\operatorname{span}(\mathbf{1}_T) = \mathbb{R}$, so only v can supply the contradiction), and $\Phi_T \equiv 0$ would force $Qv \in \operatorname{span}(e_{j_0})$ for *every* Q—impossible since $v \neq 0$ and $O(d)$, $d \ge 2$, is transitive on directions. Either way $\Phi_T \not\equiv 0$, so $\Phi_T^{-1}(0)$ has Haar measure zero, and so does the finite union. □

Remark 2. *Both hypotheses are needed. If clipping is inactive ($c > \max_j h_j$) then $R_c(h) = h$ and $Q^\top R_{c'}(Qh) = h$ for* every *Q as soon as $c' > \max_j (Qh)_j$—the proposition's conclusion is false without conditioning. The statement is careful about the conditioning event: under iid coordinates with $\mathbb{P}(h_j \le c) = p$, it has probability $1 - p^d - (1-p)^d < 1$, so no "for almost every h" claim is available, and none is claimed.*

Proof of Proposition 1. For each coordinate, since $d_j > 0$,

$$\big(R_{c'}(Dh)\big)_j = \min(d_j h_j, c') = d_j \min\big(h_j, c'/d_j\big), \tag{13}$$

$$\text{hence} \qquad \big(D^{-1} R_{c'} D\, h\big)_j = \min\big(h_j, c'/d_j\big)\ : \tag{14}$$

a per-channel clip at levels $c'_j = c'/d_j$. Two maps $h \mapsto \min(h_j, a_j)$ and $h \mapsto \min(h_j, b_j)$ on $\mathbb{R}$ are equal iff $a_j = b_j$ (evaluate at any $x > \max(a_j, b_j)$). Therefore

$$D^{-1} R_{c'} D = R_c \iff c'/d_j = c\ \ \forall j \iff d_1 = \cdots = d_d = c'/c, \tag{15}$$

and for heterogeneous D no c' exists. □

Corollary 1 (Effective per-channel quantiles)**.** *Let channel j have marginal law F_j on the ID calibration data and let the re-calibrated pooled percentile be $c'(D)$. Then channel j is clipped at effective quantile $F_j\big(c'(D)/d_j\big)$ of its own law. If the channels are exchangeable with common law F (so that in the original basis the pooled percentile c_p clips every channel at quantile $\approx p$), then after rescaling the effective quantiles $F\big(c'(D)/d_j\big)$ spread across j; for the two-channel case $d_1 = e^\sigma, d_2 = e^{-\sigma}$ the spread at the pooled-matching choice is $F\big(c'/e^{-\sigma}\big) - F\big(c'/e^{\sigma}\big) = \Theta(\sigma)$ by a first-order expansion (positive: F is non-decreasing and $c'/e^\sigma < c'/e^{-\sigma}$). Without exchangeability the original basis is already heterogeneous in quantile terms—which is exactly why Proposition 1 is phrased at the level of operators, where the statement is unconditional.*

Proof of Proposition 2. The per-channel p-percentile is positively homogeneous: order statistics scale by λ and so does every convex combination of them, so the threshold fitted on the rescaled channel satisfies

$$\hat{c}_j\big(\{d_j h_{i,j}\}_i\big) = d_j \hat{c}_j \qquad \text{and hence} \qquad \min\big(d_j h_j,\ d_j \hat{c}_j\big) = d_j \min\big(h_j, \hat{c}_j\big) \tag{16}$$

coordinatewise, which is the claim for D. For a permutation P the channels—activations, statistics, and thresholds—are permuted together, and $\min$ commutes with the relabeling. With the head transformed as $W_D = W D^{-1}$, the post-shaping logits satisfy

$$W_D\, PC(Dh) + b = W\, D^{-1} D\, PC(h) + b = W\, PC(h) + b \tag{17}$$

identically. □

Proof of Proposition 3. Equivariance applied to $Q = -I$ gives $T(0) = 0$ (only 0 is fixed by all of $O(d)$, $d \geq 2$). Let ρ be the signed permutation swapping coordinates 1 and 2. For h supported on $\{e_1, e_2\}$, equivariance $T(\rho h) = \rho T(h)$ read in coordinate 1 gives

$$\varphi_1(h_2) = T(\rho h)_1 = \big(\rho\, T(h)\big)_1 = T(h)_2 = \varphi_2(h_2) \qquad \forall h_2 \in \mathbb{R}, \tag{18}$$

so $\varphi_1 = \varphi_2$ on all of $\mathbb{R}$; all coordinates give a common $\varphi := \varphi_1 = \cdots = \varphi_d$. Equivariance under $-I$ gives φ odd, so $\varphi(0) = 0$. Now let ω rotate the $(1,2)$-plane by $45°$: $(\omega h)_1 = (h_1 - h_2)/\sqrt{2}$. For $h = (a, b, 0, \dots)$,

$$\varphi\Big(\tfrac{a-b}{\sqrt{2}}\Big) = T(\omega h)_1 = (\omega\, T(h))_1 = \tfrac{\varphi(a)-\varphi(b)}{\sqrt{2}} \qquad \forall a, b \in \mathbb{R}. \tag{19}$$

Taking $b = 0$ (and $\varphi(0) = 0$) yields the scale recursion

$$\varphi\big(a/\sqrt{2}\big) = \varphi(a)/\sqrt{2}; \tag{20}$$

multiplying the general identity by $\sqrt{2}$ and using this homogeneity gives $\varphi(a - b) = \varphi(a) - \varphi(b)$, and with $b \mapsto -b$ and oddness,

$$\varphi(a + b) = \varphi(a) + \varphi(b) : \tag{21}$$

φ is additive and continuous, hence $\varphi(x) = kx$. So $T(h)_j = kh_j$: a uniform rescaling, which clips nothing. □

Proof of Theorem 2. ($\Leftarrow$) If $T(0) = 0$ and $T(h) = \alpha(\|h\|)h$ with α as stated, equivariance is immediate ($\|Qh\| = \|h\|$), and T is continuous away from 0; at 0,

$$\|T(h) - T(0)\| = \|h\|\,\big|\alpha(\|h\|)\big| \to 0 \tag{22}$$

by the limit condition. ($\Rightarrow$) $T(0) = 0$ as above. Fix $h \neq 0$ and let $G_h = \{Q \in O(d) : Qh = h\}$, which acts on $h^\perp$ as the full $O(d-1)$ ($d \geq 2$). For $Q \in G_h$, $T(h) = T(Qh) = QT(h)$, and the fixed subspace of that action is $\mathrm{span}(h)$, so

$$T(h) = \alpha(h)\, h, \qquad \alpha(h) = \frac{\langle T(h), h\rangle}{\|h\|^2}, \tag{23}$$

with α continuous on $\mathbb{R}^d \setminus \{0\}$. If $\|h'\| = \|h\|$, pick Q with $h' = Qh$; then

$$\alpha(h')\, h' = T(h') = T(Qh) = Q\, T(h) = \alpha(h)\, h', \tag{24}$$

so $\alpha(h') = \alpha(h)$: α is a function of $r = \|h\|$ alone, continuous on $(0, \infty)$. Finally

$$\|T(h)\| = r\,\big|\alpha(r)\big| \to \|T(0)\| = 0 \quad \text{as } r \downarrow 0 \tag{25}$$

by continuity of T, which is the stated limit condition. (For $d = 1$, $O(1) = \{\pm 1\}$ and equivariance reduces to oddness: $T(h) = \alpha(|h|)h$ with $\alpha \in C(0, \infty)$, $r\alpha(r) \to 0$.) □

Proposition 5 (Pseudo-inverse conjugation)**.** *Let $\Sigma \succeq 0$. **(i)** For every $Q \in O(d)$: $\mathrm{pinv}(Q\Sigma Q^\top) = Q\,\mathrm{pinv}(\Sigma)\,Q^\top$. **(ii)** If $\Sigma \succ 0$ and A is invertible: $(A\Sigma A^\top)^{-1} = A^{-\top}\Sigma^{-1}A^{-1}$. **(iii)** For singular Σ, exact affine covariance fails* even when the transformed covariance preserves the eigenspace exactly*: take $\Sigma = \mathrm{diag}(\frac{3}{2}, 0)$ and the shear $A = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ (which acts within the null space of Σ), so $A\Sigma A^\top = \Sigma$ identically; yet $A^{-\top}\Sigma^+A^{-1} = \frac{2}{3}\left(\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right) \neq \Sigma^+ = \mathrm{diag}(\frac{2}{3}, 0)$, so the radius $\sqrt{(Ah)^\top\Sigma^+(Ah)}$ does not reduce to $\sqrt{h^\top\Sigma^+h}$. Consequently, the fixed-cutoff eigendecomposition estimator used by Maha-Radial is exactly orthogonal-equivariant for every $\Sigma \succeq 0$; for rank-deficient Σ the guarantee degrades* strictly *to the orthogonal one—neither rank stability nor eigenspace preservation rescues affine covariance—and exact affine equivariance requires $\Sigma \succ 0$ (part **(ii)**).*

Proof. (i) Let $\Sigma = U\Lambda U^\top$ be an orthogonal eigendecomposition with $\Lambda \succeq 0$ diagonal. Then $Q\Sigma Q^\top = (QU)\Lambda(QU)^\top$ with QU orthogonal, so this is an orthogonal eigendecomposition of the rotated matrix; applying pinv (invert the strictly positive eigenvalues, keep the eigenvectors) gives

$$\mathrm{pinv}\big(Q\Sigma Q^\top\big) = (QU)\,\Lambda^+\,(QU)^\top = Q\,\mathrm{pinv}(\Sigma)\,Q^\top. \tag{26}$$

The eigenvalue multiset is unchanged, so a relative cutoff selects the same eigenvalues in both bases. (ii) Direct verification:

$$\big(A\Sigma A^\top\big)\big(A^{-\top}\Sigma^{-1}A^{-1}\big) = A\,\Sigma\,A^\top A^{-\top}\,\Sigma^{-1}\,A^{-1} = I, \tag{27}$$

and likewise on the left. (iii) is a direct computation: $A\Sigma A^\top = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right) = uu^\top$ with $u = (1, 1)^\top$, so $\mathrm{pinv}(A\Sigma A^\top) = uu^\top/\|u\|^4 = \frac{1}{4}\left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, whereas $A^{-\top}\Sigma^+A^{-1} = \mathrm{diag}(1, 0)$. (The failure in (iii) is why the caveat in §3.3 restricts the affine claim to stably-ranked Σ.) □

Proposition 6 (Shrinkage covariance: what survives at $\lambda > 0$)**.** *Let $\hat{\Sigma} \succeq 0$ and $\Sigma_\lambda = (1-\lambda)\hat{\Sigma} + \lambda I$ with $\lambda \in (0, 1]$. **(i)** Orthogonal equivariance is exact for every λ: $Q\,\Sigma_\lambda(\hat{\Sigma})\,Q^\top = \Sigma_\lambda(Q\hat{\Sigma}Q^\top)$ for all $Q \in O(d)$. **(ii)** $\Sigma_\lambda \succ 0$, so the inverse exists and Proposition 5(ii) applies within the shrunk family; affine covariance of the* shrunk *estimator fails by exactly $\lambda(I - AA^\top)$: $\Sigma_\lambda(A\hat{\Sigma}A^\top) - A\,\Sigma_\lambda(\hat{\Sigma})\,A^\top = \lambda(I - AA^\top)$. **(iii)** Under a positive diagonal $D = \mathrm{diag}(d_j)$ with $\min_j d_j \le 1 \le \max_j d_j$ (which for the zero-centered log-scalings $d_j = e^{\sigma s_j}$ holds with probability exactly $1 - 2^{1-d}$, since it fails only if all s_j share a sign), the mismatch is the diagonal matrix $\lambda(I - DD^\top)$, of operator norm $\lambda \max_j |1 - d_j^2| \le \lambda\big(\mathrm{cond}(D)^2 - 1\big)$.*

Proof. (i) Using $QQ^\top = I$,

$$Q\,\Sigma_\lambda(\hat{\Sigma})\,Q^\top = (1-\lambda)\,Q\hat{\Sigma}Q^\top + \lambda\,QQ^\top = (1-\lambda)\,Q\hat{\Sigma}Q^\top + \lambda I = \Sigma_\lambda\big(Q\hat{\Sigma}Q^\top\big). \tag{28}$$

(ii) Every eigenvalue of Σ_λ is at least $\lambda > 0$, so $\Sigma_\lambda \succ 0$ and the cutoff never truncates. For the identity, expand both sides:

$$\Sigma_\lambda\big(A\hat{\Sigma}A^\top\big) = (1-\lambda)\,A\hat{\Sigma}A^\top + \lambda I, \tag{29}$$

$$A\,\Sigma_\lambda(\hat{\Sigma})\,A^\top = (1-\lambda)\,A\hat{\Sigma}A^\top + \lambda\,AA^\top, \tag{30}$$

and subtract. (iii) Set $A = D$. Under the straddle assumption $\max_j |1-d_j^2| = \max\big(1-d_{\min}^2,\, d_{\max}^2 - 1\big)$. Since $d_{\min} \le 1$, $d_{\max}^2 - 1 \le d_{\max}^2/d_{\min}^2 - 1$; and $1 - d_{\min}^2 \le d_{\max}^2/d_{\min}^2 - 1$ because $d_{\max}^2/d_{\min}^2 + d_{\min}^2 \ge 1/d_{\min}^2 + d_{\min}^2 \ge 2$ (the latter by AM–GM, using $d_{\max} \ge 1$). Hence

$$\big\|I - DD^\top\big\|_{\mathrm{op}} = \max_j \big|1 - d_j^2\big| \le \mathrm{cond}(D)^2 - 1. \tag{31}$$

□

B Implementation details

B.1 Protocol, in full

Every experiment: (1) train the classifier; (2) extract ID calibration / ID test / OOD test features once, in float64; (3) for each of 100 sampled A (including the identity) transform features and head, *re-fit* every method's threshold on the transformed ID calibration set, and record metrics; (4) verify $\max_x \|W_A h_A(x) - Wh(x)\|_\infty < 10^{-13}$ (else abort). Step (4) is the key sanity check: all comparisons are between provably identical classifier functions.

B.2 Normalization layers and the natural group

Which parts of G_{nat} survive depends on the normalization in the architecture. (i) *Plain ReLU*: all of G_{nat}, per hidden layer, exactly. (ii) *BatchNorm* (Ioffe & Szegedy, 2015) with frozen inference statistics: a positive diagonal $D = \mathrm{diag}(d_j)$ acting on a hidden channel is absorbed by the normalization's own parameters—rescale the running mean to $d_j\mu_j$, the affine gain to γ_j/d_j (the variance buffer is untouched), and the frozen-statistics output is bit-identical—so G_{nat} remains an exact symmetry of the function family, realized through BN's parameters instead of the weights; permutations act by relabeling $(\gamma, \beta, \mu, \sigma)$. (iii) *LayerNorm* (Ba et al., 2016): the answer depends on the hook. For *pre-LN* features, normalization across channels commutes with permutations ($\mathrm{LN}(Px) = P\,\mathrm{LN}(x)$) but not with heterogeneous per-channel diagonals—only the global scale survives ($\mathrm{LN}(\alpha x) = \mathrm{LN}(x)$)—so the natural group reduces to permutations $\oplus$ global scaling. For *post-LN* features $h = \gamma \odot \hat{x} + \beta$, a positive diagonal D acting on h is absorbed into the block's own learnable parameters, $\gamma' = D\gamma$, $\beta' = D\beta$ (with the head rescaled $W \to WD^{-1}$): exactly as for BatchNorm, the full G_{nat} is a native symmetry of the same model class. The ViT hook of Appendix C.17 is post-LN, so its rescaling measurements are exact-symmetry tests. The natural-tier results (Propositions 1, 2) apply verbatim to plain-ReLU and BatchNorm networks, and their rescaling half weakens for LayerNorm networks; the orthogonal tier is an external design requirement in all three cases.

B.3 Models and training

Table 2 summarizes the architectures (one representative per input family; small-scale models trained from scratch with 3 seeds). The dimension trend of §4.3 retrains each family from scratch at every width $d \in \{8, 16, 24, 32, 48, 64, 96, 128, 256\}$ (two seeds)—the grid covers genuinely retrained classifiers, not truncated features.

Table 2: Architectures and training recipes. GAP = global average pooling.

Setting	Architecture	Training	ID acc.
Digits (0–4 vs 5–9)	MLP $64 \to 64 \to d{=}32$, ReLU	Adam 10^{-3}, batch 128, $\leq$ 50 epochs, early stop (patience 8)	98.5%
MNIST (far & near)	conv($1 \to 16$, $3{\times}3$)–ReLU–pool–conv($16 \to 32$, $3{\times}3$)–ReLU–pool–linear($1568 \to d{=}64$)–ReLU–head	AdamW 10^{-3}, batch 256, 10 epochs	98.9–99.7%
CIFAR-10 (far & near)	ResNet-18-lite: stem $3 \to 16$; stages $[16, 32, 64, 128]{\times}2$ BasicBlocks, strides $1, 2, 2, 2$; GAP $\to d{=}128$	SGD 0.05, momentum 0.9, wd $5{\cdot}10^{-4}$, batch 256, 30 epochs, pad-4 crop + hflip	80.8–84.8%
Pretrained heads (§4.4)	frozen ImageNet ResNet-18 ($d{=}512$) / ResNet-101 ($d{=}2048$) + linear head on CIFAR-10	head only; 3 head seeds	—
ImageNet (§4.4)	torchvision ResNet-50 IMAGENET1K_V1 ($d{=}2048$), original 1000-way head	none (used as trained)	76.1% top-1

ImageNet-scale data (§4.4). ID = ImageNet-1k val (Russakovsky et al., 2015) through a shorter-side-256 JPEG parquet mirror (top-1 75.8% on the 12k subsample), split 2,000 calibration + 10,000 eval; OOD = OpenOOD (Yang et al., 2022; Zhang et al., 2023) curated splits, listed with sizes in Table 3; oversize sets are deterministically subsampled with seed 0; decoding is the standard Resize-256 / CenterCrop-224 / ImageNet normalization. We anchor to the V1 checkpoint because the 2022 IMAGENET1K_V2 behaves anomalously for the raw energy score at $T{=}1$ on iNaturalist (0.50 AUROC: the log-sum-exp concentration penalty cancels the max-logit gap; 0.77 at $T{=}0.25$)—V1 is the checkpoint family of the original ReAct evaluation.

Table 3: ImageNet-scale OOD benchmark: evaluation sets and sizes. "Eval" = images scored (deterministic seed-0 subsampling from "Full set").

Set	Role	Images	
		Eval	Full set
ImageNet-1k val	ID	10,000 (+2,000 cal)	50k (12k mirror)
iNaturalist	OOD	10,000	10,000
SSB-hard	OOD	10,000	$\approx$ 49k
NINCO	OOD	5,878	5,878
OpenImages-O	OOD	10,000	17,632
Textures	OOD	5,640	5,640
Places365	OOD	10,000	36,500

B.4 Detector implementations

All methods read penultimate activations extracted once in float64; every fitted quantity is re-estimated on the transformed ID calibration split (size $\min(1000, |\text{test}|/3)$, carved from ID test, never trained on) after each reparameterization. Table 4 gives the exact rules.

Remark 3 (DICE's symmetry tier)**.** *Official DICE is exactly equivariant under the* natural *group when re-fit: under $h \to hD$, $W \to WD^{-1}$ (positive diagonal D) the contribution matrix is unchanged entrywise, $V'_{cj} = (w_{cj}/d_j)\,(d_j\,\mathbb{E}[h_j]) = V_{cj}$, so mask, masked head, and score are all preserved—verified numerically to 5×10^{-15} (its re-calibrated rescaling residuals are exactly 0, like PC-ReAct's). It is* not *basis-agnostic: under a rotation the ranking object $w_{cj}\mathbb{E}[h_j]$ is a product of two different linear combinations of the old coordinates, and the selected subset changes*

Table 4: Detector implementations. h = penultimate activation, d = its dimension, p=90 unless stated; ε is a fixed small constant.

Detector	Rule	Fitted on ID calibration	Source
REACT	$R_c(h)_j = \min(h_j, c)$, c = p-th percentile of all pooled activations (official protocol; positive-only variant gives the same conclusions)	1 scalar	(Sun et al., 2021)
PC-REACT	$\min(h_j, \hat{c}_j)$, $\hat{c}_j$ = same percentile per channel	d scalars	this work
ASH-B	per sample (Djurisic et al., 2023, Alg. 2): t = p-th percentile of the sample's activations, $s = \sum_j h_j$ *before* pruning; $h_j < t \to 0$; the n survivors are all set to s/n	— (stateless, on-the-fly)	(Djurisic et al., 2023)
DICE	mask the head: $V_{cj} = w_{cj}\,\mathbb{E}_{\mathrm{ID}}[h_j]$; keep the k=$\lfloor 0.3\,Cd \rfloor$ largest entries of V in M; score = energy of $(M \odot W)h + b$	mask M, re-fit per transform	(Sun & Li, 2022)
ViM	$o = W^{\dagger}b$ (the source's Eq. (1) origin; cancels the bias exactly when $b \in \mathrm{col}(W^{\top})$); P/R = top-D/trailing eigenvectors of $(H-o)^{\top}(H-o)$, D=1000 if d>1500 else $d/2$; score = $\log \sum_c e^{l_c} - \alpha \lVert R^{\top}(h-o) \rVert$, α = Eq. (6) of the source	(o, P, α), re-fit per transform	(Wang et al., 2022)
Radial-ReAct	$T_\tau(h) = h\,\min\!\big(1, \frac{\tau}{\lVert h \rVert_2 + \varepsilon}\big)$, τ = p-th percentile of $\lVert h \rVert$	1 scalar	this work
Maha-Radial	$\mu + \min\!\big(1, \frac{\tau}{r(h)+\varepsilon}\big)(h-\mu)$, $r(h)$ = Mahalanobis radius (fixed-cutoff inverse)	(μ, Σ, τ)	this work

wholesale—on CIFAR-10/SVHN the mask overlap between identity and rotated bases is 0.08*, the chance level of two independent* 30% *masks. DICE therefore sits in the same tier as PC-*REACT *(natural-symmetric, re-calibrable) rather than the geometric tier, consistent with its large rotation swing in Table 17.*

Scores and reparameterizations. Energy $-T \log \sum_k e^{z_k/T}$ with T=1, reported so that higher = ID; MSP and MaxLogit in the ablations. Haar rotations Q via QR with sign correction (Mezzadri, 2007); rescalings $D = \mathrm{diag}(e^{\sigma s})$ with $s \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, 30 draws per $\sigma \in \{0, 0.25, 0.5, 0.75, 1, 1.25, 1.5\}$.

B.5 COMPUTE AND EXACTNESS BOOKKEEPING

Compute. Small-scale experiments (including pretrained-backbone extraction) run in $<$ 3 GPU-hours on one RTX 4090 with $<$ 600 MB of data; the ImageNet-scale benchmark adds $\approx$ 10 minutes of extraction across four RTX 4090s plus $\approx$ 3 CPU-hours of sweeps, with $\approx$ 8 GB of labeled, size-verified downloads.

Exactness bookkeeping. Logit preservation max error 3.0×10^{-14} (rotations) / 1.4×10^{-13} (rescalings). Raw-energy and Radial-ReAct across-rotation AUROC std $\leq 2\times10^{-16}$ everywhere; Maha-Radial std $\leq 3\times10^{-16}$ with the fixed-cutoff estimator (before that fix, naive `inv` produced $O(10^{-3})$ spurious variation on rank-deficient banks—the pitfall of §3.3).

The covariance inverse caveat, in full. Naively calling inv is quietly wrong under rotation: Σ of post-ReLU features is typically singular, and after rotation the singularity becomes numerical noise on which inv returns garbage ($O(1)$ radius errors in our check) rather than raising. The fixed-cutoff eigendecomposition (relative cutoff $10^{-12}\lambda_{\max}$) is deterministic under orthogonal conjugation. We flag this because it would corrupt any naive re-implementation of Mahalanobis-style rules under reparameterization.

C ADDITIONAL RESULTS

C.1 THE SYMMETRY HIERARCHY: FULL TABLE

Same ReLU architecture • group $G_{\text{nat}} = \{PD\}$: per-unit positive rescaling, unit relabeling • realized *natively* by the model class (positive homogeneity)
A detector that is not G_{nat}-equivariant is not well defined on the model class itself: respect it, or explain the violation. → **PC-REACT**

stronger demand

Same classifier function • group: any invertible A (a change-of-basis linear map before the head) • a *design* requirement, not an architecture fact
Basis-agnosticism: the detector should be a function of the classifier function alone. → **Radial-ReAct**, **Maha-Radial**

Figure 5: **Two levels of function preservation.** *Top:* the natural group of the architecture, under which the reparameterized network is the same model, not an inserted one—violations here (e.g. REACT's global threshold, Proposition 1) are deployment-facing. *Bottom:* arbitrary invertible basis changes preserve only the classifier *function*; invariance to them is a design desideratum, and the rotation experiments (§4.2) show coordinate-wise detectors are not even well defined at this level.

Table 5 backs the taxonomy summary of §4.2: worst-case (over 5 tasks × 3 seeds; 100 bases each) AUROC range across rotations, and worst-case re-calibrated FPR_{95} residual at σ=1.5 (30 draws). Coordinate-wise detectors are rotation-sensitive (PC-REACT least); only per-channel calibration is exactly flat under rescaling; geometric detectors (ViM, kNN, Radial-ReAct) are exactly flat under rotations but not under non-orthogonal rescaling; only Maha-Radial is flat under both. Appendix Tables 20 and 17 use the same protocol.

Table 5: **The symmetry hierarchy is strict.** Worst-case (over 5 tasks × 3 seeds; 100 bases each) AUROC range across rotations, and worst-case re-calibrated FPR_{95} residual at σ=1.5 across tasks (30 draws). Coordinate-wise detectors are rotation-sensitive (PC-REACT least, Propositions 2–3); only per-channel calibration is exactly flat under rescaling; geometric detectors (ViM, kNN, Radial-ReAct) are exactly flat under rotations but ViM/kNN are not under non-orthogonal rescaling; only Maha-Radial is flat under both. (kNN's AUROC is itself weak on our small models—below chance on Digits—but its invariance behavior is what the taxonomy predicts.)

		rot. range (pts)		
Family	Detector	max	mean	scal. residual (pts)
coordinate-wise	REACT (global p-percentile)	9.4	4.5	9.6
	PC-REACT (per-channel)	2.6	1.1	**0.0**
	DICE (per-channel bias)	29.1	16.7	**0.0**
	ASH (top-k prune)	74.8	49.2	50.8
affine	ViM (re-fit PCA residual)	**0.0**	**0.0**	20.7
	kNN (k=50, Euclidean)	**0.0**	**0.0**	34.0
	Radial-ReAct	**0.0**	**0.0**	0.3
	Maha-Radial	**0.0**	**0.0**	**0.0**

C.2 THE DIMENSION TREND: FULL TABLE

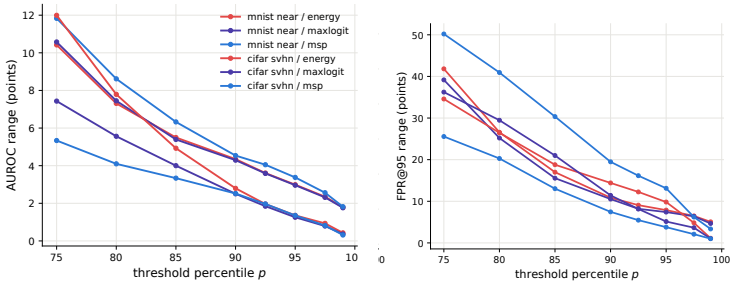
Table 6 gives the per-dimension entries behind the dimension trend of §4.3 and Figure 11(c,f): REACT AUROC-range / FPR_{95}-range (points, 50 bases, p=90 re-calibrated) across the full penultimate-dimension grid. Marks, per column: **bold** = largest swing, underlined = second largest, green = smallest (most stable); cell shading encodes magnitude against the strong-effect bars of Figure 11(c,f) (amber ≥ 8 points, red ≥ 15).

Table 6: **Dimension law.** REACT AUROC-range / FPR_{95}-range (points, 50 bases, p=90 re-calibrated) vs penultimate dimension. Marks, per column: **bold** = largest swing, underlined = second largest, green = smallest (most stable). Cell shading encodes magnitude against the strong-effect bars of Figure 11(c,f): amber ≥ 8 points, red ≥ 15.

	MNIST near		MNIST far		CIFAR vs SVHN		CIFAR near	
d	AUROC	FPR_{95}	AUROC	FPR_{95}	AUROC	FPR_{95}	AUROC	FPR_{95}
8	**23.3**	37.4	**11.4**	33.9	8.2	10.4	**5.2**	2.3
16	20.9	**53.0**	8.7	**48.4**	**15.6**	12.5	4.9	1.4
24	11.4	30.6	5.9	47.3	7.7	**13.2**	4.4	**2.4**
32	10.2	31.2	7.1	41.4	4.4	11.6	3.5	2.1
48	8.2	21.2	1.9	7.4	9.1	4.7	2.7	1.6
64	5.9	14.1	2.1	13.1	7.3	5.7	3.1	1.2
96	4.4	13.6	0.7	4.3	4.5	4.2	2.6	0.8
128	3.8	7.7	0.7	4.6	3.6	8.6	2.7	1.4
256	2.7	6.0	0.4	3.2	1.8	4.1	2.1	1.0

C.3 THE PERCENTILE TREND: FIGURE AND TABLE

Figure 6 and Table 7 give the per-percentile swings behind the percentile trend of §4.2 (REACT, 50/100 bases, energy score, seed 0); the MSP and MaxLogit score variants follow the same monotone trend.



(a) AUROC range. (b) FPR_{95} range.

Figure 6: **The percentile trend.** REACT's AUROC range (a) and FPR_{95} range (b) across 100 bases, as a function of the clipping percentile p (p=90 is REACT's standard setting). Monotone in p for every task/score combination; the more of the activation distribution the clipper modifies, the more the hidden basis leaks into the OOD decision.

Table 7: **Percentile law:** REACT AUROC range across 50 bases (points) vs threshold percentile p (energy score, seed 0).

Task	p=75	80	85	90	92.5	95	97.5	99
MNIST near	10.4	7.3	5.5	4.3	3.6	3.0	2.4	**1.8**
CIFAR vs SVHN	12.0	7.8	4.9	2.8	2.0	1.4	0.9	**0.4**

C.4 CALIBRATION-SET SIZE AND BOUNDED CONDITION NUMBERS

Table 8 varies the ID calibration-set size over a $50\times$ range ($n_{\mathrm{cal}} \in \{100, \ldots, 5000\}$; calibration drawn from the front of the pooled ID features, evaluation always on the last 5,000 held-out ID rows, disjoint from every draw; 30 bases, 3 seeds): REACT's across-basis AUROC range is flat in n_{cal} in every task—the swing is a property of the coordinate system, not threshold noise from a small calibration set. Table 9 repeats the rescaling stress test with *bounded* condition numbers ($\mathrm{cond}(D) = \kappa$ exactly by affine log-renormalization, $\kappa \in \{2, \ldots, 32\}$): even at κ=2 the re-calibrated FPR_{95} residual is already $+0.5$/$+1.1$ points (MNIST near / CIFAR vs SVHN) and grows monotonically with

κ; frozen thresholds degrade faster ($+12.0$ points at κ=32, MNIST near). Deployment-plausible rescalings already carry the artifact.

Table 8: **Not a calibration artifact.** REACT's across-basis AUROC range (points; 30 bases, p=90 re-calibrated, mean over 3 seeds) vs. ID calibration-set size: flat over a $50\times$ range in every task.

Task	n_{cal}=100	n_{cal}=250	n_{cal}=500	n_{cal}=1000	n_{cal}=2000	n_{cal}=5000
CIFAR-10 (0–4 vs 5–9)	2.3	2.3	2.3	2.3	2.3	2.3
CIFAR-10 vs SVHN	4.3	4.1	4.1	4.1	4.1	4.0
Digits (0–4 vs 5–9)	3.2	3.1	3.1	3.1	3.1	3.1
MNIST vs Fashion	2.3	2.4	2.5	2.5	2.5	2.5
MNIST (0–4 vs 5–9)	5.9	5.8	5.7	5.7	5.6	5.7

Table 9: **Bounded condition numbers.** REACT FPR_{95} residual vs. the identity basis (points; mean over 30 draws $\times$ 3 seeds) under D with $\text{cond}(D) = \kappa$ exactly. Re-calibration shrinks but never removes the residual; frozen thresholds degrade monotonically worse.

Task	thresholds	κ=2	κ=4	κ=8	κ=16	κ=32
MNIST (0–4 vs 5–9)	re-calibrated	0.5	1.2	1.9	2.5	3.0
MNIST (0–4 vs 5–9)	frozen	0.9	2.8	5.5	8.7	12.0
CIFAR-10 vs SVHN	re-calibrated	1.1	2.7	3.8	4.3	4.6
CIFAR-10 vs SVHN	frozen	1.5	4.0	6.0	7.5	8.4

C.5 POSITIVE RESCALING: FULL SWEEP TABLE

Table 10 gives the full σ-sweep behind Figure 11(a,b,d,e) (REACT, p=90; $\pm$ is std over 30 draws of D; cond(D) is the mean condition number). Blue rows: "recal." re-fits the threshold on rescaled ID activations; red rows: "frozen" keeps the original-basis threshold.

Table 10: **Positive rescaling on MNIST vs Fashion-MNIST** (REACT, p=90; $\pm$ is std over 30 draws of D). cond(D) is the mean condition number. Blue rows: "recal." re-fits the threshold on rescaled ID activations; red rows: "frozen" keeps the original-basis threshold.

Threshold	σ	cond(D)	AUROC	FPR_{95}
recal.	0	1.0	0.97±0.00	0.17±0.00
recal.	0.25	3.1	0.97±0.00	0.17±0.02
recal.	0.5	10.3	0.97±0.01	0.19±0.04
recal.	0.75	38.4	0.97±0.01	0.19±0.04
recal.	1	129.8	0.96±0.01	0.21±0.06
recal.	1.25	587.7	0.96±0.01	0.22±0.05
recal.	1.5	1290.5	0.96±0.01	0.22±0.07
frozen	0	1.0	0.97±0.00	0.17±0.00
frozen	0.25	3.1	0.97±0.00	0.18±0.02
frozen	0.5	10.3	0.96±0.01	0.23±0.08
frozen	0.75	38.4	0.95±0.01	0.28±0.08
frozen	1	129.8	0.93±0.03	0.34±0.11
frozen	1.25	587.7	0.94±0.02	0.33±0.11
frozen	1.5	1290.5	0.93±0.03	0.37±0.11

C.6 THE IMAGENET-SCALE BENCHMARK: FULL TABLE

Table 11 reports the complete per-set results behind §4.4 (untouched ResNet-50 V1, the classifier family REACT was designed for; 2,000 ID calibration, 10,000 ID eval, OpenOOD curated splits; 20 bases): identity-basis AUROC, across-basis ranges, and re-calibrated rescaling residuals for all eight detectors, including official ASH-B's identity gains (mean 0.837, best of all eight; 0.954 on

iNaturalist) set against its 20–77-point across-basis collapses (worst 0.19), PC-REACT's +2.7 mean, and ViM's flat-but-sample-starved behavior summarized there.

Table 11: **ImageNet-scale OOD benchmark** (untouched ResNet-50 V1, the classifier family REACT was designed for; 2,000 ID calibration, 10,000 ID eval, OpenOOD curated splits; 20 bases). AUROC_I at the identity basis (bold = best per set); green cells are exact zeros (bit-identical outputs across bases); rot.rng / rot.max are the mean / worst-set AUROC range across the 20 bases, and scale is the worst re-calibrated FPR_{95} residual at σ=1.5. Shaded rows are exactly $O(d)$- or affine-invariant detectors.

	iNat	SSB-h	NINCO	OpenO	Text.	Places	across 6 sets			
Method	AUROC_I	AUROC_I	AUROC_I	AUROC_I	AUROC_I	AUROC_I	mean	rot.rng	rot.max	$\text{scale}_{\sigma=1.5}$
Energy	0.869	0.662	0.744	0.853	0.830	0.771	0.788	**0.0**	**0.0**	**0.0**
ReAct	0.948	**0.691**	0.761	0.878	0.874	**0.815**	0.828	4.5	9.1	28.7
PC-REACT	0.929	0.633	0.755	0.887	0.920	0.764	0.815	3.1	6.7	**0.0**
ASH	**0.954**	0.657	**0.777**	**0.902**	0.942	0.793	**0.837**	46.5	76.6	70.2
DICE	0.917	0.649	0.703	0.841	0.872	0.734	0.786	5.6	12.5	**0.0**
ViM	0.825	0.540	0.746	0.819	**0.964**	0.624	0.753	**0.0**	**0.0**	1.3
Radial-ReAct	0.869	0.662	0.744	0.854	0.833	0.771	0.789	**0.0**	**0.0**	5.1
Maha-Radial	0.921	0.662	0.724	0.736	0.825	0.482	0.725	**0.0**	**0.0**	7.2

Table 12: 99 **Haar bases, REACT at** p=90 **re-calibrated** (ImageNet ResNet-50). Identity AUROC (I), median over bases, 5/95% quantiles, worst draw, and $\Pr_Q[\text{AUROC}_{\text{REACT}}(Q) < \text{AUROC}_{\text{energy}}]$. The median basis sits at the raw-energy level on every set: the identity basis is the best draw, not a typical one.

OOD set	I	median	q_{05}	q_{95}	worst	$\Pr[\text{ReAct}<E]$
iNat	0.948	0.863	0.858	0.867	0.853	1.00
SSB-h	0.691	0.660	0.657	0.663	0.655	0.90
NINCO	0.761	0.745	0.742	0.748	0.741	0.35
OpenO	0.878	0.851	0.849	0.853	0.846	0.78
Text.	0.874	0.828	0.826	0.830	0.825	0.91
Places	0.815	0.772	0.768	0.774	0.765	0.30
mean	0.828	0.786	0.783	0.789	0.781	0.71

C.7 CASE STUDIES: FROM BASIS TO VERDICT

The aggregate statistics of §4.2–§4.4 compress 10,000 images into one number per transform; Figures 7 and 8 unpack the same measurement at the level of individual images, on the ImageNet bank of §4.4 (untouched ResNet-50, OOD = iNaturalist). Figure 7 selects, by REACT's own behavior, eight images with clear-cut trained-basis verdicts—four OOD images *caught* at the trained basis (identity percentile ≈ 0) but *missed* under all 19 random bases, and four ID images *accepted* at the trained basis (percentile 8–15%) but *false-alarmed* under all 19—and traces each image's score percentile among ID test scores across the same 20 bases for seven detectors (per-basis re-calibration throughout; the plotted decision rule is the paper's standard 5% threshold). The rows separate exactly as the hierarchy predicts: REACT flips on every image, PC-REACT moves less, ASH-B scatters over nearly the full percentile range (its verdict on the *same* image ranges from the 1st to the 99th percentile depending on the basis alone), while ViM, Maha-Radial, and raw energy hold one value to machine precision. Figure 8 follows one caught OOD image through REACT's pipeline at the trained basis and at the basis that flips it: the re-calibrated threshold c lands on similar activation profiles yet clips a different fraction of channels (6.5% vs. 7.7%), the shaped logits of the largest classes move by multiple units, and the energy score crosses the decision threshold—the same function, a different verdict. Figure 9 shows the dependence one level further up, at the shaped classifier's *prediction*: the raw logits are basis-invariant (top-1 = "daisy" at all 20 bases), REACT's shaped top-1 spans 2 classes across the bases, PC-REACT 2, and official ASH-B's spans 8 ("vacuum", "potpie", "earthstar"...)—the hidden basis does not just move the safety score, it corrupts the semantics of the shaped read-out.

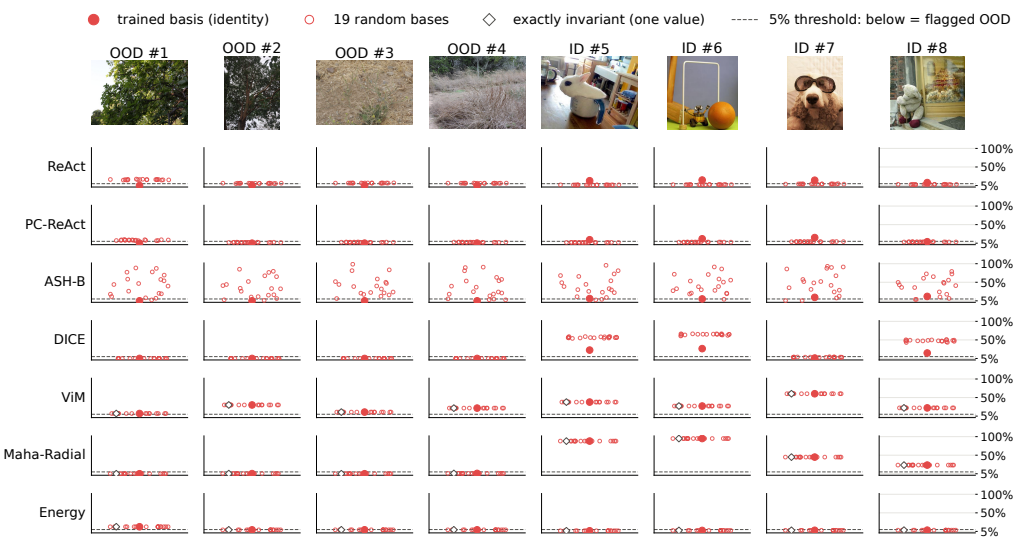


Figure 7: **Verdict-flip matrix.** Four OOD (caught at the trained basis, missed under all 19 random bases by REACT) and four ID images (accepted at the trained basis, false-alarmed under all 19), selected on the ImageNet/iNaturalist bank by REACT's behavior. Each cell: the image's score percentile among the 10,000 ID test scores across the 20 bases of the paper's sweep (filled = trained basis, open = random bases; re-calibrated per basis); dashed line = the 5% decision threshold, below which the image is flagged as OOD. Coordinate-wise rows flip and scatter (ASH-B across nearly the whole range); invariant rows (ViM, Maha-Radial, energy) hold one value.

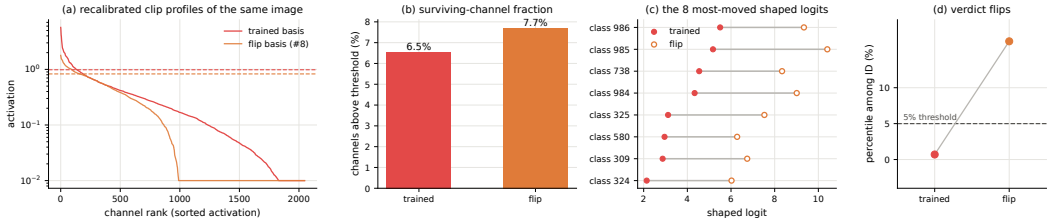


Figure 8: **Mechanism trace for one image.** (a) Sorted penultimate activations of the same OOD input at the trained basis and at the basis that flips its verdict, with the re-calibrated REACT thresholds c (dashed; log scale). (b) Fraction of channels surviving the clip. (c) The eight shaped logits that move the most between the two bases. (d) The resulting score percentile among ID images: the same classifier function, a different verdict.

C.8 THE ADVERSARIAL BASIS SEARCH

The adversarial basis search of §4.2 optimizes the rotation directly: $Q = \mathrm{Cayley}(\mathrm{Skew}(\Omega))$ stays orthogonal by construction, the re-calibrated clip-logit-energy pipeline is piecewise-smooth in Ω, and Adam minimizes a pairwise logistic surrogate of AUROC. The attack runs under a *transfer protocol*: it optimizes on one half of the test sets and reports AUROC on the held-out half it never saw. Per task, REACT falls from 0.951 to 0.297 (Digits), 0.952 to 0.025 (MNIST near), 0.971 to 0.003 (MNIST far), 0.871 to 0.028 (CIFAR/SVHN)—inversion, not degradation; at the same adversarial bases, raw energy and Maha-Radial are unchanged to four decimals (as their invariance demands) and PC-REACT degrades only mildly (0.83 on CIFAR/SVHN at its worst).

C.9 BENEFIT AND FRAGILITY: FIGURE

Figure 10 backs §4.5: per-point scatter of identity-basis gain against AUROC swing (a) and worst-case degradation (b), all settings and methods, with the shared legend strip.

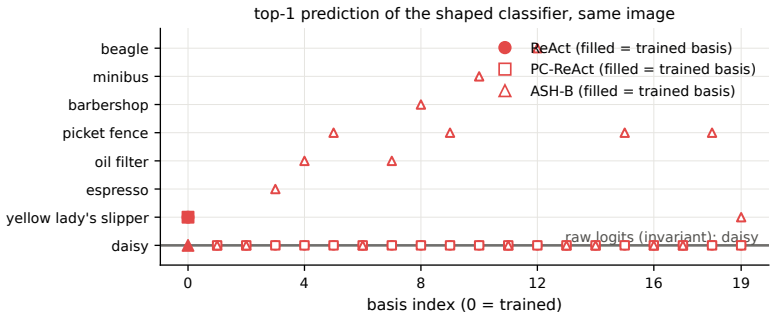
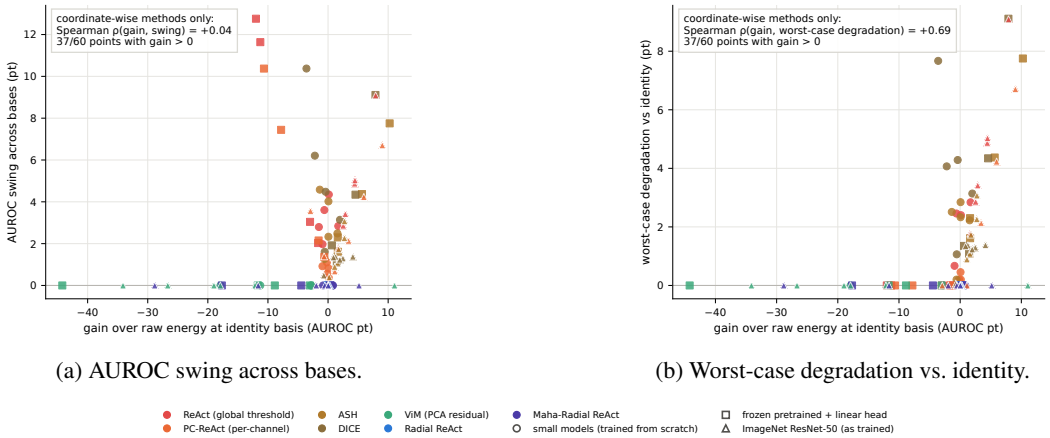


Figure 9: **The basis corrupts the shaped prediction, not just the score.** Top-1 class of the shaped classifier for the OOD image of Figure 8, across the same 20 bases (filled = trained basis). Raw logits give one class at every basis; REACT's shaped top-1 spans 2 classes, PC-REACT 2, and ASH-B's spans 8 of 20.



(a) AUROC swing across bases.

(b) Worst-case degradation vs. identity.

Figure 10: **Benefit vs. basis sensitivity, all settings and methods.** Each point is a (configuration, OOD set, method); x = gain over raw energy at the identity basis (AUROC points), y = AUROC swing across random bases (a) / worst-case degradation vs. identity (b). Markers: circles = small models trained from scratch; squares = frozen pretrained backbone + linear head; triangles = untouched ImageNet ResNet-50. In (b), coordinate-wise methods with positive gain dominate the upper region—effectiveness and downside fragility co-occur (Spearman ρ=+0.34 across the coordinate-wise family, +0.97 for REACT alone; −0.04 for the raw swing in (a), which also counts favorable bases)—while the exactly-invariant detectors (radial, Maha-Radial, ViM) lie on the y=0 axis at every gain level.

C.10 THE SHRINKAGE DIAL FOR MAHA-RADIAL AT $n_{\text{cal}} < d$

Table 13 sweeps $\Sigma_\lambda = (1-\lambda)\hat{\Sigma} + \lambda I$ on the rank-deficient regimes: the untouched ImageNet ResNet-50 (d=2048; $n_{\text{cal}} \in \{250, 1000, 2000\}$ with Textures and NINCO OOD) plus the frozen ResNet-101/SVHN bank of §4.4—seven banks in total. Three facts, matching Proposition 6 exactly. **(i) Orthogonal equivariance is free.** The across-rotation AUROC range is 0.000 points for every bank at every λ—even at rank 249/2048. **(ii) The affine residual stays small.** Across all 49 (bank, λ) conditions, the σ=1 rescaling range never exceeds 3 points (worst 2.95, at the sphere limit λ=1). **(iii) Accuracy is a dial with a genuine optimum—and the trap is a tiny-but-nonzero λ.** At λ=10^{-4} the shrinkage inflates near-null directions and can be catastrophic (0.166 AUROC on ResNet-101/SVHN, below chance), whereas the rank-truncated λ=0 estimator is protected by its cutoff; at moderate λ the estimator recovers and often *improves* sharply: Textures at n_{cal}=2000 goes from 0.825 (λ=0) to 0.970 (λ=10^{-3})—+14.5 points, above every detector in Table 11—and the sphere limit λ=1 is safe everywhere (its ResNet-101/SVHN score, 0.907, is that bank's best). Practical guidance: keep λ=0 with the fixed cutoff when exact affine covariance is required and the rank is stable; otherwise select $\lambda \in [10^{-3}, 10^{-1}]$ (where a Ledoit–Wolf choice falls), which keeps

exact orthogonal equivariance, bounds the affine bias by Proposition 6(iii), and is often the most accurate configuration in its own right.

Table 13: **The shrinkage dial** for Maha-Radial when $n_{\text{cal}} < d$ (d=2048 throughout; identity-basis AUROC on four representative banks of the seven measured; "worst" columns are maxima over all seven). Rotation equivariance is exact at every λ (Proposition 6(i)); the rescaling residual stays ≤ 3 points; accuracy has an interior optimum, and tiny nonzero λ is the failure mode to avoid.

		identity-basis AUROC$_I$				worst over all 7 banks	
λ	rank	Textures n=2000	Textures n=1000	NINCO n=2000	R101/SVHN n=1000	rot. range	scale range
0	$n_{\text{cal}}-1$	0.825	0.960	0.724	0.864	**0.00**	1.23
0.0001	2048	0.957	0.700	0.771	0.166	**0.00**	1.09
0.001	2048	0.970	0.929	0.780	0.523	**0.00**	2.29
0.01	2048	0.960	0.969	0.782	0.828	**0.00**	1.08
0.1	2048	0.918	0.947	0.757	0.887	**0.00**	2.18
0.3	2048	0.885	0.910	0.750	0.899	**0.00**	1.99
1	2048	0.843	0.843	0.746	0.907	**0.00**	2.95

C.11 Practical recommendations

Table 14 summarizes which invariant shaper to deploy given the symmetry requirement; all are training-free and one line of code.

Table 14: **Practical recommendations.** Which invariant shaper to deploy, given the symmetry requirement. All are training-free and one line of code.

If you require	Use	What you get
Only the architecture's own symmetry G_{nat} (plain-ReLU or BatchNorm net; coordinate semantics meaningful)	PC-ReAct	exact under permutations $\oplus$ positive diagonals; still rotation-sensitive (Proposition 3)
Basis-agnosticism (the detector should be a function of the classifier function)	Radial-ReAct	exact under $O(d)$; small residual under non-orthogonal D
Affine geometry, with covariance stably estimated ($n_{\text{cal}} \gtrsim d$, or shrinkage $\lambda > 0$)	Maha-Radial (Σ_λ)	exact affine when λ=0, $\hat{\Sigma} \succ 0$; exact $O(d)$ with λ-bounded affine bias otherwise (Proposition 6)

C.12 Bootstrap confidence intervals

Table 15 reports the complete per-seed across-basis statistics for ReAct (100 bases per classifier seed, all five tasks). Every one of the fifteen 95% bootstrap CIs on the across-basis std excludes zero by a wide margin (e.g. $[1.42, 1.84]$ points on MNIST near-OOD, seed 1): the rotation sensitivity is a property of the classifier function, not sampling noise. The bold entry in each task block marks the most sensitive seed.

C.13 Pre- vs post-ReLU penultimate features, and where the hook sits

Table 16 compares the basis-sensitivity when features are read before rather than after the final ReLU. The effect *grows* on pre-ReLU features (e.g. 4.9 vs 3.8 AUROC-range points on MNIST near-OOD), ruling out non-negativity of the activations as the cause; Radial-ReAct is exactly flat in both readings (shaded rows).

Before or after pooling? For the linear read-outs used by every architecture in this paper (global average pooling into the head), the distinction is vacuous for our transform families: average pooling is a fixed linear map applied per spatial position, so it commutes with any orthogonal Q or diagonal

Table 15: Per-seed across-basis statistics (REACT, p=90, 100 bases each; RS = std across bases in points; CI from 1000 bootstrap resamples; bold = most sensitive seed per task).

Task	seed	mean	RS (pts)	CI (RS)	5–95% (pts)
Digits (0–4 vs 5–9)	0	0.96	0.7	[0.6, 0.9]	[94.8, 97.1]
	1	0.96	1.0	[0.8, 1.1]	[93.6, 97.0]
	2	0.94	**1.5**	[1.2, 1.8]	[91.0, 95.7]
MNIST vs Fashion	0	0.98	0.3	[0.3, 0.4]	[97.2, 98.2]
	1	0.97	0.7	[0.6, 0.8]	[96.0, 98.1]
	2	0.94	**0.8**	[0.7, 1.0]	[92.4, 94.8]
MNIST (0–4 vs 5–9)	0	0.95	1.0	[0.9, 1.1]	[93.2, 96.5]
	1	0.90	**1.6**	[1.4, 1.8]	[87.5, 93.2]
	2	0.91	1.1	[0.9, 1.4]	[89.1, 92.6]
CIFAR-10 vs SVHN	0	0.88	0.6	[0.5, 0.6]	[87.7, 89.5]
	1	0.83	0.6	[0.5, 0.8]	[82.2, 83.8]
	2	0.77	**0.9**	[0.8, 1.1]	[75.6, 78.3]
CIFAR-10 (0–4 vs 5–9)	0	0.69	**0.5**	[0.4, 0.5]	[68.5, 70.0]
	1	0.71	0.4	[0.3, 0.5]	[70.2, 71.5]
	2	0.69	0.4	[0.4, 0.5]	[68.2, 69.5]

D acting on channels ($\mathrm{avg}_i\{Qh_i\} = Q\,\mathrm{avg}_i\{h_i\}$, and likewise for D). Rotating the pre-pool feature map and rotating the pooled representation generate the same family of reparameterizations; the substantive choice of hook is therefore before vs. after the nonlinearity—which Table 16 varies—and before vs. after a normalization layer, where Remark 1 explains which symmetries survive.

Table 16: Identity-basis AUROC and across-basis AUROC range (points) for post- vs pre-ReLU penultimates; bold marks the larger of the two ranges. The sensitivity is not an artifact of non-negativity.

Task	Method	post AUROC	post range	pre AUROC	pre range
MNIST near	ReAct	0.95	**3.8**	0.92	4.9
MNIST near	Radial-ReAct	0.95	0.0	0.92	0.0
MNIST far	ReAct	0.98	**1.7**	0.96	3.0
MNIST far	Radial-ReAct	0.98	0.0	0.96	0.0

C.14 THE COORDINATE-WISE FAMILY AND THE GEOMETRIC CONTROL

Dropping per-sample-ness (allowing T or the score to see the whole calibration set) escapes Theorem 2; the natural members are geometric—ViM's PCA-residual score and Mahalanobis-type distances, covariant by construction (a re-fit PCA basis rotates with the features). Table 17 extends the rotation and rescaling measurements to ASH and DICE (coordinate-wise, like REACT) and ViM (geometric). The family claim is direct: ASH and DICE inherit the basis sensitivity—ASH-B most severely, up to 75 AUROC-range points on the small models (77 at ImageNet scale), DICE up to 29—while ViM with re-fit PCA is exactly flat over rotations; bold marks the best identity-basis AUROC per task, green cells mark exact-zero rescaling residuals (official DICE's are exactly zero: Remark 3).

C.15 GEOMETRIC BASELINES, THE CLIP-IMPACT AXIS, AND THE FULL-RANK REPAIR

Geometric baselines. Table 19 adds the standard class-conditional Mahalanobis score (Lee et al., 2018) (shared covariance, relative-cutoff eigendecomposition inverse, re-fit per transform), its ℓ_2-normalized variant (Müller & Hein, 2025), and fDBD (Liu & Qin, 2024) (minimum feature-space margin distance $(z_c - z_j)/\|w_c - w_j\|$ over classes j). All three are exactly flat over rotations (0.0 points; fDBD's normalizer is orthogonal-invariant by construction). The raw Mahalanobis score's re-calibrated rescaling residual is $\leq$ 2.2 points on our banks (empirically even at $n_{\mathrm{cal}}<d$,

Table 17: ASH, DICE (coordinate-wise / weight-masking) and ViM (geometric) per task: identity-basis AUROC (mean over seeds), AUROC range across 100 bases (worst seed), and re-calibrated FPR_{95} residual at σ=1.5 (worst task). ASH/DICE behave like REACT (sensitive, ASH-B most by far); ViM is exactly flat over rotations but not under rescaling.

Task	Method	AUROC (identity)	rot. range (pts)	scal. residual (pts)
Digits (0–4 vs 5–9)	ASH	0.59	48.5	13.8
Digits (0–4 vs 5–9)	DICE	**0.88**	26.9	**0.0**
Digits (0–4 vs 5–9)	ViM	0.82	0.0	9.0
MNIST (0–4 vs 5–9)	ASH	0.72	54.7	9.9
MNIST (0–4 vs 5–9)	DICE	0.82	26.1	**0.0**
MNIST (0–4 vs 5–9)	ViM	**0.92**	0.0	13.6
MNIST vs Fashion	ASH	0.89	51.8	50.8
MNIST vs Fashion	DICE	0.91	13.1	**0.0**
MNIST vs Fashion	ViM	**0.95**	0.0	20.7
CIFAR-10 vs SVHN	ASH	**0.90**	70.8	40.9
CIFAR-10 vs SVHN	DICE	0.87	11.0	**0.0**
CIFAR-10 vs SVHN	ViM	0.79	0.0	3.6
CIFAR-10 (0–4 vs 5–9)	ASH	0.65	20.5	1.2
CIFAR-10 (0–4 vs 5–9)	DICE	0.68	6.2	**0.0**
CIFAR-10 (0–4 vs 5–9)	ViM	**0.69**	0.0	0.5

where Proposition 5(iii) grants no guarantee—the class-mean differences happen to lie in the retained eigenspace); the normalized variant is norm-based and inherits the orthogonal tier's failure: up to $+33.7$ re-calibrated FPR_{95} points, exactly the tier Radial-ReAct occupies. Accuracy tracks per-class calibration mass: 2,000 ImageNet calibration samples give ~ 2 per class (the normalized variant averages 0.733 over the six sets, strong on Textures at 0.967 and weak on Places at 0.533), while the 100-per-class pretrained heads reach 0.976 (ResNet-101/SVHN). **The clip-impact axis.** Define $I = \mathbb{E}_x \left\|(h - R_c(h))W^\top\right\|_2 / \left\|hW^\top\right\|_2$: the share of the logit signal the pooled-percentile clip rewrites at the trained basis. Across our 41 from-scratch banks $I \in [0.11, 0.30]$ at every width; the frozen ImageNet-pretrained backbones sit at $I = 0.52$–0.56 and the untouched ResNet-50 at $I = 0.41$—a clean separation with no overlap. Within the bounded-impact regime the across-basis swing tracks d (the concentration mechanism of §4.3); the high-impact pretrained regime re-opens large swings at d=512 and d=2048, reconciling §4.3 with §4.4. **Full-rank repair.** With n_{cal}=10,000 $\geq d$ on the untouched ResNet-50, Maha-Radial's re-calibrated rescaling residual is 0.00 FPR_{95} points on all six OOD sets (ten draws at σ=1.5), confirming Proposition 5(ii) end-to-end, and mean AUROC *rises* from 0.725 (n_{cal}=2,000, rank-deficient) to 0.814: the rank caveat is a small-calibration artifact, not a property of the estimator.

Covariate-shift control (ImageNet-C-lite). We corrupt the first 2,000 ID-eval images (gaussian noise σ=25, blur r=2, pixelate $\times\frac{1}{4}$; severity $\approx$ 3), re-extract the ResNet-50 features, and report each detector's false-OOD rate at the clean 5%-TPR threshold (Table 18). The energy family treats corruptions as ID: raw energy flags 9.7–22.7%, REACT stays within 0.1–4.6 points of energy, and Radial-ReAct matches energy *exactly* (it reduces to energy at this scale); ASH-B is the worst shaper (28.1% on blur). The covariance/subspace detectors, however, flag essentially every corrupted image (100% for Maha-Radial and ViM): their ID-manifold estimate is tight enough that covariate shift itself reads as anomalous—an honest cost of the geometric tier that basis invariance does not mitigate, and a deployment consideration orthogonal to the symmetry hierarchy (corrupted inputs are *different inputs*, not a reparameterization).

C.16 PRETRAINED BACKBONES: FULL GRID

Table 20 gives the complete pretrained-backbone results behind §4.4: both backbones (ResNet-18, d=512; ResNet-101, d=2048), both OOD sets (SVHN, CIFAR-100), all eight detectors, identity-basis AUROC plus rotation range and recalibrated rescaling residual. Shaded rows are the exactly rotation-invariant detectors; green cells mark exact zeros; bold marks the best AUROC per task block.

Table 18: **Covariate-shift control:** false-OOD rate (%) on corrupted ID images at the clean 5%-TPR threshold ($2{,}000$ images, severity ≈ 3). Energy-family shapers track raw energy; the geometric tier flags nearly everything.

Detector	noise	blur	pixelate
Energy	9.7	22.7	13.7
REACT	9.8	22.0	18.3
PC-REACT	12.7	17.5	12.0
ASH-B	10.6	28.1	21.6
Radial-ReAct	9.7	22.7	13.7
Maha-Radial	100.0	100.0	100.0
ViM	100.0	100.0	100.0

Table 19: **Geometric baselines** (identity AUROC by regime; rotation range and worst re-calibrated FPR_{95} rescaling residual). All are exactly rotation-flat; the normalized variant's norm-based score pays the orthogonal tier's rescaling residual.

Detector	identity AUROC			rot. range	rescale res.
	small	pretrained	ImageNet	(worst)	(worst, re-cal.)
Mahalanobis (Lee et al., 2018)	0.700	0.587	0.607	0.0	2.2
Mahalanobis++ (Müller & Hein, 2025)	0.899	0.846	0.733	0.0	33.7
fDBD (Liu & Qin, 2024)	0.872	–	0.785	0.0	–

C.17 A LAYERNORM BACKBONE: VIT-B/16

Table 21 repeats the ImageNet benchmark of §4.4 on torchvision's ViT-B/16 (IMAGENET1K_V1, $d{=}768$ post-LN representation, $2{,}000$ calibration / $10{,}000$ ID eval / OpenOOD splits; 20 bases, $p{=}90$): the coordinate-wise family's basis dependence transfers to a transformer backbone, and the geometric detectors stay exactly rotation-flat, exactly as on ResNets. The hook here is the *post-LN* representation, for which the full natural group is an exact native symmetry (γ, β absorb any positive diagonal; Appendix B.2), so the rescaling measurements are exact-symmetry tests, not worst-case ones. Two notes: ASH-B's per-sample binarization is a poor detector on LN features (identity 0.518) *and* swings 27 points; and Radial-ReAct's rescaling residual of 12.8 points (vs. ≤ 0.3 on the ReLU models at the same protocol) is the hierarchy's separation made visible—a purely orthogonal-equivariant shaper need not track the natural group, and on this backbone it does not—while Maha-Radial remains exactly flat under both transforms.

C.18 OTHER SETTINGS

Figures 12–20 give the remaining per-task rotation panels (both metrics, all five tasks) and scaling curves, one task per figure at full width; percentile tables under MSP/MaxLogit scoring follow the same trends; the full per-run CSVs and generating scripts are in the supplementary repository.

D ADDITIONAL FIGURES

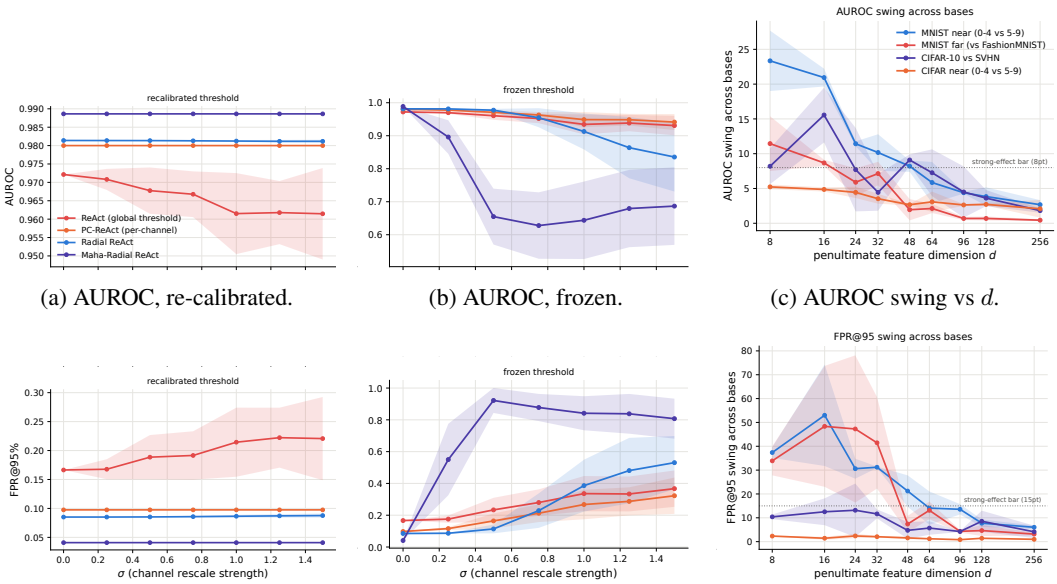
Checklist of claims → where each is established. Bit-identical logit preservation under absorbed reparameterizations: §4.1; non-invariance construction and genericity: Theorem 1 and Proposition 4; rotation/rescaling sweeps (all methods, all tasks): Table 1, appendix Tables 10, 5, and Figures 2 and 11; percentile, calibration-size, bounded-rescaling, and adversarial-basis ablations: §4.2 and appendix Figure 6, Tables 7, 8, 9; dimension trend and pre-/post-ReLU hook: §4.3 and appendix Tables 6, 16; per-image verdict flips and mechanism trace: §4.4 and appendix Figures 7, 8; pretrained backbones and the ImageNet-scale benchmark: appendix Tables 20, 11; shrinkage dial: Table 13.

Table 20: Frozen ImageNet-pretrained backbones with a linear head on CIFAR-10 (§4.4); same protocol as Table 5. Note Maha-Radial's exact rotation flatness everywhere, with the rescaling residual appearing exactly when $n_{\text{cal}} < d$ (ResNet-101), as Proposition 5(iii) predicts.

Task	Method	AUROC (identity)	rot. range (pts)	scal. residual (pts)
ResNet-101 / CIFAR-100	ASH	**0.89**	53.0	23.8
ResNet-101 / CIFAR-100	DICE	0.65	7.2	**0.0**
ResNet-101 / CIFAR-100	Maha-Radial	0.88	**0.0**	12.1
ResNet-101 / CIFAR-100	Energy	0.87	**0.0**	**0.0**
ResNet-101 / CIFAR-100	PC-REACT	0.86	2.2	**0.0**
ResNet-101 / CIFAR-100	Radial-ReAct	0.87	**0.0**	0.6
ResNet-101 / CIFAR-100	ReAct	0.84	4.0	5.9
ResNet-101 / CIFAR-100	ViM	0.76	**0.0**	1.0
ResNet-101 / SVHN	ASH	**0.98**	97.1	60.6
ResNet-101 / SVHN	DICE	0.98	2.3	**0.0**
ResNet-101 / SVHN	Maha-Radial	0.87	**0.0**	15.2
ResNet-101 / SVHN	Energy	0.91	**0.0**	**0.0**
ResNet-101 / SVHN	PC-REACT	0.83	7.6	**0.0**
ResNet-101 / SVHN	Radial-ReAct	0.91	**0.0**	0.5
ResNet-101 / SVHN	ReAct	0.80	11.6	8.2
ResNet-101 / SVHN	ViM	0.75	**0.0**	1.6
ResNet-18 / CIFAR-100	ASH	0.81	41.4	15.4
ResNet-18 / CIFAR-100	DICE	0.65	6.7	**0.0**
ResNet-18 / CIFAR-100	Maha-Radial	0.82	**0.0**	**0.0**
ResNet-18 / CIFAR-100	Energy	0.82	**0.0**	**0.0**
ResNet-18 / CIFAR-100	PC-REACT	0.82	1.6	**0.0**
ResNet-18 / CIFAR-100	Radial-ReAct	0.82	**0.0**	0.2
ResNet-18 / CIFAR-100	ReAct	0.80	3.6	2.4
ResNet-18 / CIFAR-100	ViM	**0.83**	**0.0**	0.9
ResNet-18 / SVHN	ASH	0.95	87.7	57.1
ResNet-18 / SVHN	DICE	**0.98**	5.1	**0.0**
ResNet-18 / SVHN	Maha-Radial	0.66	**0.0**	**0.0**
ResNet-18 / SVHN	Energy	0.82	**0.0**	**0.0**
ResNet-18 / SVHN	PC-REACT	0.72	11.0	**0.0**
ResNet-18 / SVHN	Radial-ReAct	0.82	**0.0**	0.1
ResNet-18 / SVHN	ReAct	0.69	15.8	1.3
ResNet-18 / SVHN	ViM	0.67	**0.0**	0.4

Table 21: ViT-B/16, six OpenOOD sets, 20 bases: identity AUROC (mean over sets), rotation AUROC-range (mean / worst set), worst re-calibrated FPR_{95} residual at σ=1.5.

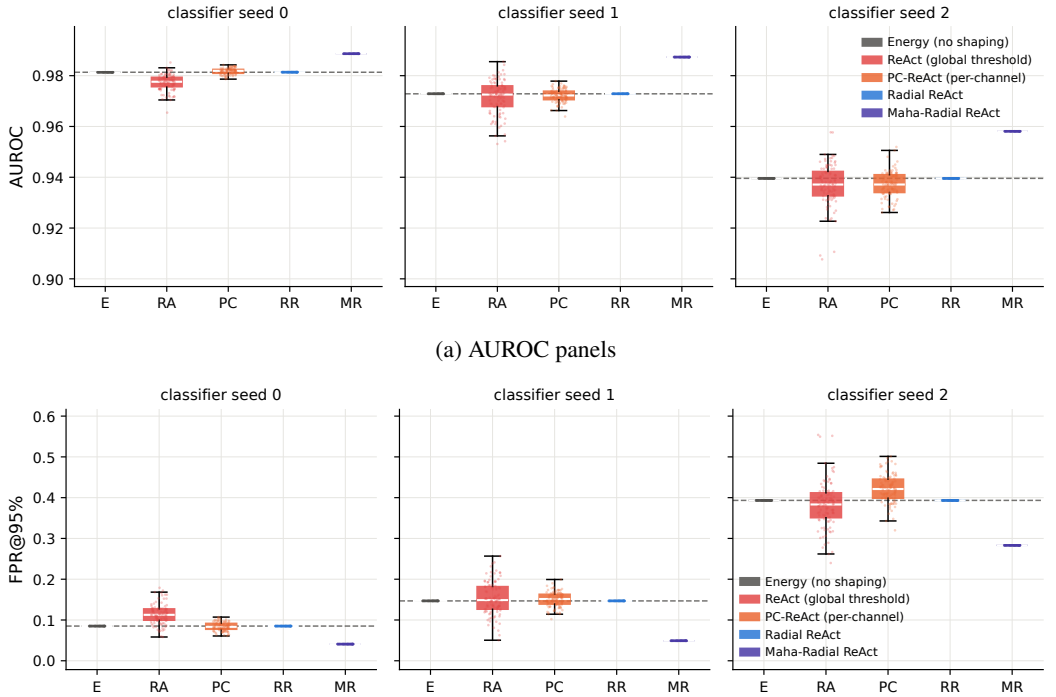
Detector	I	rot. range (mean)	(worst)	rescale residual
Energy	0.693	0.0	0.0	–
ReAct	0.773	3.8	4.8	3.4
PC-REACT	0.767	3.6	4.9	0.0
ASH	0.518	17.0	26.9	1.8
DICE	0.681	9.2	15.3	0.0
ViM	0.815	0.0	0.0	6.2
Radial-ReAct	0.700	0.0	0.0	12.8
Maha-Radial	0.805	0.0	0.0	0.0



(a) AUROC, re-calibrated. (b) AUROC, frozen. (c) AUROC swing vs d.

(d) FPR_{95}, re-calibrated. (e) FPR_{95}, frozen. (f) FPR_{95} swing vs d.

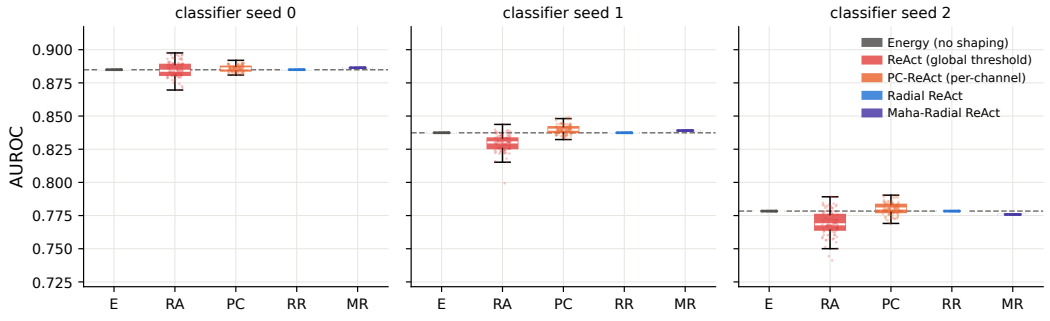
Figure 11: **Rescaling and dimension sweeps** (top row AUROC, bottom row FPR_{95}). (a,b,d,e): positive rescaling $D{=}\mathrm{diag}(e^{\sigma s})$ on MNIST far-OOD (30 draws per σ), re-calibrated (a,d) / frozen (b,e) thresholds. (c,f): REACT's swings across 50 bases vs penultimate dimension (2 seeds; dotted: strong-effect bars 8/15 points). PC-REACT is exactly flat under re-calibration (Proposition 2); global REACT carries monotone residuals (Proposition 1); smaller heads, larger swings.



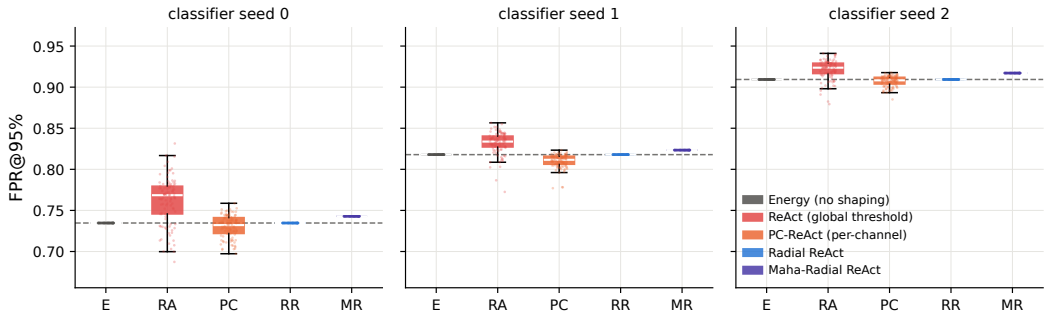
(a) AUROC panels

(b) FPR_{95} panels

Figure 12: **Rotation sensitivity, MNIST vs Fashion-MNIST** — same protocol as Figure 2: 100 hidden bases per classifier seed, $p{=}90$ re-calibrated, energy score.

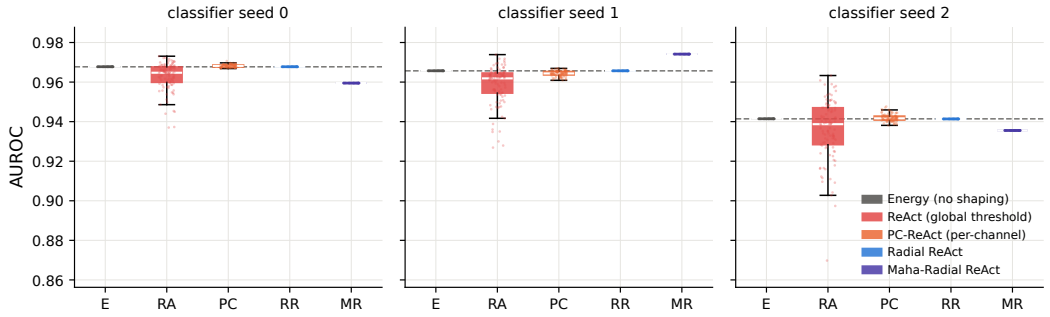


(a) AUROC panels

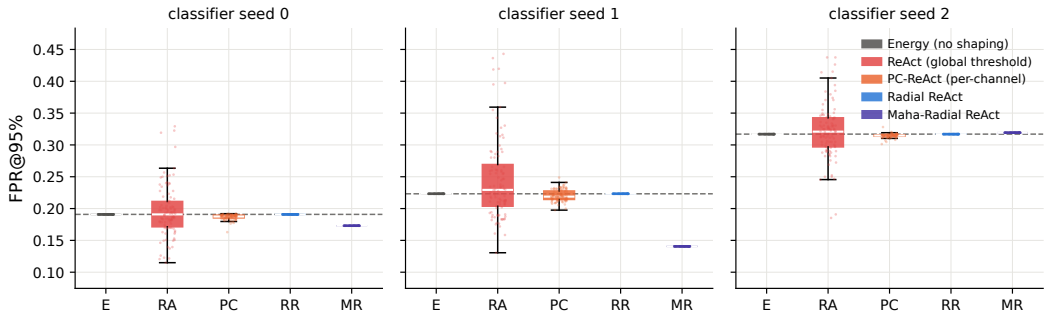


(b) FPR_{95} panels

Figure 13: **Rotation sensitivity, CIFAR-10 vs SVHN** — same protocol as Figure 2: 100 hidden bases per classifier seed, p=90 re-calibrated, energy score.

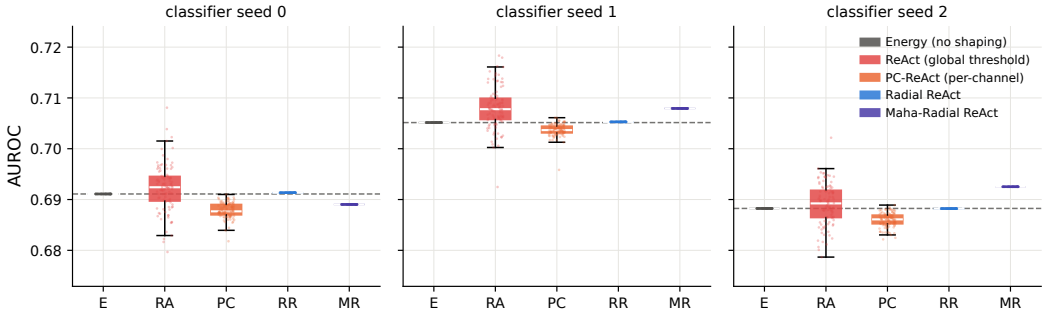


(a) AUROC panels

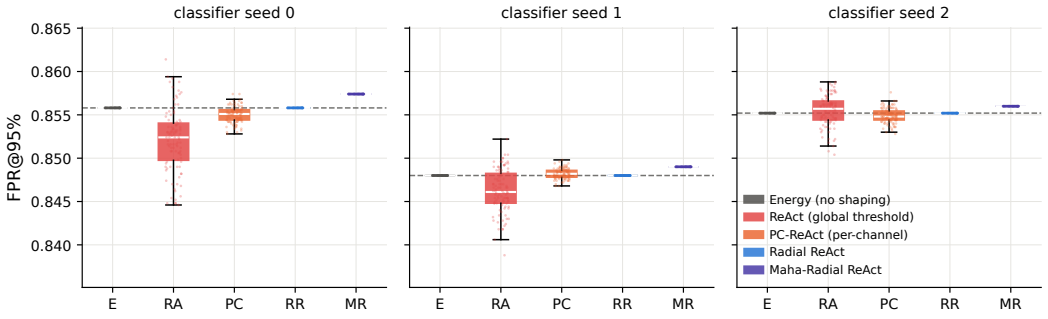


(b) FPR_{95} panels

Figure 14: **Rotation sensitivity, Digits (0–4 vs 5–9)** — same protocol as Figure 2: 100 hidden bases per classifier seed, p=90 re-calibrated, energy score.



(a) AUROC panels



(b) FPR_{95} panels

Figure 15: **Rotation sensitivity, CIFAR-10 (0–4 vs 5–9)** — same protocol as Figure 2: 100 hidden bases per classifier seed, p=90 re-calibrated, energy score.

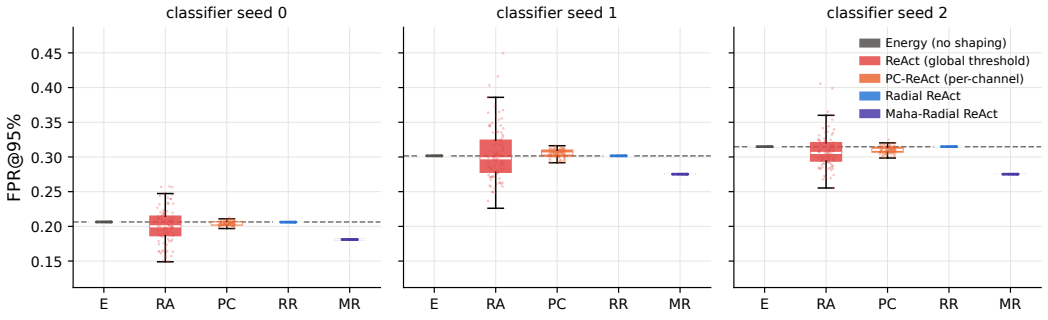


Figure 16: **Rotation sensitivity in FPR_{95} for the main-text task** (MNIST near-OOD; AUROC panels in Figure 2). The FPR_{95} swings are larger than the AUROC swings (up to 31.4 points).

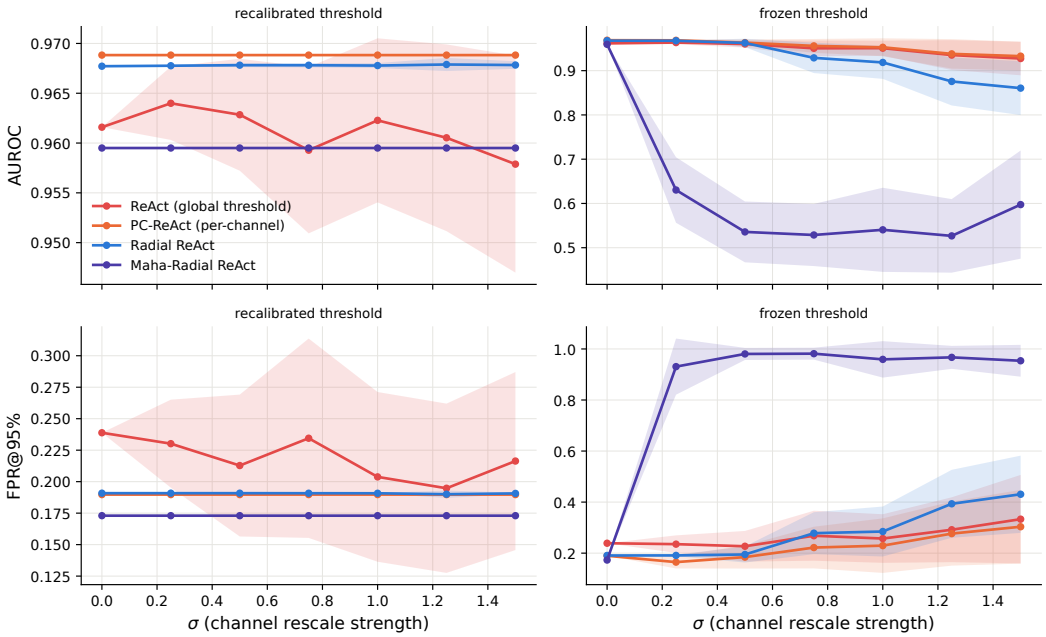


Figure 17: **Positive rescaling sweep, Digits** (mean ± std over 30 draws per σ; top row AUROC, bottom FPR_{95}; left re-calibrated, right frozen). MNIST vs Fashion-MNIST is Figure 11(a,b,d,e).

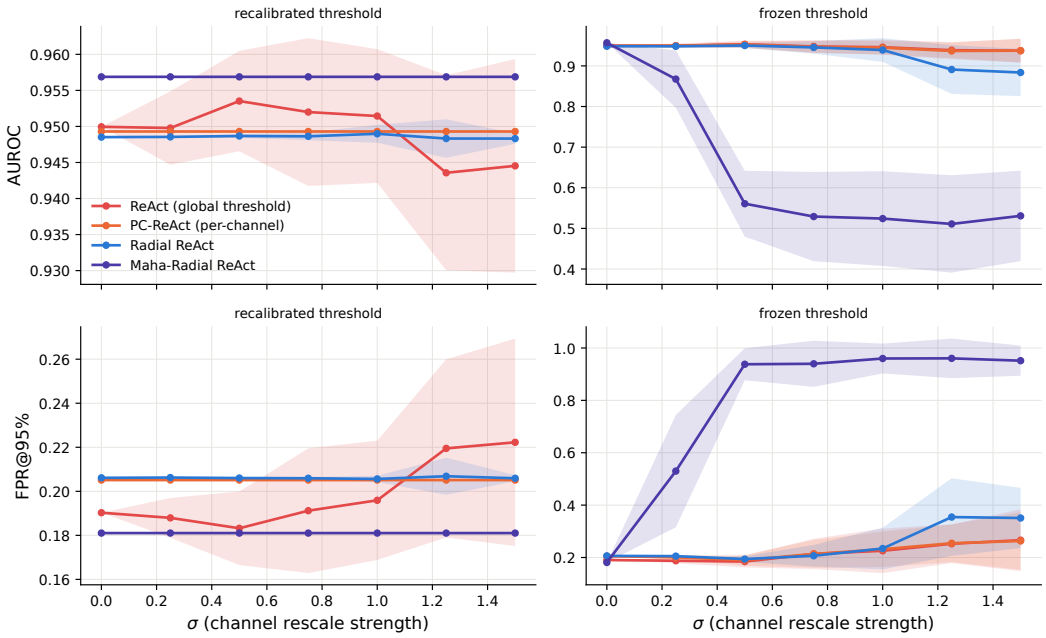


Figure 18: **Positive rescaling sweep, MNIST near-OOD** (mean ± std over 30 draws per σ; top row AUROC, bottom FPR_{95}; left re-calibrated, right frozen). MNIST vs Fashion-MNIST is Figure 11(a,b,d,e).

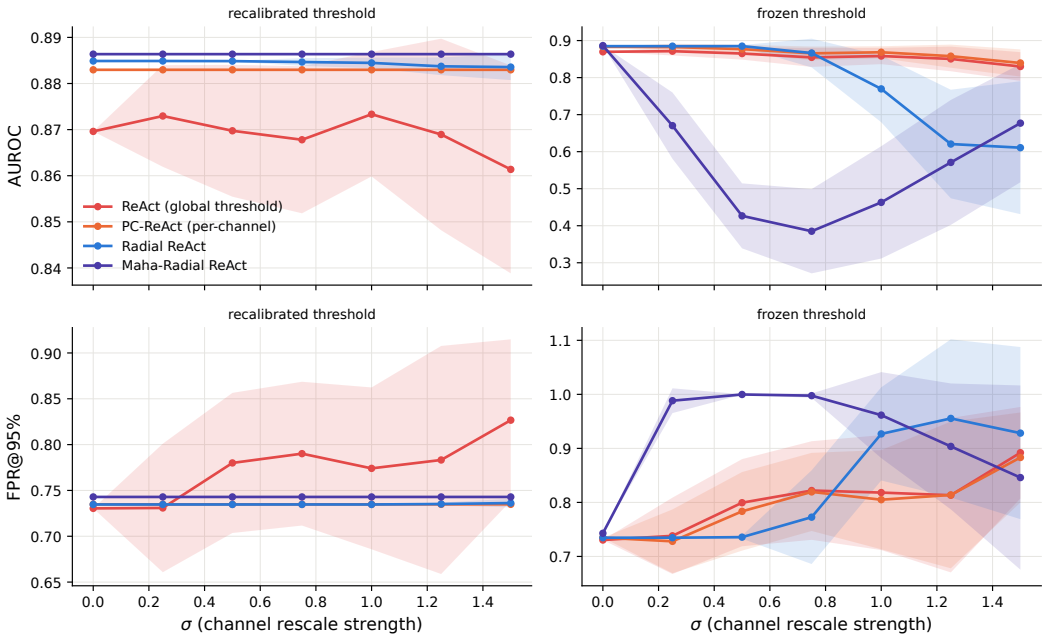


Figure 19: **Positive rescaling sweep, CIFAR-10 vs SVHN** (mean ± std over 30 draws per σ; top row AUROC, bottom FPR_{95}; left re-calibrated, right frozen). MNIST vs Fashion-MNIST is Figure 11(a,b,d,e).

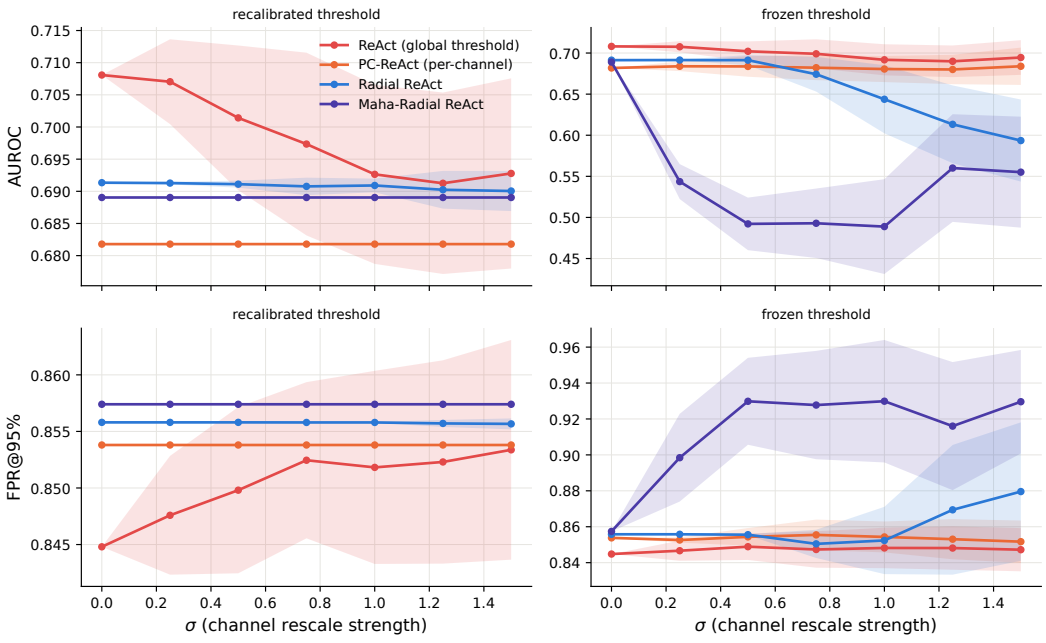


Figure 20: **Positive rescaling sweep, CIFAR-10 (0–4 vs 5–9)** (mean ± std over 30 draws per σ; top row AUROC, bottom FPR_{95}; left re-calibrated, right frozen). MNIST vs Fashion-MNIST is Figure 11(a,b,d,e).