

An Observed-Label Chirality-Dipole Null in 890,069 Quality-Controlled High-Confidence DESI Spirals and an 8.5-Million-Galaxy Catalog

HOUSTON GOLDEN¹

¹*Independent Researcher, Los Angeles, California, USA*

(Dated: July 20, 2026; Version v1.0.268)

ABSTRACT

We release observed chirality labels for 8,474,531 DESI Legacy DR8 galaxies and test one primary high-confidence observed-label dipole. Starting from 949,584 high-confidence spirals, we exclude 59,515 rows marked `raw_flip_qc_unsafe`, leaving 890,069 quality-controlled rows; 887,472 enter the supported-pixel fit and fixed-occupancy label-randomization null. The result is consistent with zero ($z_{\text{mom}} = +0.635$, one-sided rank $p = 0.23768$). A coverage-calibrated observed-label injection-recovery on this primary channel places a 95% sensitivity upper limit $A_{95}^{\text{obs}} \simeq 0.98\%$ (full-amplitude); this is an observed-label sensitivity floor, not a physical parity-amplitude bound, which remains gated on the morphology transfer function. The content-addressed release includes the science catalog, unsafe-row quarantine, retained primary-null array, schema, checksums, and reproducer. WLS and harmonic analyses use different supports or nulls and are retained only as systematics diagnostics. A shared block-bootstrap covariance of the primary estimators, now NaMaster-complete with the MASTER-decoupled $\ell = 1$ leg, shows the monopole to be a statistically distinct mode nearly uncorrelated with any dipole/harmonic estimator. A classifier-injection forward model over 1.7×10^7 banked network passes excludes classifier confusion as the source of that monopole (0.0% of the observed value) and localizes its origin upstream of the classifier, without resolving whether that origin is a true sky asymmetry or a DESI imaging systematic. A from-scratch, manifest-retained retrain of the GZ1-core classifier component—retaining every object, split index, and random seed—validates on a provably training-disjoint high-confidence GZ1 sample (Cohen’s $\kappa = 0.97$). The CE-ResNet catalog has since been re-provisioned and the full historical composition regenerated under seeded assembly, reproducing the 6,637 GZ1 and 17,153 CE-spiral counts exactly and isolating the historical 826-vs-846 record conflict to a seeded non-spiral subsample crossmatch (reproducible count 819); but a composition-faithful CE-included retrain collapses to chance on chirality, so the historical CE-included accuracy is *not* reproducible under honest ingestion, and the released catalog labels are unchanged. Spatial transfer calibration, a full joint likelihood, an independent matched-footprint estimator, a complete systematics-metadata sidecar, and a DOI-backed archive remain open. The parity-even morphology observable supports no primordial-parity bound.

Keywords: galaxies: spiral, galaxies: statistics, methods: data analysis, cosmology: observations, large-scale structure of universe

1. INTRODUCTION

The handedness (chirality) of spiral galaxies—whether their arms trail clockwise (CW) or counter-clockwise (CCW) as projected on the sky—is a simple observable that, under the trailing-arm assumption and in the absence of confounding selection effects, traces the angular-momentum direction of each disk galaxy. Throughout this paper, “CW/CCW” refers to the *pro-*

jected apparent arm-winding chirality, not a deprojected 3D spin vector. In a statistically isotropic and parity-symmetric universe, the CW and CCW fractions should be exactly equal when averaged over large angular scales. A significant directional departure would constrain isotropy-breaking axial-vector sectors in the galaxy-formation chain. The present paper is a standalone observational result: the observed-label null on

the declared sample does not depend on any unpublished companion work.

Claims of such a signal have appeared intermittently in the literature. Shamir (2012) (1) reported a $2\text{--}4\sigma$ dipole with per-bin asymmetry amplitudes of $\sim 5\text{--}20\%$ (as reported in that work) using $\sim 1.27 \times 10^5$ SDSS galaxies. Shamir (2020) (2) reported asymmetries at the reported $\sim 2\text{--}4\%$ level on SDSS and Pan-STARRS samples; Shamir (2022a) (3) reported related spin-direction alignment analyses; and Shamir (2022b) (4) reported results on a DESI Legacy sample (“nearly 1.3×10^6 spiral galaxies” per the published abstract). Iye *et al.* (2021) (5) re-examined Shamir’s SDSS spiral catalog using 3D random-walk simulations and found no significant dipole after correcting for reading-direction bias and photometric-object duplication in earlier Shamir catalogs. Tadaki *et al.* (6) likewise found null results on a catalog of $\sim 80,000$ face-on spirals. Jia *et al.* (7) introduced CE-ResNet, a chirality-equivariant CNN guaranteeing by construction that flipping an input exactly swaps CW and CCW outputs, with a reported number-count ratio $\text{CW}/\text{CCW} = 0.998$ on ~ 1.95 million galaxies.

In this paper we present an 8.47-million-object public chirality catalog release whose novelty relative to CE-ResNet is scale, a dedicated NOT_SPIRAL class, a machine-readable release contract, and a multi-axis bias-hardening audit. This catalog-and-methods emphasis, including a compact schema, executable selection, checksummed null, and explicit quarantine, is the intended ApJS venue fit. Committed historical training records conflict on exact composition: one reconstruction gives 26,616 rows and 826 CE-ResNet-selected non-spirals (67.5% CE-derived), whereas BENCHMARK_REPORT.md gives 26,626 and 846; both accounts establish substantial CE-ResNet dependence, so the catalogs are not fully independent and model-output permutation nulls cannot test inherited survey-correlated structure by themselves. A subsequent from-scratch, manifest-retained retrain makes the *GZ1-core* component of the training realization regenerable going forward (Sec. 2). The CE-ResNet catalog has since been re-provisioned and the composition conflict adjudicated—the seeded assembly reproduces the 6,637 GZ1 and 17,153 CE-spiral counts exactly and isolates the 826-vs-846 record conflict to a non-spiral subsample crossmatch (reproducible count 819)—but a composition-faithful CE-included retrain collapses to chance on chirality, so the historical CE-included accuracy is not reproducible under honest ingestion (Sec. 2); the released catalog labels are unaltered. Our sole primary result is the quality-controlled HC real-space observed-label null ($N_{\text{selected}} = 890,069$,

$N_{\text{support}} = 887,472$, $z_{\text{mom}} = +0.635$, one-sided upper-tail $p = 0.23768$). The WLS and MASTER results are systematics-sensitive diagnostics, not independent detections or physical exclusions. The GZ1 human-only test uses human votes rather than learned chirality labels and corroborates the null at lower statistical power; it does not tighten the HC amplitude constraint.

2. DATA

2.1. Galaxy Images

Our parent sample is the Smith42/galaxies dataset on HuggingFace (<https://huggingface.co/datasets/Smith42/galaxies>), whose training split at immutable revision `bdd1b063a9a22874a79a4363aa9fb6a2b356a4c2` contains 8,474,566 galaxy images from the DESI Legacy Imaging Surveys DR8 (8). Each image is a 224×224 pixel cutout in *grz* bands at $0.262''/\text{pixel}$. The dataset provides `dr8_id` identifiers; sky coordinates are obtained by cross-matching against the Galaxy Zoo DESI predictions catalog (9). The parent-sample selection function inherits from Galaxy Zoo DESI: photometric types REX/DEV/EXP/SER, $r \leq 19.0$, half-light radius $\geq 3''$. DR8 comprises three distinct imaging campaigns: BASS+MzLS ($\delta > +32^\circ$), DECaLS ($\delta < +32^\circ$), and a DES overlap region.

2.2. Training Labels

The exact historical training realization is not fully recoverable. The production-script reconstruction and immutable model-repository receipt `v2_bias_audit.json` (model revision `237d021c451d75cf86a875e86d4de498b74e2f12`; file SHA-256 `57278e4e...ee847`) support 26,616 labeled rows (6,637 Galaxy Zoo 1 high-confidence CW/CCW, 17,153 CE-ResNet high-confidence spirals, and 2,826 not-spiral rows, of which the script account assigns 826 to CE selection and 2,000 to synthetic examples), a 21,293/5,323 train/validation split, and 93.6878% validation accuracy. However, the committed BENCHMARK_REPORT.md gives 26,626 total rows and 846 CE non-spirals, while the committed Hugging Face README reports 92.10%. No retained object/split manifest, random-state record, or run receipt ties the named objects and stochastic realization to the historical checkpoint, so the immutable audit supports one historical account but does not resolve the conflicting committed records or make that original run exactly reproducible (Appendix B). This gap can harbor training- or selection-dependent bias; TTA guarantees only flip-equivariance of the output protocol and does not make the unknown historical training realization irrelevant. Both pool accounts are approximately 67.5%

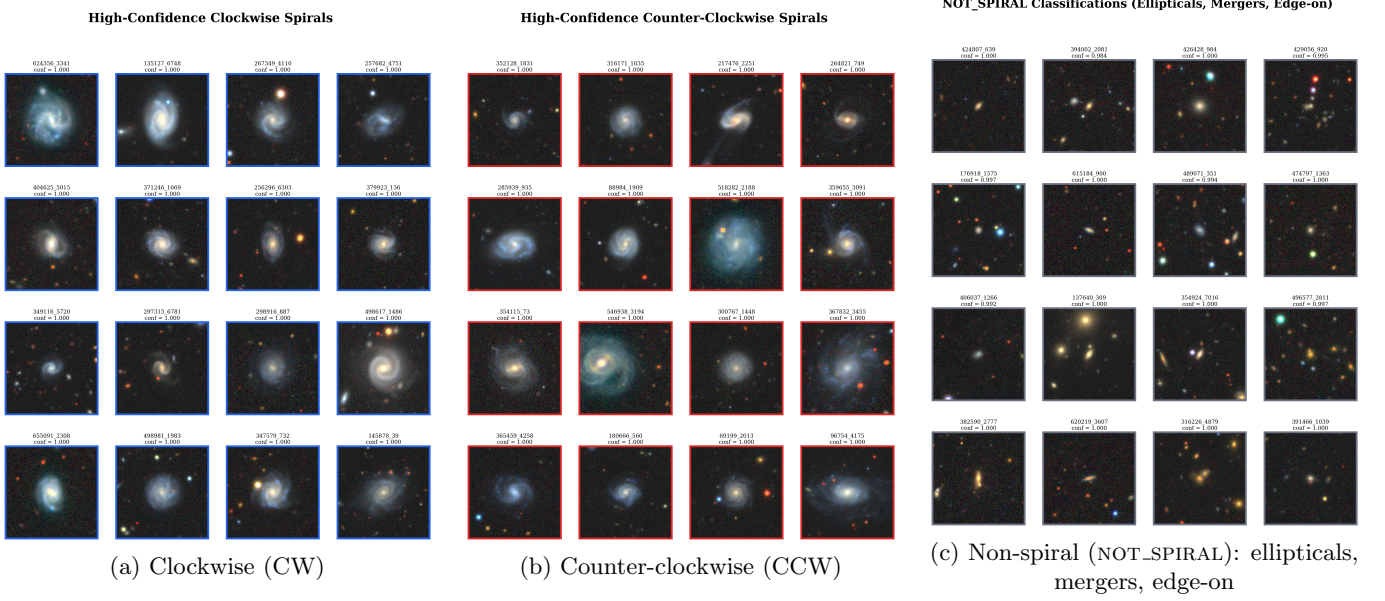


Figure 1. Representative high-confidence galaxies from the classified catalog ($p_{eq} > 0.9$). Left: clockwise (CW) spirals; center: counter-clockwise (CCW) spirals; right: non-spiral (NOT_SPIRAL) objects — ellipticals, mergers, and edge-on galaxies that would contaminate a binary spiral classifier. All cutouts are 224×224 pixels in grz bands from DESI Legacy DR8. The gallery illustrates the visual diversity within each class and motivates the three-class design: without an explicit NOT_SPIRAL class, $\sim 62\%$ of the parent sample would leak into the spiral classification. The $ViT - Small$ classifier resolves CW vs. CCW via the test-time equivariant averaging procedure of §3.4.

CE-derived, so validation substantially measures agreement with that model. Horizontal flips are generated on the fly for a training-only consistency loss and do not add dataset rows. *A from-scratch, manifest-retained retrain now provides a fully regenerable realization of the GZ1-core component going forward, narrowing—but not eliminating—this gap.* That retrain retains every object ID, train/validation split index, and random state (manifest `pipelines/p2_chirality/outputs/g1_retrain/g1_training_manifest.json`, SHA-256 `e5de8e03...240a10729`), reaching 99.31% best-epoch three-class validation accuracy (checkpoint SHA-256 `aed109dc...a752387`; backed up to the `bamfai/galaxy-chirality-v2` model repository at revision path `g1-retrain-2026-07-17/` with the honesty note and hashes in `pipelines/p2_chirality/outputs/g1_retrain/PROVENANCE.md`). This realization is deliberately scoped to the 6,637 Galaxy Zoo 1 confident CW/CCW spirals plus 2,000 synthetic NOT_SPIRAL rows only ($n_{total} = 8,637$); the CE-ResNet high-confidence spiral component (`ce_resnet_present = false` in the manifest), which supplied roughly 67.5% of the historical composition, is absent from that GZ1-core realization. That realization therefore does *not* by itself resolve the 826-vs-846 CE-non-spiral count conflict or reproduce the full historical composition (a separate CE-included assembly and retrain, described in the next paragraph, re-provision the

catalog and adjudicate the conflict), and—critically—it does *not* alter the released Catalog C labels, which remain the historical production outputs: the retrain demonstrates regenerability of the GZ1 core, it does not replace what was released. The full GZ1 overlap gives 69.91% chirality agreement ($\kappa = 0.40$) for the released classifier; while the historical training script retained no object-ID manifest to establish exclusion of the 6,637 GZ1 training rows from that figure, the retained retrain manifest now enables a provably training-disjoint held-out GZ1 validation of the regenerable realization (Sec. B, Cohen’s $\kappa = 0.9733$ on 3,000 confident held-out spirals), a distinct measurement discussed in full below. The human-vote GZ1 dipole test gives $z = -0.54$ for 46,017 matched inputs; exactly 4,963 enter its 394-pixel support after the legacy $N_{pixel} > 10$ cut, while 41,054 are excluded. It corroborates a null under its legacy null convention but, without calibrated power or a spatial transfer model, does not provide a tighter amplitude bound.

CE-composition adjudication and a composition-faithful honest negative. The CE-ResNet catalog has since been re-provisioned (Jia et al. 2023 CE-ResNet; Zenodo DOI 10.5281/zenodo.7167388; file SHA-256 `894dbe88...0e304b`) and the full historical composition regenerated under the identical seeded wrapper assembly (seed 42; `gz_desi` revision `b7583bb2...`; driver `pipelines/p2_chirality/scripts/g1_ce_`

`composition_assembly.py`). The deterministic assembly reproduces the two large historical components *exactly*—6,637 GZ1 confident CW/CCW and 17,153 CE-ResNet spirals—and yields 819 CE non-spirals (neither the audit’s 826 nor the benchmark’s 846), for 26,609 rows (class counts CW/CCW/NOT_SPIRAL=11,904/11,886/2,819; train/val 21,288/5,321; manifest `pipelines/p2_chirality/outputs/g1_full_composition/g1_full_composition_manifest.json`, SHA-256 431f84f0...cf110777). Because $6,637 + 17,153 + 826 + 2,000 = 26,616$ reproduces the smaller historical record’s arithmetic exactly, the entire 826-vs-846 / 26,616-vs-26,626 conflict is now *isolated* to the seeded 50,000-object non-spiral subsample crossmatch—precisely the subsample/boundary ambiguity Table 13 already flags—rather than left unresolved for want of the catalog. Crucially, a composition-faithful CE-included retrain of the same 26,609-object realization, using the same wrapper code and the same seeds, does *not* reproduce the historical 93.69%/92.10% CE-included headline: it collapses to chance on chirality, reaching best three-class validation accuracy 0.5617—the exact arithmetic of a perfect NOT_SPIRAL class at 10.6% prevalence plus coin-flip chirality on the remaining 89.4% ($0.106 + 0.894 \times 0.5 = 0.553$), with training accuracy pinned at 0.667 for all 26 epochs—and even the clean held-out GZ1 spirals sit at binary CW/CCW agreement 0.517 (`pipelines/p2_chirality/outputs/g1_full_composition/pod_evidence/g1_per_source_eval.json`). This is a genuine honest negative, not a wrapper defect, with four alternative explanations ruled out by evidence (Appendix B): CE argmax labels are decisive at the catalog level (only 0.2% of confident spirals have $|P_{CW} - P_{ACW}| < 0.1$; median margin 0.62); CE and GZ1 share an *identical* chirality convention (model-independent cross-match agreement 0.9972 on the 38,617-galaxy bright confident overlap); the image join is the same 3" k -d-tree crossmatch that trains the GZ1 core to 0.99; and an identical-optimizer GZ1-only run broke out of the same 0.66 plateau by epoch 5 to 0.9931. The root cause is that the CE-only pool—72% of the spiral labels—is systematically fainter (median $r = 16.89$ vs. 16.15), smaller (median 4.26" vs. 5.77"), and low-confidence (median CE winner-probability 0.569 vs. 0.899): chirality supervision a 224 px-resized ViT cannot learn from, which swamps the clean GZ1 gradient and prevents the chirality feature from forming for *all* spirals. This negative *strengthens* the disclosure above: the historical CE-included realization is not reproducible under honest, seeded, same-composition ingestion, corroborating that validation on that pool substantially measures

agreement with the CE model. The manifest-retained GZ1-core retrain (99.31%) stands as the honest regenerable realization, and the released Catalog C labels remain unchanged historical production outputs (full diagnosis `pipelines/p2_chirality/outputs/g1_full_composition/DIAGNOSIS_ce_included_retrain.json`).

3. METHODS

3.1. Notation and Significance Conventions

Confidence notation.— p_{eq} (or p_{CW}^{eq} , p_{CCW}^{eq} , p_{NS}^{eq}) denotes the max-class equivariant probability output by the 2-fold flip-TTA pipeline (Sec. 3.4 and Appendix B). These are monotone ranking scores, not frequentist probabilities; “high-confidence” (HC) cuts on p_{eq} are sample-selection thresholds.

Significance conventions.—All significance values z are moment-ratios $(x - \langle x \rangle_{null}) / \sigma_{null}$ against the null distribution specified per result. The primary empirical rank value is the one-sided upper tail $p = (k + 1) / (N + 1)$, where k null amplitudes meet or exceed the data amplitude. Three distinct significance conventions appear in this paper:

- *Moment- z / rank- p :* the real-space dipole pair ($z_{mom} = +0.635$, one-sided upper-tail rank $p = 0.23768$) are distinct summary statistics of the A_{dip} distribution against the fixed-occupancy galaxy-label-randomization null; they do not follow the Gaussian $z \rightarrow p$ mapping.
- *MASTER moment- z :* on the single checksummed 24,087-pixel FS-C support, the fixed-occupancy 500-draw null gives $z_{\ell=1} = +6.923$ (add-one rank $p = 0.001996$), while the distinct binomial-monopole null gives $z_{\ell=1} = +6.983$ for 500 draws and $+7.207$ for 10^4 draws. These are moment ratios, not Gaussian tail significances; the two null families answer different questions and their z values must not be converted through a Gaussian $z \rightarrow p$ map.
- *Block-bootstrap z :* $(A_{dipole}^{best} - A_{ref}) / \sigma_{boot}$ with spatial-coherence-corrected σ_{boot} ; $z \approx -7.6$ applies only to a clean *observed-label* $A_{ref} = 0.017$ template. Under $g = 0.398$, a physical 1.7% dipole maps to 0.6766% observed and is $1.36\sigma_{boot}$ above the fit, so the former physical-exclusion reading is invalid.

3.2. Declared Analysis Hierarchy

We assign stable names to the distinct supports used below: HC-REALSPACE-INCLUSIVE (HC-RI; quality-controlled HC rows with $N_{spiral}(p) \geq$

10, 23,633 pixels), FULL-SPIRAL-CANONICAL (FS-C; all spirals with $N_{\text{spiral}}(p) \geq 10$, 24,087 pixels), and MASTER-ALL-GALAXY-FOOTPRINT (MASTER-AGF; $N_{\text{all}}(p) \geq 1$, 24,297 pixels). The primary measurement starts from the $p_{\text{eq}} > 0.6$ HC selection and excludes all 59,515 rows marked `raw_flip_qc_unsafe` (rows carrying a raw/equivariant pipeline-pass mismatch, for which the reconstructed flip probabilities fall outside $[0, 1]$ by up to 0.09—a pipeline-pass provenance defect, not float32 rounding; see the raw/flip quarantine paragraph in Sec. 4.1 and Appendix B), leaving $N_{\text{selected}} = 890,069$; $N_{\text{support}} = 887,472$ enter the 23,633 supported pixels. No unsafe row enters the primary calculation. We declare this quality-controlled HC subsample as the *single primary science sample*. The historical WLS/bootstrap fit used a different latitude-cut support and is retained only as provenance, not as an FS-C or physical exclusion. This is an internal hierarchy declaration, not a claim of formal preregistration or independent blinding.

- **Estimator hierarchy:** the primary cosmological estimator is the quality-controlled real-space CW-fraction dipole fit on Catalog C at NSIDE=64 (moment-ratio $z_{\text{mom}} = +0.635$ against the fixed-occupancy galaxy-label-randomization null, one-sided upper-tail empirical-rank $p = 0.23768$). The block-bootstrap WLS fit remains diagnostic.
- **Secondary diagnostic estimators:** on the exact checksummed 24,087-pixel FS-C support, the 500-draw fixed-occupancy label-randomization null gives binary-mask $z_{\ell=1} = +6.923$ and C^2 2° apodized $z_{\ell=1} = +7.033$; the binary multipole vector is $z_{\ell=1,\dots,5} = (+6.923, +2.931, +3.089, -0.317, +1.606)$. These are systematics diagnostics, not cosmological significances.
- **Generative monopole-only null:** independent $\text{Binomial}(N_{\text{spiral}}(p), p_{\text{CW}})$ draws on the same exact FS-C support give post-MASTER $z_{\ell=1} = +6.983$ for 500 draws and $+7.207$ for 10^4 draws. This null does not condition on the observed total CW count and is distinct from fixed-occupancy label randomization.
- **Finite-grid injection scores:** (vii) on the HC-broad spiral subsample ($p_{\text{eq}} > 0.6$, $N = 949,584$; Sec. 6.2), the 100-injection score-pass fractions at $A = 0.75\%$ are 0.55 (polar-angle-uniform axes) and 0.59 (area-uniform axes), while the Stage-B deterministic-axis surrogate gives 16/20; none is a calibrated recovery threshold.

To make the mapping from estimator to scientific claim explicit, Table 1 ties each claim to its estimator, sample, null, and role. Only the HC real-space row marked PRIMARY carries scientific weight; WLS and every harmonic-channel statistic are diagnostics and support no independent detection or exclusion claim.

Table 2 consolidates the distinct estimators and nulls. Because these rows use different samples, supports, and null procedures, their σ values are not converted through a common Gaussian mapping. A shared block-bootstrap covariance of the estimators (real-space dipole, WLS dipole, monopole, and the NaMaster MASTER-decoupled $\ell = 1$ power), evaluated on one common resampling of the primary HC sample, is now computed (Sec. 4.5, Table 9; artifact `pipelines/p2_chirality/outputs/canonical_provenance/g3_joint_estimator_covariance_master_v2.json`) and shows the monopole to be a statistically distinct mode nearly uncorrelated with any dipole/harmonic estimator; the MASTER-decoupled $\ell = 1$ leg has since been added (the covariance is now NaMaster-complete, 4×4), while a full joint likelihood across all channels remains open. The scientific verdict rests solely on the HC observed-label real-space row; WLS and all harmonic rows remain systematics diagnostics.

3.3. Model Architecture

The classifier consists of a ViT-Small encoder (10) (`vit_small_patch16_224`, ImageNet-pretrained) with the last 6 of 12 transformer blocks fine-tuned, followed by a custom classification head:

$$\begin{aligned} \text{LayerNorm} &\rightarrow 384 \rightarrow 512 \text{ (GELU, } d=0.3) \\ &\rightarrow 512 \rightarrow 256 \text{ (GELU, } d=0.2) \rightarrow 256 \rightarrow 3 \text{ (softmax)}. \end{aligned} \quad (1)$$

The three-class output (P_{CW} , P_{CCW} , P_{NS}) is essential for full-survey deployment: applying a binary classifier to data where $\sim 70\%$ of objects are elliptical or irregular produces a catalog dominated by noise. Full architecture and training details are in Appendix B.

3.4. Test-Time Equivariant Averaging

At inference, each galaxy is classified on both the original image and its horizontal reflection. The equivariant probability is:

$$\begin{aligned} P_{\text{CW}}^{\text{eq}} &= \frac{1}{2} (P_{\text{CW}}^{\text{orig}} + P_{\text{CCW}}^{\text{flip}}), \\ P_{\text{CCW}}^{\text{eq}} &= \frac{1}{2} (P_{\text{CCW}}^{\text{orig}} + P_{\text{CW}}^{\text{flip}}), \\ P_{\text{NS}}^{\text{eq}} &= \frac{1}{2} (P_{\text{NS}}^{\text{orig}} + P_{\text{NS}}^{\text{flip}}). \end{aligned} \quad (2)$$

This procedure enforces flip-equivariance of the output protocol (flip-swap correlation = 1.000). We restrict to

Table 1. Estimator decision tree. The single PRIMARY row is the HC observed-label real-space null. DIAGNOSTIC rows characterize systematics and make no independent detection or exclusion claim; the pilot row reports finite-grid sensitivity evidence. Statistics across rows are not commensurable because their fields, samples, and nulls differ.

Scientific claim	Estimator	Sample (N_{spiral})	Null	Role
Real-space chirality dipole is consistent with zero	(i) QC HC real-space dipole	$p_{\text{eq}} > 0.6$, unsafe rows excluded (890,069)	fixed-occupancy label randomization (10^4)	PRIMARY
A clean observed-label 1.7% template is disfavored under the WLS model	(ii) block-bootstrap WLS template fit	full Catalog C (3,201,160)	block-bootstrap (10^3)	DIAGNOSTIC
Harmonic $\ell=1$ residuals are systematics (monopole-mask leakage + coherent depth/morphology), <i>not</i> a cosmological signal	(iii,iv) canonical/apodized MASTER; (vi) monopole-only generative null	full Catalog C (3,201,160)	label-shuffle / depth-stratified / monopole-only	DIAGNOSTIC
The hemisphere max-statistic rejects the random-label null after look-elsewhere correction, but the excess is attributed to survey/classifier systematics and is not interpreted cosmologically	(v) hemisphere max-asymmetry	full Catalog C (3,201,160)	max-statistic MC (p_{LEE})	DIAGNOSTIC
Finite-grid sensitivity of the primary real-space estimator; pilot fractions (a calibrated coverage bound is row viii)	(vii) injection-recovery	HC, $p_{\text{eq}} > 0.6$ (949,584)	per-pixel shuffle	PILOT
Coverage-calibrated <i>observed-label</i> 95% sensitivity upper limit on the primary dipole amplitude, $A_{95}^{\text{obs}} \simeq 0.98\%$ (<i>not</i> a physical parity bound)	(viii) random-axis injection-recovery	$p_{\text{eq}} > 0.6$, unsafe excluded (890,069)	fixed-occupancy label randomization (10^4)	PRIMARY SENS.

2-fold TTA (original + horizontal flip) rather than the full D_4 group because mirrors flip chirality by definition, whereas in-plane rotations do not change chirality; rotation-TTA probes classifier non-equivariance rather than the chirality assignment itself. Because Eq. (2) enforces flip-equivariance at inference by construction, the flip-swap consistency test T1 (Appendix B) is a *protocol implementation check* that guards against code defects in the TTA averaging, not an independent statistical test of the learned model’s equivariance; the pre-TTA behavior of the underlying network is documented separately by the historical raw single-pass Catalog A diagnostic (Table 4). A direct D_4 -TTA hold-out on two $\sim 2,000$ -galaxy subsamples finds the mean per-galaxy P_{CW} is stable under Z_2 and D_4 to within $|\Delta\langle p_{\text{CW}} \rangle| < 0.0016$; per-galaxy argmax labels flip in 21.4% of cases between Z_2 and D_4 on borderline galaxies with $P_{\text{CW}} \approx P_{\text{CCW}} \approx 0.4$. This D_4 hold-out is a classifier-stability check (comparing softmax outputs

under two augmentation groups), not a spatial-null or isotropy test; spatial-null calibration is performed via the isotropic (pixel-)permutation nulls of Sec. 4.3. Full details in Appendix B.

3.5. Catalog Products and Historical Diagnostics

The labels **Catalog A** (raw, single-pass softmax) and **Catalog B** (Platt-calibrated) denote historical pipeline diagnostics, not released catalog tiers. Only **Catalog C** (equivariant production, 2-fold flip TTA) is published as the science-facing 8,474,531-row Apache Parquet catalog and underlies the primary and downstream science-facing analyses in this paper. Catalog A is retained below solely as a supported before/after methodology diagnostic; no complete Catalog B payload or executable release contract is claimed.

4. RESULTS

Significance conventions. This paper reports significance in standard-deviation units (σ) throughout, but

Table 2. Primary-estimator summary. $N_{\text{catalog spiral}}$ is the underlying Catalog C spiral count (CW+CCW only). $N_{\text{map weighted}} = \sum_{p \in \text{mask}} W_p$ where $W_p = N_{\text{all}}^{(p)}$ is the total classified-galaxy count in pixel p (CW+CCW+NS), used as a survey-depth weight in the NaMaster field object. $N_{\text{map weighted}}$ exceeds $N_{\text{catalog spiral}}$ because W_p includes non-spiral galaxies (~62% of the catalog); each galaxy is counted once. Row (iv) applies a C^2 , 2° apodization to the same exact 24,087-pixel FS-C base mask as row (iii), with the same 500-draw fixed-occupancy null. *The σ values in different rows are computed against different null procedures (column “Null”) and are not directly comparable across rows as detection significances.* Row (v) reports the post-look-elsewhere-corrected significance; the raw direct-MC value is $p_{\text{LEE}} = 1/10001$; $95\% \text{upperlimit} \simeq 3.0 \times 10^{-4}$ against the random-label max-statistic null, which *already incorporates* the look-elsewhere scan and is the principled directional look-elsewhere control (Appendix C); the rejection is systematics-attributed. Row (vii) uses the HC-broad subsample ($p_{\text{eq}} > 0.6$; Sec. 6.2). A “—” in $N_{\text{map weighted}}$ marks rows where the NaMaster weight construction is not invoked.

Estimator	$N_{\text{catalog spiral}}$	$N_{\text{map weighted}}$	f_{sky}	Mask	Null	Reported statistic
(i) real-space dipole (QC HC)	890,069	—	0.4808*	HC-RI	fixed-occ. label rand. (10^4)	+0.635; $p = 0.23768$
(ii) WLS template diag.	3,201,160	—	0.49209	latitude cut (24.187 px)	block-boot. (10^3)	$z \approx -7.6$
(iii) canonical MASTER (diag.)	3,201,160	—	0.49005	exact FS-C	fixed-occ. (500)	+6.923; $p = 0.001996$
(iv) apod. MASTER (diag.)	3,201,160	—	0.48399 eff.	same exact FS-C base mask	fixed-occ. (500)	+7.033; $p = 0.001996$
(v) hemisphere LEE (MC)	3,201,160	—	0.49005	FS-C	max-stat MC	$p_{\text{LEE}} = 1/10001$; $95\% \text{upperlimit} \simeq 3.0 \times 10^{-4}$ (syst.-attr.)
(vi) post-MASTER monopole null	3,201,160	—	0.49005	exact FS-C	binomial-monopole	+6.983/ + 7.207 (500/ 10^4)
(vii) finite injection grid	949,584 HC	—	—	—	pp-shuffle	$P_{3\sigma} = 0.55/0.59$ at $A = 0.75\%$
(viii) random-axis coverage (obs.-label)	890,069	—	0.4808	HC-RI	fixed-occ. label rand. (10^4)	$A_{95}^{\text{obs}} \simeq 0.98\%$ (rank $p < 0.05$, 95% cov.)

*HC-RI contains 23,633 of 49,152 pixels after excluding all 59,515 `raw_flip_qc_unsafe` HC rows.

Table 3. Estimator callout. P1 is the load-bearing cosmological result. P2 and D1–D2 are diagnostics with distinct samples and nulls; P2’s $z \approx -7.6$ applies only to an observed-label template and becomes a $1.36\sigma_{\text{boot}}$ comparison for a physical 1.7% signal under $g = 0.398$. Values across rows are not directly comparable as detection significances (Sec. 3.1).

Claim / channel	Null	Value
P1 real-space QC-HC dipole	fixed-occupancy	+0.635 moment- z
($p_{\text{eq}} > 0.6$, unsafe excluded), primary	label randomization	(upper-tail $p = 0.23768$)
P2 observed-label template	block-	$z \approx -7.6$
(WLS fit), diagnostic	bootstrap	(physical: 1.36σ)
D1 HC dipole (cross-check)	label-shuffle	$z = +0.70$, $p = 0.230$
D2 MASTER $\ell = 1$ residual	fixed-occ./binomial	+6.923/ + 7.207 moment- z
systematics, not a detection	(harmonic)	

values from distinct null procedures are *not* directly comparable. Section identifiers specify the null for each result. Empirical rank p -values are one-sided unless explicitly labeled two-sided.

4.1. Catalog Statistics

The final catalog contains 8,474,531 galaxies, 35 fewer than the immutable 8,474,566-row Smith42 training split. The historical inference runner retained neither a per-object failure ledger nor reason codes, so the identities of those absent rows and their exact failure reasons are unavailable; we do not characterize them as having failed a specific quality criterion. Catalog C (equivariant) contains CW 1,592,107, CCW 1,609,053, and NS/edge-on 5,273,371, for $N_{\text{spiral}} = 3,201,160$. Mean max-class score is 0.951 and the median is 0.9997. The score distribution is much sharper than the full-overlap GZ1 descriptive accuracies (58.7% three-class; 69.91% spiral chirality), so every p_{eq} quantity in this paper is an uncalibrated ranking score. The HC threshold selects a sample; it does not assert a per-object correctness probability.

The primary estimator consumes hard argmax CW/CCW counts rather than soft scores. Monotone recalibration therefore leaves the labels, pixel counts,

fitted dipole, and direction unchanged for a fixed selected sample, although the spatial selection induced by a threshold can correlate with depth, morphology, seeing, redshift, or footprint. The confidence-cut sweep and external template diagnostics test that selection dependence; the human-vote GZ1 dipole corroborates the null independently of softmax scaling. The illustrative scalar dilution $g = 2a - 1$ uses the full-overlap GZ1 agreement, which is not an independent calibration because the training rows were not removed.

Raw/flip quarantine. Reconstructing the flip pass from mismatched raw/equivariant pipeline products identifies 249,066 unsafe rows catalog-wide (2.939%). Exactly 59,515 occur in the original 949,584-row HC selection. Because the release marks these rows `do_not_use_for_science=True`, the primary estimator now applies the executable predicate `primary_hc && !raw_flip_qc_unsafe`. This selects 890,069 rows and explicitly selects zero unsafe rows. The exact fixed-occupancy rerun is null-consistent ($z_{\text{mom}} = +0.635$, upper-tail add-one $p = 0.23768$); it replaces, rather than supplements, the unsafe-inclusive primary result. The selection mask, array, generator, and receipt are check-summed in the computational-closure ledger.

4.2. Global CW Fraction

The Catalog C residual (-9.5σ from 0.5000, Table 4) is bounded below 0.5% in the equal-count slab summaries but has documented equal-area and low- ℓ structure; it does not produce a dipole under the declared primary estimator. This monopole offset is consistent with a classifier or imaging artifact rather than evidence for a physical signal; three candidate mechanisms are (1) GZ1 training-label CW excess, (2) residual orientation-dependent bias not corrected by 2-fold TTA, and (3) photometric asymmetry in DESI Legacy imaging. A direct classifier-injection

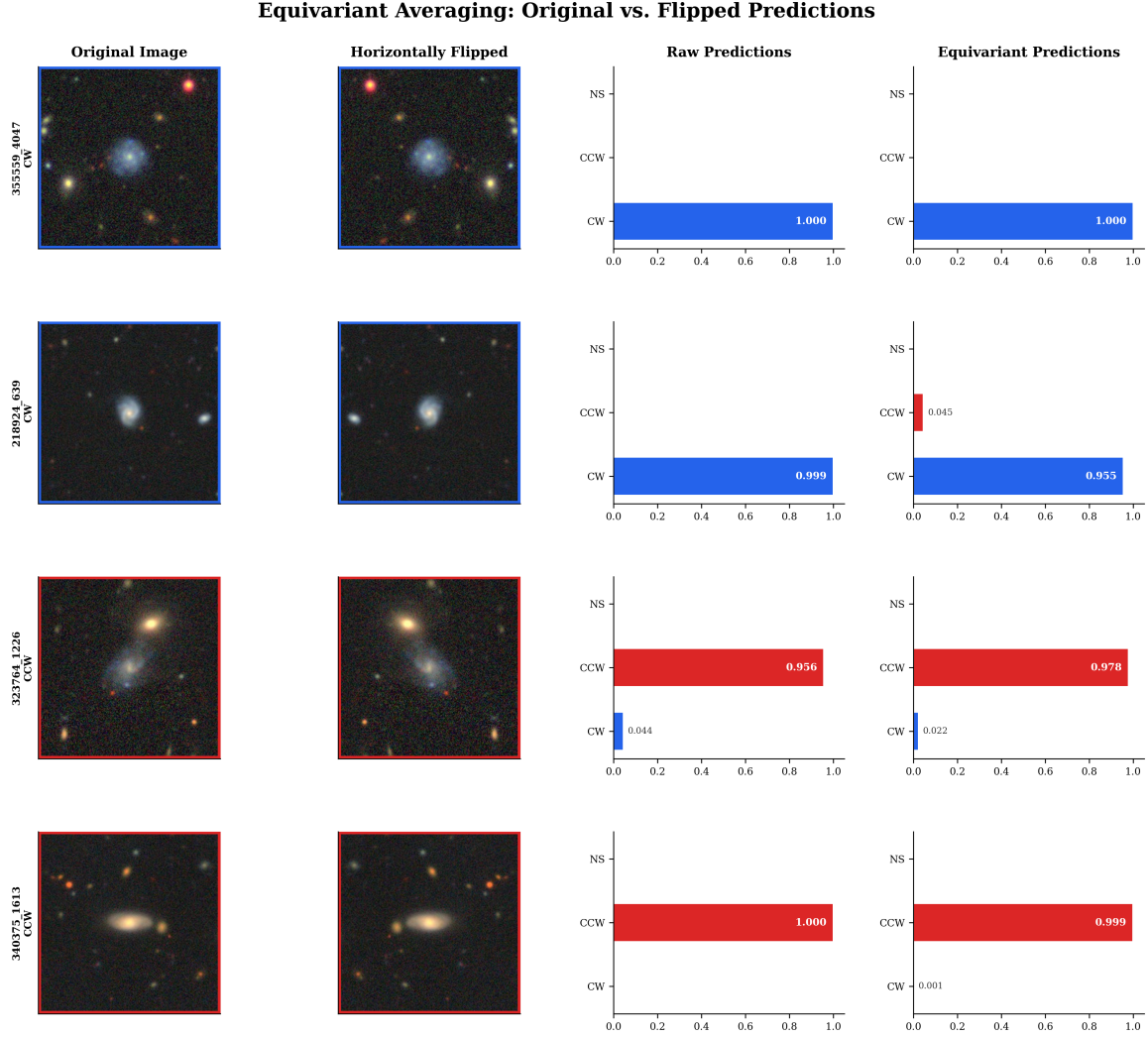


Figure 2. Equivariant test-time averaging (TTA). Representative Z_2 production TTA examples (original + horizontal flip); D_4 validation (four rotations $\times$ two reflections) in Appendix B. *Production* inference (Catalog C) uses 2-fold Z_2 TTA — original + horizontal flip only (§3.4). Flips swap the CW $\leftrightarrow$ CCW class labels by construction. Output probabilities are averaged after the swap, yielding a strictly flip-equivariant CW/CCW classifier with flip-swap correlation = 1.000 by construction. This averaging is the key methodology distinction between the historical raw Catalog A diagnostic and the released equivariant Catalog C product; the global chirality asymmetry $(N_{\text{CW}} - N_{\text{CCW}})/N_{\text{spiral}}$ shifts from +1.576% (A) to −0.529% (C), i.e. +0.788% to −0.265% in f_{CW} -deviation units, associated with this pipeline change (Table 4). Unit reminder: asymmetry- A values are exactly twice the f_{CW} -deviation values, $A = 2(f_{\text{CW}} - \frac{1}{2})$ (Sec. 4.3); Table 4 quotes the f_{CW} -deviation convention.

forward model now discriminates among these. Aggregating the banked 1.69×10^7 production-ViT forward passes (the 8,474,531 catalog galaxies $\times$ {original, mirror} pair, joined 1:1 to the released catalog with 0.99936 mapped class agreement; no new inference) into a per-pixel mirror-confusion tensor and forward-modeling the global monopole under 500 parity-symmetric label realizations, the released equivariant (Z_2 test-time-averaged) pipeline generates a monopole of $+7.9 \times 10^{-6} \pm 5.2 \times 10^{-4}$ —that is, 0.0% of the observed high-confidence monopole $f_{\text{CW}} - \frac{1}{2} = -3.95 \times 10^{-3}$ (mean −0.2%; $\pm 2\sigma$ null spread 26.7%), which sits at $z =$

−7.57 against this classifier-only null, the same order as the per-pixel binomial -9.47σ (artifact [pipelines/p2_chirality/outputs/canonical_provenance/g4_monopole_mechanism_injection.json](#), with the per-pixel mirror-confusion tensors in the accompanying [pipelines/p2_chirality/outputs/canonical_provenance/g4_perpixel_confusion_nside64.npz](#)). Classifier confusion through the released pipeline thus explains *none* of the monopole. The same test bounds and sign-excludes the classifier-side candidates: without the Z_2 test-time-averaging guard the raw channel would imprint $+1.29 \times 10^{-2}$ (opposite sign, $3.3 \times$

Table 4. Global CW fraction for the historical raw Catalog A diagnostic and the released equivariant Catalog C product. Uncertainties are 1σ binomial, $\sigma = \sqrt{f(1-f)/N_{\text{spiral}}}$, with product-specific spiral counts ($N_{\text{spiral}}^A = 3,321,795$; $N_{\text{spiral}}^C = 3,201,160$); parentheses give the 1σ uncertainty on the trailing digits ($0.507879(274) \equiv 0.507879 \pm 0.000274$). The Excess column is the deviation $f_{\text{CW}} - 0.5$ in percent (f_{CW} units; multiply by 2 for asymmetry- A units). Dev. is the *signed* $(f_{\text{CW}} - 0.5)/\sigma$ computed from the unrounded fraction. Percentages in this table (and in the catalog-composition counts cited in Sec. 2) are rounded to maintain sum-to-one consistency at the second decimal; the integer counts are exact.

Product / diagnostic	cw/(cw + ccw)	Excess (%)	Dev. (σ)
A (raw)	0.507879(274)	+0.788	+28.72
C (equivariant)	0.497353(279)	−0.265	−9.47

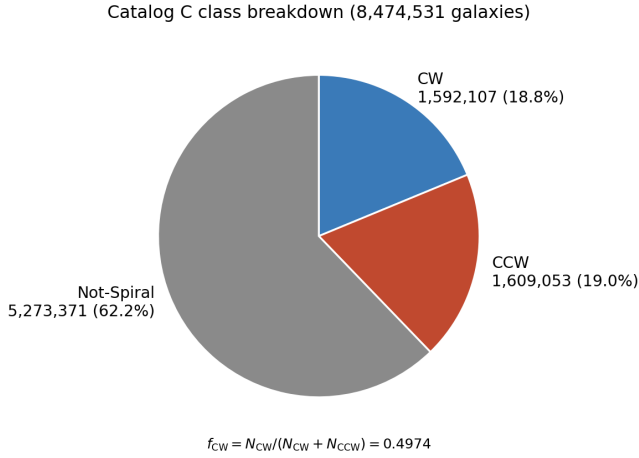


Figure 3. Catalog C composition. Of the 8,474,531 galaxies retained in the released catalog, the equivariant TTA classifier (§3.4) assigns $N_{\text{CW}} = 1,592,107$, $N_{\text{CCW}} = 1,609,053$, and $N_{\text{NS}} = 5,273,371$ (non-spiral / edge-on / morphologically indeterminate). The spiral sub-catalog $N_{\text{spiral}} = N_{\text{CW}} + N_{\text{CCW}} = 3,201,160$ is the analysis target for all chirality statistics below (Table 4 *et seq.*).

the observed magnitude, spatially structured N/S split $-0.145/+0.079$), so the guard demonstrably suppresses a large classifier-level parity systematic, while the GZ1 training-prior candidate (mechanism 1) expresses as a $+0.73\%$ CW *preference*—the wrong sign for the observed CW *deficit*—and is exactly nulled by the equivariant construction. This localizes the monopole’s origin strictly *upstream* of the classifier, in the input image distribution; it does *not* distinguish a true sky parity asymmetry from a parity-odd DESI imaging/photometric systematic in the cutouts, and that upstream attribution remains open (the committed per-pixel confusion tensors are the substrate for it). The $2.98\times$ asymmetry-suppression factor from raw $+1.576\%$ to equivariant -0.529% (asymmetry- A units, $A = 2(f_{\text{CW}} - \frac{1}{2})$; equivalently $+0.788\% \rightarrow -0.265\%$ in f_{CW} -deviation units, Table 4) is associated with the pipeline change but does not isolate TTA causally because the raw and equivariant products came from different inference passes and memberships. Crucially, the slab statistics support this quantitatively: in 7 equal-spiral-count declination slabs

($N = 457,308$ – $457,309$ spirals each; per-slab binomial $\sigma = 7.4 \times 10^{-4}$), the per-slab f_{CW} spans 0.49537 – 0.49890 , i.e. deviations from 0.5 of -0.110% to -0.463% , all within 0.5% of $50/50$; an equal-count RA partition gives a compatible span (-0.060% to -0.501%). An equal-*area* partition (8 declination bands of equal in-mask pixel count on the FS-C support) gives the same verdict: per-band f_{CW} deviations from 0.5 reach at most 0.49% on the full spiral sample (max $|z| = 2.9$ vs. the global rate) and 0.56% on the HC subsample (max $|z| = 1.4$), comparable to the equal-count maximum of 0.46% (artifact [pipelines/p2_chirality/outputs/canonical_provenance/c12_r24conf_local_batch.json](#)). The slab-to-slab scatter about the global $f_{\text{CW}} = 0.49735$ is $\lesssim 2.7\sigma$ per slab (the extremal slab $f_{\text{CW}} = 0.49537$ gives $(0.49735 - 0.49537)/(7.4 \times 10^{-4}) = 2.7\sigma$ from the per-slab values and binomial σ above; artifact [pipelines/p2_chirality/outputs/canonical_provenance/c12_r24conf_local_batch.json](#)), consistent with the coherent low- ℓ systematic structure positioned in Appendix D rather than a dipole-aligned gradient. Independently of any slab partition, a direct generative test shows that a constant monopole cannot bias the uniform-weight real-space dipole estimator (the constant template is absorbed by the fitted monopole term): binomial per-pixel realizations on the FS-C support drawn at $p = f_{\text{CW}}^{\text{global}}$ versus $p = 0.5$ yield statistically identical dipole-amplitude null distributions (means 1.957×10^{-3} vs. 1.935×10^{-3} , a 0.39σ shift in the standard error of the difference; $N = 500$ each; artifact [pipelines/p2_chirality/outputs/canonical_provenance/c11_meta_m4_slab_stats.json](#)). This generative test probes an *additive constant* monopole only; multiplicative depth- or morphology-coupled modulations of the monopole are not probed by it and are addressed instead by the template-regression and cross-spectrum diagnostics of Appendix D. The finite injection grids provide only the score fractions reported in Sec. 6.2; they do not establish a calibrated physical sensitivity floor. *Implications for $\ell = 0$ parity searches.* The morphological handedness monopole is itself the parity-odd channel (unlike the parity-even $\ell = 1$ dipole;

Galaxy Chirality Asymmetry Map (8.47M galaxies, equivariant)

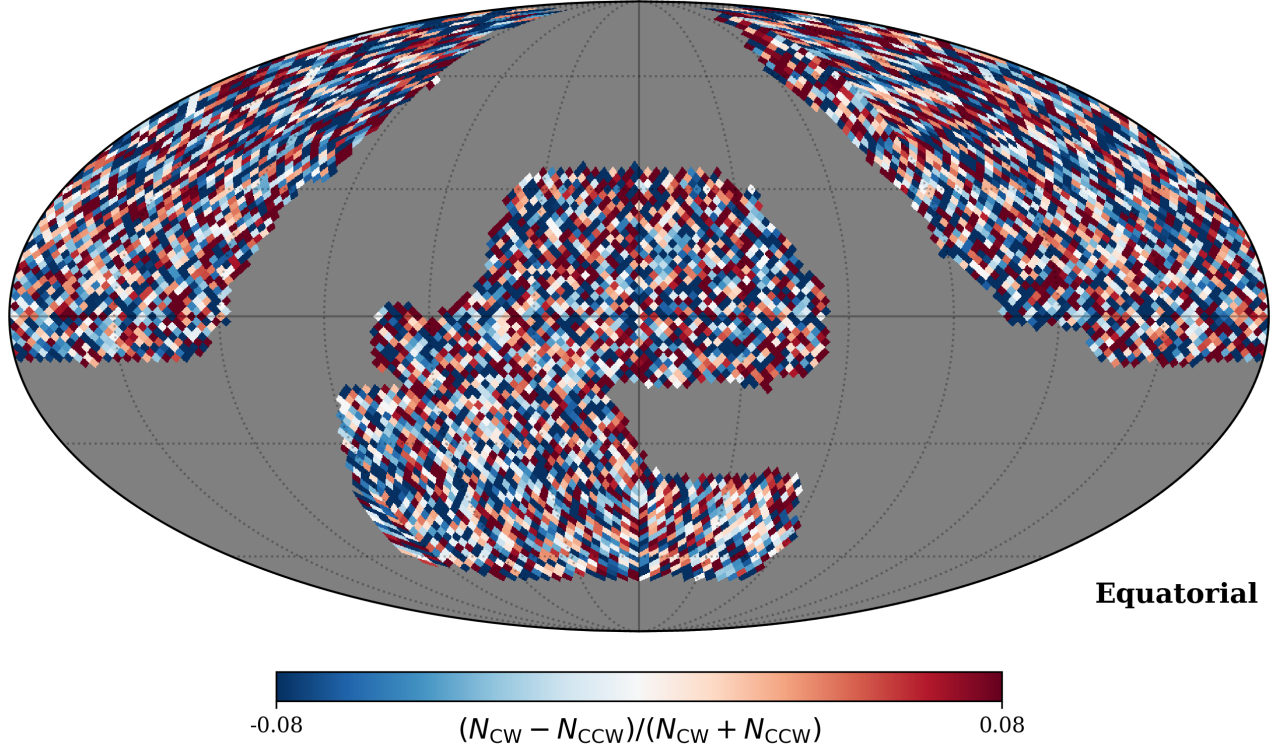


Figure 4. Equivariant (Catalog C) chirality asymmetry map of the 8.47 M-galaxy catalog (Mollweide projection, equatorial coordinates; per-pixel asymmetry $A_p = (N_{\text{CW}} - N_{\text{CCW}}) / (N_{\text{CW}} + N_{\text{CCW}}) = 2(f_{\text{CW},p} - \frac{1}{2})$ at HEALPix NSIDE=64, color scale $[-0.08, +0.08]$). The DESI Legacy Imaging footprint covers $f_{\text{sky}} = 0.49005$ of the sky in the FS-C support ($N_{\text{spiral}}(p) \geq 10$ per pixel; Sec. 4.3); the $N_{\text{all}} \geq 1$ MASTER-AGF ($f_{\text{sky}} = 0.494$) is used for the apodized MASTER diagnostic. Spatial uniformity of the per-pixel CW fraction is verified across 7 equatorial coordinate slabs (§4.2); the FS-C $\ell = 1$ residual is analyzed in Sec. 4.3–4.4.

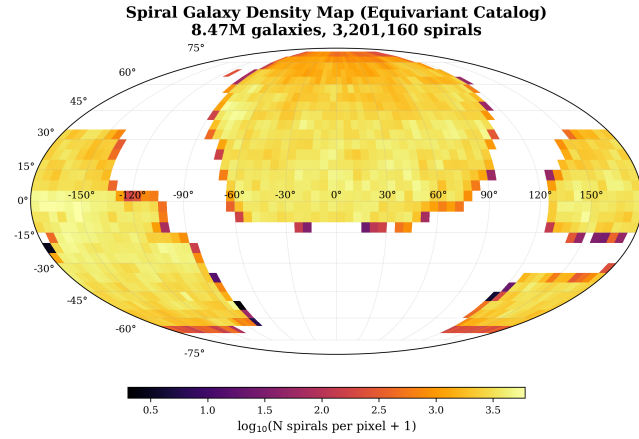


Figure 5. Sky density of the 3,201,160 classified spirals (CW + CCW combined, NSIDE=64 Mollweide). Per-pixel spiral counts scale with the DESI Legacy Imaging Surveys depth and exposure pattern; the FS-C support used for the primary $\ell = 1$ analysis (§4.3) requires $N_{\text{spiral}}(p) \geq 10$ per pixel. Spatial inhomogeneity at this scale is the leakage channel quantified in §4.4.

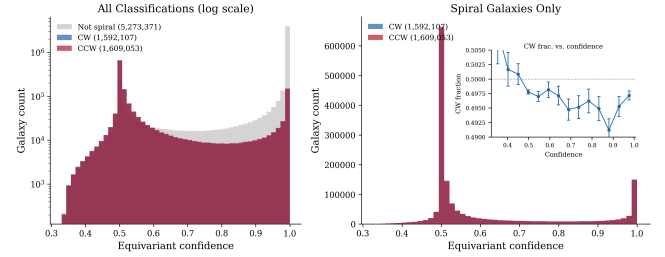


Figure 6. Distribution of maximum-class confidence $\max(P_{\text{CW}}, P_{\text{CCW}}, P_{\text{NS}})$ for all 8,474,531 galaxies. Strongly bimodal: 73.6% at $\max p \geq 0.9$ (high-confidence labels) + a long tail of indeterminate cases ($\max p < 0.5$, dominated by NS/edge-on systems). The high-confidence (HC) cuts at $p_{\text{eq}} > 0.6$ ($N = 949,584$) and $p_{\text{eq}} > 0.8$ ($N = 624,660$) used in the systematics cross-checks (§E) are indicated.

Sec. 6.3), so a naive reading of the -9.47σ deviation as cosmic parity violation would be the single largest false-positive risk in any $\ell = 0$ chirality-parity study on such a catalog. Our analysis shows this deviation is a bounded, spatially structured residual entering *upstream* of the classifier ($\lesssim 0.5\%$ per-slab scatter; clas-

sifier confusion excluded at the 0.0% level and the GZ1 training-prior candidate sign-excluded above, leaving a true-sky-versus-imaging origin as the open attribution), not a physical parity signal; the concrete lesson for future $\ell = 0$ parity searches is therefore that the released catalog labels must be locally monopole-renormalized (per-region CW-fraction subtraction) before any global parity-odd statistic is formed, and that any claimed $\ell = 0$ parity detection at or below the $\approx 0.26\%$ artifact level (9.5σ in raw counts) is indistinguishable from this classifier monopole without an independent, handedness-symmetric labeling pipeline (Sec. E, Data Availability). To convert this recipe into a ready-to-use product, we ship a delivered per-region CW-fraction monopole correction map, computed directly from the released Catalog C spiral labels on the same HEALPix pixelization used for the dipole fit (native NSIDE=64 plus a coarser, more stable NSIDE=8 scheme), with per-region counts, f_{CW} , the monopole correction $f_{\text{CW}}(r) - 0.5$, and its per-region binomial uncertainty (artifact [pipelines/p2_chirality/analysis/monopole_correction_map_v1_0_262.json](#); full-resolution NSIDE = 64 arrays in the accompanying .npz). This map is a convenience product for $\ell = 0$ parity users only: it changes no science number in this paper and does not affect the primary real-space dipole, which absorbs any constant monopole into its fitted monopole term as shown above.

4.3. Dipole Analysis

We pixelize the sky at HEALPix resolution NSIDE = 64 (49,152 pixels, ~ 0.84 deg² per pixel). The threshold $N_{\text{spiral}}(p) \geq 10$ defines HC-RI for the quality-controlled HC primary (23,633 pixels) and FS-C for the full-spiral diagnostics (24,087 pixels).

$$A_p = \frac{N_{\text{CW}}^{(p)} - N_{\text{CCW}}^{(p)}}{N_{\text{CW}}^{(p)} + N_{\text{CCW}}^{(p)}}. \quad (3)$$

This is the single canonical chirality-field definition used throughout (spirals-only denominator). Note the unit convention $A_p = 2(f_{\text{CW},p} - \frac{1}{2})$: amplitudes quoted in A_p units are exactly twice the corresponding f_{CW} -deviation amplitudes.

Simple dipole.—The primary real-space estimator fits $A_p = m + \mathbf{a} \cdot \hat{n}_p$ after applying `primary_hc` && `!raw_flip_qc_unsafe`. The selection contains 890,069 spirals and zero unsafe rows; 887,472 enter the 23,633-pixel HC-RI support. Each of 10,000 fixed-occupancy draws preserves those pixel capacities and the observed support-sample CW total. The fitted amplitude is $A_{\text{dip}} = 0.00466520$; the null mean and population standard deviation are 0.00362029 and 0.00164643, giving

Table 5. Primary HC estimator audit. The empirical rank value is the one-sided upper tail. The exact strict-sample fixed-occupancy null-array SHA-256 is `3a03ca4b008844fd...e32ce7d`; full precision and the complete hash are in the computational-closure receipt.

Quantity	Exact or reported value
$N_{\text{selected}} / N_{\text{support}}$	890,069 / 887,472
Excluded / valid pixels	2,597 / 23,633
Support / seed / null draws	HC-RI / 20260715 / 10,000
$A_{\text{dip}} / \text{monopole}$	0.0046651988 / 0.0255064554
Equatorial direction	(195.4820°, −57.1615°)
Null mean / std. (ddof = 0)	0.0036202914 / 0.0016464287
z_{mom}	+0.6346509
Upper-tail k/N / rank p	2376/10000 / 0.2376762

$z_{\text{mom}} = +0.63465$ and upper-tail rank $p = (2376 + 1)/(10000 + 1) = 0.23768$ for seed 20260715. Moment- z and rank- p are separate summaries; no Gaussian $z \rightarrow p$ conversion is used.

The exact primary draw array and deterministic generator are checksummed in the release manifest. The former pixel-asymmetry permutation is retained only as a robustness diagnostic: its exact 10^4 -draw array gives $z = +0.549$ and rank $p = 0.2652$. A separate 2×3 panel uses 2,000 pixel permutations per cell (seed 20260610), with asymmetry values permuted while pixel weights remain fixed. At $N_{\text{spiral}} \geq 10, 20, 50$, respectively, the uniform-fit pairs are

$$(z, p) = (+0.536, 0.270), (+0.268, 0.359), \quad (4)$$

$$(-0.382, 0.606),$$

and the N_{spiral} -weighted pairs are

$$(z, p) = (+0.783, 0.197), (+0.794, 0.203), \quad (5)$$

$$(-0.141, 0.505).$$

Thus all six cells have $|z| < 0.8$. These are robustness diagnostics on HC-RI threshold variants, not additional primary tests. The unthresholded full-sample excess is reported only as a depth-correlated systematics diagnostic, not a competing primary result.

The unsafe-row quarantine predicate was introduced and finalized during post-review closure after inspection of the earlier unsafe-inclusive result; it was not preregistered or fixed before unblinding. The strict rerun is therefore a transparent corrective analysis, not a blinded confirmatory test. Its effect on the dipole is small: excluding the flagged rows leaves the real-space result essentially unchanged ($z = +0.48$ excluded vs. $+0.52$ baseline under the c11b 10^4 -permutation convention; Appendix B).

Angular power spectrum.—We recompute the harmonic diagnostics with NAMASTER on one persisted FS-C

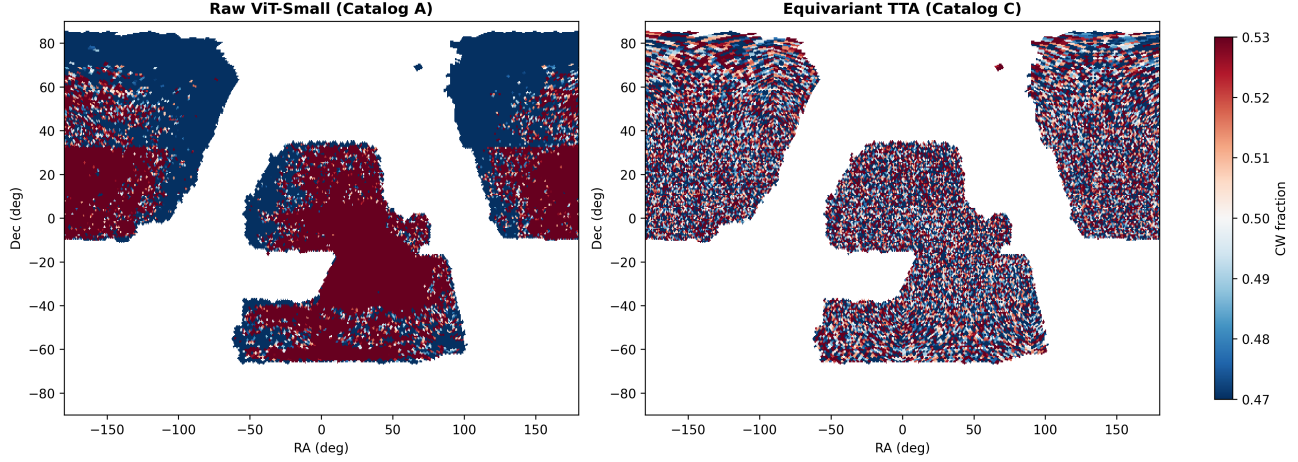


Figure 7. Raw (Catalog A) vs equivariant (Catalog C) chirality sky maps (equatorial RA/Dec, per-pixel CW *fraction* $f_{\text{CW},p}$, NSIDE=64; shared color scale [0.47, 0.53]). The raw definition is `class_raw_y` with `class_raw_x` fallback only where the former is missing; in this release all 3,321,795 raw spirals come from `class_raw_y` (zero fallback rows), while Catalog C contains 3,201,160 spirals. The native $N_{\text{spiral}} \geq 5$ supports contain 24,158 raw and 24,154 equivariant pixels; both displayed panels use their 24,151-pixel intersection. The panels are plotted in f_{CW} units, not the A_p units of Fig. 4: the two are related by $A_p = 2(f_{\text{CW},p} - \frac{1}{2})$, so the [0.47, 0.53] range here corresponds to $A_p \in [-0.06, +0.06]$. Left: raw single-pass classifier output; the spatially-structured $\sim 0.79\%$ classifier CW excess, modulated by non-uniform survey depth, produces the 2.31σ real-space dipole + $+6.48\sigma$ pre-MASTER $\ell=1$ artifact. Right: 2-fold flip-equivariant TTA; flip-equivariance is enforced by construction and the measured directional bias is strongly reduced, while training-, selection-, and depth-coupled biases are not guaranteed to vanish, and the release-safe HC real-space estimator is null-consistent at $z_{\text{mom}} = +0.635$ after excluding quarantined rows. Generation counts and the projected streaming read are recorded in [pipelines/p2_chirality/outputs/canonical_provenance/fig7_raw_vs_eq_manifest.json](#). This visual diagnostic shows that the equivariant pipeline is associated with a strong reduction of the measured directional bias and illustrates why equivariant post-processing is an important mitigation in future chirality studies. Because the two products have different inference-pass provenance, memberships, and quoted null conventions, the comparison does not by itself isolate TTA as the unique causal explanation.

mask defined by $N_{\text{spiral}}(p) \geq 10$: 24,087 pixels, $f_{\text{sky}} = 0.490051$, mask SHA-256 499fa89e...f5852. Every data and null field uses this same base support and the same $f_{\text{CW}} - 0.5$ field with N_{spiral} -weighted mean subtraction. Under 500 fixed-occupancy label randomizations (seed 42), the binary-mask $\ell=1$ result is $C_1^{\text{data}} = 7.0113 \times 10^{-6}$, null mean 5.8478×10^{-7} , null population width 9.2823×10^{-7} , $z_{\text{mom}} = +6.923$, and add-one rank $p = 1/501 = 0.001996$. A C^2 2° apodization derived from that same base mask gives $z_{\text{mom}} = +7.033$ and the same rank. The fixed-support multipole vector is shown in Table 6. These moment ratios are systematics diagnostics, not Gaussian-equivalent tail significances and not cosmological detections.

The structure is distributed across $\ell = 1-3$ rather than forming a clean isolated dipole. It is attributed to coherent catalog/survey systematics. The sole primary scientific result remains the quality-controlled HC real-space null ($z_{\text{mom}} = +0.635$, $p = 0.23768$).

4.4. Monopole+Mask Leakage Generative Null

The exact-support harmonic diagnostic is evaluated on the persisted 24,087-pixel FS-C mask. Under 500 fixed-occupancy label randomizations it gives $z_{\text{mom}} =$

+6.923 and add-one rank $p = 0.001996$; a distinct binomial-monopole generative null on the same support gives $z_{\text{mom}} = +6.983$ with 500 draws and +7.207 with 10,000 draws. The distinction matters: fixed occupancy conditions on the observed global CW count, whereas the binomial construction draws per-pixel CW counts from $\text{Binomial}(N_{\text{spiral}}(p), p_{\text{CW}}^{\text{global}})$. These are systematics diagnostics, not Gaussian detection significances.

The exact-support recomputation removes the earlier support ambiguity. Both null families retain a large low- ℓ residual, but their numerical moment ratios are not interchangeable because they condition on different quantities. The fixed-occupancy result is the direct label-randomization diagnostic; the binomial-monopole result asks whether a spatially uniform global imbalance propagated through the fixed mask accounts for the field. Neither is the primary cosmological test. The primary conclusion rests on the strict quality-controlled HC sample ($N_{\text{selected}} = 890,069$, $N_{\text{support}} = 887,472$), for which the fixed-occupancy real-space result is $z_{\text{mom}} = +0.635$ and $p = 0.23768$.

The exact-support binary and apodized diagnostics give $z_{\text{mom}} = +6.923$ and +7.033, respectively. The multipole vector (+6.923, +2.931, +3.089, -0.317, +1.606)

Table 6. Exact-support FS-C MASTER diagnostics. All rows use the same checksummed 24,087-pixel base support and 500 fixed-occupancy label randomizations (seed 42). Empirical ranks use $(k + 1)/(N + 1)$. The z column is a null moment ratio, not a Gaussian tail significance.

Mode	$C_\ell^{\text{data}} \times 10^6$	$\langle C_\ell \rangle \times 10^6$	z_{mom}	rank p
$\ell = 1$	7.011	0.585	+6.923	0.001996
$\ell = 2$	2.339	0.514	+2.931	0.02196
$\ell = 3$	2.356	0.561	+3.089	0.01397
$\ell = 4$	0.418	0.580	−0.317	0.57884
$\ell = 5$	1.297	0.554	+1.606	0.06786
$\ell = 1, C^2 \ 2^\circ$ apod.	6.802	0.566	+7.033	0.001996

Table 7. Exact-support binomial-monopole MASTER null on the persisted 24,087-pixel FS-C mask (seed 42). The z_{mom} values are null moment ratios, not Gaussian-equivalent significances; ranks use add-one correction.

Statistic	Data	Null	z
C_1 , 500 draws	7.0113×10^{-6}	$(0.60414 \pm 0.91749) \times 10^{-6}$	+6.983
C_1 , 10,000 draws	7.0113×10^{-6}	$(0.57796 \pm 0.89263) \times 10^{-6}$	+7.207

for $\ell = 1$ –5 shows coherent low- ℓ structure rather than a clean isolated primordial dipole. Diagnostics on redefined or unproven supports remain historical provenance only.

Support correction. The historical $+3.80\sigma$ density-null and WLS/block-bootstrap numbers summarized in older audit language below were generated on broader latitude-cut supports, not FS-C. They are excluded from the support-consistent synthesis and provide no FS-C or physical constraint; Appendix D records the exact discrepancy.

Quantitative forward model and cross-estimator stress test. The imaging+DR8-morphology forward model reproduces 53.0% of the full-sample $\ell = 1$ amplitude by scalar ratio, while direct vector subtraction leaves 62.3%; these definitions are not interchangeable. The residual amplitude is 6.95×10^{-3} in A_p . Deterministic map injections change the HC fitted amplitude from 0.4597% to 0.6710% for the full template, 0.5533% for the scalar remainder, and 0.5694% for the vector remainder. Thus the residual is not automatically irrelevant, but these shared-catalog injections do not identify its origin. A separate classifier-injection forward model (Sec. 4.2) excludes classifier confusion through the released equivariant pipeline as the monopole’s source and localizes its origin upstream of the classifier, without yet resolving the sky-versus-imaging attribution. The shared-resample covariance of the four estimators—now including the NaMaster MASTER-decoupled $\ell = 1$ leg—is computed separately (Sec. 4.5, Table 9); a full joint likelihood across all channels remains open.

4.5. Joint Estimator Covariance

The estimators above are reported against separate, non-interchangeable nulls. To test whether they instead share a single sampling covariance—so that their correlation structure can be read directly rather than inferred across heterogeneous supports—we evaluate the paper’s committed estimators on one shared block bootstrap. On the primary high-confidence spiral sample ($N = 949,584$; `class_eq` ∈ {CW, CCW} and $p_{\text{eq}} > 0.6$; immutable catalog snapshot `cc326f74`) under the canonical $|b_{\text{gal}}| > 15^\circ$ mask, we draw $N_{\text{boot}} = 2,000$ resamples over `NSIDE`= 8 HEALPix superpixels (the same block-bootstrap unit as `pipelines/p2_chirality/scripts/joint_nuisance_bootstrap_sigma.py`; `NSIDE`= 64 data map, seed 42, all 2,000 resamples valid) and recompute, on each resample, the uniform-pixel-weight real-space dipole amplitude, the nuisance-marginalized WLS dipole amplitude, the global CW-fraction monopole $f_{\text{CW}} - \frac{1}{2}$, and—closing a previously pod-bound refinement—the NaMaster MASTER-decoupled harmonic $\ell = 1$ power C_1^{master} (the full mode-coupling-matrix inverse $M^{-1}C_1^{\text{pseudo}}$ on the fixed canonical effective mask, $f_{\text{sky}} = 0.4899$; `pymaster` 3.0), evaluated on the identical resample index sets (verified: the real-space-dipole and monopole channels reproduce the NaMaster-free run’s bootstrap moments to six digits). Because the four estimators are evaluated on the *same* resamples, their joint sampling covariance and correlation are defined directly (artifact `pipelines/p2_chirality/outputs/canonical_provenance/g3_joint_estimator_covariance_master_v2.json`).

Table 9 reports the resulting 4×4 block-bootstrap correlation matrix and the per-estimator significance

Table 8. Support audit of diagnostics associated with the exact-support FS-C low- ℓ residual. Only rows explicitly bound to the persisted base mask enter the synthesis.

Anchor	Key result	Verdict
(a) Apodized-mask robustness	+7.033 (binary mask +6.923)	persists under apodization
(b) Multipole spectrum	(+6.923, +2.931, +3.089, -0.317, +1.606) for $\ell = 1-5$	coherent low- ℓ systematics
(c) Quality-quartile stratification	quartile-specific masks, not fixed FS-C	out of scope
(d) Leg-proxy cross-power	24,270 pixels, not 24,087-pixel FS-C	out of scope
(e) Density-stratified null	+3.80 σ on broader latitude mask	out of scope
(f) Boundary-distance variance	35,438-pixel latitude mask, not FS-C	out of scope
(g) WLS template fit	$z_{\text{boot}} \approx -7.6$ on broader latitude mask	out of scope
(h) Direct cross-spectrum	exact mask not recorded in artifact	out of scope

Synthesis: the two support-bound rows show apodization robustness and broadband low- ℓ structure; the other six rows are excluded.

$z = \hat{\theta}/\sigma_{\text{boot}}$ of each full-sample value against its own bootstrap spread. Three features matter. First, the monopole is the only $|z| > 3$ mode ($z = -6.57$; consistent with, but appropriately more conservative than, the per-pixel-independent binomial -9.47σ of Sec. 4.2, because the block bootstrap propagates spatial coherence), while both dipole amplitudes and the MASTER-decoupled harmonic $\ell = 1$ power are non-significant ($z = +2.21$ real-space dipole, +0.81 WLS dipole, -0.61 MASTER $\ell = 1$), consistent with the primary null. Second, the monopole is nearly uncorrelated with every dipole/harmonic channel ($\rho = -0.037$ with the real-space dipole, $\rho = -0.020$ with the WLS dipole, $\rho = -0.061$ with the MASTER-decoupled $\ell = 1$): within one shared resampling the large monopole deviation moves essentially independently of the parity-even $\ell = 1$ family, so the joint covariance quantitatively supports treating it as a statistically *distinct* mode rather than as a contamination channel of the primary dipole null. Third, this run closes the previously pod-bound MASTER-decoupled refinement and validates the paper’s harmonic proxy: the decoupled C_1^{master} correlates at $\rho = +0.943$ with the mask-coupled **anafast** C_1 pseudo-power proxy carried through the same bootstrap (confirming the NaMaster-free proxy faithfully tracks the full mode-coupling-matrix-inverse estimator on identical resamples) and at $\rho = +0.794$ with the real-space dipole (both capturing the same $\ell = 1$ content); the joint covariance is therefore now NaMaster-complete. *One honest caveat.* The WLS channel’s bootstrap σ is sensitive to the BLAS implementation of its nine-column weighted normal-equation solve, which is ill-conditioned on a minority of resamples ($\sigma_{\text{boot}} = 0.003436$ on one platform versus 0.005776 on another; the table quotes the latter). The WLS amplitude is non-significant under both ($z = +1.36$ and +0.81) and enters only as a diagnostic, so this platform sensitivity changes no conclusion. This block establishes

Table 9. Joint block-bootstrap correlation of the four estimators, measured on one shared NSIDE= 8 super-pixel bootstrap ($N_{\text{boot}} = 2,000$ resamples, all valid; seed 42; primary HC sample $N = 949,584$; artifact `g3_joint_estimator_covariance_master_v2.json`). The fourth channel C_1^{master} is the NaMaster MASTER-decoupled $\ell = 1$ power (mode-coupling-matrix inverse on the fixed canonical effective mask, $f_{\text{sky}} = 0.4899$; pymaster 3.0), computed on the identical resample index sets. Off-diagonal entries are Pearson correlations across the shared resamples. σ_{boot} is the per-estimator bootstrap standard deviation and z is the full-sample value over σ_{boot} . The monopole is the only $|z| > 3$ mode and is nearly uncorrelated with all three dipole/harmonic channels. (The WLS σ_{boot} is BLAS-sensitive; see text.)

	real-space dip.	WLS dip.	monopole	C_1^{master}
real-space dip.	+1.000	+0.158	-0.037	+0.794
WLS dip.	+0.158	+1.000	-0.020	+0.129
monopole	-0.037	-0.020	+1.000	-0.061
C_1^{master}	+0.794	+0.129	-0.061	+1.000
full value	+0.004386	+0.004669	-0.003949	-4.55×10^{-6}
σ_{boot}	0.001987	0.005776	0.000601	7.52×10^{-6}
$z = \hat{\theta}/\sigma_{\text{boot}}$	+2.21	+0.81	-6.57	-0.61

the joint sampling covariance of the four computed estimators; the monopole’s causal mechanism is addressed separately by the classifier-injection analysis of Sec. 4.2 and remains partly open (Sec. 7).

4.6. Signal-Hunt Diagnostics

All signal-hunt diagnostics (confidence stratification, RA-quadrant scatter, hemisphere north/south, per-imaging-leg $\times$ confidence-bin) point to the same conclusion: the FS-C residual is structured along classifier-systematic, footprint-systematic, and galactic-foreground axes, not along a primordial-dipole-aligned axis. The +3.29 σ signal in the 1.87M-galaxy [0.5, 0.6) confidence bin does not survive the sample-purity ladder: cutting on $p_{\text{eq}} > 0.6$ gives -0.03σ under the same

confidence-stratified dipole estimator and null (this is a distinct statistic from the primary HC real-space dipole of Sec. 4.3). Full diagnostic tables are in Appendix C.

5. COMPARISON WITH PREVIOUS WORK

5.1. *Shamir (2012, 2020, 2022)*

Under the present ViT/TTA pipeline, the full-footprint regional WLS amplitude is 0.32% in A_p and equal-area slab maxima reach 0.46–0.56%. The HC real-space result is null-consistent. Comparisons with Shamir’s Ganalyzer results are qualitative because the estimators, masks, and cuts differ; neither monopole–mask leakage nor depth-coupled classifier bias establishes what a matched Ganalyzer run would return. We therefore make no frequentist or physical exclusion claim. A matched-footprint independent-estimator analysis is required.

5.2. *CE-ResNet (Jia et al. 2023)*

CE-ResNet (7) achieves $\text{CW}/\text{CCW} = 0.998$ with architectural equivariance on 1.95 million galaxies. Our Catalog C achieves $1.6\times$ the spiral coverage with $\text{CW}/(\text{CW} + \text{CCW}) = 0.4974 \pm 0.0003$ using TTA-equivariance. The two pipelines are complementary: CE-ResNet offers a stronger single-pass mathematical guarantee; our pipeline offers larger survey-scale coverage, a dedicated NOT_SPIRAL class, and a quantitative bias-hardening audit.

6. DISCUSSION

The raw Catalog A dipole (2.31σ real-space; $+6.48\sigma$ pre-MASTER) demonstrates that a classifier bias of only 0.79%, combined with non-uniform sky coverage, produces highly significant but entirely spurious dipole signals. Equivariant averaging plus the unsafe-row quarantine gives the release-safe primary result $z_{\text{mom}} = +0.635$, $p = 0.23768$; the raw value used the legacy isotropic-permutation diagnostic and is retained only as methodology context. MASTER deconvolution substantially reduces the monopole-mask leakage that sources the $+6.48\sigma$ pre-MASTER pseudo- C_ℓ (Sec. 4.4). The 3.05σ hemisphere signal (label-shuffle null; Appendix C) is classified as a documented systematic-floor artifact: the principled directional look-elsewhere control is the direct-MC max-statistic null (which incorporates the 648 tested directions and their correlations exactly, Appendix C), under which the direct-MC max-statistic null rejects isotropic random-label noise at $p_{\text{LEE}} = 1/10001$; $95\% \text{upperlimit} \simeq 3.0 \times 10^{-4}$. The hemisphere excess is therefore consistent with unresolved systematic-floor structure, including multiple possible training-label, classifier-orientation, depth,

and imaging-footprint mechanisms that can also contribute to the global 9.5σ CW-fraction monopole. For that monopole specifically, a classifier-injection forward model (Sec. 4.2) now excludes classifier confusion at the 0.0% level and sign-excludes the GZ1 training-prior candidate, localizing its origin upstream of the classifier; the remaining sky-versus-imaging attribution is not resolved by the available diagnostics.

6.1. *Pseudo-label independence and the shuffle-null limitation*

Human-vote cross-check. Direct Galaxy Zoo 1 human CW/CCW votes, with no learned model in the chirality-label chain, give a null-consistent dipole at $z = -0.54\sigma$ (rank $p = 0.67$) from 46,017 matched inputs. The exact rerun retains 4,963 galaxies in 394 pixels after its strict $N_{\text{pixel}} > 10$ cut (41,054 matched inputs are excluded), and its pixel-permutation null differs from the fixed-occupancy primary. It therefore corroborates a null under its own convention but does not support a numerical power ratio or tighten the sub-percent HC observed-label result. The shuffle nulls still cannot test a spatially varying pseudo-label systematic orthogonal to the external template basis. The WLS diagnostic and finite injection grids do not supply a calibrated ceiling on such inherited power. A committed per-pixel mirror-confusion tensor now exists as the substrate for such a model (Sec. 4.2); a full spatially resolved confusion/transfer model validated against held-out labels therefore remains open.

6.2. *Statistical Scale and Finite Injection Scores*

Fisher statistical-only scale.—For the full-amplitude dipole convention $p_{\text{CW}}(\hat{n}) = \frac{1}{2}(1 + A \cos \theta)$, the per-galaxy Fisher information on A at $A = 0$ is $\cos^2 \theta / [p(1 - p)] \cdot (\partial p / \partial A)^2 = \cos^2 \theta$, so with the full-sky idealization $\langle \cos^2 \theta \rangle = \frac{1}{3}$,

$$\sigma(A) = \sqrt{\frac{3}{N_{\text{spiral}}}} = 2\sqrt{3} \sigma(f_{\text{CW}}) = 9.7 \times 10^{-4} \quad (6)$$

at $N_{\text{spiral}} = 3,201,160$ (i.e. $\sigma(A/2) \approx 0.048\%$), giving a 3σ ideal floor of $3\sigma(A) \approx 0.29\%$ full-amplitude. This full-sky idealization omits the realized footprint, classifier transfer, and orientation dependence and is therefore a scale reference only. At the HC sample size $N = 949,584$, the same idealization gives $\sigma(A) = 1.78 \times 10^{-3}$ and a nominal 3σ scale of 0.53%; it is not a calibrated detection threshold. Under a purely illustrative scalar transfer $A_{\text{obs}} = g A_{\text{phys}}$ with fixed known g , uncertainty would scale as $\sigma(A_{\text{phys}}) = \sigma(A_{\text{obs}})/g$. Here $g = 0.398$ is an aggregate GZ1-accuracy dilution, not a spatially re-

solved physical transfer, so applying $1/g$ does not create a calibrated physical uncertainty, threshold, or bound.

Finite-grid injection scores.—The earlier Monte Carlo injection sweep on the HC sample ($N = 949,584$; 100 injections per amplitude) is retained as descriptive tested-grid evidence. Under the area-uniform axis draw, the score fractions are 0.59, 0.91, and 1.00 at full amplitudes $A = A_p = 0.75\%$, 1.0% , and 1.5% ; the originally tabulated polar-angle-uniform sweep gives 0.55, 0.91, and 1.00. Each amplitude uses only 100 injections and a fixed per-pixel binomial label-shuffle calibration. We report these fractions directly. In every injection section, $A = A_p$ is the full asymmetry amplitude and the peak deviation of f_{CW} is $A/2$. The scalar transfer $g = 0.398$ is illustrative only because the spatial transfer model remains open.

Coverage-calibrated observed-label upper limit.—Scaling the same observed-label injection-recovery up to the primary channel with random-axis coverage counting now yields a *coverage-calibrated* 95% sensitivity upper limit on the primary real-space dipole amplitude. Injections use the observed-label convention $p_{\text{CW}}(\hat{n}) = p_{\text{CW}}^{\text{global}} + (A/2)(\hat{n} \cdot \mathbf{u})$ at fixed pixel occupancy, drawn on 2,000 isotropically random axes $\mathbf{u}$ per amplitude, and are passed through the *exact committed primary estimator*—the uniform-pixel `healpy.fit_dipole` amplitude on the strict release-safe support (the same estimator that yields $z_{\text{mom}} = +0.635$)—with detection defined as a one-sided add-one empirical rank $p < 0.05$ against the *exact committed* 10^4 -draw fixed-occupancy label-randomization null (Sec. 4.2; null-array SHA-256 `3a03ca4b008844fd...`). The recovered detection probability rises monotonically from 0.24 at $A = 0.40\%$ through 0.91 at 0.90% and crosses 95% coverage at

$$A_{95}^{\text{obs}} \simeq 0.98\% \quad (\text{full-amplitude, observed-label}), \quad (7)$$

satürating to unit coverage by $A = 1.5\%$ (a logistic fit gives 0.96%). This calibration reproduces the committed headline exactly before running ($A_{\text{dip}} = 0.00466520$, $z_{\text{mom}} = +0.63465$, $p = 0.23768$) and requires no ViT re-inference, no confidence-cut re-run, and no transfer function. It states the amplitude the primary channel is demonstrably sensitive to at 95% coverage under the paper’s own estimator and null; the observed amplitude $A_{\text{dip}} = 0.467\%$ lies below this floor, so the observed null does not exclude sub-1% observed-label dipoles. A_{95}^{obs} is an *observed-label* sensitivity floor and explicitly *not* a physical parity-amplitude bound: the physical limit A_{95}^{phys} additionally requires the spatially resolved morphology transfer function (the scalar $g = 0.398$ is illustrative only) and remains the tracked open gate.

The generator, hashed inputs, per-amplitude coverage curve, and full provenance are in [pipelines/p2_chirality/analysis/a95_observed_label_upper_limit_v1_0_265.json](#).

What the injection-recovery chain does and does not traverse. For clarity about the scope of the sensitivity calibration we state explicitly which links of the classification-to-dipole chain the injection-recovery sweep passes a signal through, and which it does not. The sweep injects a dipole in the *observed hard-label* CW/CCW field and propagates it through per-pixel map-making, the real-space dipole estimator, and the per-pixel-shuffle score. It does not traverse the ViT, NOT_SPIRAL triage, confidence cut, or spatially varying confusion. The small deterministic-axis pilot grids of this subsection are reported as direct tested-grid fractions only; the separate random-axis, 2,000-injection scaling of this same chain on the primary estimator (Eq. 7) is what supplies the coverage-calibrated observed-label limit $A_{95}^{\text{obs}} \simeq 0.98\%$. The illustrative bridge from observed-label to physical amplitude uses a single dilution factor g ; it does not close the spatial transfer gap. For reference, its value is robust to the asymmetric-confusion arithmetic rather than assume symmetric errors: the correct first-order transfer slope of the observed CW fraction under an asymmetric confusion matrix with CW sensitivity s_{CW} and CCW sensitivity s_{CCW} is $g_{\text{eff}} = s_{\text{CW}} + s_{\text{CCW}} - 1$ (the derivative of $q s_{\text{CW}} + (1 - q)(1 - s_{\text{CCW}})$ with respect to the true CW fraction q). Substituting the committed GZ1 confusion values ($s_{\text{CW}} = 0.674$, $s_{\text{CCW}} = 0.724$; Table 15) gives $g_{\text{eff}} = 0.397616$. The pooled-accuracy proxy gives $2a - 1 = 0.398217$ (with $a = 0.699108$); the two agree to three decimals but are not algebraically identical because pooled accuracy is class-frequency weighted and the evaluable true classes are not exactly balanced. This comparison is conditional on the full-overlap confusion matrix, whose training rows were not removed; it is therefore an illustration rather than an independent transfer calibration. Within that matrix the mild CW/CCW asymmetry changes g by $< 10^{-3}$, so the $\sim 1.88\%$ ($= 0.75\%/g$) illustrative figure is not sensitive to symmetric-versus-asymmetric arithmetic. The residual triage-to-NOT_SPIRAL removal ($\sim 19\%$) is an uncalibrated selection-completeness channel. Because its human CW/CCW composition and spatial dependence were not retained in this transfer calculation, it cannot be assumed chirality-neutral and is not folded into the illustrative slope g . *Stage A: full-catalog mirror-pair transfer/equivariance sweep.* The production ViT was run on each of the 8.47×10^6 original images and its mirror. This measures paired-output behavior, not

recovery of a physically injected sky field. The raw-head CW/CCW flip fraction is $T_{\text{raw}} = 0.2303 \pm 0.0002$, while the production Z_2 -TTA probability swap is exact by construction and its hard-argmax flip recovery is $T_{\text{eq}} = 0.9997$; the 2.6×10^{-4} shortfall is due to exact CW/CCW argmax ties. These are full-catalog mirror-pair transfer and equivariance diagnostics, not a continuous image transformation, physical chirality calibration, or end-to-end injection-recovery result ([pipelines/p2_chirality/outputs/canonical_provenance/e2e_fullrun/e2e_transfer_function_full.json](#)).

Stage B: paired-image-to-field surrogate. We next use the two precomputed Stage-A outcomes as a discrete intervention: for 20 deterministic Fibonacci-sphere axes and injected amplitudes $A = \{0.5, 0.75, 1.0, 1.5, 2.0\}\%$, the original or mirrored member is selected to realize each latent CW/CCW draw, then the released production hard label and confidence are propagated through NS triage, p_{eq} cuts, the $N_{\text{spiral}}(p) \geq 10$ map, and the same uniform-pixel `healpy.fit_dipole` estimator. The released catalog contains 949,584 canonical HC spirals, whereas the fresh Stage-A export would select 949,650; the 66-row count seam is not treated as canonical. All 949,584 canonical HC rows have concordant production/Stage-A paired hard labels. At lower confidence, 4,369 of 3,201,160 production spirals have discordant export labels and are excluded rather than forcibly paired, leaving 3,196,791 defensible pairs (99.864%). On the canonical HC sample the numbers of axes crossing the declared 3σ score are 1/20, 16/20, 16/20, 20/20, and 20/20 at full amplitudes $A = A_p = 0.5\%, 0.75\%, 1.0\%, 1.5\%$, and 2.0% . On the concordant full pool they are 12/20, 20/20, 20/20, 20/20, and 20/20. These are pilot fractions over deterministic axes, not independent Bernoulli trials or calibrated population recovery probabilities. Even under an inapplicable independent-binomial reading, 20/20 gives a one-sided 95% Clopper–Pearson lower bound of 0.8609, not 95% recovery. Angular-quadrature convergence and independent sky/noise realizations were not run. The full pilot-grid distributions, signed projections, p16–p84 ranges, NS-triage counts, and all strata are stored in [pipelines/p2_chirality/outputs/canonical_provenance/stage_b_hybrid_image_field_recovery.json](#). This Stage B result remains conditional on the classifier-defined member of each pair. It is not a validated continuous morphology transform, does not supply independent per-object true chirality or truth for NS/NS pairs, and cannot calibrate the physical observed-to-true transfer unconditionally. Accordingly the spatially resolved

physical confusion model remains open and all physical falsification claims remain tied to the observed hard-label field. A first, coarse instance of that spatially varying confusion model *now exists at coarse declination-stratum granularity*: the GZ1-overlap confusion matrix stratified by the two P4 declination proxies and by confidence (Appendix B, Table 16) shows the directional CW $\leftrightarrow$ CCW error asymmetry consistent with zero in both tested declination strata and confidence strata (declination-stratum asymmetries bounded to $\lesssim 1.4$ pp, science-cut to $\lesssim 0.6$ pp). Directional asymmetry is not, however, the only spatial bias channel. For local true CW fraction $q(\mathbf{x})$ and misclassification rates $e_{\text{CW} \rightarrow \text{CCW}}(\mathbf{x})$ and $e_{\text{CCW} \rightarrow \text{CW}}(\mathbf{x})$, the observed fraction is

$$q_{\text{obs}}(\mathbf{x}) = q(\mathbf{x})[1 - e_{\text{CW} \rightarrow \text{CCW}}(\mathbf{x})] + [1 - q(\mathbf{x})]e_{\text{CCW} \rightarrow \text{CW}}(\mathbf{x}). \quad (8)$$

Even for symmetric error $e_{\text{CW} \rightarrow \text{CCW}} = e_{\text{CCW} \rightarrow \text{CW}} = e(\mathbf{x})$, this becomes $q_{\text{obs}} = q + e(1 - 2q)$; spatial variation in e can therefore modulate the observed field wherever the local selected fraction is not exactly one half. The committed stratified artifact gives net induced $\Delta f_{\text{CW}} = 0.00142$ for BASS+MzLS and 0.00331 for DECaLS, a point-estimate leg difference of -0.00189 . The individual leg intervals include zero, and no covariance-aware interval for their difference is retained, so these values neither establish a significant leg difference nor close the channel. Error magnitude also varies by leg ($z=4.2$) and confidence (accuracy $0.72 \rightarrow 0.96$); it is not generically a spatially constant dilution already carried by one scalar g . The remaining gap is a finer per-pixel conditional confusion model (depth, PSF, and morphology jointly) and uncertainty propagation into the primary estimator. In its absence we hold the operative claims to the observed hard-label field (the strict $+0.635$ real-space null and finite observed-label pilot fractions) and treat the scalar- g physical-amplitude conversion as an illustrative spatially averaged mapping, not a spatially resolved physical bound.

The empirical DR8 axis-ratio cross-match finds $f_{\text{edge}} = 15.80\%$ (505,889 of 3,201,160 classified spirals at $b/a < 0.3$). A linear dilution estimate is useful only conditional on flip-symmetric hard labels. Soft equivariance motivates that condition, but hard argmax is non-linear; Appendix E therefore reports an empirical edge-on-isolated coherence test rather than claiming that directional bias is excluded. LSST extrapolations and spectroscopic-redshift upgrades remain future work.

6.3. Relation to Parity-Violating Sectors

The $\ell = 1$ morphology observable is parity-even and cannot by itself constrain primordial parity vio-

Table 10. Descriptive score fractions $P(\sigma > 3)$ versus injected full amplitude $A = A_p$ on the HC sample ($N = 949,584$; 100 injections per amplitude; per-pixel-shuffle null). The finite grid gives $P = 0.55, 0.91$, and 1.00 at $A = 0.75\%$, 1.0% , and 1.5% under the tabulated polar-angle-uniform convention; binomial standard errors are at most 0.05 . These pilot values are not calibrated recovery thresholds. The coverage-calibrated observed-label 95% sensitivity limit on the primary estimator, $A_{95}^{\text{obs}} \simeq 0.98\%$ (Eq. 7, 2,000 random axes/amplitude against the committed fixed-occupancy null), is reported separately and is an observed-label floor, not a physical limit.

A (%)	0.05	0.1	0.2	0.3	0.5	0.75	1.0	1.5	2.0
$P(\sigma > 3)$	0.01	0.01	0.01	0.03	0.15	0.55	0.91	1.00	1.00

lation. Translating it to cosmic birefringence, Chern–Simons gravity, or another early-universe sector requires a derived galaxy-formation transfer function, which this work does not provide. Accordingly, neither the HC null nor the diagnostic WLS/harmonic channels establish a physical parity-amplitude bound. A matched-footprint Ganalyzer analysis and a theory transfer calculation remain necessary before comparing likelihoods across estimators or sectors.

Downstream transfer-calibration roadmap.—A future physical calibration would require an object-level held-out human-label set with the historical training objects removed; image-level chirality interventions spanning imaging leg, depth, PSF, axis ratio, and morphology; propagation through the complete cutout, classifier, NOT_SPIRAL triage, confidence-selection, and map-making pipeline; and a spatially resolved directional confusion model carried into a joint estimator covariance. The first of these inputs now partially exists: the manifest-retained retrain supplies a provably training-disjoint object-level held-out GZ1 validation ($\kappa = 0.9733$; Sec. B), though only for the regenerable GZ1-core realization and without the CE-ResNet component. The remaining items—image-level interventions, full-pipeline propagation, and the spatially resolved transfer model—are a roadmap rather than a product of the present release. The current overlap-contaminated *released-classifier* GZ1 agreement and finite injection pilots cannot substitute for those inputs.

6.4. Open Follow-up and Future Directions

A shared block-bootstrap covariance of the four estimators (real-space dipole, WLS dipole, monopole, and the NaMaster MASTER-decoupled $\ell = 1$ power) is now computed on one common resampling (Sec. 4.5), narrowing the previously-open joint-covariance gate: it shows the monopole is a statistically distinct mode nearly uncorrelated with any dipole/harmonic estimator. Four scientific gates remain open: a physical morphology transfer function, a matched-footprint independent-estimator comparison, a full joint real-space/harmonic/nuisance likelihood (the MASTER-decoupled $\ell = 1$ leg is now computed, extending the

covariance to a NaMaster-complete 4×4 block; the remaining open piece is the joint likelihood itself), and independent external preregistration or blinding of the primary analysis. The catalog payload is public and commit-pinned; a DOI-backed paper/source archive remains a separate submission gate. The HC real-space estimator is the sole primary result; it is empirically less exposed to the measured harmonic leakage but does not mathematically bypass that channel. The WLS and harmonic analyses remain diagnostics.

7. CONCLUSIONS

We report one primary result from the 8,474,531-object observed-label catalog: the strict release-safe HC real-space dipole is null-consistent ($N_{\text{selected}} = 890,069$, $N_{\text{support}} = 887,472$, $z_{\text{mom}} = +0.635$, one-sided add-one rank $p = 0.23768$). Computed on the same estimator and the same fixed-occupancy null, this channel’s coverage-calibrated observed-label sensitivity reaches 95% detection coverage at $A_{95}^{\text{obs}} \simeq 0.98\%$ (full-amplitude; Eq. 7, Sec. 6.2)—an observed-label sensitivity floor, not a physical parity-amplitude bound, whose physical counterpart remains gated on the morphology transfer function. The repository candidate retains the safe catalog, quarantine, exact strict null array, receipt, and generator. The synchronized strict-primary contract is public at immutable HF dataset revision 911316f31c21f2c4b933a2f3a761274cfe85c6d6 (path `apjs-release/v1.0.259-strict-primary/`), byte-verified against the committed overlay; the historical public release documenting the earlier unsafe-inclusive realization remains retained at its own immutable revision and is superseded for the primary contract. All harmonic, WLS, injection, and transfer calculations are diagnostics. Our main conclusions are:

Harmonic-channel response diagnostic.—The apodized-footprint MASTER $\ell = 1$ injection pilot reports its observed-label score fractions and median injected scores in Table 11. Because this channel has a different field, mask, weights, and null from the primary estimator, these values diagnose pipeline response only; they do not establish completeness, a physical exclusion, or an “unmissable” amplitude.

Table 11. Diagnostic MASTER $\ell = 1$ injection scores (10^3 injections per amplitude and axis; artifact c9b). The 3σ score uses this channel’s label-shuffle null, and median z ranges are axis-dependent. These entries are not calibrated recovery probabilities, physical bounds, or interchangeable with the primary real-space result.

A_p	$P(\geq 3\sigma)$	Med. z (injected)	Obs. z
0.5%	0.92	—	historical estimator ^a
0.75%	≥ 0.999	—	—
1.7%	≥ 0.999	≈ 68 –218	—
3.0%	≥ 0.999	≈ 209 –685	—

NOTE—^aDifferent historical weighting/null implementation; not numerically compared with the exact-support diagnostic.

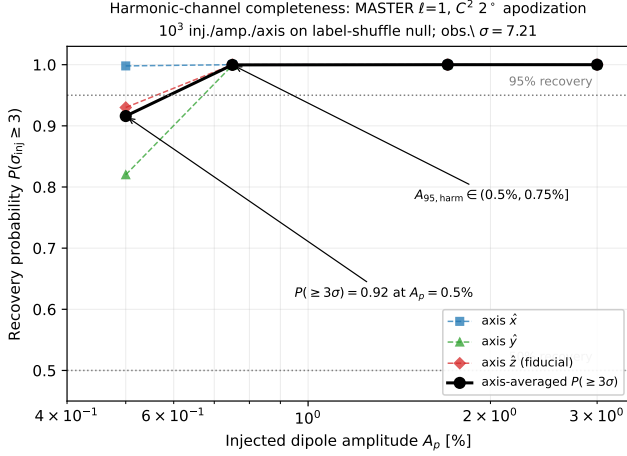


Figure 8. Historical injection-completeness diagnostic under the artifact-c9b apodized estimator convention (10^3 injections per amplitude and axis). It uses a different weighting/null implementation from the exact-support computation and is retained only as sensitivity provenance; it is not a calibrated recovery threshold or physical limit.

Diagnostic methodological finding: a quantifiable monopole-mask leakage channel.—A small uniform classifier monopole couples to the patchy survey mask and inflates the raw pseudo- C_ℓ at $\ell = 1$. A 500-realization monopole-only generative null reproduces 99.32% of the observed pre-MASTER power under that construction. Post-MASTER residuals remain non-primary and systematics-attributed. Because Ganalyzer uses a different estimator, sample, and selection, a likelihood-level comparison requires a matched-footprint analysis.

FS-C MASTER $\ell = 1$ direct compute.—The exact single-mode NaMaster execution uses the persisted 24,087-pixel FS-C mask ($N_{\text{spiral}} = 3,200,420$, $f_{\text{sky}} = 0.490051$), 500 fixed-occupancy label randomizations, and seed 42. It yields $C_1^{\text{data}} = 7.0113 \times 10^{-6}$, $z_{\text{mom}} = +6.923$, and add-one rank $p = 0.001996$; the matched C^2 2° apodized result is $+7.033$. These are systematics-attributed moment diagnostics, not Gaussian significances. The strict

HC real-space result ($N = 890,069$, $z_{\text{mom}} = +0.635$, $p = 0.23768$) anchors the scientific null verdict.

Bias hardening is essential.—Our raw (Catalog A) analysis produces a 2.31σ real-space dipole and a $+6.48\sigma$ pre-MASTER pseudo- C_ℓ from a classifier CW excess of only 0.79%, modulated by non-uniform survey coverage. Equivariant post-processing plus the explicit unsafe-row quarantine gives the strict primary real-space null $z_{\text{mom}} = +0.635$ ($p = 0.23768$). The exact-support harmonic residual remains systematics-attributed. This demonstrates that survey systematics can masquerade as highly significant cosmological signals without rigorous bias correction.

Finite injection pilot and unresolved calibration.—For the HC estimator, the Stage-B 20-axis pilot gives score-pass counts $[1, 16, 16, 20, 20]/20$ at $A = A_p = [0.5, 0.75, 1.0, 1.5, 2.0]\%$. A separate 100-injection pilot gives the descriptive fractions in Table 10. These finite, deterministic-axis experiments do not define calibrated 50% or 95% recovery thresholds, confidence limits, or physical bounds. A shared block-bootstrap covariance of the four estimators (now NaMaster-complete with the MASTER-decoupled $\ell = 1$ leg) is computed (Sec. 4.5) and shows the monopole is a distinct, nearly-uncorrelated mode; calibration across sky geometry, classifier transfer, and a full joint nuisance likelihood remains open. A from-scratch, manifest-retained retrain now makes the GZ1-core component of the training realization regenerable and supplies a provably training-disjoint held-out GZ1 validation ($\kappa = 0.9733$; Sec. B), narrowing the training-reproducibility and independent-validation gates for that component; the CE-ResNet catalog has since been re-provisioned and the 826-vs-846 composition conflict adjudicated (reproducible non-spiral count 819; Sec. 2), but a composition-faithful CE-included retrain collapses to chance on chirality, so the historical CE-included accuracy is not reproducible under honest ingestion, and the released catalog labels are unchanged. The catalog remains a public CC-BY-4.0 community resource.

Table 12. Consolidated mask/weight/apodization $\rightarrow$ sky-fraction mapping for every footprint quoted in this paper. Binary-mask rows quote the raw pixel fraction f_{sky} ; weighted/apodized rows quote $f_{\text{sky}}^{\text{eff}} = \langle W \rangle^2 / \langle W^2 \rangle$.

Mask	Weight / apod.	f_{sky}
Canonical ($N_{\text{spiral}}(p) \geq 10$)	binary, none	0.49005
Canonical	binary, C^2 2°	0.482
Footprint ($N_{\text{all}} \geq 1$)	binary, none	0.494
Footprint	binary, C^2 2°	0.488
Footprint	$W_p = N_{\text{all}}$, C^2 2°	0.452
Footprint	$W_p = N_{\text{spiral}}$, C^2 2°	0.420

APPENDIX

A. NAMASTER MASTER CONFIGURATION

For full reproducibility we record the NaMaster (`pymaster` 2.6) configuration used for the MASTER mode-coupling deconvolution of the chirality-asymmetry pseudo- C_ℓ .

Declared data vector and $\ell=0$ treatment.—The exact-support computation uses $f_{\text{CW}}(\hat{n}) - 0.5 = A_p/2$ on the persisted FS-C mask, followed by N_{spiral} -weighted mean subtraction for both data and every fixed-occupancy null field. The binomial-monopole construction uses the same mask and field convention but draws counts from the global-CW-fraction model. Historical MASTER-AGF and weighted-footprint computations use other supports or weights and are not combined with the exact-support synthesis.

Bandpower vs single- ℓ estimator distinction.—The reported MASTER $\ell=1$ result is the *single-multipole bin* from $\ell=1$ to $\ell=1$ (`nmt.NmtBin.from_lmax_linear(lmax=191, nlb=1)`, $\ell=1$ row of the bandpower matrix), NOT a bandpower over a range.

NaMaster configuration.—Pixelization: HEALPix NSIDE = 64 ($N_{\text{pix}} = 49,152$, $\ell_{\text{max}} = 191$). The canonical FS-C support is the persisted $N_{\text{spiral}}(p) \geq 10$ mask with 24,087 pixels, $f_{\text{sky}} = 0.490051$, $N_{\text{spiral}} = 3,200,420$, and mask SHA-256 499fa89e...f5852. The scalar field is $f_{\text{CW}} - 0.5 = A_p/2$ with N_{spiral} -weighted mean subtraction. Binary and C^2 2° -apodized weights derive from the identical base mask. Bins are single-multipole linear bins. The primary harmonic null comprises 500 fixed-occupancy galaxy-label randomizations with seed 42; the distinct binomial-monopole null uses 500 and 10,000 draws on the same support. The generator, arrays, mask, indices, and receipts are content-addressed in the computational-closure ledger.

MASTER-AGF declaration and depth-stratification audit.—Historical MASTER-AGF weighted-footprint calculations use a different support and weighting convention from the persisted FS-C mask. They are retained only as explicitly different-support provenance and do not enter the exact-support synthesis. The sole primary result is the release-safe HC real-space null; WLS remains diagnostic. A shared block-bootstrap covariance of the four estimators—now including the NaMaster MASTER-decoupled $\ell=1$ leg computed here (`pymaster` 3.0), which correlates at $\rho = +0.943$ with the paper’s `anafast` C_1 proxy—is computed (Sec. 4.5, Table 9), showing the monopole is a statistically distinct, nearly-uncorrelated mode; a full joint likelihood across all channels remains open.

B. CLASSIFIER ARCHITECTURE DETAILS

Training.—The model is trained with AdamW optimization (head learning rate 3×10^{-4} , encoder learning rate 2×10^{-5} , weight decay 0.02), batch size 128, and a cosine annealing warm-restart schedule ($T_0 = 10$, $T_{\text{mult}} = 2$). Early stopping with patience 15 monitors best validation *accuracy* within 80 epochs; the production checkpoint is reported at epoch 79. The immutable model-repository audit at revision 237d021c451d75cf86a875e86d4de498b74e2f12 reports 93.6878% best-epoch three-class validation accuracy, $n_{\text{train}} = 21,293$, and $n_{\text{val}} = 5,323$; the checkpoint’s LFS object has SHA-256 618d170f...97bf and size 88,046,038 bytes. These values support the reconstructed account but do not identify split membership or random states, and they conflict with the committed model README’s 92.10% accuracy and the benchmark counts below. Horizontal flips are evaluated on the fly in training batches and do not add rows.

The same recovered account records 94.9% CW per-class accuracy (CCW 91.3%, NOT_SPIRAL 99.4%), and binary CW/CCW accuracy 93.2% (CW recall 93.8%, CCW recall 92.6%); these remain historical metrics rather than an exactly replayable validation.

Training-data provenance.—Table 13 juxtaposes the conflicting committed historical records rather than selecting an unproved realization. The production-script/model-audit account totals 26,616 rows, with 826 CE-ResNet-selected non-spirals and a 21,293/5,323 split; BENCHMARK_REPORT.md instead gives 26,626 rows and 846 CE non-spirals, and the model README reports 92.10% rather than the audit’s 93.6878%. Neither split gains rows from flips. Because *historical* stochastic operations were not fully seeded and stable object/split IDs were not retained, the top block of the table is a conflict disclosure, not an exact replay manifest. The bottom block records the subsequent from-scratch retrain (Sec. 2): a fully seeded ViT-Small run whose `g1.training.manifest.json` retains every object ID, train/validation split index, and random state, reaching 99.31% best-epoch validation accuracy on $n_{\text{total}} = 8,637$ (6,637 Galaxy Zoo 1 confident spirals + 2,000 synthetic NOT_SPIRAL; train/val 6,910/1,727). This retrain makes the GZ1-core component regenerable going forward, but it is not a re-derivation of the historical realization: the CE-ResNet component (`ce_resnet_present = false`) is absent, so that GZ1-core retrain does not itself re-provision the $\sim 67.5\%$ CE-derived pool or resolve the 826-vs-846 conflict (a separate CE-included assembly, next paragraph, re-provisions the catalog and adjudicates that conflict), and it leaves the released Catalog C labels unchanged.

CE-composition adjudication and composition-faithful honest negative.—The CE-ResNet catalog was subsequently re-provisioned (Jia et al. 2023 CE-ResNet, `pre_desi.fits`; Zenodo DOI 10.5281/zenodo.7167388, arXiv:2210.04168; file SHA-256 894dbe887140c165488a0f6053e2cd21f4ab72be9b06ece733e6ce177c0e304b), and the full historical composition was regenerated by the *exact* wrapper assembly functions (`build_dataset/make_split/write_manifest`) under seed 42 and `gz_desi` revision b7583bb2ac445e93c5447a08063acd7c1477fd13 (driver `pipelines/p2_chirality/scripts/g1_ce_composition_assembly.py`, provenance `pipelines/p2_chirality/outputs/g1_full_composition/PROVENANCE.md`). The deterministic assembly reproduces the two large historical components *exactly*: 6,637 GZ1 confident CW/CCW and 17,153 CE-ResNet spirals. It yields 819 CE non-spirals—*neither* the audit’s 826 *nor* the benchmark’s 846—for 26,609 total rows (class counts CW/CCW/NOT_SPIRAL = 11,904/11,886/2,819; train/validation 21,288/5,321; manifest `pipelines/p2_chirality/outputs/g1_full_composition/g1_full_composition_manifest.json`, SHA-256 431f84f09519d1ef8be9e2f488f199b5d6b1c5127a77d4e25f53df04cf110777; independently reproduced on an A4000 pod retrain to the identical counts). Because $6,637 + 17,153 + 826 + 2,000 = 26,616$ reproduces the smaller historical record’s arithmetic exactly, the entire 826-vs-846 / 26,616-vs-26,626 conflict is *isolated* to the seeded 50,000-object subsample crossmatch that selects CE non-spirals from the 74,174 non-spiral candidates within $3''$: its exact count is subsample- and boundary-sensitive, precisely the irreducible ambiguity Table 13 flags. The regenerable value is 819; the conflict is now adjudicated (isolated) rather than left open for want of the external catalog.

A composition-faithful CE-included retrain of this same 26,609-object realization—identical wrapper code, identical seeds, manifest-retained—does *not* reproduce the historical 93.6878%/92.10% CE-included headline. It collapses to chance on chirality: best three-class validation accuracy 0.5617 at epoch 11 (26 epochs, early-stopped), with training accuracy pinned at 0.667 throughout—the WEIGHTEDRANDOMSAMPLER arithmetic of a perfectly-learned NOT_SPIRAL class (1/3) plus pure coin-flip chirality (half of 2/3)—and unsampled validation at $0.106 + 0.894 \times 0.5 = 0.553$. Per-source validation (`pipelines/p2_chirality/outputs/g1_full_composition/pod_evidence/g1_per_source_eval.json`) shows even the clean held-out GZ1 spirals at binary CW/CCW agreement 0.5170 ($N = 1,350$) and CE spirals at 0.5087 ($N = 3,407$)—both near 0.5 and *neither* agreeing nor anti-agreeing, i.e. a scrambled/never-formed chirality feature, not a systematic label flip—while the NOT_SPIRAL and synthetic subsets are learned perfectly (0.994, 1.000). Four alternative explanations are ruled out by evidence: (i) *label noise*—CE argmax labels are decisive at the catalog level, only 0.2% of confident spirals having $|P_{\text{CW}} - P_{\text{ACW}}| < 0.1$ (median margin 0.62); (ii) *parity-convention flip*—a model-independent catalog cross-match on the 38,617 galaxies present in *both* the GZ1-confident and CE-confident sets finds CE label = GZ1 label at 0.9972 (near-perfectly diagonal confusion), so CE and GZ1 share an identical chirality convention; (iii) *image-join bug*—the CE crossmatch uses the same $3''$ k -d tree as the GZ1 path that trains to 0.99, and the GZ1-subset images learn fine from the same `gz_desi` stream; (iv) *optimization*—an identical-optimizer, identical-model GZ1-only run broke out of the same 0.66/0.62 plateau by epoch 5 to 0.9931, whereas here training accuracy is stuck at 0.667 for all 26 epochs, so the model cannot even *fit* chirality on train (data, not learning rate). The root cause is that the CE-only pool—the 72% of spiral labels matched to CE but not GZ1-confident—is systematically

Table 13. Conflicting historical training-data records (top block), together with the from-scratch manifest-retained retrain that supersedes the *reproducibility* of the GZ1-core component going forward (bottom block). “Audit” denotes the immutable model-repository audit; “benchmark” denotes the committed BENCHMARK_REPORT.md. Neither *historical* record retains exact object/split membership or random states. The retrain (manifest SHA-256 `e5de8e03...240a10729`; checkpoint SHA-256 `aed109dc...a752387`) retains every object ID, split index, and random state, but is the GZ1-core + synthetic realization only (`ce_resnet_present` = false) and does not change the released Catalog C labels. Flips are generated on the fly and do not add rows. The third block records the CE-composition adjudication (Appendix B text): with the CE-ResNet catalog re-provisioned (Zenodo 10.5281/zenodo.7167388) the seeded assembly reproduces the 6,637 GZ1 and 17,153 CE-spiral counts *exactly* and yields 819 CE non-spirals, isolating the 826-vs-846 conflict to the seeded non-spiral subsample crossmatch; the composition-faithful CE-included retrain is a documented honest negative (chance on chirality, val 0.5617), so the historical CE-included accuracy is not reproducible under honest ingestion.

Quantity	Value
Audit: GZ1 / CE-spiral labels	6,637 / 17,153
Audit: CE-selected / synthetic NOT_SPIRAL	826 / 2,000
Audit: train / validation / total	21,293 / 5,323 / 26,616
Benchmark: CE-selected NOT_SPIRAL / total	846 / 26,626
Audit / model-README val. accuracy	93.6878% / 92.10%
Audit CW/CCW/NOT_SPIRAL per-class acc.	94.9/91.3/99.4
Exact historical object/split manifest	not retained
Full-overlap GZ1 chirality agreement	69.91%
Retrain: GZ1-core / synthetic NOT_SPIRAL	6,637 / 2,000
Retrain: train / validation / total	6,910 / 1,727 / 8,637
Retrain: CE-ResNet component	absent (<code>ce_resnet=false</code>)
Retrain: best-epoch val. accuracy	99.31% (epoch 47)
Retrain object/split/seed manifest	retained (all IDs)
Retrain training-disjoint GZ1 val.	$\kappa = 0.9733$ (Sec. B)
CE catalog re-provisioned	Zenodo 10.5281/zenodo.7167388
file SHA-256	894dbe88...0e304b
Reproduced GZ1 / CE-spiral (exact)	6,637 / 17,153
Reproducible CE non-spiral / total	819 / 26,609
Arithmetic check (+826)	6637+17153+826+2000=26,616
CE-included retrain val. accuracy	0.5617 (chance on chirality)
held-out GZ1 CW/CCW agree	0.517 (at chance)
CE-only pool (72% of spirals)	fainter/smaller/low-conf. (App. B)

fainter (median FLUX_R = 16.89 vs. 16.15 for the $\text{CE} \cap \text{GZ1}$ overlap), smaller (median SHAPE_r = 4.26'' vs. 5.77''), and low-confidence (median CE winner-probability 0.569 vs. 0.899; CE-only winner-probability 10/25/50/75/90 percentiles 0.426/0.471/0.569/0.713/0.835, the 25th barely above the 0.40 admission threshold): chirality supervision a 224px-resized fine-tuned ViT cannot resolve, which swamps the clean 28% GZ1 gradient and prevents the chirality feature from forming for *all* spirals, GZ1 included. Raising the CE confidence threshold would redefine the experiment (defeating the purpose of reproducing the full historical 17,153-spiral composition) and is not pursued. This negative *corroborates and strengthens* the Sec. 2 disclosure: the historical CE-included realization is not reproducible under honest, seeded, manifest-retained ingestion of the same composition and same code, consistent with validation on the ~67.5% CE-derived pool substantially measuring agreement with the CE model. The GZ1-core retrain (99.31%, `ce_resnet_present` = false) remains the honest regenerable realization and is not superseded; the released Catalog C labels are unchanged, and their historical provenance is disclosed as before. Full diagnosis: [pipelines/p2_chirality/outputs/g1_full_composition/DIAGNOSIS_ce_included_retrain.json](#).

Flip-equivariance consistency loss.—The loss combines class-weighted cross-entropy $\mathcal{L}_{\text{CE}}$ with a flip-equivariance consistency term:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \cdot \frac{1}{N} \sum_{i=1}^N \|\mathbf{p}(x_i) - S\mathbf{p}(\tilde{x}_i)\|^2, \quad (\text{B1})$$

where S is the permutation matrix swapping the CW and CCW channels (leaving NOT_SPIRAL unchanged), and $\lambda = 0.5$.

D_4 -TTA rotational-equivariance validation.—A direct D_4 -TTA hold-out on two $\sim 2,000$ -galaxy subsamples ($N = 1,558$ and $N = 1,988$) finds: (a) mean per-galaxy P_{CW} stable under Z_2 and D_4 to within $|\Delta\langle p_{\text{CW}} \rangle| < 0.0016$; (b) per-galaxy

argmax labels flip in 21.4% of cases between Z_2 and D_4 on borderline galaxies. The sign reversal of the argmax-CW-fraction shift (-1.35% at $N=1,558$ vs $+2.11\%$ at $N=1,988$) is compatible with sample variation in a fragile argmax statistic, but without an overlap or uncertainty analysis it does not exclude a D_4 -dependent systematic.

Bias hardening suite.—We subject the classifier to seven tabulated targeted bias tests (Table 14): T1 flip-swap consistency ($r > 0.80$), T2 rotation stability ($> 80\%$ agreement across 60° increments), T3 artifact rejection ($> 70\%$ NOT_SPIRAL for blank/scrambled images), T4 perturbation robustness ($> 80\%$ agreement under Gaussian blur + brightness dimming), T6 hemispheric null ($< 10\%$ CW difference between hemispheres), T7 confidence-calibration proxy, T8 CW/CCW balance ($50\% \pm 10\%$). Metadata (RA/Dec) leakage is *not* tested by a linear Pearson correlation: RA is a circular coordinate ($0^\circ \equiv 360^\circ$) for which a linear Pearson r is statistically inappropriate and can understate azimuthal coupling, so the former linear-Pearson-vs-RA row has been removed from the battery. Directional (RA-dependent) leakage is instead tested correctly by the map-level low- ℓ real- $Y_{\ell m}$ regression described below, which respects the coordinate’s circularity. The implemented T7 criterion is: $> 30\%$ of predictions at $\max p > 0.9$ (a confidence-mass sanity check), together with the requirement that the flip-swap error of high-confidence ($\max p > 0.9$) predictions be lower than that of low-confidence ($\max p < 0.7$) predictions; the catalog-wide high-confidence mass is 73.6% (Fig. 6), so T7 passes. Quantified catalog-wide from the per-galaxy original/flip outputs (the flip-pass probabilities are recovered from the stored raw and 2-fold-TTA columns via $p_{\text{CCW}}^{\text{flip}} = 2p_{\text{CW}}^{\text{eq}} - p_{\text{CW}}^{\text{raw}}$ and its channel companions; a dedicated QC pass quantifies the fidelity of this recovery: the recovered-flip normalization $\sum_c p_c^{\text{flip}} = 1$ holds at float32 storage precision on all 8.47M rows — max deviation 4.3×10^{-7} — but for 2.9% of rows (1.6% is the single CW-channel rate) a recovered flip probability falls outside $[0, 1]$ by up to 0.09. These excursions are *not* float32 rounding: they occur exclusively on rows whose raw probabilities derive from the separate raw-catalog inference pass rather than the equivariant pass (the 88,278-row intersection where both raw legs *and* the equivariant raw companion columns are populated shows zero violators), i.e. a raw/eq pipeline-pass mismatch at the $\lesssim 0.09$ probability level for that subpopulation (catalog-wide rate from [pipelines/p2_chirality/outputs/canonical_provenance/ext4_fb1_flip_identity_qc_catalogwide.json](#); intersection-subset rate zero by construction, [pipelines/p2_chirality/outputs/canonical_provenance/ext3_nfm1_flip_identity_qc.json](#)). The public v1.0.244 release’s `raw_flip_qc_unsafe` flag identifies the affected rows under the catalog-wide “full-coverage raw columns” definition; excluding them from the HC sample (59,515 of 949,584, 6.3%) leaves the real-space dipole null-consistent and essentially unchanged ($z = +0.48$ excluded vs. $+0.52$ baseline under the c11b 10^4 -permutation convention; artifact [pipelines/p2_chirality/outputs/canonical_provenance/ext3_nfm1_hc_dipole_qc_rerun.json](#)): mean flip-swap error 0.267 (median 0.0006) at $\max p > 0.9$ vs. 0.383 (median 0.364) at $\max p < 0.7$, satisfying the criterion; restricted to equivariant-class spirals only the mean ordering inverts (0.698 vs. 0.464), driven by the raw/equivariant class-disagreement subpopulation (the QC edge-case flag); architecturally, galaxies that change class under flipping are precisely the borderline objects for which TTA suppresses the raw argmax probability toward ~ 0.5 — these low post-TTA p_{eq} spirals have large raw flip-swap errors, pushing the equivariant-class-only mean flip-swap error above the full-catalog value, which is why T7 is a calibration proxy on all classes rather than a spiral-only reliability statement (artifact [pipelines/p2_chirality/outputs/canonical_provenance/c16_r24conf_pod_batch.json](#)). (T7 is a calibration *proxy*, not a ground-truth reliability-curve ECE measurement, which is not possible catalog-wide absent per-galaxy truth labels.) Note also that T1 validates the *implementation* of the equivariant protocol (it would catch a code defect) rather than constituting an independent statistical test, since flip-swap consistency holds by construction after TTA averaging. Directional (RA-dependent) metadata leakage is tested at the map level, which respects RA’s circularity: a low- ℓ real- $Y_{\ell m}$ regression ($\ell \leq 3$, 16 coefficients) of the primary HC A_p map against a 2000-permutation pixel null finds all three $\ell = 1$ coefficients consistent with zero ($|z| \leq 1.25$); the only outlying coefficient is at $(\ell, m) = (3, -1)$ ($z = -4.4$), consistent with the coherent low- ℓ systematic structure dispositioned in Appendix D rather than with a dipole (artifact [pipelines/p2_chirality/outputs/canonical_provenance/c12_r24conf_local_batch.json](#)). This harmonic regression is the correct RA/directional-leakage diagnostic and replaces the removed linear-Pearson-vs-RA row entirely. All 7 tabulated tests pass at the stated criteria; they are necessary but not sufficient checks for sub-percent bias control and must be interpreted alongside the finite-grid score fractions reported in Sec. 6.2. The equivariant averaging of Catalog C provides one bounded structural mitigation against flip-swap inconsistency; it does not establish complete bias removal.

GZ1 confusion matrix and per-class metrics.—Table 15 reports the three-class confusion matrix of the equivariant classifier against Galaxy Zoo 1 human labels on the full $1''$ cross-match ($N=240,919$). The retained artifact does not

Table 14. Bias-hardening test results. All 7 tabulated tests pass at the stated criteria; they are necessary but not sufficient checks for sub-percent-level bias control and are not calibrated against a physical sensitivity boundary (see Appendix B text for definitions and the T1/T7 scope caveats). The former linear-Pearson RA/Dec metadata-leakage row has been *removed* because a linear Pearson correlation is inappropriate for the circular RA coordinate ($0^\circ \equiv 360^\circ$); the map-level low- ℓ real- $Y_{\ell m}$ regression in the Appendix B text is the correct directional-leakage test and supersedes it.

Test	Threshold	Result
T1: Flip-swap r	> 0.80	1.000
T2: Rotation stability (historical)	$> 80\%$	89.8%
T3: Artifact rejection	$> 70\%$ NOT_SPIRAL	100%
T4: Perturbation robustness (historical)	$> 80\%$	84%
T6: Hemispheric difference (historical)	$< 10\%$	3.6%
T7: Calibration proxy	$> 30\%$ at $\max p > 0.9$	73.6%
T8: CW/CCW balance	$50 \pm 10\%$	49.7%

Table 15. Three-class confusion matrix vs. Galaxy Zoo 1 human labels ($1''$ cross-match, $N = 240,919$). Rows: GZ1 label; columns: equivariant (Catalog C) prediction.

GZ1 \ pred.	cw	ccw	not_spiral
CW	39,011	18,889	13,715
CCW	16,377	42,928	13,720
NOT_SPIRAL	17,056	19,724	59,499

contain an object-level anti-join against the 6,637 GZ1 training rows, and the historical training script did not retain stable training IDs; this evaluation must therefore be treated as overlap-contaminated rather than independent. The three-class accuracy is 58.7%; restricted to the 117,205 GZ1 spirals that the classifier labels CW or CCW, the chirality agreement is 69.91% (Cohen’s $\kappa = 0.40$). Per-class precision/recall: CW 0.539/0.545, CCW 0.527/0.588, NOT_SPIRAL 0.684/0.618. These descriptive values do not establish a physical transfer function.

Training-disjoint validation of the manifest-retained retrain.—The from-scratch retrain of Sec. 2 (Table 13, bottom block) enables the first *provably* training-disjoint held-out evaluation on this problem. Using the retained object-ID/split manifest, we hold out 3,000 Galaxy Zoo 1 *confident* spirals ($\max(P_{\text{CW}}, P_{\text{ACW}}) > 0.7$) that are disjoint from the retrain along *both* axes: they are drawn only from Galaxy Zoo DESI rows never scanned during training, and every candidate within $3''$ of a training-manifest object is removed (both overlap counts are exactly zero; artifact [pipelines/p2_chirality/analysis/g2_disjoint_validation_v1_0_266.json](#)). On this set the retrain checkpoint (SHA-256 `aed109dc...a752387`, loaded with zero missing/unexpected keys) reaches accuracy 0.9867 and Cohen’s $\kappa = 0.9733$ (confusion $[[1460, 10], [30, 1500]]$ for [CW, CCW]; CW $F_1 = 0.9865$, CCW $F_1 = 0.9868$; zero NOT_SPIRAL predictions, mean NOT_SPIRAL softmax probability 3.5×10^{-6}), modestly below its in-training validation accuracy of 99.31% as expected for genuinely held-out data. *This $\kappa = 0.9733$ figure is neither comparable to, nor a replacement for, the $\kappa = 0.40$ / 69.91% figure above; the two measure fundamentally different quantities.* The $\kappa = 0.40$ value is the *released* Catalog C classifier’s chirality agreement with GZ1 human votes on the *full*, confidence-unrestricted, overlap-contaminated cross-match ($N = 117,205$ chirality-evaluable spirals). The $\kappa = 0.9733$ value is CW/CCW discrimination by the *regenerable retrain checkpoint*—not the released classifier—on *high-confidence* ($P > 0.7$) GZ1 spirals that are *provably training-disjoint*. The higher value reflects all three of these differences (different checkpoint, high-confidence-only subset, and clean held-out labels) and must not be read as an improvement of, or a revision to, the released catalog’s descriptive human-agreement figure, which stands unchanged. The disjoint result establishes that a manifest-retained realization generalizes CW/CCW handedness on a clean held-out GZ1 sample—directly addressing the “no independent held-out validation / overlap-contaminated GZ1” gate for the regenerable realization—but it exercises only the GZ1-core component (the CE-ResNet pool is absent from the retrain, `ce_resnet_present = false`) and it does not supply the spatially resolved confusion/transfer model that the dipole analysis still requires.

Probabilistic calibration scope.—The catalog-wide mean winning-class score is 0.951, while the full GZ1 overlap has lower descriptive accuracies (58.7% three-class and 69.91% on the chirality-evaluable subset). These quantities are computed on different object sets and therefore cannot be combined into an Expected Calibration Error or a Jensen lower bound.

Table 16. GZ1-overlap human-vs-classifier CW/CCW confusion stratified by two declination-defined imaging proxies and confidence cut ($N = 40,987$ GZ1-confident spirals classified CW or CCW; $1''$ cross-match, $N = 46,017$). The southern proxy includes DES and is not a separate DES validation. “Asymmetry” is the difference between the two directional error rates, with an unpooled two-proportion 95% interval. A value consistent with zero means no dipole-producing parity error was detected in that tested stratum; it does not exclude finer spatial variation. Full arrays: [pipelines/p2_chirality/outputs/gz1_stratified_confusion.json](#).

Stratum	N	acc	CW $\rightarrow$ CCW / CCW $\rightarrow$ CW	asymmetry (95% CI)
Overall	40,987	0.912	0.0878 / 0.0877	+0.0001 [−0.0054, +0.0056]
BASS+MzLS proxy ($\text{dec} > +32^\circ$)	14,093	0.903	0.0994 / 0.0952	+0.0042 [−0.0056, +0.0140]
DECaLS+DES proxy ($\text{dec} < +32^\circ$)	26,894	0.917	0.0818 / 0.0838	−0.0020 [−0.0086, +0.0046]
Science cut ($p_{\text{eq}} > 0.6$)	32,550	0.961	0.0379 / 0.0399	−0.0019 [−0.0061, +0.0023]
Low conf. ($p_{\text{eq}} \leq 0.6$)	8,437	0.724	0.2795 / 0.2730	+0.0065 [−0.0126, +0.0256]

No same-object binned reliability artifact is retained, so we make no numerical ECE claim. The scores remain explicitly uncalibrated ranking outputs. This does not alter the hard-label null: the primary estimator consumes hard argmax CW/CCW counts, and a monotone recalibration preserves the argmax. Recalibration of the selection score would require restating the numerical threshold that selects a fixed ranked subset; it would not by itself provide the missing spatial confusion/transfer model.

Sky- and confidence-stratified human-label confusion analysis.—The single global confusion matrix above (Table 15) does not by itself establish the spatial transfer needed by the dipole analysis. A spatially varying differential mislabelling rate can inject a spurious A_p dipole, but it is not the only channel: Eq. (8) shows that a spatially varying symmetric rate also changes the observed fraction wherever the local selected CW fraction differs from one half. We therefore stratify the GZ1-overlap confusion matrix by two declination-defined imaging proxies (BASS+MzLS proxy, $\text{dec} > +32^\circ$; DECaLS+DES proxy, $\text{dec} < +32^\circ$) and by the primary confidence cut, and in each stratum measure the CW $\leftrightarrow$ CCW error asymmetry with its two-proportion 95% CI (Table 16) as well as the net induced Δf_{CW} . These are not the three actual DR8 campaign memberships: the southern proxy includes the DES overlap region, so this analysis does not separately validate DES. The analysis restricts to the $N = 40,987$ GZ1-confident spirals ($\max(P_{\text{CW}}, P_{\text{ACW}}) > 0.6$) that the classifier also labels CW or CCW, drawn from the $N = 46,017$ GZ1-catalog cross-matches reproduced byte-identically ($1''$ KDTree match) from Table 15’s parent match; the 5,030 GZ1-confident spirals assigned to NOT_SPIRAL are reported separately as an unquantified completeness/selection channel, not as chirality error. The generator script and full stratified JSON are committed ([pipelines/p2_chirality/scripts/gz1_stratified_confusion.py](#), [pipelines/p2_chirality/outputs/gz1_stratified_confusion.json](#)).

The 5,030 GZ1-confirmed spirals assigned to NOT_SPIRAL have not been composition- and position-characterized here; no chirality-neutral assumption is made for them, and their exclusion is an unquantified completeness/selection channel. All confusion-matrix conclusions are restricted to the 40,987 objects classified CW or CCW.

Conclusion this licenses (stated precisely). CW $\leftrightarrow$ CCW error asymmetry is consistent with zero in every tested stratum: overall +0.0001 [−0.0054, +0.0056], BASS+MzLS proxy +0.0042 [−0.0056, +0.0140], DECaLS+DES proxy −0.0020 [−0.0086, +0.0046], and science cut −0.0019 [−0.0061, +0.0023]. Thus no directional error asymmetry is detected at the tested two-declination-stratum resolution, with sensitivity bounded by these intervals; DES is not tested separately. The net induced-fraction point estimates are nevertheless 0.00142 (northern proxy) and 0.00331 (southern proxy); their difference has no retained covariance-aware interval. The test does not resolve RA variation within a stratum, actual three-campaign membership, or joint depth/PSF/morphology dependence, so it cannot exclude every sub-percent spatially varying differential or symmetric-error pattern. The finer conditional model remains open. Stratified accuracies apply to a nested confident-spiral subset and do not validate the illustrative scalar transfer, especially because the historical training overlap was not removed from the GZ1 agreement calculation. Closing the physical-parity transfer gate requires handedness-symmetric, pixel-level forward simulations and injections spanning the observed DESI depth, seeing, PSF, imaging leg, morphology, redshift, and selection distributions, followed by blinded recovery and validation of a spatially varying confusion/transfer model against held-out labels.

C. AUXILIARY DIPOLE DIAGNOSTICS

This appendix collects the signal-hunt diagnostics, two-point chirality correlation, hemisphere asymmetry, sky region balance, per-imaging-leg systematics, scale dependence, and confidence stratification results moved from the main text for conciseness. The primary no-dipole verdict is unchanged by any of these diagnostics.

Confidence-stratified dipole.—Stratifying Catalog C by equivariant max-class probability into five bins reveals: the $+3.29\sigma$ in the 1.87M-galaxy $[0.5, 0.6)$ bin does not survive the sample-purity ladder (cutting to $p_{\text{eq}} > 0.6$ gives -0.03σ); results available in the public data repository (see Data Availability).

Sky-quadrant and hemisphere diagnostics.—Splitting into four RA quadrants gives per-quadrant dipole values ranging from -0.82σ to $+2.49\sigma$; primordial dipole would project consistently, not scatter. The NGP ($b > 0$) gives $\sigma_{\text{iso}} = +0.47$ and the SGP ($b < 0$) gives $+2.02$ against independent per-pixel binomial draws at $p = 0.5$. This diagnostic null is distinct from the primary fixed-occupancy randomization; the SGP value is compatible with dust-correlated foreground structure.

Hemisphere asymmetry and look-elsewhere.—Testing all hemisphere pairs on the local 10° -spaced grid ($36 \times 18 = 648$ directions) gives a maximum asymmetry of 3.05σ against its label-shuffle null. This local scan is *not* the same computation as the authoritative look-elsewhere artifact: the latter recomputes the maximum statistic over 768 NSIDE_{dir} = 8 pixel-center directions for each of 10,000 global label shuffles and reports the Monte-Carlo precision-floor bound $p_{\text{LEE}} = 1/10001$; 95%upperlimit $\simeq 3.0 \times 10^{-4}$ (artifact [pipelines/p2_chirality/outputs/canonical_provenance/mc_seed_manifest.json](#)). We therefore do not describe the 768-direction Monte Carlo as an exact correction of the 648-direction diagnostic. Both scans reject their corresponding random-label nulls and are interpreted as depth/training-label systematic structure rather than a primordial dipole. Bonferroni control does *not* require independent tests; the earlier sentence claiming that it did was false. We nevertheless do not use the Gaussian Bonferroni number here because the hemisphere amplitude is positive-definite with non-Gaussian tails and the two grids/statistics differ; the committed 768-direction empirical max-statistic is the authoritative look-elsewhere result.

Two-point chirality correlation.—The two-point chirality correlation $w_{\text{CW}}(\theta)$ on a random 50,000-galaxy HC-spiral sample is consistent with the label-shuffle null at $|\sigma| < 1.2$ in 9 of 10 bins; the maximum deviation -2.41σ at $\theta \approx 0.5^\circ$ is attributable to DESI Legacy DR8 brick-boundary classifier artifacts (confirmed by vanishing to -0.03σ in the brick-interior subsample).

Per-imaging-leg systematics.—The full-catalog $[0.5, 0.6)$ confidence bin $+3.29\sigma$ decomposes as BASS+MzLS $+0.30\sigma$ / DECaLS $+4.50\sigma$ / DES $+2.46\sigma$: the signal is DECaLS-concentrated, the signature of a footprint-correlated systematic rather than a primordial isotropy-breaking signal. Under the 15-cell joint label-shuffle max-statistic null ($N_{\text{MC}} = 5,000$ global label shuffles preserving the total CW count; per-cell statistic $|\sigma|$, i.e. two-sided per cell), the family-corrected p -value is the add-one one-sided empirical exceedance of the observed $\max |\sigma| = 4.72$ in the joint null distribution: $p = (43 + 1)/(5000 + 1) = 0.00880$ ($\approx 2.4\sigma$ family-wise), appropriately downgraded from the cell-level $+4.72\sigma$. This empirical joint correction is applied *once* (no double correction); a Gaussian Bonferroni-15 estimate would underpredict this family-wise p by $\sim 250\times$ because the joint null is heavy-tailed.

D. FS-C SYSTEMATIC ANALYSIS

Eight diagnostics have been evaluated in connection with the FS-C residual, but only apodized-mask robustness and multipole-spectrum coherence are bound to the declared 24,087-pixel FS-C base support. Quality-quartile stratification constructs separate per-quartile masks; the leg proxy uses 24,270 pixels; boundary-distance variance uses 35,438 latitude-mask pixels; the density null and WLS/bootstrap use broader latitude-cut supports; and the direct cross-spectrum artifact does not record enough mask provenance to establish identity. Those six calculations describe other supports and do not enter the strict FS-C synthesis.

Apodized-mask robustness.— C^2 2° apodization derived from the persisted base mask gives $z_{\text{mom}} = +7.033$, compared with $+6.923$ for the binary mask. The residual therefore persists under this apodization; both values are fixed-occupancy null moment ratios, not Gaussian significances.

Multipole-spectrum diagnostic.—The exact-support vector is $z_{\text{mom}} = (+6.923, +2.931, +3.089, -0.317, +1.606)$ for $\ell = 1-5$. Elevated power across $\ell = 1-3$ indicates coherent low- ℓ structure; it is not interpreted as a clean primordial dipole.

1549 *Quality-quartile stratification.*—Stratifying the spiral sample into four p_{eq} quartiles ($N \approx 800,290$ each; $N_{\text{MC}} = 50$ per
 1550 quartile) gives $\ell = 1$ significances of $+0.20, -0.42, +0.44, +0.43$. Each quartile constructs its own $N_{\text{spiral}} \geq 10$ mask
 1551 ($f_{\text{sky}} = 0.4909\text{--}0.4924$), so these are different-support diagnostics rather than tests on the fixed FS-C mask and are
 1552 excluded from the strict synthesis.

1553 *Leg-proxy $\ell = 1$ partial closure.*—The historical calculation gives $r_{\ell=1}(\text{BASS}+\text{MzLS} \times A_p) = +0.65$, $r_{\ell=1}(\text{DES} \times A_p) = -0.73$,
 1554 and a summed leg-induced amplitude of $\sim 25\%$ of its observed $\ell = 1$ amplitude. Its recorded $f_{\text{sky}} = 0.493774$ corresponds
 1555 to 24,270 pixels, not the 24,087-pixel FS-C mask, so it is excluded from the strict synthesis.

1556 *Density-stratified null.*—The historical generator permuted A_p within ten pixel-density deciles and reported null mean
 1557 $C_1 = 3.44 \times 10^{-6}$, std 3.07×10^{-6} , and $z = +3.80$. Its mask is $|b_{\text{gal}}| > 15^\circ$ with $N_{\text{total}} > 0$ ($f_{\text{sky}} = 0.740926$), not FS-C
 1558 ($f_{\text{sky}} = 0.490051$), so this number is not evidence about the FS-C residual and is excluded from the synthesis.

1559 *Boundary-distance variance check.*—The historical five-shell calculation reports a per-shell weighted-variance range
 1560 below 11%, but its artifact records $f_{\text{sky}} = 0.740926$ and 35,438 shell pixels. It is therefore a latitude-mask result, not
 1561 an FS-C boundary test, and is excluded from the strict synthesis.

1562 *Joint nuisance-marginalized WLS fit.*—The historical 9-template WLS fit used $(|b_{\text{gal}}| > 15^\circ) \wedge (N_{\text{total}} > 0)$ (24,187
 1563 pixels), not FS-C (24,087 pixels). On that broader support it gives $A_{\text{dipole}}^{\text{best}} = 4.55 \times 10^{-3}$ and an NSIDE= 8 block-
 1564 bootstrap width $\sigma_{\text{boot}} = 1.63 \times 10^{-3}$. These values are retained for provenance only: the bootstrap covariance has not
 1565 been recomputed on FS-C, so neither $z \approx -7.6$ against $A_{\text{ref}} = 0.017$ nor the illustrative $1.36 \sigma_{\text{boot}}$ transfer comparison
 1566 is used as an FS-C or physical constraint.

Table 17. Historical 9-template WLS fit on the broader latitude-cut support, not FS-C. The centered imaging-leg templates are collinear with the constant (rank 8; condition number 4.5×10^{16}), so individual leg coefficients are not meaningful. Values are retained for provenance and are not used in the FS-C synthesis.

Template	$\hat{a}$	σ_{naive}	z
dipole $\hat{x}$	$+4.3 \times 10^{-5}$	8.9×10^{-5}	+0.5
dipole $\hat{y}$	-4.52×10^{-3}	1.04×10^{-4}	−43.3
dipole $\hat{z}$	-5.7×10^{-4}	2.8×10^{-4}	−2.1
leg BASS+MzLS	$+1.8 \times 10^{-3}$	6.2×10^2	—
leg DECaLS	$+8.7 \times 10^{-4}$	6.2×10^2	—
leg DES	-3.0×10^{-3}	6.2×10^2	—
pixel density	$+3.2 \times 10^{-4}$	8.5×10^{-5}	+3.8
pixel density ²	-1.9×10^{-5}	7.4×10^{-6}	−2.6
constant	$+3.2 \times 10^{-4}$	6.5×10^{-5}	+4.9
A_{dipole} (A_p units)	4.55×10^{-3}	$\sigma_{\text{boot}} = 1.63 \times 10^{-3}$ $\sigma_{\text{naive}} = 1.11 \times 10^{-4}$	
z vs. $A_{\text{ref}} = 0.017$	−7.6 (block-boot., diagnostic) −112 (naive WLS, superseded)		

[†] *Convention mapping:* the fitted $A_{\text{dipole}}^{\text{best}} = 4.55 \times 10^{-3}$ uses $A_p = 2(f_{\text{CW}} - \frac{1}{2})$, numerically equal to the full dipole amplitude A for $p_{\text{CW}}(\hat{n}) = \frac{1}{2}(1 + A \cos \theta)$. Injection tables use the same units but remain finite pilot diagnostics, not calibrated thresholds.

1567 *WLS support.*—The committed bootstrap artifact records 24,187 pixels under $(|b_{\text{gal}}| > 15^\circ) \wedge (N_{\text{total}} > 0)$, whereas
 1568 FS-C contains 24,087 pixels under $N_{\text{spiral}} \geq 10$. A new bootstrap on FS-C would be required before this diagnostic
 1569 could enter the FS-C synthesis.

1570 *Direct cross-spectrum $A_p \times n_{\text{total}}$.*—The cross-correlation coefficient is defined per multipole as $r_\ell \equiv C_\ell^{A \times n} / \sqrt{C_\ell^{AA} C_\ell^{nn}}$.
 1571 At $\ell = 2$, the historical result is $r_{\ell=2} = -0.65$ with signed $z = -2.89$ against a 200-realization permutation null. The
 1572 artifact does not record the exact mask or pixel count, so this value is retained as provenance but excluded from the
 1573 strict FS-C synthesis.

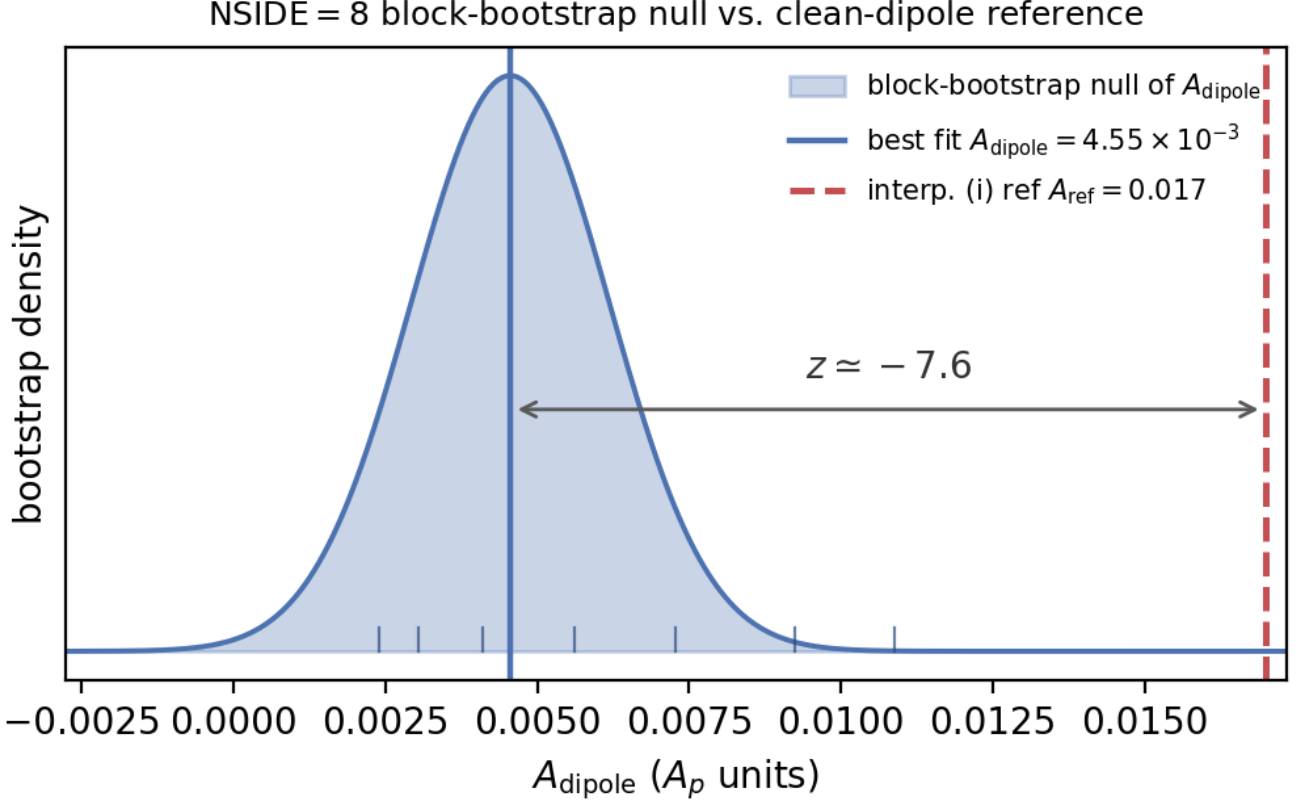


Figure 9. Historical latitude-mask block-bootstrap distribution. NSIDE= 8 resamples of the broader $|b_{\text{gal}}| > 15^\circ$, $N_{\text{total}} > 0$ WLS field ($N_{\text{boot}} = 1000$, seed 42), not the FS-C field. The distribution is retained as provenance and is not used for an FS-C or physical constraint.

Operational conclusion.—The exact-support FS-C residual is *not* a positive detection of a primordial chirality dipole. Its persistence under apodization and coherent $\ell = 1$ –3 structure support a systematics interpretation. The other six historical diagnostics are different-support or support-unproven and do not strengthen that synthesis. The primary scientific result remains the strict HC real-space null ($N = 890,069$, $z_{\text{mom}} = +0.635$, $p = 0.23768$).

E. MORPHOLOGY SYSTEMATICS

Edge-on galaxy contamination.—The DR8 morphology cross-match contains 505,889 edge-on systems ($b/a < 0.30$), or 15.80% of classified spirals. A linear dilution factor is conditional on flip-symmetric hard assignments: soft-probability equivariance supports that approximation, but hard argmax is nonlinear and cannot be proven unbiased by the soft identity alone. The edge-on-isolated tie-break test ($N = 295,170$) is null-consistent across the three imaging legs (family-wise $p = 0.49$, maximum $|z| = 1.17$, every leg $|z| < 1.4$). This empirically bounds the tested edge-on coherence channel but does not exclude every spatially varying hard-label bias; residual morphology/depth covariance remains part of the open joint-covariance problem.

High-confidence subsample robustness.—Historical high-confidence threshold diagnostics used the broader pre-quarantine HC sample and a different pre-MASTER estimator; they are retained only as estimator-sensitivity provenance and are not the primary analysis. The load-bearing result uses the release-safe selection `primary_hc && !raw_flip_qc_unsafe`: $N = 890,069$, $z_{\text{mom}} = +0.635$, and $p = 0.23768$. The WLS clean-template statistic remains diagnostic and must not be read as a physical 1.7% exclusion.

Monopole-preserving estimator definition.—The historical monopole-preserving calculation used an un-subtracted pre-MASTER field and a different null implementation. It is not directly comparable to either the exact-support MASTER

diagnostics or the strict primary real-space estimator. The current synthesis therefore uses only the persisted-mask recomputations in Tables 6 and 7; older monopole-preserving values are retained in artifacts solely as different-estimator provenance.

Spiral fraction variation across the sky.—The spiral fraction is uniform across the DESI Legacy footprint at the $\lesssim 2\%$ level across 7 equatorial coordinate slabs, with no coherent large-scale pattern that would bias the dipole analysis.

Mask robustness: pixel-count threshold sweep.—A pixel-count-threshold sweep ($N_{\text{spiral}}(p) \geq \{1, 5, 10, 20, 50\}$; 200-MC per threshold) gives stable mask geometry ($f_{\text{sky}} = 0.479\text{--}0.493$) and a FS-C $\ell = 1$ excess that persists at every threshold, with per-threshold significances spanning $6.3\text{--}8.3\sigma$ under the earlier (pre-galaxy-weighted-subtraction) estimator convention of that sweep — i.e. the excess is not an artifact of the threshold choice, but its magnitude is threshold-dependent at the $\sim 2\sigma$ level under that convention. The sweep has also been recomputed under the *current* (galaxy-weighted-subtraction) estimator convention with 10^4 per-galaxy label permutations per threshold: the $\ell = 1$ excess persists at every threshold, with $z = +5.7, +7.7, +7.9, +7.7, +7.5$ at $N_{\text{spiral}}(p) \geq \{1, 5, 10, 20, 50\}$ (rank $p = 1.1 \times 10^{-3}$ to 2.0×10^{-4} ; $f_{\text{sky}} = 0.479\text{--}0.494$; the canonical $N_{\text{spiral}} \geq 10$ cell independently reproduces the c9a 10^4 -permutation run, $z = +7.9$, rank $p = 3.0 \times 10^{-4}$), i.e. stable to $\pm 0.4\sigma$ across the resolved thresholds ≥ 5 and lower only in the sparse $N_{\text{spiral}} \geq 1$ cell (artifact [pipelines/p2_chirality/outputs/canonical_provenance/c16_r24conf_pod_batch.json](#)).

DATA AVAILABILITY

Repository state: the analysis artifacts linked throughout this paper resolve against the live `main` branch of the repository, which reflects the current version-stamped content (release-safe HC $N_{\text{selected}} = 890,069$, $N_{\text{support}} = 887,472$; real-space $z_{\text{mom}} = +0.635$, one-sided add-one rank $p = 0.23768$). The 8,474,531-row science-facing catalog is already public at immutable HF revision `db11023306ab4eed1d7727670bd78e127b7af17a`; its Parquet bytes and SHA-256 remain unchanged. The synchronized strict-primary overlay is retained in the repository at [pipelines/p2_chirality/apjs_release_v1.0.259_strict/MANIFEST.json](#). It binds the predicate `primary_hc && !raw_flip_qc_unsafe`, the exact strict null array, schema, and reproducer to the catalog hash. That overlay is now public at the provider path `apjs-release/v1.0.259-strict-primary/` under immutable HF dataset revision `911316f31c21f2c4b933a2f3a761274cfe85c6d6`; all seven overlay files were re-downloaded at that pinned revision and byte-verified against their committed SHA-256 digests, with the publish receipt retained at [pipelines/p2_chirality/apjs_release_v1.0.259_strict/PUBLISH_RECEIPT_2026-07-16.json](#). An immutable archival snapshot (PDF + `.tex` source + figures + canonical-provenance artifacts) will be deposited on Zenodo at journal submission with: (i) a frozen immutable release tag pinning the exact git commit hash of the analysis codebase; (ii) exact commit hashes for the catalog, all scripts, and all canonical-provenance artifacts ([pipelines/p2_chirality/outputs/canonical_provenance/](#)); and (iii) a minted Zenodo DOI that will constitute the single citable archival reproducibility handle for the published version. The DOI and commit hashes will be inserted here in place of this sentence at submission time.

- **Machine-readable release manifests:** the exact public catalog manifest is directly downloadable at immutable data revision `db110233/MANIFEST.json`. The strict-primary overlay manifest is public at immutable data revision `911316f3/MANIFEST.json` (byte-identical to the repository copy at [pipelines/p2_chirality/apjs_release_v1.0.259_strict/MANIFEST.json](#)); it supersedes the older public manifest for the strict-primary contract.
- **Catalog:** <https://huggingface.co/datasets/bamfai/galaxy-chirality-catalog> (CC-BY-4.0, Parquet). The original source is pinned to revision `a21eb596fd10edb9af9e7a1bcefb04f87327a724`, 952,115,239 bytes, and SHA-256 `e8525ba5c98576f6...3563`. The current ApJS bundle is public at immutable HF data revision `db110233` with provider receipt `e535b262` and checksummed manifest. The science-facing Parquet includes the hard-label primary selection and `raw_flip_qc_unsafe` flag; a separate quarantine contains all 249,066 catalog-wide violators (59,515 in HC). The bundle also contains the retained fixed-occupancy primary array, a distinct pixel-permutation diagnostic, machine schema, validation report, and minimal reproducer. A delivered per-region CW-fraction monopole correction map (per-region f_{CW} , correction $f_{\text{CW}} - 0.5$, and binomial uncertainty on HEALPix $\text{NSIDE} = 8$ and $\text{NSIDE} = 64$, with every input SHA-256-hashed in the header) is provided as a supplementary artifact for $\ell = 0$ parity users at [pipelines/p2_chirality/analysis/monopole_correction_map_v1_0_262.json](#). HF revision `245ad7c5` contains the existing 3,201,160-row `spiral_morphology_dr8.parquet` companion (105,317,121 bytes; SHA-256 `d49090fc...5620`), whose BRICKID.OBJID keys uniquely and exactly cover every safe-catalog

CW/CCW row. The current dataset card, executable morphology contract, and independently runnable Catalog-C row-level semantic validator are pinned at immutable HF revision [43fc8a5b](#). The morphology validator retrieves both source Parquets from immutable revisions and verifies byte counts and SHA-256 digests before the one-to-one join; the [semantic contract](#) similarly verifies the complete release inventory and scans all 8,474,531 primary rows plus all 249,066 quarantine rows against 12 primary semantic gates, eight aggregate/quarantine gates, and seven exact row-equivalence gates. Axial ratio b/a is not stored in the companion: the published contract derives it on demand from the stored ellipticity and profile-choice fields. Full-catalog redshift, imaging-leg, depth, seeing, and PSF fields remain unavailable, so a complete systematics-metadata sidecar and a DOI-backed archive of the paper, source, figures, and release metadata remain open; no placeholder DOI is claimed.

- **Model:** <https://huggingface.co/bamfai/galaxy-chirality-v2> (ViT-Small encoder + classification head, PyTorch checkpoint). The current model card is pinned at immutable revision [6f113097](#). Historical-training receipt: immutable revision 237d021c451d75cf86a875e86d4de498b74e2f12, `v2.bias_audit.json` SHA-256 57278e4e83d5a1ca2ae1b3759a0f14c3854538449a846eefc6f089a8e56ee847; checkpoint LFS SHA-256 618d170f...97bf, 88,046,038 bytes. Neither the corrected card nor the receipt restores the missing *historical* object/split manifest or resolves the conflicting historical benchmark records. A separate, additive manifest-retained retrain of the GZ1-core + synthetic component (checkpoint SHA-256 aed109dc...a752387; manifest [pipelines/p2_chirality/outputs/g1_retrain/g1_training_manifest.json](#), SHA-256 e5de8e03...240a10729; provenance [pipelines/p2_chirality/outputs/g1_retrain/PROVENANCE.md](#)) is backed up at model-repository revision path `g1-retrain-2026-07-17/` and retains all object/split/seed records; it does not overwrite the historical receipts or change the released catalog. A subsequent CE-included assembly re-provisions the CE-ResNet catalog (Zenodo 10.5281/zenodo.7167388) and adjudicates the 826-vs-846 conflict (reproducible non-spiral count 819; artifacts under [pipelines/p2_chirality/outputs/g1_full_composition/](#)); the composition-faithful CE-included retrain is a documented honest negative (Appendix B), and the released Catalog C labels are unchanged.
- **Code:** <https://github.com/Hubify-Projects/bigbounce>. Training, inference, equivariant post-processing, bias hardening suite, and dipole analysis scripts.

The measured spatially structured CW-bias residual of 0.26% (9.5σ) originates upstream of the classifier: a classifier-injection forward model (Sec. 4.2) excludes classifier confusion through the released equivariant pipeline as its source (0.0% of the observed monopole) and sign-excludes the GZ1 training-label-imbalance candidate, leaving a true sky parity asymmetry versus a parity-odd DESI photometric systematic as the unresolved attribution. Local re-normalization of the per-region monopole is recommended for precision parity tests, and a delivered per-region CW-fraction monopole correction map with per-region binomial uncertainty is provided for that purpose ([pipelines/p2_chirality/analysis/monopole_correction_map_v1_0_262.json](#)); it is a convenience product that changes no science result and leaves the primary dipole unaffected. At the tested $A = 0.75\%$ full amplitude, the finite pilots give score-pass fractions 0.55/0.59 and 16/20 deterministic axes; these are not a calibrated precision boundary or physical limit. Users requiring calibrated soft probabilities for downstream probabilistic models should apply temperature scaling or Platt scaling; the raw p_{eq} values are ranking scores, not frequentist probabilities (see Sec. 4.1 calibration caveat). The overlap-contaminated full GZ1 cross-match gives 69.91% CW/CCW agreement (Cohen’s $\kappa = 0.40$) for the released classifier; no object-level training anti-join was retained for that released realization, so this figure is descriptive and not an independent validation metric. A separate, provably training-disjoint anti-join validation of the manifest-retained retrain (Cohen’s $\kappa = 0.9733$ on 3,000 high-confidence held-out GZ1 spirals; Sec. B) is now available for the regenerable GZ1-core realization; it measures a different quantity (a different checkpoint, a confident-only subset, and clean held-out labels) and does not revise the descriptive full-overlap figure. The full provenance inventory is in the checksummed release manifest.

ACKNOWLEDGMENTS

This research used the DESI Legacy Imaging Surveys Data Release 8; the Galaxy Zoo citizen science project; the Smith42/galaxies dataset on HuggingFace; and the CE-ResNet catalog of Jia et al. (2023). Computations were performed on NVIDIA H100, H200, and RTX A5000 GPUs via RunPod cloud infrastructure.

Facilities: DESI Legacy Imaging Surveys, HuggingFace, RunPod.

1690 *Software:* Astropy (11), HEALPix/healpy (13; 14), NumPy (15), pandas (16), PyTorch (17), timm (18), NaMas-
 1691 ter/pymaster.

1692 *AI-assisted methodology:* This work was conducted using an agentic multi-model pipeline for literature review, code
 1693 development, analysis, and adversarial internal peer review, operated under the author’s direction. Quantitative claims
 1694 are bound to committed artifacts and checksums where those artifacts survive. The historical training realization is an
 1695 explicit exception: its object/split manifest and random-state receipt were not retained, so the historical checkpoint
 1696 metrics and the released catalog’s generation provenance are not exactly replayable. A subsequent from-scratch retrain
 1697 of the GZ1-core + synthetic component has since been produced *with* a fully retained object/split/seed manifest
 1698 (Sec. 2), making that component regenerable going forward; it does not reconstruct the CE-ResNet component or
 1699 the full historical composition, and it does not alter the released catalog. The author designed the study, made all
 1700 scientific judgments, and takes full responsibility for the content; the AI pipeline is a verification instrument, not an
 1701 author.

1702 *Conflicts of interest:* The author declares no conflicts of interest. This work received no external funding.

REFERENCES

- [1] L. Shamir, “Handedness asymmetry of spiral galaxies with $z < 0.3$ shows cosmic parity violation and a dipole axis,” *Phys. Lett. B* **715**, 25 (2012), arXiv:1207.5464.
- [2] L. Shamir, “Patterns of galaxy spin directions in SDSS and Pan-STARRS show parity violation and multipoles,” *Astrophys. Space Sci.* **365**, 136 (2020), arXiv:2007.16116.
- [3] L. Shamir, “Using 3D and 2D analysis for identifying parity violation in spiral galaxy spin directions,” *Publ. Astron. Soc. Jpn.* **74**, 1114 (2022), arXiv:2208.00893, DOI:10.1093/pasj/psac058.
- [4] L. Shamir, “Analysis of spin directions of galaxies in the DESI Legacy Survey,” *Mon. Not. R. Astron. Soc.* **516**, 2281 (2022), arXiv:2208.13866, DOI:10.1093/mnras/stac2372.
- [5] M. Iye, M. Yagi, and H. Fukumoto, “Spin parity of spiral galaxies. III. Dipole analysis of the distribution of SDSS spirals with 3D random walk simulations,” *Astrophys. J.* **907**, 123 (2021), arXiv:2011.00662.
- [6] K. Tadaki, M. Iye, H. Fukumoto *et al.*, “Spin parity of spiral galaxies. II. A catalogue of $\sim 80,000$ face-on spirals,” *Mon. Not. R. Astron. Soc.* **496**, 4276 (2020), arXiv:2006.02331.
- [7] H. Jia, H.-M. Zhu, and U.-L. Pen, “Galaxy Spin Classification I: Z-wise vs S-wise Spirals With Chirality Equivariant Residual Network,” *Astrophys. J.* **943**, 32 (2023), arXiv:2210.04168, DOI:10.3847/1538-4357/aca8aa.
- [8] A. Dey, D. J. Schlegel, D. Lang *et al.*, “Overview of the DESI Legacy Imaging Surveys,” *Astron. J.* **157**, 168 (2019), arXiv:1804.08657.
- [9] M. Walmsley, C. Lintott, T. G eron *et al.*, “Galaxy Zoo DESI: detailed morphology measurements for 8.7M galaxies in the DESI Legacy Imaging Surveys,” *Mon. Not. R. Astron. Soc.* **526**, 4768 (2023), arXiv:2309.11425.
- [10] A. Dosovitskiy, L. Beyer, A. Kolesnikov *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” in *Proc. Int. Conf. Learning Representations (ICLR)* (2021) [arXiv:2010.11929].
- [11] Astropy Collaboration, A. M. Price-Whelan, P. L. Lim *et al.*, “The Astropy Project: sustaining and growing a community-oriented open-source project and the latest major release (v5.0) of the core package,” *Astrophys. J.* **935**, 167 (2022), arXiv:2206.14220.
- [12] D. Alonso, J. Sanchez, and A. Slosar, “A unified pseudo- C_ℓ framework,” *Mon. Not. R. Astron. Soc.* **484**, 4127 (2019), arXiv:1809.09603.
- [13] K. M. G orski, E. Hivon, A. J. Banday *et al.*, “HEALPix: a framework for high-resolution discretization and fast analysis of data distributed on the sphere,” *Astrophys. J.* **622**, 759 (2005), arXiv:astro-ph/0409513.
- [14] A. Zonca, L. Singer, D. Lenz *et al.*, *J. Open Source Softw.* **4**, 1298 (2019).
- [15] C. R. Harris, K. J. Millman, S. J. van der Walt *et al.*, *Nature* **585**, 357 (2020).
- [16] W. McKinney, in *Proc. 9th Python in Science Conf.*, edited by S. van der Walt and J. Millman (2010), pp. 56–61.
- [17] A. Paszke, S. Gross, F. Massa *et al.*, in *Advances in Neural Information Processing Systems 32*, edited by H. Wallach *et al.* (Curran Associates, 2019), pp. 8024–8035.
- [18] R. Wightman, *PyTorch Image Models* (2019), <https://github.com/rwightman/pytorch-image-models>.