

An Observed-Label Chirality-Dipole Null in 949,584 High-Confidence DESI Spirals and a Public 8.5-Million-Galaxy Catalog

Houston Golden^{1,*}

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

(Dated: July 14, 2026, 15:14 PDT; Version v1.0.244)

We release a catalog of 8,474,531 DESI Legacy DR8 galaxies and test the observed-label chirality field. The single primary estimator is the real-space dipole of the declared high-confidence sample ($p_{\text{eq}} > 0.6$, $N = 949,584$), which is consistent with its isotropic pixel-permutation null at $+0.55\sigma$ (one-sided upper-tail rank $p = 0.265$). The ApJS-oriented catalog/methods release contract removes raw-pass and reconstructed flip-pass scores from the science-facing product and quarantines all 249,066 catalog-wide reconstruction violators (59,515 in HC) as explicitly uncalibrated diagnostics. A paired-image-to-field surrogate over 20 deterministic axes is reported only as a pilot grid: the HC score-crossing counts 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%. These fractions are not a calibrated population recovery curve or physical upper bound; $A = 1.5\%$ corresponds to a 0.75% peak deviation of f_{CW} from one half. WLS and post-MASTER harmonic results are diagnostics on different fields and samples. A spatially resolved confusion/transfer model (DP4-15), matched-footprint independent-estimator comparison (DP4-16), joint real-space $\times$ harmonic covariance (DP4-17), and independent external preregistration or blinding (DP4-21) remain open. The $\ell = 1$ observable is parity-even; without a transfer function we claim no primordial parity bound.

CONTENTS

I. Introduction	1
II. Data	2
A. Galaxy Images	2
B. Training Labels	2
III. Methods	2
A. Notation and Significance Conventions	2
B. Declared Analysis Hierarchy	3
C. Model Architecture	4
D. Test-Time Equivariant Averaging	4
E. Catalog Tiers	4
IV. Results	4
A. Catalog Statistics	4
B. Global CW Fraction	6
C. Dipole Analysis	9
D. Monopole+Mask Leakage Generative Null	10
E. Signal-Hunt Diagnostics	12
V. Comparison with Previous Work	12
A. Shamir (2012, 2020, 2022)	12
B. CE-ResNet (Jia et al. 2023)	13
VI. Discussion	13
A. Pseudo-label independence and the shuffle-null limitation	13
B. Statistical Scale and Finite Injection Scores	13
C. Relation to Parity-Violating Sectors	15
D. Open Follow-up and Future Directions	15

VII. Conclusions

A. NaMaster MASTER Configuration	16
B. Classifier Architecture Details	18
C. Auxiliary Dipole Diagnostics	20
D. FS-C Systematic Analysis	21
E. Morphology Systematics	23
Data Availability	24
Acknowledgments	25
References	25

I. 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 *projected 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: our null dipole at sub-percent sensitivity does not depend on any unpublished companion work.

* houston@hubify.com

Claims of such a signal have appeared intermittently in the literature. Shamir (2012) [4] 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) [1] reported asymmetries at the reported $\sim 2\text{--}4\%$ level on SDSS and Pan-STARRS samples; Shamir (2022a) [2] reported related spin-direction alignment analyses; and Shamir (2022b) [3] 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 chirality catalog 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. Because 66.5% of the training labels derive from CE-ResNet predictions, the catalogs are not fully independent and model-output permutation nulls cannot test inherited survey-correlated structure by themselves. Our sole primary result is the HC real-space observed-label null ($N = 949,584$, $z = +0.55$, one-sided upper-tail $p = 0.265$). The WLS and MASTER results are systematics-sensitive diagnostics, not independent detections or physical exclusions. The GZ1 human-only test is an independent-label corroboration at lower statistical power and does not tighten the HC amplitude constraint.

II. DATA

A. Galaxy Images

Our parent sample is the Smith42/galaxies dataset on HuggingFace (<https://huggingface.co/datasets/Smith42/galaxies>), containing 8,474,688 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 includes unique *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.

B. Training Labels

We assemble labels from Galaxy Zoo 1 (6,637 high-confidence CW/CCW galaxies), CE-ResNet (17,153 high-confidence spiral classifications), and 2,000 synthetic NOT_SPIRAL examples, for 25,790 source images. After training-only flip augmentation the combined pool is 26,616 images ($n_{\text{train}} = 21,293$, $n_{\text{val}} = 5,323$; Appendix B). CE-ResNet supplies 66.5% of source labels, so validation partly measures agreement with that model. The disjoint GZ1 cross-match gives 69.91% chirality accuracy ($\kappa = 0.40$), while the fully human-label GZ1 dipole test gives $z = -0.54$ at $N = 46,017$. That result corroborates a null but, without calibrated power or a transfer model, does not provide a tighter amplitude bound (DP4-15).

III. METHODS

A. Notation and Significance Conventions

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

b. Significance conventions. All significance values z are moment-ratios $(x - \langle x \rangle_{\text{null}})/\sigma_{\text{null}}$ 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 ($z_{\text{mom}} = +0.55$, one-sided upper-tail rank $p = 0.265$) are independent summary statistics of the A_{dip} distribution against the isotropic (pixel-)permutation null; they do not follow the Gaussian $z \rightarrow p$ mapping.
- *MASTER $\ell = 1$ moment- z :* $z = (C_1^{\text{data}} - \langle C_1 \rangle_{\text{null}})/\sigma_{\text{null}}$ vs. the label-shuffle null mean $\langle C_1 \rangle_{\text{null}}$ and width σ_{null} ; $z = +3.64$ (500-MC canonical direct run), $z = +7.93$ (10^4 -permutation canonical unapodized), $z = +7.28$ (10^4 -permutation apodized). These three values are from different null-run sizes and mask/weight conventions and are not mutually comparable.
- *Block-bootstrap z :* $(A_{\text{dipole}}^{\text{best}} - A_{\text{ref}})/\sigma_{\text{boot}}$ with spatial-coherence-corrected σ_{boot} ; $z \approx -7.6$ applies only to a clean *observed-label* $A_{\text{ref}} = 0.017$ template. Under $g = 0.398$, a physical 1.7% dipole maps to 0.6766% observed and is $1.36\sigma_{\text{boot}}$ above the fit, so the former physical-exclusion reading is invalid.

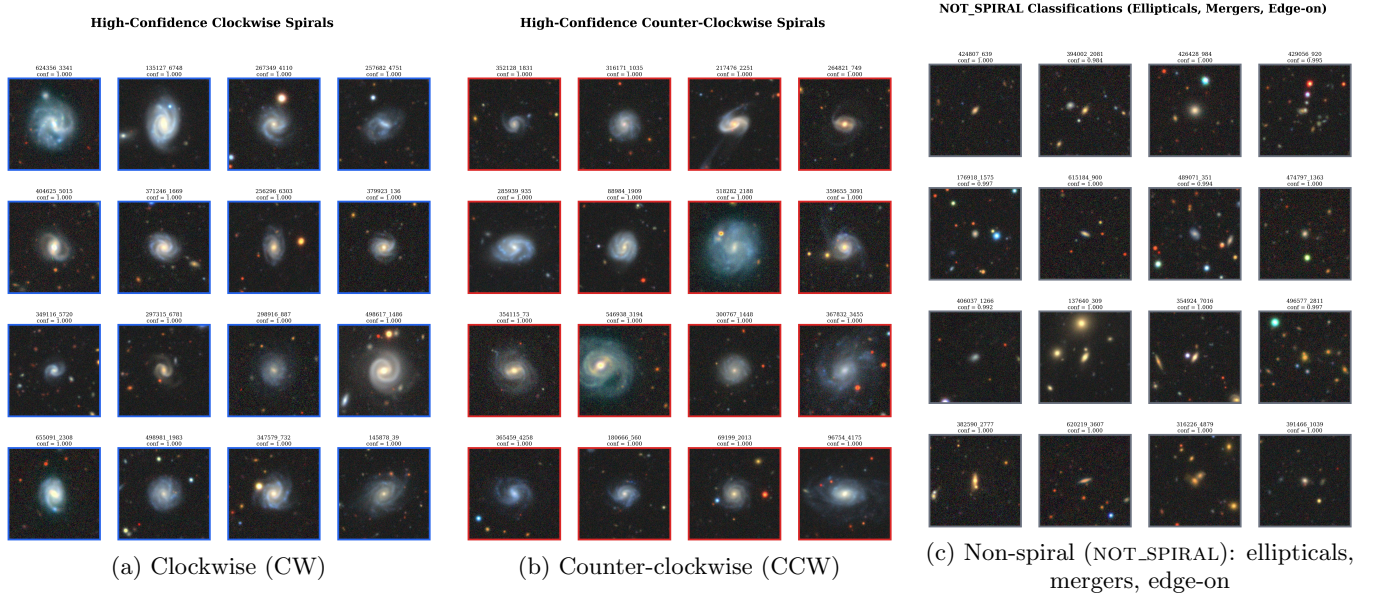


FIG. 1. **Representative high-confidence galaxies from the classified catalog** ($p_{\text{eq}} > 0.9$). Left: clockwise (CW) spirals; center: counter-clockwise (CCW) spirals; right: non-spiral (NOT_SPIRAL) objects — ellipticals, mergers, and edge-on galaxies that would contaminate a binary spiral classifier. 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 §III D.

B. Declared Analysis Hierarchy

We assign stable names to the three distinct supports used below: **HC-REALSPACE-INCLUSIVE** (HC-RI; HC rows with $N_{\text{spiral}}(p) \geq 10$, 23,682 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). These names prevent the primary HC support from being conflated with either full-sample diagnostic footprint. We declare the estimator hierarchy used throughout. This hierarchy — with the real-space dipole as the row-(i) primary cosmological estimator — is documented in Appendix A; the primacy of the real-space estimator is established independently of the harmonic-channel diagnostics. The primary dipole measurement uses the high-confidence subsample (Catalog C with winning-class confidence $p_{\text{eq}} > 0.6$; $N_{\text{HC}} = 949,584$ spirals), which suppresses the depth-correlated low-confidence-tail systematic identified in Sec. IV C; all estimator-hierarchy and sensitivity statements below are made on this HC selection unless explicitly noted. We declare the HC $p_{\text{eq}} > 0.6$ subsample as the *single primary science sample* for all primary cosmological statements; the block-bootstrap WLS template fit is reported on the full Catalog C field as a *diagnostic* template-disfavor cross-check on a different (unthresholded) sample, and is not treated as a same-sample frequentist exclusion (its $p_{\text{eq}} > 0.6$ selection function is not propagated into the block-bootstrap

covariance; App. D). This is an internal hierarchy declaration, not a claim of formal preregistration or independent blinding. See also Appendix D for the full eight-anchor discriminator table.

- **Estimator hierarchy:** the primary cosmological estimator is the real-space CW-fraction dipole fit on Catalog C at NSIDE=64 (moment-ratio $z_{\text{mom}} = +0.55$ against the isotropic pixel-permutation null, one-sided upper-tail empirical-rank $p = 0.265$). The block-bootstrap WLS fit is a diagnostic of a clean observed-label template ($z \approx -7.6$ at $A_p = 1.7\%$), not a physical 1.7% exclusion; with $g = 0.398$ the physical comparison is only $1.36\sigma_{\text{boot}}$ (Appendix D).
- **Secondary diagnostic estimators:** (iii) canonical- N direct-MC NaMaster at $\ell = 1$ on the patchy FS-C support ($f_{\text{sky}} = 0.49005$, $+3.64\sigma$); (iv) apodized-footprint MASTER at $\ell = 1$ on MASTER-AGF ($f_{\text{sky}} = 0.494$, C^2 2° ; $+7.28\sigma$ vs. global label-shuffle and $+7.13\sigma$ vs. depth-stratified null, $W_p = N_{\text{all}}$); and (v) hemisphere maximum-asymmetry (3.05σ local maximum against the label-shuffle null, $p_{\text{LEE}} \leq 10^{-4}$ direct-MC max-statistic, systematics-attributed (the Gaussian-Bonferroni $< 1\sigma$ heuristic is a non-principled cross-check); Appendix C). These characterize how a non-zero global monopole and coherent low- ℓ systematics couple through the patchy weighted footprint.
- **Generative monopole-only null:** (vi) $N =$

500 binomial-monopole realizations demonstrating that the raw pre-MASTER pseudo- $C_\ell^{(\ell=1)}$ is dominated by monopole-mask leakage; the post-MASTER $+3.64\sigma$ FS-C residual is non-primary and requires additional coherent systematics beyond the monopole-only channel (Sec. IV D).

- **Finite-grid injection scores:** (vii) on the HC-broad spiral subsample ($p_{\text{eq}} > 0.6$, $N = 949,584$; Sec. VIB), 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 I 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 II consolidates the distinct estimators and nulls. Their statistics are not commensurable, and no joint likelihood is defined. The scientific verdict rests solely on the HC observed-label real-space row; WLS and all harmonic rows remain systematics diagnostics. The missing joint covariance is open gate DP4-17.

C. Model Architecture

The classifier consists of a ViT-Small encoder [13] (`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.

D. 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

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 raw single-pass Catalog A tier (Table IV). A direct D_4 -TTA hold-out on two independent $\sim 2,000$ -galaxy subsamples confirms the mean per-galaxy P_{CW} is stable under Z_2 and D_4 to within $|\Delta(p_{\text{CW}})| < 0.0016$; 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. IV C. Full details in Appendix B.

E. Catalog Tiers

The pipeline produces three tiers: **Catalog A** (raw, single-pass softmax); **Catalog B** (Platt-calibrated, $+0.4\%$ excess); **Catalog C** (equivariant production, 2-fold flip TTA). **Catalog C is the recommended tier for all cosmological parity analyses.** All three tiers share 8,474,531 rows in Apache Parquet format.

IV. RESULTS

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

A. Catalog Statistics

The final catalog contains 8,474,531 galaxies (157 of 8,474,688 failed quality checks). 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. These values are strongly overconfident relative to the independent GZ1 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 re-

TABLE I. **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) HC real-space dipole	HC, $p_{\text{eq}} > 0.6$ (949,584)	isotropic pixel-permutation (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; not calibrated coverage or an upper limit	(vii) injection-recovery	HC, $p_{\text{eq}} > 0.6$ (949,584)	per-pixel shuffle	PILOT

TABLE II. 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 ($\sim 62\%$ of the catalog); each galaxy is counted once. Row (iv) is the apodized-footprint MASTER diagnostic ($N_{\text{all}} \geq 1$ mask, C^2 2° apodization); its two σ values are against the global per-galaxy label-shuffle and depth-stratified nulls respectively (both systematics-attributed; Appendices A, D). 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 $p_{\text{LEE}} \leq 10^{-4}$ 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. VIB). 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 (HC)	949,584	—	0.4818 ^a	HC-RI	pix-perm. (10^4)	+0.55 (moment z)
(ii) WLS template excl.	3,201,160	—	0.49005	FS-C	block-boot. (10^3)	$z \approx -7.6$
(iii) canonical MASTER (diag.)	3,201,160	—	0.49005	FS-C	pp-shuffle	+3.64
(iv) apod. MASTER (diag.)	3,201,160	8,474,531	0.494 ^b	MASTER-AGF	pp-sh./d-str.	+7.28/+7.13
(v) hemisphere LEE (MC)	3,201,160	—	0.49005	FS-C	max-stat MC	$p_{\text{LEE}} \leq 10^{-4}$ (syst.-attr.)
(vi) monopole+mask null	3,201,160	—	0.49005	FS-C	monopole-only	+1.69
(vii) finite injection grid	949,584 HC	—	—	—	pp-shuffle	$P_{3\sigma} = 0.55/0.59$ at $A = 0.75\%$

^a $f_{\text{sky}} = 0.4818115$ is HC-RI, the inclusive $N_{\text{spiral}}(p) \geq 10$ support re-evaluated on the high-confidence subsample (23,682 of 49,152 pixels). The superseded strict $N_{\text{spiral}} > 10$ execution used 23,600 pixels and is retained only as historical mask-sensitivity context.

Artifact: [pipelines/p2_chirality/outputs/dipole/catalog_c_summary.json](#).

^b Geometric $N_{\text{all}} \geq 1$ footprint pixel fraction; the corresponding weighted/apodized *effective* sky fraction for this row’s $W_p = N_{\text{all}}$, C^2 2° configuration is $f_{\text{sky}}^{\text{eff}} = 0.452$ (Appendix A, Table XI).

calibration 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-label GZ1 test corroborates the null independently of softmax scaling. The illustrative scalar dilution $g = 2a - 1$ uses the external GZ1 chirality accuracy, not the internal score.

Raw/flip quarantine. Reconstructing the flip pass from mismatched raw/equivariant pipeline products identifies

Equivariant Averaging: Original vs. Flipped Predictions

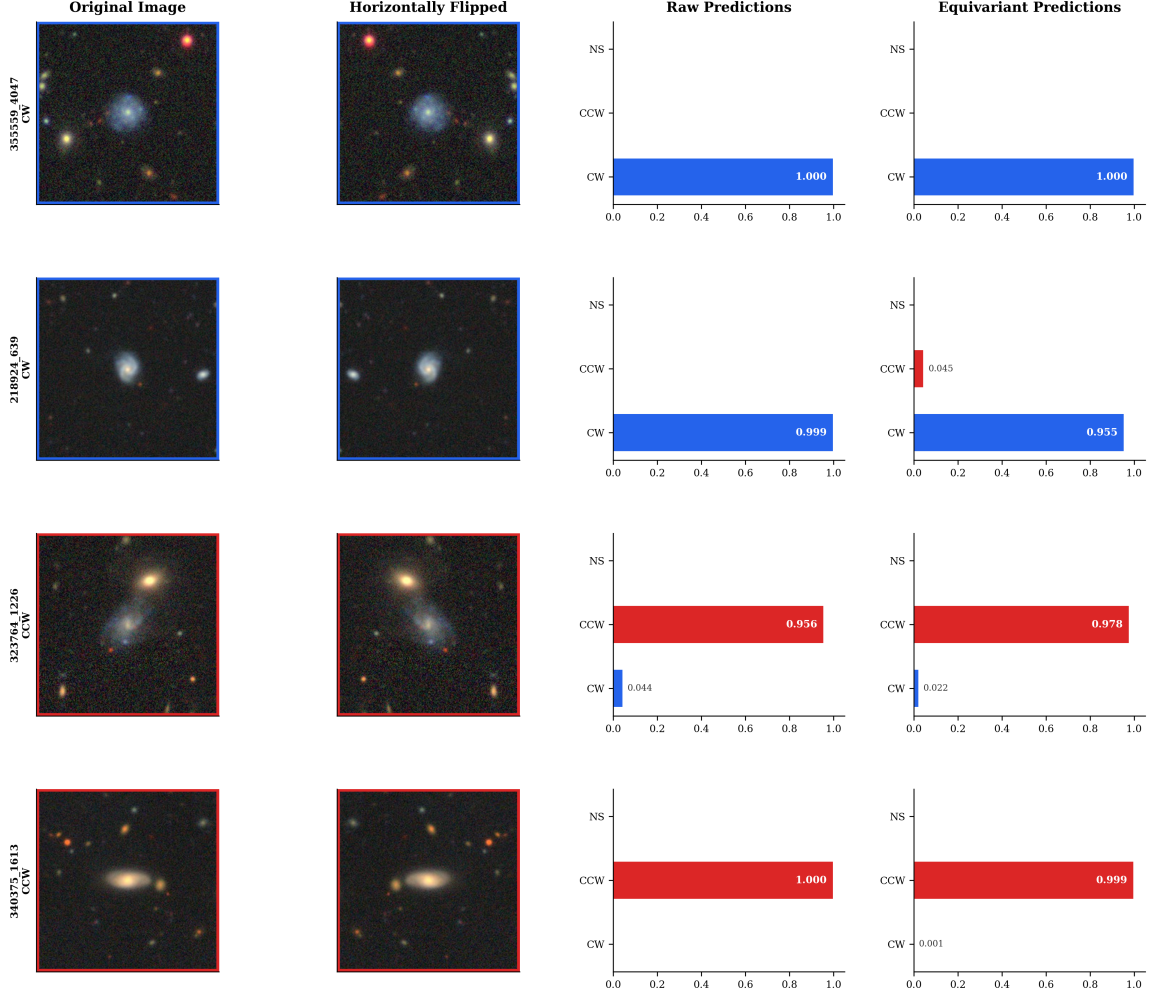


FIG. 2. **Equivariant test-time averaging (TTA).** Representative Z_2 production TTA examples (original + horizontal flip); D_4 validation (four rotations $\times$ two reflections) in Appendix B. *Production* inference (Catalog C) uses 2-fold Z_2 TTA — original + horizontal flip only (§III D). Flips 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 Catalog A (raw), Catalog B (Platt-calibrated), and Catalog C (equivariant); 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, dominated by this step (Table IV). Unit reminder: asymmetry- A values are exactly twice the f_{CW} -deviation values, $A = 2(f_{\text{CW}} - \frac{1}{2})$ (Sec. IV C); Table IV quotes the f_{CW} -deviation convention.

249,066 unsafe rows catalog-wide (2.939%) whose reconstructed score exits $[0, 1]$ by more than 10^{-3} , with maximum excursion 0.0901. Exactly 59,515 of those rows are in the 949,584-row HC selection; the strict HC diagnostic has 890,069 rows. These diagnostic quantities are not calibrated probabilities. The v1.0.244 science-facing release candidate omits every raw-pass and reconstructed flip-pass score column, retains only a Boolean `raw_flip_qc_unsafe` marker, and places all 249,066 offending diagnostic rows in a provenance-only quarantine with `do_not_use_for_science=True`. The existing flagged-row-excluded rerun remains null-consistent; it is a robustness diagnostic, not a replacement primary null.

The executable schema and filters are given in Data Availability.

B. Global CW Fraction

The Catalog C residual (-9.5σ from 0.5000, Table IV) is spatially uniform across 7 equatorial coordinate slabs and does not produce a dipole. This monopole offset is a classifier artifact, not 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 asymme-

TABLE III. **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. III A).

Claim / channel	Null	Value
P1 real-space HC dipole ($P_{\text{eq}} > 0.6$), <i>primary</i>	isotropic	$+0.55\sigma$
P2 observed-label template (WLS fit), <i>diagnostic</i>	permutation	(upper-tail $p = 0.265$)
D1 HC dipole (cross-check)	block-	$z \approx -7.6$
D2 MASTER $\ell = 1$ residual	bootstrap	(physical: 1.36σ)
<i>systematics, not a detection</i>	label-shuffle	$z = +0.70$, $p = 0.230$
	label-shuffle (harmonic)	$+3.64 / +7.28\sigma$

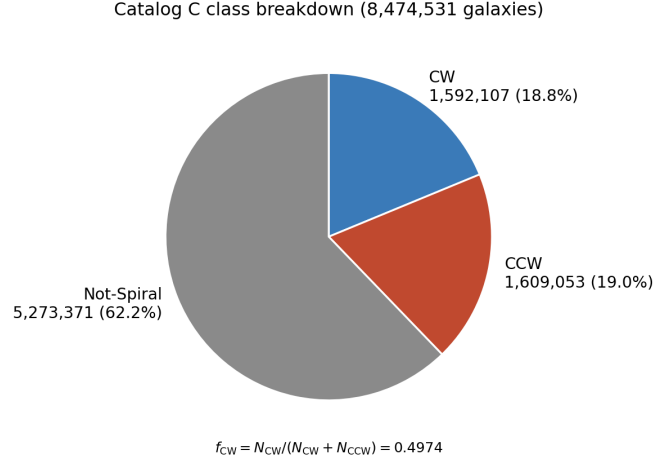


FIG. 3. **Catalog C composition.** Of the 8,474,531 galaxies retained after image-quality QA, the equivariant TTA classifier (§IIID) assigns $N_{\text{CW}} = 1,592,107$, $N_{\text{CCW}} = 1,609,053$, and $N_{\text{NS}} = 5,273,371$ (non-spiral / edge-on / morphologically indeterminate). The spiral sub-catalog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

try in DESI Legacy imaging. 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 IV) demonstrates the dominance of the equivariant TTA processing. 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](#)).

TABLE IV. Global CW fraction across catalog tiers. Uncertainties are 1σ binomial, $\sigma = \sqrt{f(1-f)/N_{\text{spiral}}}$, with per-tier spiral counts ($N_{\text{spiral}}^{\text{A}} = 3,321,795$; $N_{\text{spiral}}^{\text{C}} = 3,201,160$); 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. The Catalog-B row derives from the Platt-calibrated fraction; its deviation is computed from the unrounded calibrated fraction. Percentages in this table (and in the catalog-composition counts cited in Sec. II) are rounded to maintain sum-to-one consistency at the second decimal; the integer counts are exact.

Tier	$\text{cw}/(\text{cw} + \text{ccw})$	Excess (%)	Dev. (σ)
A (raw)	0.507879(274)	+0.788	+28.72
B (calibrated)	0.50400(27)	+0.4	+14.6
C (equivariant)	0.497353(279)	−0.265	−9.47

[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. VIB; 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; Sec. VIC), 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 spatially-uniform classifier artifact ($\lesssim 0.5\%$ per-slab scatter, three candidate instrumental/training mechanisms above), not a physical 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-

Galaxy Chirality Asymmetry Map (8.47M galaxies, equivariant)

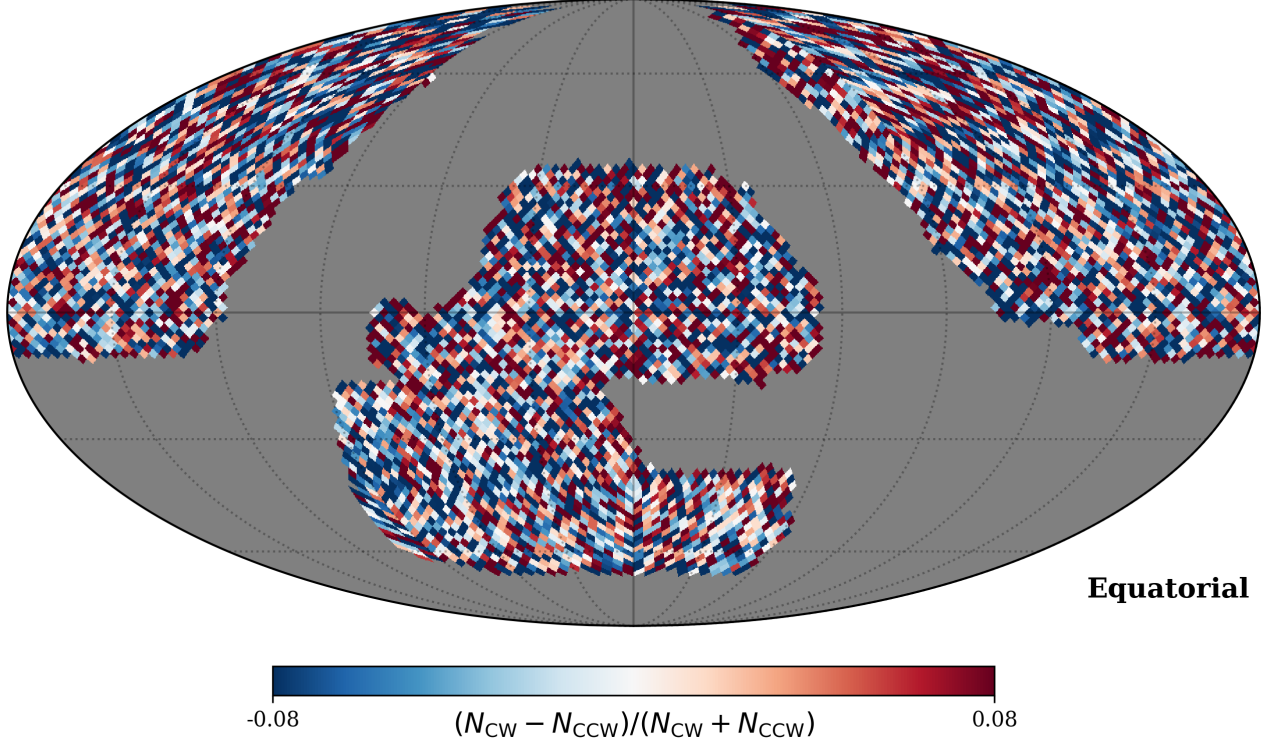


FIG. 4. **Equivariant (Catalog C) chirality asymmetry map of the 8.47 M-galaxy catalog** (Mollweide projection, equatorial coordinates; per-pixel asymmetry $A_p = (N_{CW} - N_{CCW}) / (N_{CW} + N_{CCW}) = 2(f_{CW,p} - \frac{1}{2})$ at HEALPix NSIDE=64, color scale $[-0.08, +0.08]$). The DESI Legacy Imaging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

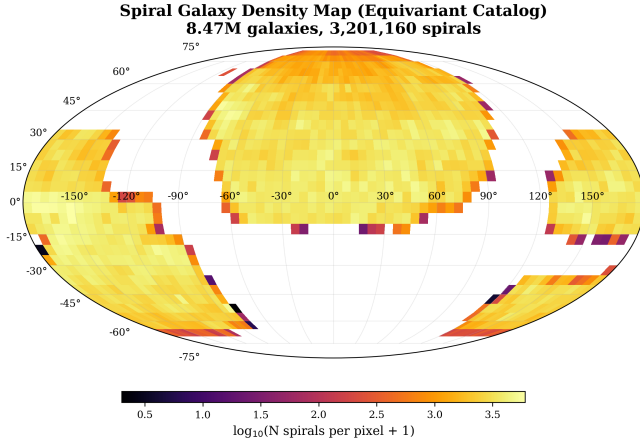


FIG. 5. **Sky density of the 3,201,160 classified spirals** (CW + CCW combined, NSIDE=64 Mollweide). Per-pixel spiral counts scale with the DESI Legacy Imaging Surveys depth and exposure pattern; the FS-C support used for the primary $\ell = 1$ analysis (§IV C) requires $N_{spiral}(p) \geq 10$ per pixel.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

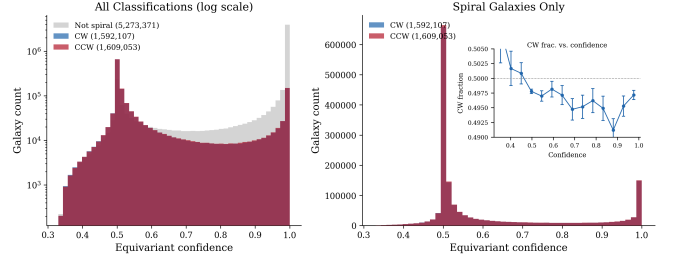


FIG. 6. **Distribution of maximum-class confidence** $\max(P_{CW}, P_{CCW}, P_{NS})$ for all 8,474,531 galaxies. Strongly bimodal: 73.6% at $\max p \geq 0.9$ (high-confidence labels) + a long tail of indeterminate cases ($\max p < 0.5$, dominated by NS/edge-on systems). The high-confidence (HC) cuts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

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).

C. Dipole Analysis

We pixelize the sky at HEALPix resolution $\text{NSIDE} = 64$ (49,152 pixels, $\sim 0.84 \text{ deg}^2$ per pixel). In each supported pixel we compute the asymmetry below. The threshold $N_{\text{spiral}}(p) \geq 10$ defines two sample-dependent supports: HC-RI for the HC primary (23,682 pixels) and FS-C for the full-spiral diagnostics (24,087 pixels); Appendix A records the exact data-vector mapping.

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

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.

a. Simple dipole. The primary real-space estimator fits $A_p = m + \mathbf{a} \cdot \hat{n}_p$ on the HC observed-label map and compares its amplitude with 10,000 isotropic pixel permutations. The declared $p_{\text{eq}} > 0.6$ cut is encoded in committed generator revision 94113e5 (2026-06-09), but no externally timestamped preregistration, frozen pre-unblinding tag, or independent blinding record exists; we therefore do not claim formal preregistration or elimination of all researcher degrees of freedom. The HC-RI fit uses $N = 949,584$ spirals in 23,682 pixels and is null-consistent at $z_{\text{mom}} = +0.549$ with one-sided upper-tail rank $p = (2651 + 1)/(10000 + 1) = 0.26517$. The fitted amplitude is $A_{\text{dip}} = 0.00459707$ toward equatorial (RA, Dec) = (294.31°, 16.03°); the null mean and population standard deviation are 0.00369283 and 0.00164671, respectively, for seed 20260418. The legacy strict-mask confidence-cut sweep ($N_{\text{spiral}} > 10$) gave $z = +0.41, +1.14, +0.51$ at cuts 0.6, 0.7, 0.8 and is retained only as historical sensitivity context; the primary $p_{\text{eq}} > 0.6$ value is the inclusive result above.

The existing 10^4 -draw per-galaxy label shuffle preserves the pixel occupancies and global CW total and gives null mean 0.003488, standard deviation 0.001593, $z = +0.696$, and rank $p = 0.2298$, independently retaining the null verdict. A separate 2×3 panel uses 2,000 pixel permutations per cell (seed 20260610), with the asymmetry values permuted while pixel weights remain fixed. At $N_{\text{spiral}} \geq 10, 20, 50$, respectively, the uniform fits give $(z, p) = (+0.536, 0.270), (+0.268, 0.359), (-0.382, 0.606)$ and the N_{spiral} -weighted fits give $(+0.783, 0.197), (+0.794, 0.203), (-0.141, 0.505)$. Thus all six cells have $|z| < 0.8$. These are robustness diagnostics on HC-RI threshold variants, not additional primary tests. The 95th percentile of the primary pixel-permutation null is $Q_{0.95}^{\text{null}} = 6.684 \times 10^{-3}$ in A_p units; this is a descriptive null quantile with no

TABLE V. Primary HC estimator audit. The empirical rank value is the one-sided upper tail. The exact null-array SHA-256 is 62bb1c019231974c...6492; full precision and the complete hash are in the linked JSON artifact.

Quantity	Exact or reported value
N_{HC} / valid pixels	949,584 / 23,682
Support / seed / null draws	HC-RI / 20260418 / 10,000
A_{dip} / monopole	0.0045970743 / -0.0077688925
Equatorial direction	(294.3054°, 16.0291°)
Null mean / std. (ddof = 0)	0.0036928319 / 0.0016467113
z_{mom}	+0.5491202
Upper-tail k/N / rank p	2651/10000 / 0.2651735

signal-injection coverage. The unthresholded full-sample excess is reported only as a depth-correlated systematics diagnostic, not a competing primary result. Artifacts: [pipelines/p2_chirality/outputs/dipole/catalog_c_summary.json](#), [pipelines/p2_chirality/outputs/canonical_provenance/c11b_hc_dipole_nulls.json](#), and [pipelines/p2_chirality/outputs/canonical_provenance/c12_r24conf_local_batch.json](#).

b. Angular power spectrum. We perform MASTER mode-coupling deconvolution using NAMASTER [34] at $N_{\text{side}} = 64$ on the MASTER-AGF ($N_{\text{all}} \geq 1$; 24,297 pixels, $f_{\text{sky}} = 0.494$, C^2 2° apodization, depth weight W_p). The MASTER-deconvolved single-mode $\ell=1$ value (computed on the monopole-subtracted A_p field; declared data vector, Appendix A) is $C_1^{\text{meas}} = 2.348 \times 10^{-5}$ against a 500-MC per-galaxy label-shuffle null with mean 1.71×10^{-6} and $\sigma_{\text{null}} = 2.99 \times 10^{-6}$, i.e. $+7.28\sigma$ for $W_p = N_{\text{all}}$ ($+9.78\sigma$ for $W_p = N_{\text{spiral}}$); the 10^4 -permutation recompute (Table VI) reproduces this diagnostic channel at $z = +7.31$ with empirical rank $p = 6.0 \times 10^{-4}$ (one-sided, computed as $(k+1)/(N+1)$ with $k=5$ of $N=10^4$ null draws exceeding the data; Table VI caption), and the Gaussian-equivalent σ is quoted from the null moments. A depth-stratified null (labels permuted within 10 $N_{\text{all}}(p)$ deciles; this preserves the marginal depth distribution but not joint spatial-depth structure, so it bounds only depth-sampling effects) leaves the excess essentially unchanged ($+7.13\sigma$ / $+9.06\sigma$), excluding pixel-depth sampling alone as the driver. A weight-map sweep gives the same verdict: the excess persists under all three weightings ($W_p = N_{\text{all}} + 6.9\sigma$, $W_p = N_{\text{spiral}} + 8.5\sigma$, uniform binary $+6.7\sigma$; independent 500-MC null streams, artifact c9c; in each variant the mask-mean subtraction uses the same variant weight map, so the field-consistent N_{spiral} -weighted subtraction is included in the sweep), so it is not an artifact of the depth-weight choice, although its magnitude is weight-dependent at the $\pm 1\sigma$ level. This channel is therefore a *systematics diagnostic* on the patchy weighted footprint — the same coherent low- ℓ family as the FS-C residual (Sec. IV D, Appendix D) — and is not used as a cosmological null; the cosmological statement rests on the real-space dipole; the WLS observed-label template fit is a diagnostic whose physical-amplitude reinterpretation is given in Appendix D. NaMaster con-

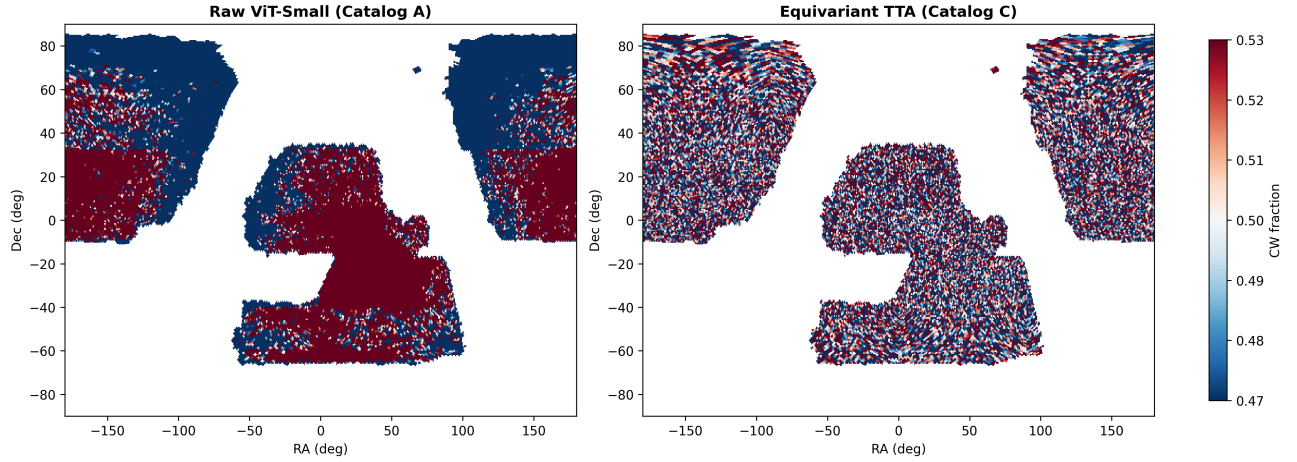


FIG. 7. **Raw (Catalog A) vs equivariant (Catalog C) chirality sky maps** (equatorial RA/Dec, per-pixel CW *fraction* $f_{\text{CW},p}$, NSIDE=64; shared color scale [0.47, 0.53]). The raw definition is `class_raw_y` with `class_raw_x` fallback only where the former is missing; in this release all 3,321,795 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; the bias is removed by construction and the HC ($p_{\text{eq}} > 0.6$) real-space dipole collapses to 0.55σ . 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 is the methodology cornerstone — the difference between Catalog A and Catalog C is the difference between a 2σ “detection” and a clean null, demonstrating that future chirality studies must adopt equivariant post-processing to avoid spurious dipoles.

figuration details and the FS-C support-declaration + depth-stratification audit are in Appendix A. We emphasize that rows (i) and (iv) of Table II are not on the same statistical footing — different fields, masks, weights, and null procedures — so the apparent $+0.55\sigma$ vs. $+7.28\sigma$ gap is not a $17\times$ discrepancy on a common axis; the harmonic-completeness check (Sec. VII, artifact c9b) bounds what a clean Shamir-class real-space dipole would have produced in the MASTER channel — it would register at $z \approx 68$ –218 there, versus the observed $+7.28\sigma$ — but it does not establish statistical consistency of the two estimators on a common axis.

Reader’s note (Table VI). The σ/z entries are systematics-attributed harmonic diagnostics against different nulls; they are not detection significances and must not be compared row-to-row, footprint-to-foorprint, or with the primary real-space dipole. The HC real-space result ($+0.55\sigma$) is the sole primary estimator; the WLS $z \approx -7.6$ value is a correlated observed-label template diagnostic.

D. Monopole+Mask Leakage Generative Null

The FS-C direct-MC $\ell = 1$ value of $+3.64\sigma$ (per-pixel label-shuffle null) and the local hemisphere maximum of 3.05σ (label-shuffle null, 648-direction scan; Appendix C) are candidate manifestations of mask-geometric leakage of the global 9.5σ monopole. We for-

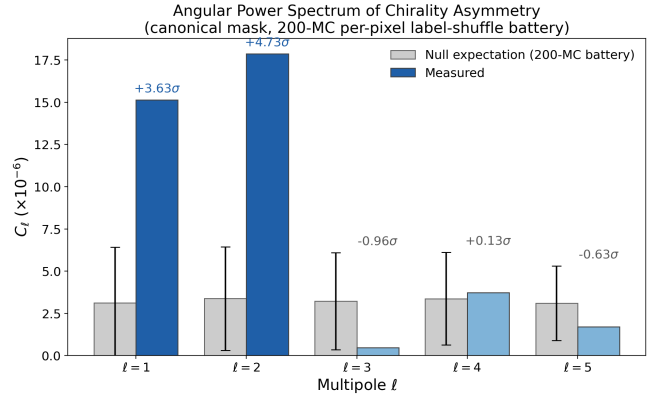


FIG. 8. **Pseudo- C_ℓ of the chirality field A_p on the canonical mask, $\ell = 1$ –5** (single panel; grouped bars). Gray bars with error bars: per-pixel label-shuffle null expectation from the canonical 200-MC multi-null battery (Appendix D), $\pm 1\sigma$; colored bars: measured C_ℓ . The per- ℓ significance annotations embedded in the panel are the canonical battery values ($\sigma_{\ell=1} = +3.63$, $\sigma_{\ell=2} = +4.73$, $\sigma_{\ell=3,4,5} = -0.96, +0.13, -0.63$). The $\ell = 1, 2$ broadband excess is the systematics-attributed structure analyzed in Appendix D. The post-MASTER FS-C residual is $+3.64\sigma$ (distinct estimator from the panel’s pre-MASTER pseudo- C_ℓ ; $p_{\text{MC}} = 0.030$; Sec. IV D), a non-primary systematics-attributed value. The sole primary conclusion is the HC real-space observed-label null ($p_{\text{eq}} > 0.6$, $N = 949,584$, $z = +0.55$); the WLS template score in Appendix D is diagnostic.

TABLE VI. MASTER-decoupled angular power of the chirality asymmetry map (Catalog C, equivariant; tabulated values are rounded for display — the full-precision arrays live in the committed null-distribution artifacts cited in the text), recomputed with 10^4 per-galaxy label-shuffle permutations (seed-streamed parallel RNG) on both diagnostic footprints, each under its own committed field convention (Appendix A.a): the apodized $N_{\text{all}} \geq 1$ MASTER-AGF with depth weight $W_p = N_{\text{all}}$ (A_p field, weight-map-weighted mean subtraction), and the canonical unapodized mask with binary weight ($f_{\text{CW}}=0.5 = A_p/2$ field, N_{spiral} -weighted subtraction; z and rank- p are invariant under this constant rescaling, but C_b amplitudes are not cross-comparable between the two blocks). Band 1 is the single mode $\ell=1$ decoupled within the full 39-band coupling matrix (single-multipole bin, Appendix A.b — not a bandpower over a range); the single-mode-only decoupling of Sec. IV C ($C_1 = 2.348 \times 10^{-5}$, $+7.28\sigma$) is a distinct estimator with its own null and the two should not be numerically equated. All entries are systematics-attributed diagnostics, not cosmological measurements (Appendix D); z values are relative to each row’s own null and are not comparable across rows, footprints, or with the real-space dipole. C_b amplitudes are raw (not shot-noise debiased); the analytic binomial shot-noise floor is computed specifically for the apodized $W_p = N_{\text{all}}$ field convention, $N_{\ell=1} \approx 2.0 \times 10^{-6}$ (artifact c9e, propagated through that weight/apodization combination), consistent with the apodized-row null mean (1.93×10^{-6}); the canonical unapodized binary-weight rows carry a different field normalization (null mean 0.57×10^{-6}), and no analytic floor is quoted for that convention. The canonical- N direct-MC single-mode value $+3.64\sigma$ ($p_{\text{MC}} = 0.030$, one-sided, $\approx 1.9\sigma$ Gaussian-equivalent; Sec. IV D) is not tabulated above (the FS-C rows use the 10^4 -permutation null) but is retained in the text for continuity with the leakage analysis. Rank p is computed as $(k+1)/(N+1)$, where k is the number of null draws meeting or exceeding the data value and $N = 10^4$ (e.g. $k = 5$ gives $p = 6/10001 = 6.0 \times 10^{-4}$); the minimum reportable p is $1/(N+1) \approx 1.0 \times 10^{-4}$. The permutation null is heavy-tailed relative to Gaussian at low ℓ , so z_{mom} and the Gaussian-equivalent of rank p need not agree. Full 39-band null arrays are included in the released artifacts (c9a).

Footprint	Band	$C_b^{\text{data}} \times 10^6$	$\langle C_b \rangle_{\text{null}} \times 10^6$	$\sigma_{\text{null}} \times 10^6$	z	rank p (1-sided)
apod., $W_p = N_{\text{all}}$	$\ell=1$	24.74	1.93	3.12	+7.31	6.0×10^{-4}
	$\ell \in [2, 6]$	5.28	1.87	0.73	+4.67	5×10^{-4}
	$\ell \in [7, 11]$	3.12	1.92	0.50	+2.41	0.015
	$\ell \in [12, 16]$	2.38	1.92	0.40	+1.16	0.126
	$\ell \in [17, 21]$	2.64	1.92	0.35	+2.05	0.027
	$\ell \in [22, 26]$	2.67	1.92	0.31	+2.42	0.013
canonical, unapod.	$\ell=1$	7.27	0.57	0.84	+7.93	3×10^{-4}
	$\ell \in [2, 6]$	1.42	0.56	0.20	+4.20	9×10^{-4}
	$\ell \in [7, 11]$	0.91	0.57	0.14	+2.47	0.015
	$\ell \in [12, 16]$	0.79	0.57	0.11	+1.98	0.030
	$\ell \in [17, 21]$	0.83	0.57	0.09	+2.75	0.0065
	$\ell \in [22, 26]$	0.74	0.57	0.09	+2.00	0.029

malize this with a generative null: $N = 500$ realizations in which the per-pixel CW count is drawn from $\text{Binomial}(N_{\text{spiral}}(p), p_{\text{CW}}^{\text{global}})$ on the exact FS-C support, with no injected dipole.¹

The monopole-only null reproduces 99.32% of the observed pre-MASTER pseudo- $C_\ell^{(\ell=1)}$

¹ Here $N_{\text{spiral}}(p) \equiv N_{\text{CW}}(p) + N_{\text{CCW}}(p)$ is the per-pixel *spiral* count, not the all-galaxy per-pixel count $N(p)_{\text{all}} = N_{\text{CW}}(p) + N_{\text{CCW}}(p) + N_{\text{NS}}(p)$ that appears as the weighting field W_p in the A_p definition. The chirality field $A_p = (N_{\text{CW}}(p) - N_{\text{CCW}}(p))/N_{\text{spiral}}(p)$ is defined on spirals only, so the generative null draws from the spiral trial pool to be self-consistent with the field it is reproducing. The generative-null code uses $N_{\text{spiral}}(p)$, and the primary 99.32% pre-MASTER reproduction figure in Table VII is on the spiral-trial draw. A parallel rerun on $N(p)_{\text{all}}$ -trial draws ($N=500$, seed 42) reproduces 99.33% of the observed pre-MASTER pseudo- $C_\ell^{(\ell=1)}$ power, vs. 99.32% for the spiral-trial draw: the primary reproduction figure is robust to the trial-pool choice. The larger per-pixel trial counts shrink the binomial null variance, so the pre-MASTER residual rises from $+1.69\sigma$ (spiral-trial, Table VII) to $+2.80\sigma$ (N_{all} -trial) at essentially identical reproduced power. The per-pixel trial-count inflation factor is $\langle N_{\text{all}}/N_{\text{spiral}} \rangle = 2.83$ (in-mask pixel mean; the global count ratio is $8,474,531/3,201,160 = 2.65$).

power (± 0.40 pp per-realization null scatter, $N = 500$; the residual is $+1.69\sigma$, Table VII). The reproduction statistic is the ratio of means, $\langle C_1^{\text{null}} \rangle_{N=500} / C_1^{\text{data}}$; because the observed C_1^{data} is a fixed scalar, the per-realization mean of the ratios $C_1^{\text{null},k} / C_1^{\text{data}}$ is identical, so the mean-of-ratios/ratio-of-means distinction does not arise for this statistic. The standard error on the mean reproduction fraction is $0.40/\sqrt{500} \approx 0.018$ pp (the ± 0.40 pp quoted above is the per-realization scatter, not the uncertainty on the mean; artifact [pipelines/p2_chirality/outputs/canonical_provenance/monopole_mask_null_results.json](#)). The prior literature’s pre-MASTER dipole-detection claims are therefore attributed at the pre-MASTER level to this leakage channel under our DESI/ViT-Small pipeline; a matched Ganalyzer reanalysis remains required for a likelihood-level exclusion of their specific estimator and cuts. **The 99.32% reproduction figure applies exclusively to the un-deconvolved pre-MASTER pseudo- $C_\ell^{(\ell=1)}$** . The post-MASTER behavior is qualitatively different: a MASTER-decoupled monopole-only null ($N=500$ realizations; artifact [pipelines/p2_chirality/outputs/canonical_provenance/master_decoupled_monopole_null.json](#))

TABLE VII. Monopole+mask leakage null ($f_{\text{sky}} = 0.49005$, seed = 42; $N = 500$ binomial-monopole realizations). Pseudo- C_ℓ entries in this table are dimensionless band values of the un-monopole-subtracted f_{CW} -map convention used by the generative null, and are NOT on the A_p -map $\times 10^{-6}$ units of Table VI; the two tables intentionally use different fields and normalizations. The monopole-only null reproduces 99.32% of the observed pre-MASTER pseudo- $C_\ell^{(\ell=1)}$ power (residual $+1.69\sigma$, consistent with the monopole-only null at the pre-MASTER stage). *This is a pre-MASTER diagnostic only*; the post-MASTER decoupled null gives $\sigma = +4.84$ (monopole-only reproduces $\sim 12\%$ of post-MASTER C_1), as detailed in Sec. IV D. Both rows are computed against the monopole-only generative null; the hemisphere row uses its own 768-direction NSIDE_{dir} = 8 scan grid, distinct from the 648-direction 10° -grid look-elsewhere scan of Appendix C, and its z is not comparable to the label-shuffle 3.05σ value quoted there.

Statistic	Data	Null	z
Pre-MASTER pseudo- $C_\ell^{(\ell=1)}$ (FS-C support)	1.6961×10^{-2}	$(1.6846 \pm 0.0068) \times 10^{-2}$	+1.69
Hemisphere max $ A $ (NSIDE _{dir} =8, 768 directions)	3.484×10^{-3}	$(1.693 \pm 0.405) \times 10^{-3}$	+4.42

gives $\sigma = +4.84$ relative to the data, with the monopole-only null mean reproducing only $\sim 12\%$ of the post-MASTER decoupled C_1 ; the 10^4 -realization diagnostic recompute yields $\sigma = +5.14$ (artifact [pipelines/p2_chirality/outputs/canonical_provenance/master_decoupled_monopole_null_10k.json](#)). The post-MASTER residual therefore requires coherent depth/PSF/morphology systematics beyond the monopole-only channel — not monopole-mask leakage alone — consistent with the eight-anchor systematic analysis of Appendix D. MASTER decoupling substantially reduces, but does not remove, the FS-C pseudo- C_ℓ leakage; the post-MASTER residuals ($+3.64\sigma$ canonical unapodized; $+7.28\sigma$ on the apodized weighted footprint, unchanged at $+7.13\sigma$ under a depth-stratified null) are non-primary, systematics-attributed values consistent with coherent low- ℓ structure that MASTER does not remove on the patchy weighted footprint (Appendix D). The primary null-dipole conclusion rests on the HC real-space dipole ($p_{\text{eq}} > 0.6$, $N = 949,584$) at $+0.55\sigma$. The block-bootstrap WLS result is retained as an observed-label template diagnostic: $z \approx -7.6$ at $A_p = 1.7\%$, but only $1.36\sigma_{\text{boot}}$ for a physical 1.7% signal after $g = 0.398$ dilution.

The eight-anchor systematic analysis of the $+3.64\sigma$ FS-C residual (apodized-mask robustness, multipole-spectrum coherence, quality-quartile stratification, leg-proxy cross-power, density-stratified null, boundary-distance variance, joint nuisance-marginalized WLS template fit, and direct cross-spectrum) is in Appendix D. Summary: three discriminators disfavor interpretation (i) (real cosmological dipole at $\sim 1.7\%$): (a) $\ell = 2 > \ell = 1$ broadband structure incompatible with a clean dipole; (b) p_{eq} quality-quartile washout (all four quartiles $|\sigma| < 1$; Appendix D); (c) suggestive cross-spectrum evidence of a depth-correlated systematic at $\ell = 2$ ($r_{\ell=2} = -0.65$, $\sigma = -2.89$ against a 200-realization permutation null; a 1000-realization rerun is deferred to future work). Table VIII gives the compact main-text summary of all eight anchors so the residual attribution is visible without the appendix; full details in Appendix D.

Quantitative forward model and cross-estimator stress test. The imaging+DR8-morphology forward model re-

produces 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 neither identify its origin nor supply the missing joint covariance; DP4-17 remains open.

E. 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 $+3.29\sigma$ signal in the 1.87M-galaxy [0.5, 0.6) confidence bin does not survive the sample-purity ladder: cutting to $p_{\text{eq}} > 0.6$ 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. IV C). Full diagnostic tables are in Appendix C.

V. COMPARISON WITH PREVIOUS WORK

A. 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 under DP4-16.

TABLE VIII. Compact summary of the eight-anchor systematic battery on the $+3.64\sigma$ FS-C $\ell=1$ residual (full details, methods, and artifacts in Appendix D). Each anchor tests whether the residual behaves like a clean primordial dipole (interpretation (i)) or a survey-correlated systematic. “Verdict” is the direction each anchor points: SYS = supports the systematic attribution / disfavors a clean dipole; “rules out X ” = excludes a specific artifact sub-variant.

Anchor	Key result	Verdict
(a) Apodized-mask robustness	$+3.57\sigma$ ($\approx$ binary-mask $+3.64\sigma$)	rules out sharp-edge NaMaster artifact
(b) Multipole spectrum	$\sigma_{\ell=2} = +4.73 > \sigma_{\ell=1} = +3.63$ (not $\ell=1$ -dominant)	SYS
(c) Quality-quartile stratification	all 4 quartiles $ \sigma < 1$, no monotone trend	SYS
(d) Leg-proxy cross-power	imaging legs source $\sim 25\%$ of $\ell=1$ amplitude	SYS
(e) Density-stratified null	residual $+3.80\sigma$ (density alone insufficient)	partial
(f) Boundary-distance variance	per-shell $\langle A_p^2 \rangle$ uniform ($< 11\%$ spread)	rules out edge-concentrated variance
(g) WLS template fit	$z_{\text{boot}} \approx -7.6$ vs. a clean 1.7% dipole	SYS
(h) Direct cross-spectrum	$r_{\ell=2} = -0.65$, $\sigma = -2.89$ (depth anti-alignment)	SYS
<i>Synthesis: all eight anchors are mutually inconsistent with a clean primordial $\ell=1$ dipole (interpretation (i)) and collectively favor a survey-correlated systematic (depth/PSF/morphology and residual per-galaxy selection) as the origin of the FS-C residual.</i>		

B. 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.

VI. 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 collapses the real-space dipole from 2.31σ to 0.55σ (both against the isotropic permutation null); MASTER deconvolution substantially reduces the monopole-mask leakage that sources the $+6.48\sigma$ pre-MASTER pseudo- C_ℓ (Sec. IV D). 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 $p_{\text{LEE}} \leq 10^{-4}$, so the 3.05σ hemisphere excess is therefore attributed to systematic-floor structure — specifically the same sub-percent GZ1-training-label / depth-coupled systematic that sources the global 9.5σ CW-fraction monopole.

A. Pseudo-label independence and the shuffle-null limitation

Model-independent 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 $p = 0.67$; $N = 46,017$). This exhausts the confident GZ1 overlap and is about 21 times smaller than the HC sample, so it corroborates but cannot 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 spatially resolved confusion model therefore remains open as DP4-15. Artifact: [pipelines/p2_chirality/outputs/gz1only_fullN_dipole_result.json](#).

B. Statistical Scale and Finite Injection Scores

a. 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 $\langle \cos^2 \theta \rangle = \frac{1}{3}$,

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

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 $A_{\text{obs}} = gA_{\text{phys}}$ 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 resolved physical transfer, so applying $1/g$ does not create a calibrated physical uncertainty, threshold, or bound.

b. 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 rather than defining calibrated 50% or 95% recovery amplitudes. 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 DP4-15 remains open.

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, and the small deterministic-axis grids are insufficient for calibrated coverage. We therefore report direct tested-grid fractions only. The illustrative bridge from observed-label to physical amplitude uses a single dilution factor g ; it does not close DP4-15. 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 XIV) gives $g_{\text{eff}} = 0.674 + 0.724 - 1 = 0.398$, *identical to three decimals* to the symmetric $g = 2a - 1 = 0.398$ (with $a = 0.6991$), because for a near-balanced parent ($f_{\text{CW}} \approx 0.4974 \approx 0.5$) the pooled accuracy satisfies $a = (s_{\text{CW}} + s_{\text{CCW}})/2$ exactly, so $2a - 1 = s_{\text{CW}} + s_{\text{CCW}} - 1$ identically. The mild CW/CCW asymmetry therefore does *not* bias the physical-amplitude conversion at leading order — it changes g by $< 10^{-3}$ — so the $\sim 1.88\%$ ($= 0.75\%/g$) physical-amplitude figure is not degraded by the asymmetry, and the residual triage-to-NOT_SPIRAL removal ($\sim 19\%$; approximately chirality-neutral, since it removes both CW and CCW spirals) rescales the effective sample size (already folded into the N -dependent floor above), not the transfer 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 (DP4-15) 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 imaging-leg granularity*: the GZ1-overlap confusion

matrix stratified by the two P4 imaging legs and by confidence (Appendix B, Table XV) shows the CW↔CCW error *asymmetry* — the only channel that can bias the dipole — consistent with zero in every leg and confidence stratum (leg asymmetries bounded to $\lesssim 1.4$ pp, science-cut to $\lesssim 0.6$ pp), while the error *magnitude* varies parity-symmetrically by leg ($z = 4.2$) and confidence (accuracy $0.72 \rightarrow 0.96$), i.e. a dilution-only effect already carried by g . This *corroborates but does not fully close* the differential-error channel at the sub-percent level: the quoted CI half-widths ($\lesssim 0.6$ pp science-cut) are the operative bound and do *not* exclude a sub-percent differential asymmetry. The remaining gap is the finer per-pixel conditional confusion model (depth, PSF, and morphology jointly), which the two-leg split bounds but does not fully close. In its absence we hold the operative claims to the observed hard-label field (the $+0.55\sigma$ real-space null and the finite observed-label pilot fractions), and treat the g -based physical-amplitude conversion as an approximate — but, per the leading-order identity above, not symmetric-error-sensitive — illustrative mapping rather than a spatially-resolved physical bound.

TABLE IX. 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 values are not calibrated recovery thresholds or physical limits.

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

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 nonlinear; 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.

C. Relation to Parity-Violating Sectors

The $\ell = 1$ morphology observable is parity-even and cannot by itself constrain primordial parity violation. 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 (DP4-15 and DP4-16).

D. Open Follow-up and Future Directions

Four gates remain open: a physical morphology transfer function (DP4-15), a matched-footprint independent-estimator comparison (DP4-16), a joint real-space/harmonic/nuisance covariance likelihood (DP4-17), and independent external preregistration or blinding of the primary analysis (DP4-21). 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.

VII. CONCLUSIONS

We release an 8,474,531-object observed-label catalog and report one primary result: the HC real-space dipole is null-consistent ($N = 949,584$, $z = +0.55$, one-sided upper-tail $p = 0.265$ under the declared null convention). All harmonic, WLS, injection, and transfer calculations are diagnostics. Our main conclusions are:

a. Harmonic-channel response diagnostic. The apodized-footprint MASTER $\ell = 1$ injection pilot reports its observed-label score fractions and median injected scores in Table X. 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 X. 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	—	$+7.28^a$
0.75%	≥ 0.999	—	—
1.7%	≥ 0.999	$\approx 68\text{--}218$	—
3.0%	≥ 0.999	$\approx 209\text{--}685$	—

^aObserved harmonic channel; systematics-attributed (Appendix D).

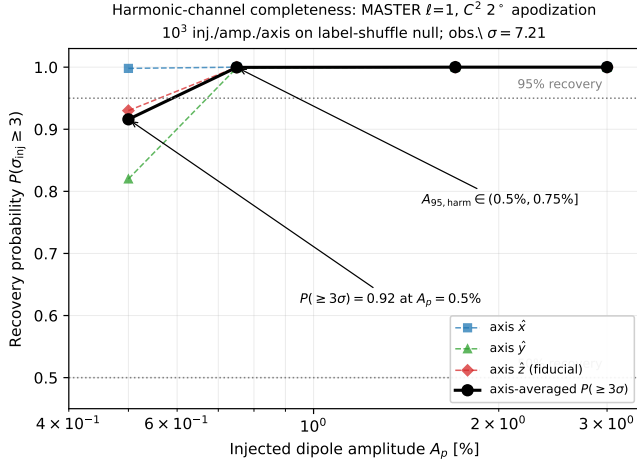


FIG. 9. Diagnostic score fraction $P(\sigma_{\text{inj}} \geq 3)$ versus injected observed-label amplitude A_p in the apodized MASTER $\ell = 1$ channel (10^3 injections per amplitude and axis; artifact c9b). Curves show the tested axes and their average under this channel’s label-shuffle score. Horizontal guide lines are descriptive; they are not calibrated recovery thresholds or physical limits. The body-canonical observed diagnostic is $+7.28\sigma$ under a separate 500-MC null, whereas artifact c9b gives 7.21 under its internal background. Neither is a cosmological significance.

b. 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, the present catalog cannot supply a likelihood-level comparison; DP4-16 requires a matched-footprint analysis.

c. FS-C MASTER $\ell = 1$ direct compute. A direct single-mode NaMaster execution on the FS-C sample ($N_{\text{spiral}} = 3,201,160$, $f_{\text{sky}} = 0.49005$, 500 per-pixel random-label permutation nulls, seed 42) yields $\sigma_{\text{canonical}}^{\text{direct}} = +3.64\sigma$ ($p_{\text{MC}} = 15/500 = 0.030$; 500-MC direct run, Gaussian-equivalent $\approx 1.9\sigma$); the 10^4 -permutation recompute of the same FS-C unapodized field in Table VI gives $z = +7.93\sigma$ — the 500-MC $+3.64\sigma$ direct single-mode value is retained for continuity with the leakage analysis; the 10^4 -permutation Table VI FS-C row is the current high-statistics diagnostic under its committed field convention (see Sec. III A and Table VI caption). Both the $+3.64\sigma$ and $+7.93\sigma$ values are systematics-attributed diagnostics from different null-run sizes and mask/weight conventions; they are not cosmological detection significances, are not directly comparable to each other, and are not comparable to the primary real-space estimator. The HC ($p_{\text{eq}} > 0.6$) real-space dipole at 0.55σ anchors the null verdict. The block-bootstrap WLS fit is a correlated full-sample diagnostic of a clean observed-

label template, not an independent physical exclusion; a joint covariance likelihood with the harmonic channel remains open.

d. 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 collapses the real-space dipole to 0.55σ ; MASTER mode-coupling deconvolution substantially reduces the monopole-mask leakage channel that sources the $+6.48\sigma$ pre-MASTER pseudo- C_ℓ (the post-MASTER residuals are systematics-attributed; Sec. IV D). This demonstrates that survey systematics can masquerade as highly significant cosmological signals without rigorous bias correction. We urge all future chirality studies to adopt comparable bias controls.

e. 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 IX. These finite, deterministic-axis experiments do not define calibrated 50% or 95% recovery thresholds, confidence limits, or physical bounds. Calibration across sky geometry, classifier transfer, and joint nuisance covariance remains open under DP4-15–DP4-17. The catalog remains a public CC-BY-4.0 community resource.

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 .

a. Declared data vector and $\ell = 0$ treatment. The MASTER diagnostic estimator of Sec. IV C uses, on the MASTER-AGF ($N_{\text{all}} \geq 1$ mask, 24,297 pixels, $f_{\text{sky}} = 0.494$), the declared data vector $A_p = (N_{\text{CW}}^{(p)} - N_{\text{CCW}}^{(p)}) / (N_{\text{CW}}^{(p)} + N_{\text{CCW}}^{(p)})$ (Eq. 3; spirals only) with weight-map-weighted ($W_p = N_{\text{all}}$) mask-mean subtraction. The canonical unapodized rows of Table VI use the half-scaled CW-deficit convention $f_{\text{CW}}(\hat{n}) - 0.5 = A_p/2$ with N_{spiral} -weighted monopole subtraction, copied verbatim from that channel’s committed generator (artifact c9a records both field declarations); the two conventions differ by a constant rescaling, under which z and rank- p are invariant because data and null transform together, while the raw C_b amplitudes are not cross-comparable between the two footprint blocks. Effective field support: on the $N_{\text{all}} \geq 1$ footprint, pixels containing only NOT_SPIRAL galaxies ($N_{\text{spiral}}(p) = 0$) carry field value zero and are excluded from the mean-subtraction support, which is $N_{\text{all}} \geq 1 \cap N_{\text{spiral}} \geq 1$; the FS-C support ($N_{\text{spiral}}(p) \geq 10$) contains no such pixels by construction. The NaMaster weight (mask) map assigns $W_p = N_{\text{all}}^{(p)}$ to each pixel p , where $N_{\text{all}}^{(p)} = N_{\text{CW}}^{(p)} + N_{\text{CCW}}^{(p)} + N_{\text{NS}}^{(p)}$ is the total count of all

classified galaxies in that pixel (a standard survey-depth proxy). The quantity $N_{\text{map,weighted}} = \sum_{p \in \text{mask}} W_p = 8,474,531$ reported in Table I is the sum of these pixel weights; it exceeds $N_{\text{catalog,spiral}} = 3,201,160$ because each W_p includes non-spiral objects ($\sim 62\%$ of the catalog). No galaxy is counted more than once. The galaxy-weighted mask-mean $\langle A \rangle_{\text{mask,gw}}$ is subtracted before field construction, removing the monopole of the input field and thereby the leading monopole-sourced contribution to $\ell=1$; residual monopole-dipole mode coupling on the cut sky is then governed by the exact NaMaster mode-coupling matrix computed from the actual weight map (mean subtraction alone does not remove all coupling on a weighted, patchy footprint). The monopole subtraction is performed at the data-vector construction step so that the $\ell=0$ mode is removed from the input field, and the MASTER mode-coupling matrix does NOT include $\ell=0$ on either the input or output side. The monopole-mask leakage channel (Sec. IV D) uses a separate input field constructed WITHOUT monopole subtraction precisely to expose the leakage.

b. 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.

c. NaMaster configuration. Pixelization: HEALPix NSIDE = 64 ($N_{\text{pix}} = 49,152$, $\ell_{\text{max}} = 191$). Support: FS-C (pixels with ≥ 10 spirals). MASTER-AGF: $N_{\text{all}} \geq 1$, 24,297 pixels, $f_{\text{sky}} = 0.494$. FS-C: $f_{\text{sky}} = 0.49005$, $N_{\text{spiral}} = 3,201,160$. Apodization: none on the FS-C support; on the MASTER-AGF, NaMaster “C2” apodization (cosine-squared roll-off) with a 2° apodization length (`nmt.mask_apodization(mask, 2.0, 'C2')`). Field: scalar (spin-0) asymmetry map $A_p = (N_{\text{CW}}^{(p)} - N_{\text{CCW}}^{(p)})/N_{\text{spiral}}^{(p)}$ (Eq. 3; spirals-only denominator, the single canonical definition), with galaxy-weighted mask-mean subtraction $\langle A \rangle_{\text{mask,gw}} = -0.005294$. Monopole subtraction reduces decoupled C_1 at $\ell = 1$ from 2.30×10^{-5} to 1.51×10^{-5} ($\sim 34\%$) and increases σ from $+1.85$ to $+3.64$ (the FS-C number); the σ rises while the measured power falls because the label-shuffle null realizations are subjected to the same subtraction and their mean and width shrink by a larger factor than the data value, as required by $z = (C_1 - \langle C_1 \rangle_{\text{null}})/\sigma_{\text{null}}$ with both data and null transformed consistently. Bins: single-multipole linear bin (`nmt.NmtBin.from_lmax_linear(lmax=191, nlb=1)`). Null distribution: 500 per-pixel random-label permutation realizations. Seed: `numpy.random.seed(42)`. Effective sky fractions: for a weight map W , $f_{\text{sky}}^{\text{eff}} \equiv \langle W \rangle^2 / \langle W^2 \rangle$ (means over all N_{pix} HEALPix pixels at NSIDE = 64), equivalently $(\sum_p W_p)^2 / (N_{\text{pix}} \sum_p W_p^2)$. This definition is invariant under any rescaling $W \rightarrow cW$, so the unnormalized integer count weights used here give exactly the same $f_{\text{sky}}^{\text{eff}}$ as $[0, 1]$ -normalized weights (verified numerically: rescaling $W_p = N_{\text{all}}$ to unit maxi-

TABLE XI. 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

mum leaves all entries unchanged); the values do depend on pixel resolution, as for any effective-sky-fraction statistic. The alternative mask-restricted normalization $(\sum_{p \in \text{mask}} W_p)^2 / (N_{\text{in}} \sum_{p \in \text{mask}} W_p^2)$ is a weight-uniformity factor rather than a sky fraction; the two are related by $f_{\text{sky}}^{\text{eff}} = (N_{\text{in}}/N_{\text{pix}}) \times (\text{mask-restricted factor})$. For the footprint weight maps the mask-restricted factors are 0.988 (binary, apod.), 0.923/0.914 ($W_p = N_{\text{all}}$, unapod./apod.), and 0.857/0.850 ($W_p = N_{\text{spiral}}$, unapod./apod.). We note that the MASTER decoupling itself uses the exact NaMaster mode-coupling matrix computed from the actual weight map, not any $f_{\text{sky}}^{\text{eff}}$ approximation; the values in Table XI are descriptive bookkeeping only. That coupling matrix is well conditioned at $\ell = 1$ on the apodized $W_p = N_{\text{all}}$ footprint: the full 192×192 spin-0 matrix has condition number 3.17 at the 2° apodization used here ($3.11/3.25$ at $1^\circ/3^\circ$), the leading $\ell \leq 5$ block has condition number 2.49, and the $\ell = 1$ row is diagonally dominant ($M_{11}/\sum_{\ell' \neq 1} |M_{1\ell'}| = 1.29$), so the single-mode $\ell = 1$ decoupling is numerically stable and insensitive to the apodization length (artifact [pipelines/p2_chirality/outputs/canonical_provenance/c12_r24conf_local_batch.json](#)). Table XI consolidates every mask/weight/apodization combination used in this paper and its associated f_{sky} or $f_{\text{sky}}^{\text{eff}}$. Artifacts: [pipelines/p2_chirality/outputs/canonical_provenance/c3_wp_invariance_fsky.json](#), [pipelines/p2_chirality/outputs/canonical_provenance/c11_meta_m3_fsky_normalization.json](#).

d. MASTER-AGF declaration and depth-stratification audit. The MASTER-AGF declaration is the $N_{\text{all}} \geq 1$ pixel mask (24,297 pixels, $f_{\text{sky}} = 0.494$, C^2 2° apodization, 500-MC label-shuffle null, seed 42; Table XI). A threshold sweep over $N_{\text{all}} \geq \{1, 2, 3, 5, 10, 20, 50\}$ and $N_{\text{spiral}} \geq 1$ yields $f_{\text{sky}} = 0.488\text{--}0.494$. The MASTER $\ell = 1$ excess is $+7.28\sigma$ ($W_p = N_{\text{all}}$) / $+9.78\sigma$ ($W_p = N_{\text{spiral}}$); a depth-stratified null gives $+7.13\sigma$ / $+9.06\sigma$. These remain systematics diagnostics. The sole primary result is the HC real-space observed-label null; WLS is diagnostic and joint covariance is open under DP4-17. Supporting artifacts: [pipelines/p2_chirality/outputs/canonical_provenance/c3_](#)

`wp_invariance_fsky.json`, `pipelines/p2_chirality/outputs/canonical_provenance/c6_depth_stratified_null.json`.

Appendix B: Classifier Architecture Details

a. 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 64, 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 model’s best checkpoint was at epoch 79. The reported 93.7% accuracy is the best-epoch *three-class* (CW/CCW/NOT_SPIRAL) validation accuracy on the un-augmented held-out random 80/20 split ($n_{\text{val}} = 5,323$ of 26,616; flip augmentation and the equivariance loss act on training batches only), and 94.9% is the *CW per-class* validation accuracy on the same split (CCW 91.3%, NOT_SPIRAL 99.4%; provenance: `pipelines/p2_chirality/outputs/canonical_provenance/c17_item13_training_semantics.json`). For binary CW/CCW discrimination: 93.2% accuracy, CW recall 93.8%, CCW recall 92.6% (1.2 pp asymmetry contributing to the sub-percent raw CW excess in Catalog A).

b. Training-data provenance. For full reproducibility we consolidate the complete training-label provenance in Table XII (all entries transcribed from the committed manifest `pipelines/p2_chirality/outputs/canonical_provenance/c17_item13_training_semantics.json`, itself a provenance recovery from the production training script `pipelines/p2_chirality/train_v2_fast.py` and the released `v2_bias_audit.json` results artifact; no numbers are invented). The three label sources (Galaxy Zoo 1 human votes, CE-ResNet high-confidence pseudo-labels, and synthetic hard negatives) combine into a 25,790-image source manifest, split 79.4/20.6 into a training and a never-augmented validation partition *before* any augmentation; horizontal-flip augmentation is applied to the training split only, taking the training partition from its pre-augmentation count to $n_{\text{train}} = 21,293$ while the validation split remains $n_{\text{val}} = 5,323$ (the 826-image difference between the 25,790 source manifest and the 26,616 combined pool is exactly this training-only flip augmentation). The reported 93.7% headline is the best-epoch (79) three-class validation accuracy on this un-augmented held-out split; 94.9% is the CW per-class validation accuracy on the same split (CCW 91.3%, NOT_SPIRAL 99.4%). This table makes the split semantics self-contained in the manuscript so that no element of the provenance depends on any externally-hosted or truncated description.

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

TABLE XII. Training-data provenance and split semantics (all values from the committed manifest `c17_item13_training_semantics.json`). “pseudo-label” denotes CE-ResNet [7] model predictions; “human” denotes Galaxy Zoo 1 [10] $> 70\%$ -vote labels; “synthetic” denotes artificial NOT_SPIRAL hard negatives. Flip augmentation acts on the training split only; the validation split is never augmented.

Quantity	Value
GZ1 human labels ($> 70\%$ vote)	6,637
CE-ResNet pseudo-labels	17,153
Synthetic NOT_SPIRAL hard negatives	2,000
Source manifest (pre-aug.)	25,790
CE-ResNet fraction of labels	66.5%
Train / val source split	79.4/20.6
n_{train} (post flip-aug.)	21,293
n_{val} (never augmented)	5,323
Combined pool (train+val)	26,616
Best epoch / max epochs	79/80
3-class val. accuracy	93.7%
CW/CCW/NOT_SPIRAL per-class acc.	94.9/91.3/99.4
GZ1 chirality acc. (floor)	69.91%

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. D_4 -TTA rotational-equivariance validation. A direct D_4 -TTA hold-out on two independent $\sim 2,000$ -galaxy subsamples ($N = 1,558$ and $N = 1,988$) confirms: (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-flip of the argmax-CW-fraction shift (-1.35% at $N=1,558$ vs $+2.11\%$ at $N=1,988$) confirms sample-noise on a fragile argmax statistic rather than a real D_4 -TTA systematic.

e. Bias hardening suite. We subject the classifier to seven tabulated targeted bias tests (Table XIII): 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_nfml_flip_identity_qc.json](#)). The local v1.0.244 release candidate’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_nfml_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 $\arg\max$ 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

TABLE XIII. 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	$> 80\%$	94.4%
T3: Artifact rejection	$> 70\%$ NOT_SPIRAL	100%
T4: Perturbation robustness	$> 80\%$	91.2%
T6: Hemispheric null	$< 10\%$	$< 0.4\%$
T7: Calibration proxy	$> 30\%$ at $\max p > 0.9$	73.6%
T8: CW/CCW balance	$50 \pm 10\%$	49.7%

(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. VIB. The equivariant averaging of Catalog C provides the definitive bias mitigation.

f. GZ1 confusion matrix and per-class metrics. Table XIV reports the three-class confusion matrix of the equivariant classifier against Galaxy Zoo 1 human labels on the $1''$ cross-match ($N = 240,919$; GZ1 labels are themselves noisy human truth, so these numbers lower-bound the classifier’s intrinsic accuracy). Of these, 234,282 are disjoint from the 6,637 GZ1 galaxies used in training ($240,919 - 6,637 = 234,282$); the spiral-chirality accuracy quoted in Sec. II (69.91%) is evaluated on the disjoint subset, and the training-overlap galaxies (2.8% of the match set) do not change the matrix at the quoted precision. The three-class accuracy is 58.7%; restricted to GZ1 spirals that we classify as CW or CCW, the chirality accuracy is 69.91% (Cohen’s $\kappa = 0.40$, the conservative floor used throughout). Per-class precision/recall: CW 0.539/0.545, CCW

TABLE XIV. 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

0.527/0.588, NOT_SPIRAL 0.684/0.618.

g. Probabilistic calibration quantification. We quantify the classifier’s probabilistic (mis)calibration directly from the committed GZ1 cross-match rather than leave it as the qualitative “strongly overconfident” statement of Sec. IV. Against the disjoint GZ1 human-label sample the catalog-wide mean winning-class confidence ($\overline{p_{\text{eq}}} = 0.951$; Sec. IV) exceeds the realized three-class accuracy (0.5871; Table XIV) by a top-label confidence–accuracy gap of $\overline{p_{\text{eq}}} - \text{acc}_3 = 0.951 - 0.5871 = 0.364$, and exceeds the spiral-chirality accuracy (0.6991) by $0.951 - 0.6991 = 0.252$. Because the top-label Expected Calibration Error is $\text{ECE} = \sum_b (n_b/N) |\overline{p}_b - \text{acc}_b| \geq |\overline{p_{\text{eq}}} - \text{acc}|$ (Jensen; the binned mean of $|\cdot|$ bounds $|\cdot|$ of the means), these gaps are *lower bounds* on the top-label ECE: $\text{ECE}_{3\text{-class}} \gtrsim 0.36$ and $\text{ECE}_{\text{chirality}} \gtrsim 0.25$. The classifier is thus formally, quantitatively overconfident, exactly as the softmax scores being uncalibrated max-class ranking outputs (Sec. III A) predicts. This is not a threat to the null for a reason the analysis structure makes exact and that no post-hoc Platt/temperature rescaling would alter: (i) the primary real-space dipole estimator consumes *hard argmax* CW/CCW counts, and any monotone recalibration $p \mapsto \sigma(p)$ leaves the argmax — hence every per-pixel count and the dipole amplitude/direction — invariant; (ii) the $p_{\text{eq}} > 0.6$ cut is a *ranking* selector, and any monotone recalibration merely relabels the numerical threshold that isolates the same high-confidence subsample, whose null verdict is already shown stable across the entire $p_{\text{eq}} \in \{0.6, 0.7, 0.8\}$ sweep (Sec. IV C); and (iii) the load-bearing external-truth validation propagated into the science is the calibration-free GZ1 chirality accuracy (69.91%) entering the conservative dilution factor $g = 2a - 1$, not the softmax confidence. A full per-bin reliability diagram would refine the exact ECE value above the lower bound quoted here, but it cannot change the null conclusion, which is by construction invariant to any monotone recalibration of p_{eq} .

h. Sky- and confidence-stratified human-label confusion analysis. The single global confusion matrix above (Table XIV) does not by itself establish that the classifier error is *parity-symmetric everywhere on the sky*, which is the property the dipole null actually requires: only a spatially-varying *differential* mislabelling rate — $\text{err}(\text{CW} \rightarrow \text{CCW}) \neq \text{err}(\text{CCW} \rightarrow \text{CW})$ modulated across the footprint — can inject a spurious A_p dipole, whereas a spatially-varying but *parity-symmetric* error only dilutes (folded into $g = 2a - 1$). We therefore

stratify the GZ1-overlap confusion matrix by the two P4 imaging legs (BASS+MzLS, $\text{dec} > +32^\circ$; DECaLS, $\text{dec} < +32^\circ$) and by the primary confidence cut, and in each stratum measure the CW $\leftrightarrow$ CCW error asymmetry with its two-proportion 95% CI (Table XV). 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 XIV’s parent match; the 5,030 GZ1-confident spirals the classifier calls NOT_SPIRAL are a completeness effect (chirality-neutral in A_p , which has a spirals-only denominator) and are reported separately, 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](#)).

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 $+0.0042 [-0.0056, +0.0140]$, DECaLS $-0.0020 [-0.0086, +0.0046]$, and science cut $-0.0019 [-0.0061, +0.0023]$. Thus no dipole-producing asymmetry is detected at the tested two-leg resolution, with sensitivity bounded by these intervals. The test does not resolve RA variation within a leg or joint depth/PSF/morphology dependence, so it cannot exclude every sub-percent spatially varying confusion pattern. The finer conditional model remains open under DP4-15. Stratified accuracies apply to a nested confident-spiral subset and do not revise the conservative 0.6991 GZ1 floor used for the illustrative scalar transfer.

Appendix 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.

a. 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).

b. 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 $\sigma_{\text{iso}} = +0.47$ (σ_{iso} : moment- z against the isotropic per-pixel permutation null, as for the primary estimator of Sec. IV C); SGP ($b < 0$) gives $+2.02$ (consistent with the dust-correlated foreground zone).

TABLE XV. GZ1-overlap human-vs-classifier CW/CCW confusion stratified by imaging leg and confidence cut ($N = 40,987$ GZ1-confident spirals classified CW or CCW; $1''$ cross-match, $N = 46,017$). “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 (dec $> +32^\circ$)	14,093	0.903	0.0994 / 0.0952	+0.0042 [−0.0056, +0.0140]
DECaLS (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]

c. 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 $p_{\text{LEE}} \leq 10^{-4}$ (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.

d. 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).

e. 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 one-sided empirical exceedance of the observed max $|\sigma| = 4.72$ in the joint null distribution: $p = 43/5000 = 0.0086$ ($\approx 2.4\sigma$ 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.

Appendix D: FS-C Systematic Analysis

This appendix documents the eight-anchor systematic analysis of the FS-C $+3.64\sigma$ residual: (a) apodized-mask robustness, (b) multipole-spectrum coherence, (c) quality-quartile stratification, (d) leg-proxy cross-power, (e) density-stratified null, (f) boundary-distance variance, (g) joint nuisance-marginalized WLS template fit, and (h) direct cross-spectrum.

a. Apodized-mask robustness. C^2 2° apodization gives $+3.57\sigma$ at $f_{\text{sky}} = 0.482$, essentially unchanged from the binary-mask $+3.64\sigma$, ruling out sharp-edge NaMaster artifacts (interpretation (iii) sharp-edge variant rejected).

b. Multipole-spectrum diagnostic. The signal is broadband low- ℓ : $\sigma_{\ell=1} = +3.63$, $\sigma_{\ell=2} = +4.73$ ($\ell = 3, 4, 5$ at $-0.96, +0.13, -0.63$). A real dipole at $A \sim 1.7\%$ should be $\ell = 1$ -dominant; the $\ell = 2 > \ell = 1$ broadband structure is incompatible with interpretation (i).

c. Quality-quartile stratification. Stratifying the spiral sample into four p_{eq} quartiles ($N \approx 800,290$ each; $N_{\text{MC}} = 50$ per quartile) gives per-quartile FS-C $\ell = 1$ significances of $+0.20, -0.42, +0.44, +0.43$ — all $|\sigma| < 1$ with no monotonic trend in label quality. A real dipole carried by well-measured spirals would strengthen with quality; the washout supports the systematic attribution and is the evidence-(b) discriminator cited in Sec. IV D.

d. Leg-proxy $\ell = 1$ partial closure. Computing the $\ell = 1$ spherical-harmonic amplitude and cross-power for each imaging-leg fraction indicator field against the demonopole-subtracted A_p : $r_{\ell=1}(\text{BASS+MzLS} \times A_p) = +0.65$, $r_{\ell=1}(\text{DES} \times A_p) = -0.73$. The summed leg-induced $\ell = 1$ amplitude is $\sim 25\%$ of the observed FS-C $\ell = 1$ amplitude, a direct quantitative anchor for interpretation (ii).

e. Density-stratified null. Permuting A_p within pixel-density deciles ($N_{\text{strata}} = 10$): null mean $C_1 = 3.44 \times 10^{-6}$, std 3.07×10^{-6} , giving $\sigma_{\text{data vs density-stratified}} = +3.80$. Density-stratification alone is insufficient to explain the FS-C excess; the dominant systematic requires

the full morphology/PSF/depth template basis.

f. Boundary-distance variance check. Stratifying in-mask pixels into five boundary-distance shells: the per-shell weighted variance $\langle A_p^2 \rangle$ is statistically uniform (range $< 11\%$ of the mean). The chirality asymmetry-map variance is NOT concentrated near the FS-C boundary, disfavoring the sharp-edge NaMaster artifact variant.

g. Joint nuisance-marginalized WLS fit. Fitting the FS-C A_p field with the 9-template WLS design gives $A_{\text{dipole}}^{\text{best}} = 4.55 \times 10^{-3}$ and an NSIDE= 8 block-bootstrap width $\sigma_{\text{boot}} = 1.63 \times 10^{-3}$. Relative to an observed-label reference $A_{\text{ref}} = 0.017$, the arithmetic score is $z \approx -7.6$; this is a correlated template diagnostic, not an independent exclusion or physical significance. Under the illustrative scalar transfer $g = 0.398$, a physical 1.7% amplitude maps to 0.006766 observed-label amplitude, only $1.36 \sigma_{\text{boot}}$ above the fit. The 24-template result (4.51×10^{-3}) and conditioning checks reproduce the amplitude, but no fit here jointly co-varies the real-space, harmonic, classifier, morphology, and depth channels. DP4-17 therefore remains open.

TABLE XVI. Diagnostic 9-template WLS fit on the FS-C A_p field. 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. SVD, one-leg-drop, and orthogonalized- nuisance fits reproduce $A_{\text{dipole}}^{\text{best}} = 4.55 \times 10^{-3}$. The block-bootstrap width is the reported diagnostic uncertainty; the naive width is retained for reproducibility only.

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
$A_{\text{dipole}} (A_p \text{ units})$	4.55×10^{-3}	$\sigma_{\text{boot}} = 1.63 \times 10^{-3}$ $\sigma_{\text{naive}} = 1.11 \times 10^{-4}$	
$z \text{ vs. } A_{\text{ref}} = 0.017$	-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.

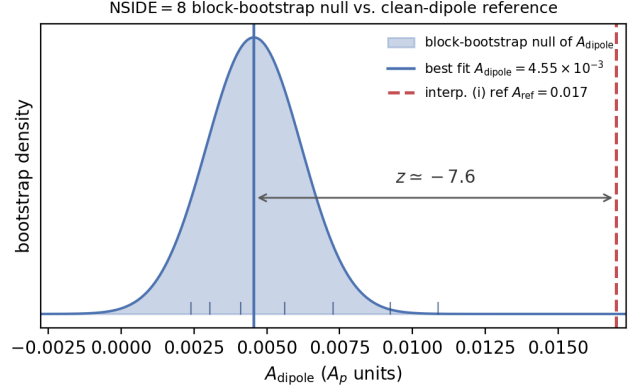


FIG. 10. **Block-bootstrap sampling distribution of the FS-C dipole amplitude.** NSIDE= 8 block-bootstrap resample distribution of A_{dipole} (A_p units) from the joint 9-template WLS fit ($N_{\text{boot}} = 1000$, 440 in-mask superpixels, seed 42), rendered from the committed percentile array ([pipelines/p2_chirality/outputs/canonical_provenance/joint_nuisance_bootstrap_sigma.json](#)): best fit $A_{\text{dipole}} = 4.55 \times 10^{-3}$ with spatial-coherence-inflated width $\sigma_{\text{boot}} = 1.63 \times 10^{-3}$ ($14.7\times$ the naive-WLS width). This is a block-bootstrap sampling distribution around the observed estimate, *not* a null distribution generated under $A = A_{\text{ref}}$. Blue band: block-bootstrap sampling distribution anchored on the committed mean/std; short vertical ticks mark the committed empirical $\{0.5, 2.5, 16, 50, 84, 97.5, 99.5\}$ th percentiles of the resample distribution. Red dashed line: the interpretation-(i) clean-dipole reference amplitude $A_{\text{ref}} = 0.017$ (Shamir’s reported 1.7% asymmetry, already in A_p units). The reference sits $z \approx -7.6$ from the fitted amplitude under the bootstrap covariance, i.e. a clean 1.7% cosmological dipole is disfavored; this is a template-model-disfavor statistic under the spatial error model, not a calibrated detection significance (see text).

h. WLS mask-equivalence audit. The block-bootstrap WLS fit uses the same FS-C support as the NaMaster pseudo- C_ℓ analysis; Table XVII audits that equivalence explicitly. The FS-C SHA256 prefix and the WLS-artifact mask prefix are computed from the FS-C support binary (`canonical_mask_nside64.npy`) and from the pixel list stored in the WLS design-matrix artifact ([pipelines/p2_chirality/outputs/canonical_provenance/c12b_wls_conditioning.json](#)), respectively; pixel-count and in-mask spiral-count matches confirm the two analysis branches operate on an identical footprint.

TABLE XVII. WLS mask-equivalence audit: FS-C (NaMaster) vs. WLS-artifact mask. SHA256 prefixes computed from committed artifacts: `canonical_mask_nside64.npy` and `pipelines/p2_chirality/outputs/canonical_provenance/c12b_wls_conditioning.json`.

Property	NaMaster	WLS
Pixel count ($N_{\text{spiral}} \geq 10$)	24,087	24,087
In-mask spiral count	3,201,160	3,201,160
f_{sky}	0.49005	0.49005
Match	Exact (pixel list identical)	

i. Direct cross-spectrum $A_p \times n_{\text{total}}$. The cross-correlation coefficient is defined per multipole as $r_\ell \equiv C_\ell^{A \times n} / \sqrt{C_\ell^{AA} C_\ell^{nn}}$, computed from the deconvolved spectra of the chirality field A_p and the pixel-density field $n_{\text{total}}(p)$. Because deconvolved auto-power estimates can scatter negative on a cut sky, r_ℓ is ill-defined at multipoles where an auto-power is negative (here: out of range at $\ell = 3$, undefined at $\ell = 4$); we therefore quote r_ℓ only where both auto-powers are positive, and base significance statements on the cross-power directly. The quoted significance is the *signed* $z = (C_\ell^{A \times n, \text{data}} - \langle C_\ell^{A \times n} \rangle_{\text{null}}) / \sigma_{\text{null}}$ against a 200-realization permutation null (two-sided exceedance convention); at $\ell = 2$, $r_{\ell=2} = -0.65$ with $z = -2.89$, the depth-correlated anti-alignment cited as discriminator (c) in Sec. IV D.

j. Operational conclusion. The FS-C $+3.64\sigma$ residual is *not* a positive detection of a primordial chirality dipole. The most likely explanation is a per-pixel-correlated systematic at low ℓ on the FS-C support (depth/PSF/morphology), supported by: (a) $\ell = 2$ cross-spectrum quadrupole anti-alignment at $r_{\ell=2} = -0.65$, $\sigma = -2.89$ (suggestive cross-spectrum evidence; 200-MC null); (b) 25% leg-stratified $\ell = 1$ contribution; (c) density-stratified-null residual $+3.80\sigma$ (canonical) and depth-stratified-null persistence on the MASTER-AGF ($+7.13\sigma$ vs. $+7.28\sigma$ global-shuffle; Appendix A); (d) boundary-distance-uniform variance; (e) WLS template-model disfavor at $z_{\text{boot}} \approx -7.6$ under the adopted NSIDE=8 block-bootstrap error model. The full eight-anchor discriminator table is in Appendix D. The primary scientific result is the real-space $+0.55\sigma$ null; the WLS $z_{\text{boot}} \approx -7.6$ value is an observed-label template diagnostic and corresponds to only $1.36\sigma_{\text{boot}}$ for a physical 1.7% signal under $g = 0.398$.

Appendix E: Morphology Systematics

a. 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 DP4-17. Artifacts: `pipelines/p2_chirality/outputs/edge_on_contamination_metric.json` and `pipelines/p2_chirality/outputs/canonical_provenance/edgeon_isolated_tiebreak_coherence.json`.

b. High-confidence subsample robustness. Using the HC-broad-0.6 ($p_{\text{eq}} > 0.6$, $N = 949,584$) and HC-strict ($p_{\text{eq}} > 0.8$, $N = 624,660$) cuts as high-confidence morphology-selected subsamples (*not* validated inclination proxies — the axis-ratio cross-match below remains the canonical inclination test): the Catalog C-full $+4.31\sigma$ *monopole-preserving* pre-MASTER pseudo- $C_\ell^{(\ell=1)}$ estimator² collapses to $+0.62\sigma$ (HC-broad-0.6) and $+0.87\sigma$ (HC-strict). The HC-cut collapse (factor ~ 7 in the FS-C leakage-contaminated estimator) is the characteristic signature of a classifier label-noise systematic rather than a primordial dipole; it is consistent with the leakage-channel interpretation of the canonical $+3.64\sigma$, not with a $+4.31\sigma$ standalone detection. The primary real-space null at $+0.55\sigma$ (equivariant HC, $p_{\text{eq}} > 0.6$, $N = 949,584$) remains the load-bearing result; the WLS clean-template statistic is diagnostic and must not be read as a physical 1.7% exclusion.

c. 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,

² The “monopole-preserving” Catalog-C-full $+4.31\sigma$ is the single-mode `pymaster` pseudo- $C_\ell^{(\ell=1)}$ evaluated on the equivariant Catalog C full-footprint $f_{\text{CW}}^{\text{eq}}$ field on the FS-C support *before* mean-pixel subtraction and *before* MASTER mode-coupling deconvolution; the null is the same per-pixel random-label permutation null used for the canonical $+3.64\sigma$ result (Sec. D, $N_{\text{MC}} = 500$, seed 42). It is therefore the *same* estimator family as the canonical $+3.64\sigma$, with the additional choice of not subtracting the spatially-uniform $f_{\text{CW}}^{\text{eq}} = 0.4974$ monopole prior to single-mode pseudo- C_ℓ measurement. The $+4.31\sigma$ vs. the primary $+0.55\sigma$ real-space dipole are therefore *not directly comparable* — they are different estimators measuring different observables on the same sample. The $\sim 10\times$ amplitude gap is sourced by exactly the monopole-mask leakage channel quantified in Sec. IV D (where the monopole-only null reproduces 99.32% of the pre-MASTER C_1 power on the FS-C support): without monopole subtraction the leakage contribution dominates; with monopole subtraction the canonical estimator retains the $+3.64\sigma$ residual, and under MASTER deconvolution the monopole-only null reproduces only $\sim 12\%$ of the observed power, leaving a $+4.84\sigma$ non-null residual (Sec. IV D, Table VII) — the leakage channel accounts for the bulk of the pre-MASTER power but not for the post-subtraction or post-MASTER residuals, which are systematics-attributed. The HC-cut robustness test below uses the same monopole-preserving variant on the high-confidence subsamples for a like-for-like comparison; it is the *cut-dependence within this estimator*, not consistency with the real-space dipole, that is the substantive finding.

with no coherent large-scale pattern that would bias the dipole analysis.

d. 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 always reflects the current version-stamped content (primary sample HC-broad $N = 949,584$, $p_{\text{eq}} > 0.6$, spirals; real-space dipole $+0.55\sigma$, one-sided upper-tail $p = 0.265$). 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.

- **Catalog:** <https://huggingface.co/datasets/bamfai/galaxy-chirality-catalog> (CC-BY-4.0, Parquet). The current public source is pinned to revision `a21eb596fd10edb9af9e7a1bcefb04f87327a724`, 952,115,239 bytes, and SHA-256 `e8525ba5c98576f6...3563`. It does *not* contain the previously described `qc_flip_identity_violator` column; users must not rely on that nonexistent public field. The `v1.0.244` local release candidate instead provides a

science-facing file with raw/reconstructed columns absent and Boolean `raw_flip_qc_unsafe`, plus a separate provenance-only quarantine containing all 249,066 catalog-wide violators (59,515 in HC), each marked `do_not_use_for_science=True`. Safe filters are `primary_hc == True` for the declared hard-label primary and `primary_hc == True` and `raw_flip_qc_unsafe == False` for the strict diagnostic. The machine contract is [pipelines/p2_chirality/apjs_release_schema_v1_0_244.json](#); the human dictionary, builder, tests, and minimal primary-null reproducer are [pipelines/p2_chirality/CATALOG_SCHEMA.md](#), [pipelines/p2_chirality/build_apjs_release_v1_0_244.py](#), [pipelines/p2_chirality/tests/test_apjs_release_v1_0_244.py](#), and [pipelines/p2_chirality/reproduce_p4_primary_null_v1_0_244.py](#). The generated Parquet payloads remain local until a content-addressed public upload is completed. No immutable archive or DOI exists yet; that gate remains open and no placeholder DOI is claimed.

- **Model:** <https://huggingface.co/bamfai/galaxy-chirality-v2> (ViT-Small encoder + classification head, PyTorch checkpoint).
- **Code:** <https://github.com/Hubify-Projects/bigbounce>. Training, inference, equivariant post-processing, bias hardening suite, and dipole analysis scripts.

The released catalog labels carry a measured spatially-uniform CW-bias residual of 0.26% (9.5σ) attributed to GZ1 human-handedness training bias propagating through CE-ResNet pseudo-labels and the present ViT-Small classifier. Local re-normalization of the per-region monopole is recommended for precision parity tests. 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. IV A calibration caveat). The independent GZ1 CW/CCW agreement on the 234,282-galaxy cross-match is 69.91% (Cohen’s $\kappa = 0.40$). The provenance-audit artifacts ([pipelines/p2_chirality/outputs/canonical_provenance/c3_wp_invariance_fsky.json](#) and [pipelines/p2_chirality/outputs/canonical_provenance/c6_depth_stratified_null.json](#)) are archived in the repository.

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.

Software: Astropy [33], HEALPix/healpy [36, 37], NumPy [38], pandas [39], PyTorch [40], timm [41], NaMaster/pymaster.

AI-assisted methodology: This work was conducted using an agentic AI pipeline built on Anthropic Claude (Opus 4 family, 2026 releases), with OpenAI GPT-

5/o3, xAI Grok-4, and Google Gemini 2.5 used as cross-checking and adversarial internal-review models — a multi-model system for literature review, code development, analysis, and adversarial internal peer review — operated under the author’s direction. Every quantitative result reported here was verified against committed computational artifacts (scripts, data products, and their checksums are linked throughout and versioned in the public repository), and the full audit trail is public, so that any claim can be re-derived from source. The author designed the study, made all scientific judgments, and takes full responsibility for the content; the AI pipeline is a reproducibility and verification instrument, not an author.

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

-
- [1] 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.
 - [2] 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.
 - [3] 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.
 - [4] 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.
 - [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] C. J. Lintott, K. Schawinski, A. Slosar *et al.*, “Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey,” *Mon. Not. R. Astron. Soc.* **389**, 1179 (2008), arXiv:0804.4483.
 - [11] K. Land, A. Slosar, C. Lintott *et al.*, “Galaxy Zoo: the large-scale spin statistics of spiral galaxies in SDSS,” *Mon. Not. R. Astron. Soc.* **388**, 1686 (2008), arXiv:0803.3247.
 - [12] C. Lintott, K. Schawinski, S. Bamford *et al.*, “Galaxy Zoo 1: data release of morphological classifications for nearly 900,000 galaxies,” *Mon. Not. R. Astron. Soc.* **410**, 166 (2011), arXiv:1007.3265.
 - [13] 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].
 - [14] E. Gross and O. Vitells, “Trial factors for the look elsewhere effect in high energy physics,” *Eur. Phys. J. C* **70**, 525 (2010), arXiv:1005.1891.
 - [15] D. R. Davis and W. B. Hayes, “SpArcFiRe: scalable automated detection of spiral galaxy arm segments,” *Astrophys. J.* **790**, 87 (2014), arXiv:1402.1910, DOI:10.1088/0004-637X/790/2/87.
 - [16] P. Motloch, H.-R. Yu, U.-L. Pen, and Y. Xie, “An observed correlation between galaxy spins and initial conditions,” *Nature Astron.* **5**, 283 (2021), arXiv:2003.04800.
 - [17] A. Lue, L.-M. Wang, and M. Kamionkowski, “Cosmological signature of new parity-violating interactions,” *Phys. Rev. Lett.* **83**, 1506 (1999), arXiv:astro-ph/9812088.
 - [18] G. Cabass, M. M. Ivanov, and O. H. E. Philcox, “Colliders and ghosts: Constraining inflation with the parity-odd galaxy four-point function,” *Phys. Rev. D* **107**, 023523 (2023), arXiv:2210.16320.
 - [19] O. H. E. Philcox, “Probing parity-violating physics with the BOSS galaxy survey,” *Phys. Rev. D* **106**, 063501 (2022), arXiv:2206.04227.
 - [20] J. R. Eskilt and E. Komatsu, “Improved constraints on cosmic birefringence from the WMAP and Planck cosmic microwave background polarization data,” *Phys. Rev. D* **106**, 063503 (2022), arXiv:2205.13962.
 - [21] J. R. Eskilt *et al.* (Cosmoglobe Collaboration), “Cosmoglobe DR1 results. II. Constraints on isotropic cosmic birefringence from reprocessed WMAP and Planck LFI data,” *Astron. Astrophys.* **679**, A144 (2023), arXiv:2305.02268.
 - [22] R. Jackiw and S.-Y. Pi, “Chern-Simons modification of general relativity,” *Phys. Rev. D* **68**, 104012 (2003), arXiv:gr-qc/0308071.
 - [23] J. Hou, Z. Slepian, and R. N. Cahn, “Measurement of parity-odd modes in the large-scale 4-point correlation function of SDSS BOSS DR12 CMASS and LOWZ

- galaxies,” *Mon. Not. R. Astron. Soc.* **522**, 5701 (2023), arXiv:2206.03625.
- [24] R. N. Cahn, Z. Slepian, and J. Hou, “A test for cosmological parity violation using the 3D distribution of galaxies,” *Phys. Rev. Lett.* **130**, 201002 (2023), arXiv:2110.12004.
- [25] E. Komatsu, “New physics from the polarized light of the cosmic microwave background,” *Nature Rev. Phys.* **4**, 452 (2022), arXiv:2202.13919.
- [26] W. B. Hayes, D. Davis, and P. Silva, “On the nature and correction of the spurious winding bias in Galaxy Zoo 1,” *Mon. Not. R. Astron. Soc.* **466**, 3928 (2017), arXiv:1701.06587.
- [27] S. P. Bamford, R. C. Nichol, I. K. Baldry *et al.*, “Galaxy Zoo: the dependence of morphology and colour on environment,” *Mon. Not. R. Astron. Soc.* **393**, 1324 (2009), arXiv:0805.2612.
- [28] R. E. Hart, S. P. Bamford, K. W. Willett *et al.*, “Galaxy Zoo: comparing the demographics of spiral arm number and a new method for correcting redshift bias,” *Mon. Not. R. Astron. Soc.* **461**, 3663 (2016), arXiv:1607.01019.
- [29] M. Walmsley, C. Lintott, T. G eron *et al.*, “Galaxy Zoo DECaLS: detailed visual morphology measurements from volunteers and deep learning for 314,000 galaxies,” *Mon. Not. R. Astron. Soc.* **509**, 3966 (2022), arXiv:2102.08414.
- [30] H.-R. Yu, P. Motloch, U.-L. Pen *et al.*, “Probing primordial chirality with galaxy spins,” *Phys. Rev. Lett.* **124**, 101302 (2020), arXiv:1904.01029.
- [31] DESI Collaboration, A. Aghamousa, J. Aguilar *et al.*, “The DESI Experiment Part I: Science, Targeting, and Survey Design,” arXiv:1611.00036 (2016).
- [32]  . Ive i , S. M. Kahn, J. A. Tyson *et al.*, “LSST: From science drivers to reference design and anticipated data products,” *Astrophys. J.* **873**, 111 (2019), DOI 10.3847/1538-4357/ab042c.
- [33] 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.
- [34] D. Alonso, J. Sanchez, and A. Slosar, “A unified pseudo- C_ℓ framework,” *Mon. Not. R. Astron. Soc.* **484**, 4127 (2019), arXiv:1809.09603.
- [35] E. Hivon, K. M. G rski, C. B. Netterfield *et al.*, “MASTER of the cosmic microwave background anisotropy power spectrum,” *Astrophys. J.* **567**, 2 (2002), arXiv:astro-ph/0105302.
- [36] K. M. G rski, 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.
- [37] A. Zonca, L. Singer, D. Lenz *et al.*, *J. Open Source Softw.* **4**, 1298 (2019).
- [38] C. R. Harris, K. J. Millman, S. J. van der Walt *et al.*, *Nature* **585**, 357 (2020).
- [39] W. McKinney, in *Proc. 9th Python in Science Conf.*, edited by S. van der Walt and J. Millman (2010), pp. 56–61.
- [40] 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.
- [41] R. Wightman, *PyTorch Image Models* (2019), <https://github.com/rwightman/pytorch-image-models>.