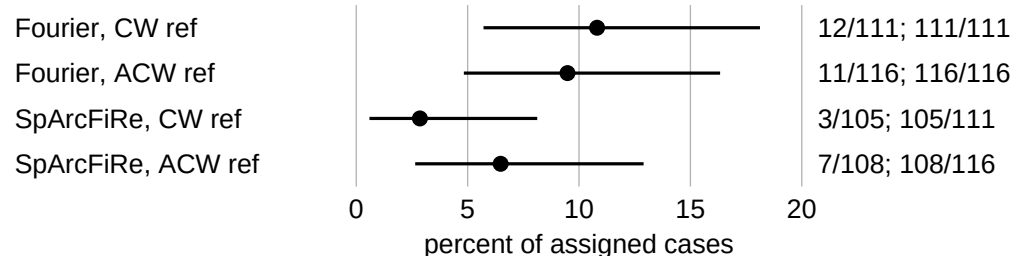


1. Agreement with reference labels, and coverage

Coverage. Of 49,211 attempted targets, the Fourier method assigned a sign to 49,007 (204 have no retained tensor). SpArcFiRe resolved 30,910 and left 18,301 unresolved. On the 30,910 targets with both labels the methods agree for 18,791 and disagree for 12,119. A sign assignment is not a quality guarantee.

Reference comparison. For cases with a strong Galaxy Zoo reference label, the chart shows how often each method gave the opposite direction among the cases it assigned. Fourier assigned every such case; SpArcFiRe abstained on some. The fractions are therefore conditional on different assigned sets and are not a fair whole-population accuracy ranking.

Limits. Reference labels are human classifications, not physical truth. DEV and EVAL were both exposed during development: neither is an untouched or representative validation set, and they are never pooled. The separate CzSL expert cohort is historical context only; it lies outside the current Fourier parent.



EVAL split: reference disagreement (opposite direction / assigned cases) with nominal Clopper-Pearson 95% intervals for these conditional counts. Notes give opposite/assigned; assigned/reference cases. CW ref rows count CW-to-ACW disagreement, ACW ref rows ACW-to-CW.

DEV split	opp./assigned	fraction	CP95	assigned/ref
Fourier, CW ref	11/69	15.9%	8.2-26.7%	69/69
Fourier, ACW ref	9/67	13.4%	6.3-24.0%	67/67
SpArcFiRe, CW ref	2/64	3.1%	0.4-10.8%	64/69
SpArcFiRe, ACW ref	2/64	3.1%	0.4-10.8%	64/67

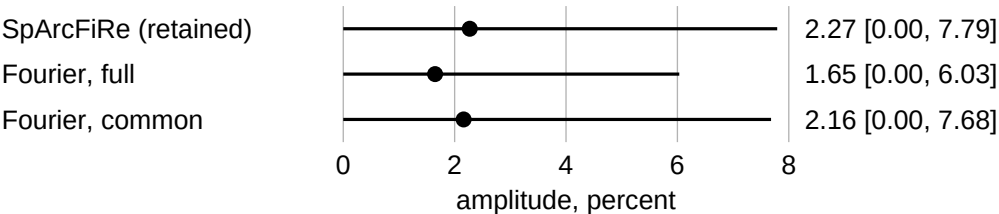
DEV split, shown separately and never pooled with EVAL. Intervals exclude reference error, the vote threshold, selection into the reference set and any transfer to the science sample. These are reference disagreements, not classifier error rates.

2. Directional tests: no significant pattern detected

Amplitude. The dipole amplitude is $2|b|$ in $A = 2p - 1$: a one-sided cosine coefficient, not a hemisphere contrast or a pooled imbalance. Fourier labels give 1.65% on full coverage (approximate 95% Wald envelope 0.00 to 6.03%) and 2.16% on the common set (0.00 to 7.68%). The unchanged SpArcFiRe result is 2.27% (0.00 to 7.79%).

Tests. Four new tests were declared before any draw: a dipole likelihood ratio and a maximum hemisphere contrast over 648 axes, each on full and common coverage, with 1,024 label shuffles per coverage set. The largest observed contrasts in A are 2.04 and 2.49 percentage points. All four Bonferroni-adjusted p -values are 1.000. The earlier SpArcFiRe test ($p = 0.684$) is separate and not part of this family.

Meaning. Under a null that fixes positions and the CW count and treats labels as exchangeable, no statistically significant directional winding pattern is detected. This is not a physical-spin or cosmic-parity result, and absence of significance does not prove exact symmetry.



Fitted dipole amplitude with approximate 95% Wald envelopes. The envelopes are asymptotic, have no guaranteed coverage and exclude classification and selection uncertainty. Best-fit axes are not established and are not shown. The SpArcFiRe row is the retained earlier result, not recomputed.

New test	observed	null >= obs	p (plus-one)	Bonferroni-4
Dipole LR, full	1.78	662/1024	0.647	1.000
Dipole LR, common	2.25	519/1024	0.507	1.000
Max contrast, full	2.04 pp	539/1024	0.527	1.000
Max contrast, common	2.49 pp	568/1024	0.555	1.000

Conditional plus-one permutation p -values with the four-test Bonferroni correction. The maximum over axes is recomputed in every shuffle, which accounts for that grid search; the likelihood ratio refits a continuous direction. The correction covers these four tests only. Full and common coverage overlap and the two algorithms are correlated. Spatial selection and classification effects are not removed. Monte Carlo intervals are in the appendix. pp = percentage points of A .

3. Where the two methods disagree, and scope

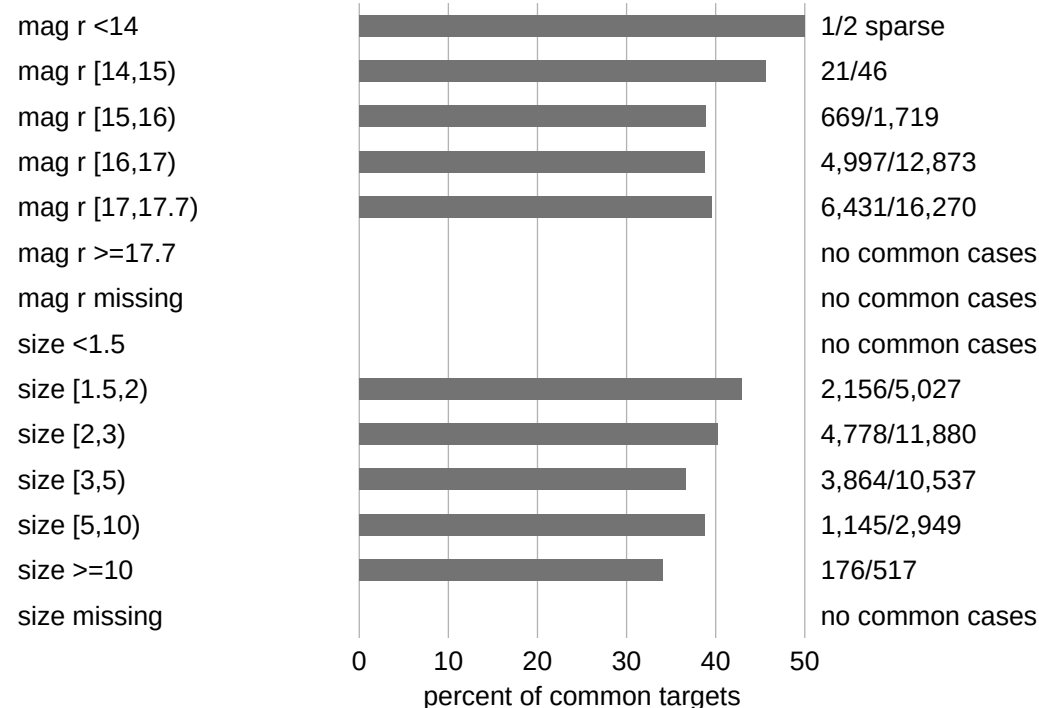
Disagreement by brightness and size. The chart gives the share of common targets on which Fourier and SpArcFiRe labels differ, in bins frozen before any winding outcome was read. These are descriptive proportions: no intervals or tests were computed and no cause is inferred from any bin. Empty bins are marked; bins with fewer than 30 common targets are flagged sparse.

Takeaway. The new reference comparison and the Fourier directional tests are complete. No statistically significant directional winding pattern is detected with either method under the conditional shuffle null, and the two methods give different labels for 12,119 of 30,910 common targets.

Scope. Nothing was reclassified and the historical SpArcFiRe test was not rerun. The optional group and cluster audit was not run in this phase; group results are unchanged.

Axis-dependent maps remain descriptive, not evidence of a preferred direction. No calibration-adjusted population asymmetry was calculated.

Checks. Independent numerical checks reproduced the tables (21 checks) and certified 2,050 saved optima with no new optimisation or draw. They confirm arithmetic, not the scientific assumptions.



Disagreement share (disagree / common, percent) by frozen r-magnitude and size bins; notes give disagree/common counts. Declination, score-magnitude and exposure-count bins and all denominators are in the appendix. Method context only, not validation of this pipeline: arxiv.org/html/2404.06617v2 and arxiv.org/abs/1610.07060.