COUPLED EQUIVARIANCE TESTS CANNOT CERTIFY WHAT ACTION LABELS MEAN

Anonymous authors
Paper under double-blind review

ABSTRACT

Equivariant outputs are used as orientations and actions by downstream programs. Do passing equivariance tests establish that a public label invokes its declared action? A shared automorphic rebinding of the input transform and output oracle reindexes every residual, preserving complete residual multisets and reindexing-invariant verdicts. On frozen REKD/C_{36} and MACE/O_{24} checkpoints, conventional tests miss 34/34 injected shared relabelings but detect 68/68 one-sided controls; independent label–state witnesses detect 34/34. MACE's mean normalized force equivariance error is 4.85×10^{-16} even as fixed-label response error increases. In a constructed REKD/HPatches matcher, the same $g \mapsto 5g$ binding in an optional orientation filter lowers held-out F1@3 px for selecting fixed candidate matches from 0.691 to 0.419 across 24 scenes; a probe-derived boundary correction restores 0.691 without retraining. Coupled consistency alone does not identify the binding needed to use an externally named action.

1 INTRODUCTION

A program requests label 1 from a rotation interface because the public declaration promises a 10° counter-clockwise turn. If both branches of an equivariance test instead use 50°, they can still agree while the program receives a response for the wrong turn under label 1. This case asks what a passing equivariance test establishes about a named action that another program must use.

This question matters when a representation supplies geometry to another component. REKD predicts orientations used for local-feature description and image matching (Lee et al., 2022); Equivariant Transporter Networks predict pick and place orientations for robotic manipulation (Huang et al., 2022). Action-aware representation learning also uses learned group actions to predict the effects of future action sequences (Keurti et al., 2023). In these applications, the representation's action relationships must connect to the conventions used to describe an image, command a gripper, or specify a future action. We ask what structural consistency establishes about that connection when the caller fixes the meaning of its action labels.

Let F be a model, x its input, and g a label in a transformation group $\mathcal{G}$. The input action α_g transforms x, while the output action ρ_g transforms its prediction. The coupled test checks

$$F(\alpha_g x) = \rho_g F(x), \qquad g \in \mathcal{G}. \tag{1}$$

Architectures and software for realizing this relation are mature (Cohen & Welling, 2016; Kondor & Trivedi, 2018; Geiger & Smidt, 2022). Evaluation samples or enumerates labels and compares the two sides' discrepancy. The result checks consistency with the actions supplied to the test; it leaves open whether those actions realize the public meanings assigned to g. This is an information gap rather than a shortage of tested angles: under a shared relabeling, even the complete residual multiset is identical.

An action label such as "turn left" has a public *binding*: the transformation it invokes. The *output oracle* computes the expected transformed prediction. A shared alternative binding can satisfy equation (1) and group laws while departing from the declaration. We hold the caller's label meaning fixed and measure the resulting discrepancy through a circular steering vector from REKD orientations and a one-step velocity increment from MACE forces. We also use the same binding in a constructed REKD/HPatches matching program: changing the public orientation decoder lowers held-out match F1, while a correction inferred from an independent probe restores it.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Figure 1 follows the $10°$ versus $50°$ case. A *grounded witness* compares the tested adapter's response to an expected state declared without either test helper, the model or shared transform code.

Our findings delimit the action-level claim supported by equivariance tests:

1. A shared automorphic relabeling reindexes approximate residuals pointwise and preserves group-law tests, limiting what a coupled test can certify about an externally declared label.
2. Matched shared and one-sided injections on two eligible released checkpoints isolate the missing information. Every shared relabeling also changes a consumer proxy evaluated under the declared, fixed label; a held-out image-matching task measures a consequence for one shared binding.
3. Independent label–state witnesses identify the shared bindings that coupled residuals miss. A probe-derived boundary correction restores the image matches without retraining. For exposed bindings within a declared finite candidate family, classical permutation-base certificates support pre-inference enforcement.

The binding faults studied in REKD, MACE, and Canonical Capsules are controlled injections; we found no actual defect in those systems.

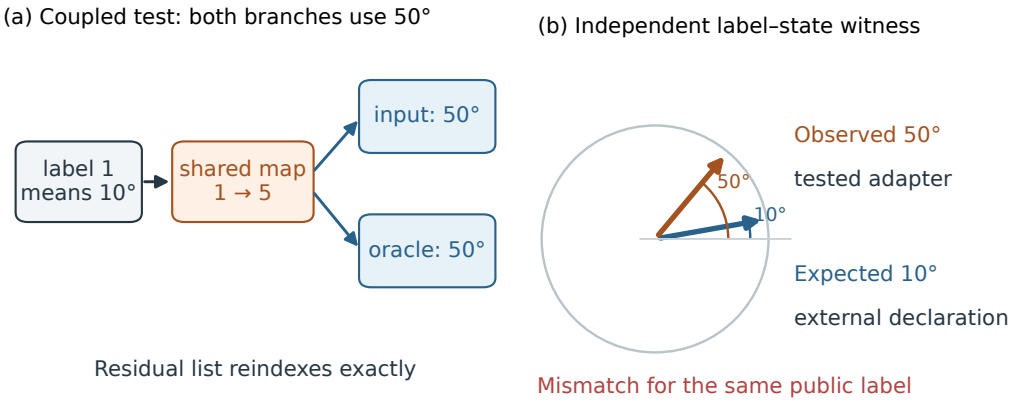

Figure 1: For the C_{36} example, label 1 promises a counter-clockwise $10°$ turn but shared relabeling $g \mapsto 5g$ executes $50°$ on both coupled-test branches (left). A probe through the tested adapter is compared with an independently declared $10°$ expected state (right).

2 What a commutation test observes

2.1 Contract, residual, and common-mode relabeling

Let $\mathcal{X}$ and $\mathcal{Y}$ be the input and output spaces. In the exact group-action statement, let $\alpha : \mathcal{G} \to \mathrm{Aut}(\mathcal{X})$ and $\rho : \mathcal{G} \to \mathrm{Aut}(\mathcal{Y})$ be actions and let $F : \mathcal{X} \to \mathcal{Y}$. An empirical evaluator instead uses indexed numerical transforms $\tilde{\alpha}_g, \tilde{\rho}_g$, which need not obey exact group laws after interpolation, cropping, or pooling. A declared binding λ is fixed outside the evaluator; it may act on a separate probe space. For statements using only indexed values, α, ρ may denote either family; exact action assumptions are stated separately. For metric d, define the ordinary residual

$$r(g, x) = d(F(\alpha_g x), \rho_g F(x)) \,. \tag{2}$$

An injected change to an implementation is a *mutant*. A *common-mode* relabeling applies the same label reassignment to both actions. For $\phi \in \mathrm{Aut}(\mathcal{G})$, define this mutant by $\alpha_g^{\phi} = \alpha_{\phi(g)}$ and $\rho_g^{\phi} = \rho_{\phi(g)}$.

Proposition 1 (reindexing and group laws). The reindexing identity extends to any label map $\phi : \mathcal{G} \to \mathcal{G}$ and any indexed transform families, including numerical ones. For every g and x,

$$r^{\phi}(g, x) = r(\phi(g), x). \tag{3}$$

For a bijective relabeling, complete enumeration gives exactly the same residual multiset—the same values with the same multiplicities, possibly in a different order. If $g \mapsto r(g, x)$ and ϕ are measurable and $\phi_* \mu = \mu$, the residual law under $g \sim \mu$ is also unchanged. This includes uniform finite-group sampling and Haar sampling on a compact group when ϕ is a continuous group automorphism. Separately, if the original transforms are exact group actions and ϕ is a group automorphism, the relabeled transforms remain exact group actions. The first conclusion does not require the second. Proofs are in Appendix A.

The complete residual tensor makes the information boundary explicit:

$$E_x(p, q) = d(F(\alpha_p x), \rho_q F(x)) \,. \tag{4}$$

The baseline reads $E_x(g, g)$; a common-mode mutant reads $E_x(\phi(g), \phi(g))$, a diagonal reindexing. Input-only and output-only controls instead read the off-diagonal entries $E_x(\phi(g), g)$ and $E_x(g, \phi(g))$. Only the shared diagonal is necessarily a reindexing; the off-diagonal entries may differ, and whether a one-sided control is detected depends on the model and observed action.

The output-level effect of a fixed external label has a direct link to the ordinary residual. Define $e^{\phi}_{\mathrm{fix}}(g, x) = E_x(\phi(g), g)$ and $\Delta_\phi(g, x) = d(\rho_{\phi(g)} F(x), \rho_g F(x))$. The reverse triangle inequality gives

$$\left| e^{\phi}_{\mathrm{fix}}(g, x) - \Delta_\phi(g, x) \right| \le r(\phi(g), x). \tag{5}$$

Under exact equivariance, fixed-label model-output error equals the separation between the two output actions. The downstream consumer metrics below are defined separately; equation (5) does not equate their numerical values with e^{ϕ}_{fix}.

Corollary 1 (certification limit). Let a test suite observe only the complete coupled residuals and truth values of the group-action laws, and let its verdict be invariant to simultaneous group reindexing. Then it returns the same verdict for (α, ρ) and (α^ϕ, ρ^ϕ) for every $\phi \in \mathrm{Aut}(\mathcal{G})$. Consequently, its observations alone cannot certify which of these public-label bindings is externally intended. This follows immediately from Proposition 1 and the preservation of identity, inverse, and composition. It does not say that every conceivable test is blind: a label-specific fact imported from an external standard can distinguish the bindings, but that fact is precisely an additional grounded observation. In the REKD experiment, T1 checks the finite label mapping against the group table; it does not test exact composition of the resampled pixel operators.

2.2 Independent witnesses

A grounded witness is a finite set

$$W = \{(g_i, u_i, v_i) : v_i = \lambda_{g_i}(u_i)\}. \tag{6}$$

Its expected states are computed without the model, input-action helper, output-oracle helper, or any shared action code; its observed states come from the tested binding. Here u_i is a probe state and v_i its declared result under label g_i. A binding ϕ passes W only if

$$\lambda_{\phi(g_i)}(u_i) = \lambda_{g_i}(u_i) \quad \text{for every } i. \tag{7}$$

To identify an unknown evaluator binding, its action on a probe must be observable; we call such a binding *exposed*. Let $K \subseteq \mathrm{Aut}(\mathcal{G})$ be a finite candidate family. A binding $\phi \in K$ answers an external label–probe query by

$$Q_\phi(g, u) = \lambda_{\phi(g)}(u). \tag{8}$$

For a query set $W = ((g_i, u_i))_i$, let $s_W(\phi)$ be the tuple of answers, called its query signature.

Proposition 2 (exposed-binding certificate). W identifies every $\phi \in K$ exactly if and only if s_W is injective on K; some finite certificate exists if and only if $\phi \mapsto \lambda \circ \phi$ is injective on K. If one shared probe u has a free orbit, so $\lambda_g(u) = \lambda_h(u)$ only for $g = h$, a label set L identifies K exactly when all restrictions $\phi|_L$ are distinct. If K is a permutation group, this is precisely a base; equivalently, L hits the moved-label set of every nonidentity element. With minimum pairwise signature distance δ under the maximum per-query metric, any observation strictly within $\delta/2$ of its exact signature decodes uniquely.

Permutation bases, their hitting-set formulation, stabilizer anchors, and the half-margin argument are inherited mathematics (Sims, 1970; Seress, 2003; Wei, 2026). The compiler implements that mathematics: it accepts only a contract written as explicit data in a prescribed format and bound to a hash; checks the group table and complete candidate family; exhausts every smaller label-set cardinality; and emits full signatures and a decoding margin. Its decoding guarantee assumes that the deployed binding belongs to that family K, or that membership has been verified separately. Validating the declared candidate table does not verify membership of an unknown deployed binding. The implementation rejects or abstains when an observation cannot be uniquely decoded within K; it is not an open-world fault detector. Appendix A gives the direct proof and exact instances.

In the original mutation audit, ϕ is injected and therefore known: T4 asks whether the independent label–state checks detect that mapping. In operational decoding, equation (8) requires an observable probe interface or extracted action mapping; it cannot recover a hidden binding from ordinary black-box predictions alone.

For the cyclic rotation group C_{36}, the literal declaration rotates a unit Cartesian probe in a fixed frame with counter-clockwise positive angle. Label 1 distinguishes all 12 candidate automorphisms, proving a minimum of one query. For the proper octahedral rotation group O_{24}, a literal generic Cartesian probe under the 24 proper signed-permutation matrices and labels 7 and 18 distinguishes all 24 candidates. Exhaustion finds no identifying singleton and 192 minimum pairs, proving a minimum of two queries. The respective minimum signature margins are $2\sin(\pi/18) = 0.3473\ldots$ and $\sqrt{3/8} = 0.6124\ldots$.

3 A GROUNDED SYMMETRY CONTRACT

Our audit separates six levels of checks, called tiers:

T0: structure. Shapes, range, finiteness, orthogonality and determinant where applicable, plus population validity.

T1: algebra. Identity, inverse, and exhaustive pairwise composition of the finite label mapping against the declared group table.

T2: direct consistency. Equation (2), using the source-declared output action.

T3: eligibility. Checkpoint integrity, outputs that have not collapsed to an uninformative constant (non-collapse), and paired evidence that the correct action beats both no action and inverse action.

T4: grounding. Deterministic checks of injected relabelings using the independent label–state checks in equation (7).

T5: consequence. A fixed-label downstream proxy holds the requested label fixed while the mutant executes $\phi(g)$.

T3 helps interpret checkpoint-dependent sensitivity and downstream consequences; it does not imply that every one-sided binding change is detectable. The deterministic T4 mapping check does not depend on checkpoint eligibility. This distinction matters for Canonical in Section 5.

Mutation matrix. For C_{36}, the common-mode family is $g \mapsto kg \bmod 36$ for all 11 nonidentity units k modulo 36. For O_{24}, it is $R \mapsto SRS^{-1}$ for each of the 23 nonidentity group elements. For every common-mode mutant we construct (i) an input-only relabeling, (ii) an output-only relabeling, and (iii) a matched non-homomorphic left shift. The anticipated first detections under the frozen test conditions are T4, T2, T2, and T1, respectively. A common-mode binding is faulty only relative to the frozen external declaration; a consumer that explicitly accepts actions up to automorphism should place it in the accepted equivalence class.

Downstream checks. A *consumer* is a downstream program using model outputs under the declared label meanings. We quantify its fixed-label response with an error matrix $C_i(p, q)$ for base object i: the interface executes p, while a reference derived from that same checkpoint's untransformed output uses the publicly declared action q. For REKD, let $D_{i,p,c,h,w}$ be the 36-channel orientation distribution after spatial registration and pooling. At each spatial cell define the *unnormalized*

circular first moment $v_{i,p}(h,w) = \sum_{c=0}^{35} D_{i,p,c,h,w}(\cos(2\pi c/36), \sin(2\pi c/36))$. The reference $v_{i,q}^{\mathrm{ref}}$ is the moment of $D_{i,0}$ after a positive cyclic channel roll by q; thus both direction and concentration are retained. With H_i the pooled spatial cells,

$$C_i^{\mathrm{R}}(p,q) = |H_i|^{-1} \sum_{(h,w)\in H_i} \|v_{i,p}(h,w) - v_{i,q}^{\mathrm{ref}}(h,w)\|_2. \tag{9}$$

For MACE, let $f_{i,p,a}$ be the predicted force on atom a after input rotation p, $m_{i,a}$ its standard atomic mass, and $a_{i,p,a} = f_{i,p,a}/m_{i,a}$. For the declared Cartesian rotation R_q, set $a_{i,q,a}^{\mathrm{ref}} = R_q a_{i,0,a}$ and $s_i = \left(n_i^{-1} \sum_a \|a_{i,0,a}\|_2^2\right)^{1/2}$. The normalized one-step, zero-initial-velocity proxy is

$$C_i^{\mathrm{M}}(p,q) = \frac{1}{n_i s_i} \sum_{a=1}^{n_i} \|a_{i,p,a} - a_{i,q,a}^{\mathrm{ref}}\|_2. \tag{10}$$

A common time step cancels from this normalized velocity-increment error. These references use transformed *baseline model outputs*, not measured orientation or physical-force targets; T5 measures fixed-label response consistency, not task or physical prediction accuracy. They are distinct from the independent, model-free label–state witnesses of T4.

For each common-mode ϕ, the paired object-level excess is $D_i^{\phi} = |\mathcal{G}|^{-1} \sum_g [C_i(\phi(g), g) - C_i(g, g)]$. We report its mean over 256 base objects and a 95% paired percentile interval from 10,000 object-clustered resamples. The pooled ratio is the median over mutants, objects, and labels of $(C_i(\phi(g), g) + 10^{-12})/(C_i(g, g) + 10^{-12})$; it is a descriptive statistic, not the detection criterion.

4 Frozen evidence design

4.1 Released systems and finite actions

REKD / C_{36}. We load the released group-36 oriented-keypoint checkpoint with strict parameter matching (Lee et al., 2022). The population comprises 256 deterministic asymmetric 192×192 analytic images. We enumerate all 36 ten-degree rotations and evaluate an interior pooled score/orientation descriptor chosen before inference to reduce raster-border effects. The direct error is normalized aligned descriptor error. Eleven nonidentity unit automorphisms are audited. The input uses nearest-neighbor raster rotation and output alignment uses bilinear rotation, cropping, and pooling. These numerical operators are not asserted to be exact invertible actions on pixel arrays; their residuals still obey Proposition 1's pointwise reindexing identity.

MACE-MP-0a / O_{24}. We load the official small `ScaleShiftMACE` checkpoint in native float64 through the released MACE calculator (Batatia et al., 2022; 2023). The population contains 256 centered configurations: 32 perturbations from each of eight Atomic Simulation Environment (ASE) molecule templates. Geometry-only rejection enforces minimum interatomic distance and asymmetry under all 24 proper signed-permutation rotations. The primary error is mean ordered per-atom force-vector error normalized by the root mean square (RMS) of the base forces; scalar-energy invariance is separate. We audit all 23 nonidentity inner automorphisms. The audited public labels denote the 24 signed-permutation matrices in a fixed Cartesian laboratory frame. If a raw numeric matrix is instead the complete contract with no external symbolic meaning, a change of enumeration or basis is a convention rather than the fault studied here. At our fixed MACE revision, official tests for forces, dipoles, polarizabilities, and magnetic forces derive both transformed inputs and expected covariant outputs from the same rotation matrix (Batatia et al., 2022). The source audit `mace_testing_practice_audit_2026-08-25.md` identifies the exact revision and test locations. These tests check consistency relative to the supplied matrix.

Canonical Capsules / O_{24} boundary. The first frozen protocol also evaluated raw ordered attention-weighted keypoints from the released multi-category unaligned checkpoint (Sun et al., 2021) on 256 asymmetric procedural point clouds. Its correct-action error did not beat no action or inverse action. We therefore exclude it from all positive mutation-coverage totals and retain it as an eligibility boundary, not as an upstream bug claim.

4.2 SEQUENTIAL REGISTRATION AND DECISIONS

The v0 source manifest and pass/fail criteria were fixed (frozen) before confirmatory REKD and Canonical inference. V0 required both systems to pass; Canonical failed, so the broad v0 claim failed. Only then was MACE selected. An eight-configuration float32 feasibility pilot was development-only and excluded. We froze a new v1 source manifest, native-float64 protocol, population seed, metrics, thresholds, and decision rule before generating or evaluating the 256-case MACE population. MACE is therefore a *prospective sequential replication*, not a replacement for the failed v0 system. The REKD result is frozen prior evidence. Both outcomes and the negative boundary are reported together. These are local, hash-bound source freezes and sequential decisions, not a third-party timestamped preregistration.

For each system, bootstrap sampling treats the base object as a cluster. We use 10,000 paired percentile bootstrap resamples. A model-dependent mutant is detected only if the 95% lower confidence bound for the error increase over the matched correct condition is above zero. Algebra and T4 grounding use exhaustive group/mapping checks and do not depend on checkpoint predictions. Common-mode conventional residuals are literal cached reindexings, so their equality is exact rather than threshold-selected. Full success thresholds, seeds, hashes, and provenance appear in Appendix B.

4.3 HELD-OUT IMAGE MATCHING

We use the same public C_{36} binding table for the two audited REKD equivariance-test branches and, in a separate HPatches matching run, for the channel-to-angle decoder of REKD's optional orientation-based match filter on full-image sequences (Lee et al., 2022; Balntas et al., 2017). The same frozen REKD and HardNet++ weights produce keypoints, descriptors, and mutual-nearest-neighbor candidate matches in every condition. The filter decodes orientation channel c as $10c$ degrees and keeps candidates within a 30° window around the modal relative orientation. The wrong interface instead decodes c as $10\phi(c)$ degrees for $\phi(c) = 5c \bmod 36$. For example, a candidate one bin (10°) from the modal relative orientation can become five bins (50°) away and cross the fixed 30° filter threshold. The legal convention sends the internal name $\phi^{-1}(c)$ through the same relabeled adapter, preserving the public angle. The correction instead infers $\hat{\phi}$ from a separate raster-marker probe of that adapter and sends $\hat{\phi}^{-1}(c)$. Neither changes model weights or descriptors.

The task uses four development scenes and 24 held-out scenes, five image pairs per scene. Its primary metric is F1@3 px for retaining geometrically correct matches from the fixed candidate set, using the supplied homographies. The held-out contrast is wrong minus normal, averaged equally over scenes; its interval uses 10,000 paired, scene-stratified bootstrap resamples. The intervention, split, metric, and analysis were locally hash-frozen before held-out inference. Appendix B.3 gives the extraction and metric details.

5 RESULTS

Before interpreting checkpoint-dependent mutation sensitivity, we check that the selected outputs exhibit the audited action.

Checkpoint eligibility. For REKD, mean correct, no-action, and inverse-action errors were 0.0900, 0.7309, and 0.7248; the lower 95% margins over correct were 0.6394 and 0.6333. For MACE, the corresponding force errors were 4.85×10^{-16}, 1.1278, and 0.7228, with lower margins 1.1144 and 0.7142. MACE's 95th percentile of relative energy-invariance error was 1.96×10^{-16} and median force RMS was 5.148 eV/Å, ruling out a zero-force collapse. Canonical's correct error (0.1164) was worse than no action (0.0685) and slightly worse than inverse action (0.1156), so its gate failed before mutation coverage was interpreted. The paired margins are shown in Appendix Figure 3.

Table 1: Frozen mutation evidence, detected/tested. T1 and T4 are deterministic mapping checks; T2 and T5 use checkpoint outputs. Canonical failed T3, so its raw mutation rows are not interpreted.

System	T3	shared T2	witness T4	input T2	output T2	nonhom. T1	proxy T5
Canonical/O_{24}	no	–	–	–	–	–	–
REKD/C_{36}	yes	0/11	11/11	11/11	11/11	11/11	11/11
MACE/O_{24}	yes	0/23	23/23	23/23	23/23	23/23	23/23

Table 1 shows the pre-specified detection pattern. The independent checks against the exact declarations detected 34/34 known common-mode mappings and produced no failures on 60 unmutated label–state checks (36 for C_{36} and 24 for O_{24}); algebra detected all 34 maps that violate the composition law. On the two eligible checkpoints, conventional direct tests detected 0/34 common-mode mutants, exactly as equation (3) predicts, but detected all 34 input-only and all 34 output-only controls. Thus the mapping witness adds the missing information rather than indiscriminately rejecting relabelings.

Fixed-label downstream consequence. Every eligible common-mode mutant changed its fixed-label consumer with a positive paired lower bound. The smallest lower bound on normalized excess error was 0.548 for REKD's circular steering vector and 0.770 for MACE's one-step force/mass velocity increment. REKD's pooled median mutant/correct error ratio was 9.31. On the same MACE checkpoint, mean normalized force equivariance residual was 4.853×10^{-16}, while the 23 shared mutants had fixed-label velocity-proxy excess errors of 0.784–0.958. Their complete coupled-residual multisets were exactly unchanged. Figure 2 puts these observations side by side for each mutant. The error scales differ: the contrast is coexistence of near-zero equivariance residual and changed fixed-label response, not a ratio between those metrics. These are controlled consumer proxies, not physical prediction or end-task accuracy.

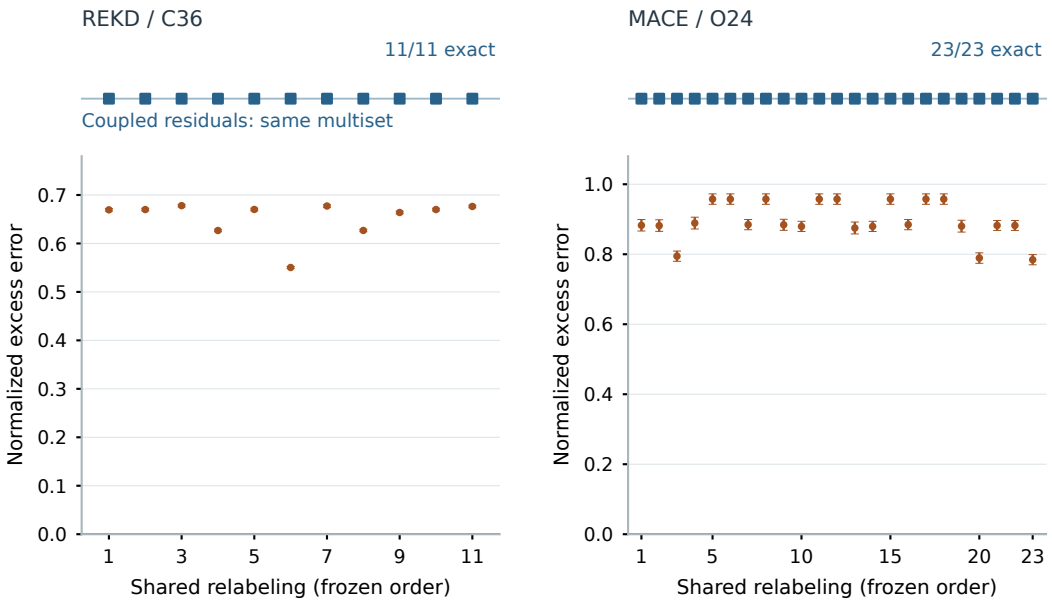

Figure 2: The same shared relabelings preserve the complete coupled-residual multiset (top strips; exact reindexing, no uncertainty interval) but increase fixed-label consumer excess error (bottom; point estimates and 95% paired object-clustered bootstrap intervals). Positions identify the frozen mutant maps. REKD and MACE have different consumer definitions and vertical scales; compare the within-system contrast, not effect magnitudes across panels.

Held-out REKD matching. Table 2 reports the 24-scene HPatches comparison. The shared $g \mapsto 5g$ binding lowers candidate-set match F1@3 px from 0.691 to 0.419: the equal-scene paired

difference is -0.272 (95% interval $[-0.307, -0.235]$). All 24 scene-level differences are negative. The wrong binding retains 61.86 geometrically correct matches per pair, versus 148.33 under the normal binding. The independent probe identifies the injected binding from the adapter's observed action; applying its inverse restores exactly the normal match selection on all 120 pairs. The legal internal renaming gives the same selection.

Table 2: HPatches matching with fixed REKD/HardNet++ features and candidate matches. Values are means over 24 held-out scenes (120 image pairs). F1@3 px uses homography-correct candidates as positives; false negatives are correct candidates removed by the filter. The two right columns are per-pair counts.

Matching condition	F1@3 px	retained	correct retained
Unfiltered candidates	0.678	263.23	160.48
Normal	0.691	217.91	148.33
Shared wrong binding	0.419	86.18	61.86
Legal internal renaming	0.691	217.91	148.33
Probe-derived correction	0.691	217.91	148.33

Independent certificate. A one-label C_{36} certificate and a two-label O_{24} certificate decode all 36 declared candidate bindings and reject all 34 injected alternatives. The three total queries tie declared-generator and greedy full-identification baselines; the 60-to-3 reduction counts queries, not contract specification. Table 5 and the candidate-family comparison are in Appendix C.1.

Downstream REKD program. In a separately frozen one-image REKD program, the independent raster witness identified the declared binding before inference. The program then loaded the official checkpoint and made one model call over all 36 labels, producing finite score and orientation maps. Under the controlled $g \mapsto 5g \bmod 36$ binding, the witness identified the relabeling and rejected it with zero model calls. On the admitted checkpoint outputs, the sorted diagonals $E(g, g)$ and $E(5g, 5g)$ of the coupled residual tensor were bitwise identical.

Authentic historical regression. Official e3nn issue 149 reported a Cartesian–irrep conversion that failed physical rotation (Chen-Cai-OSU, 2020). At the buggy parent, 13 selected internal and nearby tests passed while an independent Cartesian witness failed across 20/20 seeds; after the official repair it failed 0/20 (Geiger, 2020a;b). This is a distinct interface defect, not a shared public-label relabeling. Appendix C.3 gives the revision and reproduction details.

6 Related work and novelty boundary

Equivariance and identifiability. Equivariant architectures formalize commutation relative to supplied group actions (Cohen & Welling, 2016; Kondor & Trivedi, 2018); direct residuals are also standard in libraries such as e3nn (Geiger & Smidt, 2022). Empirical equivariance and representation equivalence were studied by Lenc & Vedaldi (2015). Wang et al. (2023) distinguish correct, incorrect, and extrinsic task equivariance, but ask whether an inductive bias matches latent task structure rather than whether an evaluation oracle binds a public label correctly. Homomorphism Autoencoders learn sensory and action representations from observed transitions and use them for future-action prediction (Keurti et al., 2023); the paired observations associate action inputs with their sensory effects. Equivariance by Contrast establishes identifiability of learned representations up to conjugacy under its assumptions (Schmidt et al., 2025); anchor–stabilizer theory gives a broad account of how external information shrinks reparameterization ambiguity (Wei, 2026). A latent change of basis can be absorbed by a matching decoder, preserving externally observed behavior. Reassigning public action labels changes which operation a caller receives unless the boundary adapter is adjusted as well. Our question concerns evidence for this correspondence between group structure and named operations: coupled residuals can remain unchanged while fixed-label use changes.

Geometric interfaces. ROS specifies body and camera axes and the sign of yaw to support component integration (Foote & Purvis, 2010); PhysFrame checks frame-type and convention inconsistencies in robotics software (Kate et al., 2021). These interfaces make external conventions explicit.

We study a shared action-label reassignment that preserves the coupled test, and measure its effect where REKD orientations enter a fixed matching filter. The e3nn issue 149 report and its two-commit repair document the historical Cartesian–irrep defect (Chen-Cai-OSU, 2020; Geiger, 2020a;b); our reproduction compares that distinct interface error with the evaluation fault studied here.

Testing semantic transformations. Pseudo-oracles compare separate implementations when no direct oracle exists (Davis & Weyuker, 1981); independently developed versions can still fail dependently (Knight & Leveson, 1986). Metamorphic testing checks cross-execution relations (Chen et al., 2018). Recent work supplies algebra-derived relations, semantic mutation, Set-Cover selection, and validity-gated testing (Li et al., 2026b;c;a;d). Classical permutation-group algorithms supply the finite generator/stabilizer ancestry (Sims, 1970; Seress, 2003); automorphism-induced non-canonicity has also been isolated in an adjacent graph-explanation setting (Xu et al., 2026). These sources supply the ingredients for finite binding checks. The post-result reconstruction in Appendix C.2 confirms that their specified runtime behavior can be assembled from those ingredients once the task is given, so we assign no algorithmic originality to the finite checker. Because that reconstruction is new post-cutoff work, it does not show that one predecessor already posed and solved the combined evaluation question. The paper's distinct result is the certification limit under shared relabeling and its measured consequence for match selection through a shared action interface; the compiler implements the independent binding check. Concurrent work characterizes a different blind spot: for a fixed, semantically correct edit-to-answer transform, a systematic prediction error is invisible exactly when the two commute, giving a joint-centralizer criterion; it also analyzes cyclic category relabeling (Khanbayov & Kurban, 2026). That work establishes the algebraic blind-spot and edit-suite completeness framing for model errors. Our evaluator-binding fault instead relabels both supplied group actions, reindexing residuals without a commutation assumption.

7 Limitations, implications, and conclusion

The empirical audit covers controlled C_{36} and O_{24} bindings. Residual reindexing applies to arbitrary indexed transforms, including numerical raster operators; preservation of exact group laws additionally requires exact actions and an automorphic relabeling. A contract that accepts labels up to automorphism should accept the corresponding relabelings; one that fixes their external meaning needs an independent declaration. An unknown deployed binding also needs an observable probe or extracted action mapping. The one- and two-label certificate minima hold for the frozen candidate families, probes, tolerance, and unit query cost; they do not reduce the full specification burden. An actual binding outside the candidate family can agree on every queried label and still be wrong on an unqueried one. In the one-label C_{36} guard, swapping labels 2 and 3 while fixing queried label 1 is admitted as identity; membership of the deployed mapping is not checked by the small certificate. A deployment claiming unconditional rejection would need an additional membership check.

The checkpoint-dependent evidence has further limits. Canonical Capsules failed its selected-output eligibility gate, and MACE was chosen after the failed v0 protocol. The fixed REKD consumer used one image and one controlled wrong binding; it demonstrated pre-inference rejection. The held-out matching experiment uses a constructed shared adapter at REKD's optional orientation filter. The e3nn regression is an authentic but different interface failure. A source-only screen of 88 issues found no exact-class case, which does not estimate its prevalence. Implementation independence also cannot rule out a shared misunderstanding of the declaration.

Across the controlled REKD and MACE bindings, the shared relabelings leave even fully enumerated coupled-residual multisets unchanged but increase fixed-label consumer-proxy error. The MACE contrast occurs alongside numerical-precision-level equivariance residuals; tighter tolerance or more labels in the same coupled test cannot supply the missing label information. In held-out REKD/HPatches matching, one shared binding also reduces correct match retention; a probe-derived correction restores the original selection without retraining. Independent label–state observations identify the mismatched bindings. These results separate consistency of a geometric representation from compatibility with a caller's action convention. When public action labels govern downstream use, the binding needs its own measurement.

AI USE STATEMENT

Generative AI tools (OpenAI Codex) were used to propose and red-team the research question; help formalize mathematical claims and draft proofs; design the frozen protocols and experiments; implement and test audit software; assist with experiment orchestration, including code that deterministically generated procedural audit populations, and result interpretation; search for, summarize, and format related literature; generate scientific figures; and draft and edit the manuscript and artifact documentation. AI output was not treated as an authority: cited claims were checked against primary sources, mathematical claims were reduced to executable finite-group tests or explicit proofs, and generated code was exercised by unit tests, frozen source manifests, exact hashes, and independent result recomputation. No generative model supplied checkpoint predictions, confirmatory samples, labels, or hand-selected result rows. The authors take responsibility for the final content, including AI-assisted text, claims, code, and artifacts.

ETHICS STATEMENT

This work uses public research software, released checkpoints, and HPatches images under their upstream terms, generates synthetic/procedural audit populations, and involves no human subjects or personal data. We avoid attributing injected evaluation mutants to upstream authors and report an out-of-domain ineligible result as an applicability boundary rather than a model defect. The main foreseeable risk is overstating a controlled test-adequacy result as evidence of natural fault prevalence; the claim boundaries above are intended to prevent that use.

REPRODUCIBILITY STATEMENT

The accompanying anonymous materials provide stored checkpoint statistics, literal contracts and their compiler, per-pair matching masks, analysis code, and figure sources. The README gives commands to replay the reported statistics and finite binding checks without model inference. No checkpoints or HPatches images are redistributed; their upstream sources and exact hashes are listed separately. The appendix gives the proofs, protocols, and input settings; Appendix D describes the artifact's coverage.

REFERENCES

Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5173–5182, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Balntas_HPatches_A_Benchmark_CVPR_2017_paper.html.

Ilyes Batatia, Dávid Péter Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/4a36c3c51af11ed9f34615b81edb5bbc-Abstract-Conference.html.

Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, Xavier R. Advincula, Mark Asta, Matthew Avaylon, William J. Baldwin, et al. A foundation model for atomistic materials chemistry. *arXiv preprint arXiv:2401.00096*, 2023. URL https://arxiv.org/abs/2401.00096.

Tsong Yueh Chen, Fei-Ching Kuo, Huai Liu, Pak-Lok Poon, Dave Towey, T. H. Tse, and Zhi Quan Zhou. Metamorphic testing: A review of challenges and opportunities. *ACM Computing Surveys*, 51(1):4:1–4:27, 2018. doi: 10.1145/3143561.

Chen-Cai-OSU. e3nn issue #149: Equivariance test. GitHub issue, 2020. URL https://github.com/e3nn/e3nn/issues/149. Opened 27 November 2020.

Taco S. Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 2990–2999, 2016. URL https://proceedings.mlr.press/v48/cohenc16.html.

Martin D. Davis and Elaine J. Weyuker. Pseudo-oracles for non-testable programs. In *Proceedings of the ACM '81 Conference*, pp. 254–257. Association for Computing Machinery, 1981. doi: 10.1145/800175.809889.

e3nn contributors. e3nn 0.1.0. Git tag 2dc6848690033b958b699fa498c51466fc2a67d2, 2020. URL https://github.com/e3nn/e3nn/tree/0.1.0.

Tully Foote and Mike Purvis. Standard units of measure and coordinate conventions, 2010. URL https://github.com/ros-infrastructure/rep/blob/master/rep-0103.rst. ROS Enhancement Proposal 103.

Mario Geiger. Fix #149. e3nn commit fe34b205e53b53cad5ac9a063b3a826c949d091b, 2020a. URL https://github.com/e3nn/e3nn/commit/fe34b205e53b53cad5ac9a063b3a826c949d091b. 27 November 2020.

Mario Geiger. Fix. e3nn commit b6d9b51fbcd2fa0270a95d5ab328ff19cf61cbe8, 2020b. URL https://github.com/e3nn/e3nn/commit/b6d9b51fbcd2fa0270a95d5ab328ff19cf61cbe8. Completed repair, 27 November 2020.

Mario Geiger and Tess Smidt. e3nn: Euclidean neural networks. *arXiv preprint arXiv:2207.09453*, 2022. URL https://arxiv.org/abs/2207.09453.

Haojie Huang, Dian Wang, Robin Walters, and Robert Platt. Equivariant transporter network. In *Proceedings of Robotics: Science and Systems*, 2022. doi: 10.15607/RSS.2022.XVIII.007. URL https://www.roboticsproceedings.org/rss18/p007.html.

Sayali Kate, Michael Chinn, Hongjun Choi, Xiangyu Zhang, and Sebastian Elbaum. PHYSFRAME: Type checking physical frames of reference for robotic systems. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pp. 45–56, 2021. doi: 10.1145/3468264.3468608.

Hamza Keurti, Hsiao-Ru Pan, Michel Besserve, Benjamin F. Grewe, and Bernhard Schölkopf. Homomorphism AutoEncoder – learning group structured representations from observed transitions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 16190–16215. PMLR, 2023. URL https://proceedings.mlr.press/v202/keurti23a.html.

Rasul Khanbayov and Hasan Kurban. Consistency has a computable blind spot: A commutation theory of label-free reliability for vision-language figure reading, 2026. URL https://arxiv.org/abs/2608.05675. Version 2, 7 August 2026.

John C. Knight and Nancy G. Leveson. An experimental evaluation of the assumption of independence in multi-version programming. *IEEE Transactions on Software Engineering*, SE-12(1): 96–109, January 1986. doi: 10.1109/TSE.1986.6312924.

Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2747–2755, 2018. URL https://proceedings.mlr.press/v80/kondor18a.html.

Jongmin Lee, Byungjin Kim, and Minsu Cho. Self-supervised equivariant learning for oriented keypoint detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4847–4857, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Lee_Self-Supervised_Equivariant_Learning_for_Oriented_Keypoint_Detection_CVPR_2022_paper.html.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 991–999, 2015. doi: 10.1109/CVPR.2015.7298701.

Meng Li, Xiaohua Yang, Jie Liu, and Shiyu Yan. Minimum complete MR subsets under semantic-mutation fault models: A support-set domination boundary. *arXiv preprint arXiv:2606.08269*, 2026a. URL https://arxiv.org/abs/2606.08269. Preprint.

Meng Li, Xiaohua Yang, Jie Liu, and Shiyu Yan. NOETHER: A constructive framework for metamorphic pattern discovery from operator algebras. *arXiv preprint arXiv:2605.17390*, 2026b. URL https://arxiv.org/abs/2605.17390. Preprint.

Meng Li, Xiaohua Yang, Jie Liu, and Shiyu Yan. A semantic mutation metric for metamorphic relation adequacy in scientific computing programs. *arXiv preprint arXiv:2605.17437*, 2026c. URL https://arxiv.org/abs/2605.17437. Preprint.

Meng Li, Xiaohua Yang, Jie Liu, and Shiyu Yan. Domain-validity-gated metamorphic testing of scientific ML surrogates. *arXiv preprint arXiv:2606.17529*, 2026d. URL https://arxiv.org/abs/2606.17529. Preprint.

Tobias Schmidt, Steffen Schneider, and Matthias Bethge. Equivariance by contrast: Identifiable equivariant embeddings from unlabeled finite group actions. *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2510.21706. arXiv:2510.21706.

Ákos Seress. *Permutation Group Algorithms*, volume 152 of *Cambridge Tracts in Mathematics*. Cambridge University Press, 2003. doi: 10.1017/CBO9780511546549.

Charles C. Sims. Computational methods in the study of permutation groups. In John Leech (ed.), *Computational Problems in Abstract Algebra*, pp. 169–183. Pergamon Press, Oxford, 1970.

Weiwei Sun, Andrea Tagliasacchi, Boyang Deng, Sara Sabour, Soroosh Yazdani, Geoffrey Hinton, and Kwang Moo Yi. Canonical capsules: Self-supervised capsules in canonical pose. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2012.04718.

Dian Wang, Jung Yeon Park, Neel Sortur, Lawson L. S. Wong, Robin Walters, and Robert Platt. The surprising effectiveness of equivariant models in domains with latent symmetry. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=P4MUGRM4Acu.

Yuan-Hao Wei. Anchor–stabilizer theory for identifiable representation learning: Generalized anchors as symmetry breakers. *Preprints*, 2026. doi: 10.20944/preprints202606.0188.v1. Not peer reviewed.

Xin Xu, Siru Tao, and Kaizhen Tan. Automorphism-induced non-canonicity in top-k explanations of graph neural networks. *arXiv preprint arXiv:2607.26344*, 2026. URL https://arxiv.org/abs/2607.26344. Preprint.

A Proofs and formal scope

Proof of Proposition 1. By definition, for any indexed transform families (including numerical families that are not exact actions),

$$\begin{aligned} r^{\phi}(g,x) &= d\big(F(\alpha_g^{\phi}x), \rho_g^{\phi}F(x)\big) \\ &= d\big(F(\alpha_{\phi(g)}x), \rho_{\phi(g)}F(x)\big) = r(\phi(g),x). \end{aligned}$$

If F is exactly equivariant for (α,ρ), substituting $\phi(g)$ into equation (1) proves exact equivariance for $(\alpha^{\phi},\rho^{\phi})$. For a finite group and bijective ϕ, $\{r^{\phi}(g,x) : g \in \mathcal{G}\}$ and $\{r(g,x) : g \in \mathcal{G}\}$ are the same multiset. More generally, suppose $g \mapsto r(g,x)$ and ϕ are measurable and $\phi_*\mu = \mu$. Then, for every measurable set A,

$$\begin{aligned} \Pr_{g\sim\mu}[r^{\phi}(g,x) \in A] &= \Pr_{g\sim\mu}[r(\phi(g),x) \in A] \\ &= \Pr_{h\sim\mu}[r(h,x) \in A]. \end{aligned}$$

Thus $r^{\phi}(g,x) \overset{d}{=} r(g,x)$. For independent, identically distributed partial samples from μ, the residual vectors are equal in law; hence means, empirical quantiles, uniform-threshold decisions, and bootstrap summaries using the same resampling rule have the same sampling law. They need not match on the same realized, unreindexed labels. Literal equality holds for complete enumeration or explicitly reindexed samples.

Group-law preservation under the stronger assumption. Now assume α and ρ are exact group actions and ϕ is a group automorphism. For $g, h \in \mathcal{G}$,

$$\alpha_g^{\phi}\alpha_h^{\phi} = \alpha_{\phi(g)}\alpha_{\phi(h)} = \alpha_{\phi(g)\phi(h)} = \alpha_{\phi(gh)} = \alpha_{gh}^{\phi}.$$

Automorphisms also preserve the identity and inverses, so the remaining group laws follow. The same proof applies to ρ. Inversion is not called an automorphism of a general non-abelian group; anti-automorphisms require the opposite group and a reversed composition oracle.

Eligibility does not imply one-sided sensitivity. For an exact counterexample, take $G = C_{36}$, $\omega = e^{2\pi i/36}$, $\alpha_g z = \omega^g z$, $F(z) = z^2$, and $\rho_g y = \omega^{2g} y$. The model is nonconstant and exactly equivariant. On generic nonzero inputs and labels, correct action beats no action and inverse action. Yet the automorphism $\phi(g) = 19g \bmod 36$ obeys $F(\alpha_{\phi(g)} z) = F(\alpha_g z)$ and $\rho_{\phi(g)} = \rho_g$ for every g: the output action has a nontrivial kernel. Both one-sided residuals remain unchanged despite action eligibility. Our observed 68/68 one-sided detections are empirical outcomes for the selected checkpoints, not consequences of the eligibility gate.

Witness condition. Equation (7) follows directly from comparing the mutated execution $\lambda_{\phi(g_i)}(u_i)$ with the independently stored expected state $v_i = \lambda_{g_i}(u_i)$. The residual ambiguity is the set of common-mode relabelings that also leave every witness outcome unchanged. Code independence does not eliminate an erroneous human declaration.

Proof of Proposition 2. Two candidates are distinguishable by W exactly when their answer tuples differ, so simultaneous exact identification is equivalent to injectivity of s_W. If $\phi \mapsto \lambda \circ \phi$ is not injective, two candidates give the same answer to every possible query and no certificate exists. Conversely, if it is injective, each of the finitely many candidate pairs differs at some label and probe; choosing one distinguishing query per pair gives a finite certificate. Under a common free probe, equality of answers at g is equivalent to equality of the candidate images of g. Thus L identifies all candidates exactly when their restrictions to L are distinct. For a permutation group this means the pointwise stabilizer of L is trivial, the definition of a base (Seress, 2003); equivalently, L intersects every nonidentity moved-label set. Finally, if two exact signatures are at distance at least δ, the triangle inequality prevents one observation from lying strictly within $\delta/2$ of both.

Exact frozen minima. For C_{36}, no empty query set can distinguish 12 candidates, while a free probe at label 1 returns a distinct state for every unit multiplier; hence the minimum is one. For O_{24}, the identity label is fixed by every candidate. For any nonidentity label g, conjugation by g fixes g and is nonidentity because the declared proper octahedral group has trivial center, so no singleton is a base. Labels 7 and 18 generate the group; the intersection of their centralizers is therefore its trivial center, so the pair is a base. Complete literal-table enumeration independently finds 192 minimum pairs. Here the probe is $u = (1/4, -1/2, 3/4)$, and the declared rotations are $R_7(x, y, z) = (-y, x, z)$ and $R_{18}(x, y, z) = (y, z, x)$. The emitted signatures give separations $2\sin(\pi/18)$ and $\sqrt{3/8}$. These proofs establish minimum cardinality only under the frozen contracts and candidate families; checking all 60 declared label states was not minimal.

B GROUPS, POPULATIONS, AND FROZEN GATES

B.1 GROUPS AND MUTATION DENOMINATORS

For $C_{36} = \mathbb{Z}/36\mathbb{Z}$, automorphisms are multiplication by a unit modulo 36. The twelve units are $\{1, 5, 7, 11, 13, 17, 19, 23, 25, 29, 31, 35\}$; excluding identity leaves eleven. Each matched non-homomorphic control is a nonzero left shift by the same parameter. For O_{24}, we enumerate the 24 proper 3×3 signed-permutation matrices in one fixed order. Conjugation by each nonidentity S gives the 23 common-mode inner automorphisms, and left multiplication by S gives the matched non-homomorphic maps. Exhaustive equivalence checks leave all declared denominators unchanged.

B.2 POPULATION AND METRIC DETAILS

REKD. The seed is 20260919. Images combine unequal analytic blobs, oriented line segments, and a labeled asymmetric motif. Candidates with normalized pixel distance below 0.20 to any

nonidentity C_{36} transform are redrawn. Model transforms use the source-faithful nearest-neighbor rotation; output maps are registered with a negative-angle bilinear rotation, center-cropped to 128×128, pooled to 8×8, and orientation-renormalized. The bootstrap seed is 20260923.

MACE. The seed is 20261003. We use 32 samples from each of H_2O, NH_3, CH_4, C_2H_6, C_6H_6, CH_3OH, HCOOH, and CH_3CHO. Centered coordinates receive independent 0.08 Å Gaussian noise, a uniform scale in $[0.97, 1.03]$, and a Haar-random proper base rotation, followed by recentering. Candidates are redrawn if any interatomic distance is below 0.45 Å or if species-matched normalized distance to a nonidentity O_{24} rotation is below 0.15. Ten candidates were redrawn for asymmetry and none for minimum distance. The accepted minima were 0.1593 and 0.7172 Å, respectively. The bootstrap seed is 20261007.

The model is the official MACE-MP-0a small `ScaleShiftMACE`: 3,847,696 parameters, 71 state entries, 6 Å cutoff, and 89 supported elements. It is loaded in checkpoint-native float64 with MACE 0.3.17, e3nn 0.4.4, ASE 3.29, PyTorch 2.11, no compilation, and no accelerated alternative backend. The confirmatory collector made 6,144 calls. The excluded development pilot used eight configurations, float32, and seed 8675309; none of its values enters a table, interval, or decision.

B.3 Held-out REKD matching protocol

The HPatches full-image sequences provide homographies for five pairs from each six-image scene (Balntas et al., 2017). A SHA-256 ordering of scene names, separately within illumination and viewpoint scenes, selected two development and twelve held-out scenes from each type before image inference. The task uses REKD source commit `49865de1870ed9ddf584bd74e3e223b8e0912760`, the released checkpoint identified in Appendix D, and source HardNet++ weights `9fe80a335fe72d1225fdd1ae0943418b12489f95014bd51dddc9819d6ffa33c7`. We extract 512 keypoints at one native-resolution pyramid level, sample their 36-channel orientation argmax, crop source 32-pixel patches at position and scale, and compute fixed HardNet++ descriptors. The source mutual nearest-neighbor matcher produces one candidate set per image pair. All conditions apply the source orientation filter to that same set; only the public channel-to-angle binding changes.

For each candidate, the supplied homography warps its reference keypoint into the target image. A candidate is correct if its reprojection distance is at most 3 px. Against this fixed candidate set, each filtered condition has $\mathrm{F1@3} = 2TP/(2TP + FP + FN)$, where FN counts geometrically correct candidates removed by the filter. We report the five-pair mean within each scene, then the equal-scene mean and a paired percentile interval from 10,000 resamples (seed 20260926). Each resample draws scenes with replacement separately within the illumination and viewpoint strata, retaining all five pairs and all conditions for each sampled scene. The development incidence gate passed on all 20 pairs; the held-out set contains 120 pairs. The separate raster-marker probe decoded $\phi(g) = 5g \bmod 36$ using declared label 1 and its inverse corrected the boundary adapter. A later public-boundary specificity check decoded the legal composed interface as identity.

Across held-out illumination and viewpoint scenes, wrong-minus-normal F1 differences were -0.229 and -0.315, respectively. At the image-pair level, 113 differences were negative, six zero, and one positive. The wrong filter retained fewer correct matches despite a mean match precision of 0.631 versus 0.618 under the normal filter; F1 captures the resulting retention loss.

B.4 Frozen decisions

For each model-dependent contrast, the paired unit is the base object and a mutant is detected only when the object-clustered 95% bootstrap lower bound for mutant-minus-correct error is strictly positive. The conventional common-mode test must detect at most 5%; grounding and both one-sided tests at least 95%; the non-homomorphic test must detect 100%. The baseline grounded false-positive rate must not exceed 1%. Direct eligibility requires positive lower bounds for both no-action-minus-correct and inverse-action-minus-correct. MACE also requires median force RMS at least 0.01 and the 95th percentile of relative energy error at most 10^{-10}. The consumer test must detect at least 95%, with pooled median ratio above two. Every v1 gate passed.

C COMPLETE RESULT SUMMARY

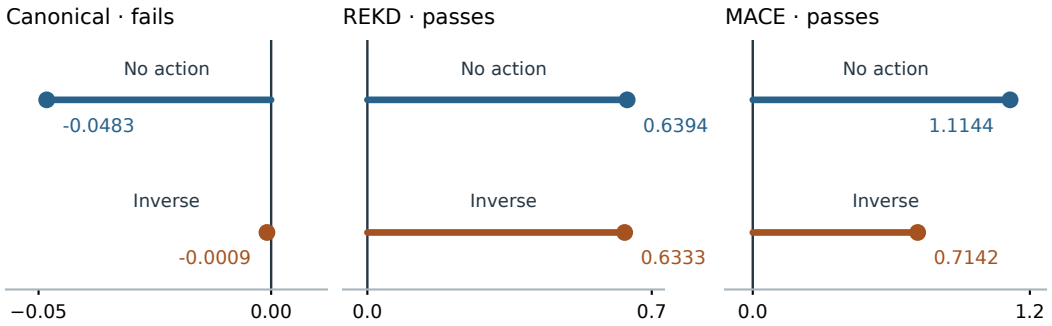

Figure 3: Checkpoint eligibility: the 95% object-clustered bootstrap lower bound for each paired control-minus-correct error. Both lower bounds must exceed zero. Each system has its own horizontal scale; only the sign relative to zero determines this gate. Canonical fails both comparisons.

Table 3: Direct-action diagnostics. Errors are normalized but use different system-specific outputs; cross-row magnitude comparisons are not meaningful.

System	Mean normalized error			95% CI lower endpoint	
	Correct	No action	Inverse	No−correct	Inverse−correct
Canonical	0.116386	0.068504	0.115570	−0.048252	−0.000902
REKD	0.090036	0.730920	0.724838	0.639436	0.633340
MACE	4.853×10^{-16}	1.127830	0.722825	1.114417	0.714207

Table 4: Consumer diagnostics for eligible systems. Ranges are across common-mode mutants; lower bounds come from 10,000 paired cluster resamples. Pooled ratio is the median of $(C_i(\phi(g), g) + 10^{-12})/(C_i(g, g) + 10^{-12})$ over mutants, objects, and labels. MACE's correct-condition error is near floating-point precision, so its 1.034×10^{12} ratio is dominated by the numerical floor and is not an effect-size estimate.

System	Detected	Excess-error range	Smallest 95% CI lower endpoint	Pooled ratio
REKD	11/11	0.550–0.678	0.548	9.314
MACE	23/23	0.784–0.958	0.770	1.034×10^{12}

Canonical's raw keypoints behaved mostly as invariant features on the frozen procedural population: 22 of 24 mean residual rows preferred the identity output label. The source action convention, checkpoint loading, and output selection were independently checked. Because the population is outside the ShapeNet surface regime and the direct gate failed as designed, we draw no conclusion about semantic-fault detection or upstream model quality from that system.

C.1 FINITE CERTIFICATE COMPARISON

Table 5: Post-result certificate comparison. "Queries" counts required label–probe queries in the independent semantic checks only. A dash means the method does not identify every candidate binding.

Method	queries	alternatives rejected	bindings decoded
All labels	60	34/34	36/36
Best singleton per group	2	32/34	–
Declared generators	3	34/34	36/36
Greedy full identification	3	34/34	36/36
Exact compiled	3	34/34	36/36

Exact external certificates. Table 5 reports the deterministic compiler audit. The one-label C_{36} certificate and two-label O_{24} certificate preserve 34/34 alternative rejection and decode the identity plus every alternative, for 36/36 candidate bindings total. The 60-to-3 reduction is certificate query count only: the literal inputs still exhaustively specify all 60 expected label states, the full group tables, and every candidate mapping. Moreover, both the obvious declared-generator and greedy baselines also use three queries. The exact method certifies optimality, validates the complete candidate family, binds the contract to a hash, and supplies a decoding margin and reusable candidate-conditional refusal logic. It offers no query-count improvement over those baselines, and we make no full-audit runtime claim. The reduced signatures decode all 36 bindings through the frozen REKD transform and MACE evaluator-harness action adapters.

C.2 PREDECESSOR-ONLY RECONSTRUCTION AUDIT

Predecessor-only reconstruction. After our implementation (the target) and its original results existed, we fixed a reconstruction study to test how readily known components could recover its behavior. An implementer in a fresh context received a neutral task of checking finite public-label bindings, public C_{36}/O_{24} contracts, and seven frozen predecessor sources, but no target code, recipe, tests, results, or the subsequently frozen D_4 contract. Separation was procedural, not externally owned or technically enforced. In the sole attempt, the comparator passed all 44 binding cases (12 C_{36}, 24 O_{24}, and 8 D_4) at independently verified minimum query counts 1/2/2 and passed all 86 specified behavioral checks. Six differences allowed in advance remained: on each contract it rejected invalid provenance where the target safely processed it, and abstained at the literal tolerance boundary where the target admitted identity. The unchanged target passed all eight D_4 bindings.

Table 6: Frozen reconstruction result. "Differences" are the two preauthorized conservative behaviors per contract, not failed behavioral checks. The D_4 holdout entry reports our unchanged implementation's result.

Contract	bindings	query count	checks	differences	holdout
C_{36}	12/12	1	26/26	2	–
O_{24}	24/24	2	38/38	2	–
D_4	8/8	2	22/22	2	8/8

The comparison is semantic rather than source or unrestricted-function equivalence. Four interface differences remain. First, comparator selection covers identity versus each nonidentity, whereas the target requires all-candidate injective signatures. Second, the comparator uses ℓ_∞ distance with a roundoff-width abstention band, whereas target decoding uses Euclidean distance with exact-match and half-margin rules. Third, comparator runtime validates the finite family supplied to it, whereas target contract loading also enforces family completeness. Fourth, the comparator counts calls to its downstream callback, whereas our implementation uses a gate within the test harness. On every contract, invalid provenance makes the comparator reject without a downstream call, whereas our implementation processes the input safely; the literal tolerance boundary yields comparator abstention but target identity admission. These limits constrain the comparison to the specified behavioral checks; they do not establish source or unrestricted-function equivalence.

C.3 HISTORICAL E3NN ISSUE 149 REPRODUCTION

The seed-0 witness error fell from 1.3889 at the last buggy parent to 2.78×10^{-16} after the completed repair. The witness failed 20/20 fixed construction seeds before repair and 0/20 afterward.

The post-result runner verifies commit and file hashes before execution. At the last buggy parent `b9631213a5286a3f5efab73f146ec0c539cd0da4`, `e3nn/o3.py` has SHA-256 `3f93a49950c4cfcb64ec9e0c73e556a8e858ba8e397dfc77020a44a1c25816ec`; that file is byte-identical at tag 0.1.0 (`2dc6848690033b958b699fa498c51466fc2a67d2`) (e3nn contributors, 2020). The first repair `fe34b205e53b53cad5ac9a063b3a826c949d091b` removes the invalid rowwise sign canonicalization but its new assertion has a multiplicity dimension error. At the completed repair `b6d9b51fbcd2fa0270a95d5ab328ff19cf61cbe8`, its SHA-256 is `b63e445d7e9399192a2ecc8722897e64ee4701bea9451d62fe1daa80bcc891bb`. The reference environment is Python 3.8.20, Torch 1.7.0+cpu, NumPy 1.24.4, SciPy 1.10.1, `lie_learn` 0.0.1.post1, SymPy 1.13.3, and pytest 7.4.4, with the real Torch Sparse 0.6.8 and Torch Scatter 2.0.5 CPU extensions. No compatibility shim, import stub, checkout edit, or numerical API replacement is used; the runner rejects a dirty source tree. The 13-node blind-spot test consists of three internal equivariance tests and ten nearby Cartesian/reduction tests and is executed at the buggy parent, not the release tag.

D ARTIFACT INTEGRITY AND REPRODUCTION

The anonymous material archive is a result-replay artifact. Its `README.md` gives the environment, commands, expected outputs, and scope; `MANIFEST.json` authenticates the included file bytes. The following paths are relative to the extracted archive:

- `checkpoint/` contains the REKD, Canonical Capsules, and MACE sufficient statistics, analysis configurations, and numerical decision projections. `scripts/replay_checkpoints.py` recomputes eligibility, mutation detection, and fixed-label consumer-proxy intervals. Canonical's failed eligibility and the original failed combined gate remain explicit.
- `contracts/` and `src/` contain the literal C_{36} and O_{24} declarations, finite compiler/decoder, evaluator adapters, and statistical functions. The compiler can synthesize and verify the one- and two-label certificates; this does not repeat checkpoint inference.
- `matching/` contains the fixed scene split, all 20 development and 120 held-out pair rows, candidate indices, inlier classifications, selection masks, recorded probe diagnosis, and analysis code. `scripts/verify_artifact.py` recomputes every saved mask's counts and F1. The scene analysis reproduces the paired intervals, including the zero homography-success contrast; homographies are not re-estimated.
- `scripts/make_figures.py` renders the manuscript figures from stored decisions with portable input and output paths.

For example, after installing the README's pinned dependencies:

```
python -B scripts/verify_artifact.py
python -B scripts/replay_checkpoints.py
python -B matching/analyze.py --task matching/heldout \
  --split held_out --output outputs/heldout.json
python -B scripts/verify_artifact.py
```

The verifier checks the artifact hashes and, when present, compares replay outputs with the stored scientific results. Original execution paths and internal authorization records were removed from the reviewer projections; numerical result fields and sufficient-statistics bytes are preserved.

`EXTERNAL_INPUTS.json` lists upstream source revisions, checkpoint retrieval pointers and exact input hashes, including the REKD checkpoint `47e9e70dfba7073bfbbe304a160e24f3bb97c82ad8a095f2c8b9f0d013b8b787`, MACE checkpoint `2ddb079cee0e131eaaf6912ba581b394551ead283e95c99cfe78c605d10b5736`, HardNet++ weights `9fe80a335fe72d1225fdd1ae0943418b12489f95014bd51dddc9819d6ffa33c7`, and full

HPatches archive `99cf7e1ca167896eb4ca2fe3d903beff6566243a3c0c8e07a13a2596648ad670`. Those assets and upstream software remain external under their original terms. The offline commands do not rerun model inference, feature extraction, the raster probe, source screening, the historical e3nn regression, or the predecessor comparator; their reported outcomes retain the scope and provenance described in the respective paper sections. Full inference replication requires the external assets, software, and compute specified in Appendix B.