Draft version July 16, 2026
Typeset using LaTeX **twocolumn** style in AASTeX7.0.1

Public-ID Recovery for a Historical DESI DR1 Anomaly List: 170 High-Coordinate-Consistency Core and 11 Lower-Confidence Positional Associations

Houston Golden[1]

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

(Dated: July 16, 2026, 17:57 PDT — v3.2.0-r10)

ABSTRACT

Anomaly searches become reusable catalogs only when rows trace public archive objects and the recovery from a declared input list is reconstructable. The primary deliverable of this work is that reusable, memory-bounded join and provenance/validation machinery itself; the 181-row DESI DR1 catalog reported here is its first instance, produced by applying the machinery to one frozen historical input list. We recover public identifiers for a frozen historical DESI anomaly list whose legacy identifiers mixed public-looking values with internal hashes. We document the available immutable clustering and BigAE lineage but do not reconstruct the unavailable production normalization or physical-feature sensitivity. Starting from 190,015 positional clusters containing a DESI member, we stream all 28,425,963 rows of the public DESI Data Release 1 (DR1) pixel-based redshift catalog in 200,000-row chunks, read 18 columns, and query a spherical-coordinate k-d tree. The predeclared parent selection is a $1''$ match to a main-survey row carrying at least one LRG, ELG, QSO, BGS_ANY, or MWS_ANY `DESI_TARGET` bit. It reproduces exactly 20,299,155 eligible DESI rows and 2,468 positional matches. Requiring the global primary redshift row leaves 2,448; requiring `ZWARN`=0 leaves 181 unique public `TARGETID` associations. These 181 rows form the v3.2.0-r2 data release audited in this v3.2.0-r10 manuscript: 170 high-coordinate-consistency core associations at or below $0.1''$ and 11 lower-confidence positional associations between $0.1''$ and $1''$. They comprise 157 Redrock GALAXY, 23 QSO, and one STAR classifications; spectral type was not a selection cut. A quality-tier column exposes this contract without silently changing the declared $1''$ list; neither tier is a secure object-identity or purity claim. Every released row and all 18 carried DESI fields were re-read from their recorded FITS row and compared exactly. The release includes the Parquet catalog, dictionary, code, waterfall, checksums, and provenance. A separately released, explicitly secondary table preserves all 2,267 warning-bearing global-primary rows without admitting them to the primary catalog. Sixteen deterministic 60–$120''$ local shifts of all 190,015 cluster positions yield 86.7 ± 14.4 parent and 76.2 ± 13.3 warning-free-primary associations within $1''$, compared with 2,468 and 181 observed. The corresponding shifted 0.1–$1''$ annulus contains 75.6 ± 13.0 warning-free-primary associations versus 11 observed, so the tail is not treated as secure candidate-level identity. The sub-$0.1''$ portion of the within-$1''$ excess is the expected self-recovery of the seed DESI members whose own coordinates define the cluster centroids; it verifies the recovery end to end rather than providing independent association evidence, and the local-shift control is informative only for the 0.1–$1''$ tail. These are reproducible follow-up lists, not validated detections or unbiased samples for anomaly-rate inference.

Keywords: catalogs — surveys — methods: data analysis — methods: statistical — galaxies: spectra — quasars: general

1. INTRODUCTION

Spectroscopic surveys make it possible to search for unusual objects at scales that cannot be inspected manually. Reconstruction models, latent-variable models, and other outlier methods have consequently been applied to SDSS and DESI spectra [2, 3, 4]. Their scientific value, however, depends on a distinction that is easy to blur: an unusual model score is a candidate-selection signal, whereas an archive-resolvable object with auditable quality metadata is a usable catalog row. Neither is, by itself, a validated astrophysical detection.

This paper addresses that distinction through a structural recovery. The earlier public *BigBounce Anomaly Catalog* (bamfai/bigbounce-anomaly-catalog) multi-survey product contained DESI sky positions, anomaly scores, and residual summaries, but most legacy identifiers were internal hashes rather than public DESI `TARGETID`s. A catalog whose coordinates cannot be joined deterministically to public spectra is difficult to verify, extend, or cite object by object. The earlier product also mixed survey-specific selection functions. Treating its aggregate row count as a homogeneous scientific sample would therefore be misleading even if every row were numerically correct.

We replace that scope with one falsifiable deliverable: a DESI DR1 science-target candidate catalog whose public identifiers and carried metadata can be rejoined exactly. The recovery uses the historical anomaly stream only as a source of coordinates and score metadata. Public identity, targeting information, primary-row status, redshift metadata, and fit warnings all come from the current public DESI DR1 redshift catalog [1]. The result is a small quality-controlled follow-up set rather than a claim about the prevalence, novelty, or physical origin of anomalous spectra.

The contributions are:

1. a memory-bounded, checkpointed join of 190,015 DESI-containing anomaly clusters to the 28.4 million-row DR1 redshift catalog;
2. an exact reproduction of the previously quoted 2,468-row main-survey science-bit cohort;
3. a declared quality waterfall yielding 181 warning-free global-primary candidates;
4. deterministic local-shift association curves for the parent and strict cohorts;
5. an exact descriptive comparison with all 2,267 warning-bearing global-primary rows;
6. a public-ID release with exact field-by-field source-row validation; and
7. a checksum-bound v3.2.0-r7 submission bundle that defines the primary, auxiliary, AAS-table, science-control, and coordinate-lineage relationships without relabeling their frozen component versions.

The primary, reusable deliverable is this machinery itself — the memory-bounded checkpointed join together with the declared quality waterfall, field-by-field validation, and checksum-bound provenance framework — and it is designed to be re-run on any future frozen anomaly or coordinate list. The 181-row DESI DR1 catalog reported below is the first instance produced by applying that framework to a single historical input list, not the endpoint of the contribution.

Section 2 defines the two inputs. Section 3 specifies the join, selection, and duplicate rules. Section 4 describes the released cohort without interpreting it as a population measurement. Sections 5 and 6 make the selection biases and reproducibility checks explicit.

2. INPUTS AND SCOPE

2.1. *Historical anomaly-cluster substrate*

The starting cluster table contains 378,480 coordinate clusters produced from eight survey inputs by a fixed $5''$ positional grouping. The immutable dataset documentation records this as "8-way $5''$ positional deduplication" [5]; the preserved executable lineage implements a friends-of-friends transitive closure with spherical neighbor pairs and union–find (pipelines/p3_anomaly_engine/pathc_positional_dedup.py). Exactly 190,015 have a survey-membership string containing `desi_dr1`; 190,005 are DESI-only clusters, eight also contain an SDSS member, and two also contain a LAMOST member. Those historical cross-survey labels are retained only as provenance columns. They do not enter the v3.2.0-r2 recovery and no non-DESI table is included in the new release.[2]

The accompanying DESI anomaly table has 195,829 rows with an internal identifier, right ascension, declination, a scalar score S, the band producing the largest residual, and three residual summaries. It is the frozen $S > 5$ canonical-S selection from the DESI-trained BigAE stream, published under CC BY 4.0 at immutable dataset commit `cdaaa03a...5b39a8` [5]. The original training lineage preserved locally identifies a 496-to-128 BigAE evaluated as the five-seed mean reconstruction MSE. That evidence does not recover the production absolute-scale normalization, spectral resampling, original object-to-spectrum mapping, or a physical-feature injection/sensitivity function. Consequently the released per-object S values cannot be numerically reproduced from the currently public

[2] Version-tag key: the internal release-candidate tags label distinct frozen components of one program and are retained for provenance rather than deleted or relabeled. v3.2.0-r2 is the audited primary data release; v3.2.0-r5 the secondary warned-primary auxiliary table; v3.2.0-r7 the checksum-bound submission bundle; and v3.2.0-r10 this manuscript. A later stage does not relabel the earlier component versions it supersedes, so a single published identifier can carry several component tags.

spectra and are carried only as frozen, uncalibrated ranking metadata; no physical feature or detection efficiency is assigned to them. This boundary is recorded in the held-out reinference and reconciliation evidence (pipelines/p3_anomaly_engine/outputs/dp3_15_heldout_reinference.json; pipelines/p3_anomaly_engine/outputs/desi_injection_recovery/RECONCILIATION_history_2026-06-28.json). The preserved upstream DESI inference code copies each anomaly row's `ra`/`dec` directly from the coadd FIBERMAP `TARGET_RA`/`TARGET_DEC` fields; a separate implementation also reads `MEAN_FIBER_RA`/`MEAN_FIBER_DEC` under distinct names. The positional-deduplication code then computes cluster `ra_mean`/`dec_mean` from the member coordinates. The exact source files and SHA-256 values are bound by the v3.2.0-r7 coordinate-lineage audit (pipelines/p3_anomaly_engine/apjs_submission_bundle_v3.2.0-r7/coordinate_lineage/COORDINATE_LINEAGE_AUDIT.json). This establishes coordinate-field lineage to public DESI target coordinates; it does not recover the unavailable production object-to-spectrum mapping or prove object identity for a positional association. The cluster positions used here are the frozen ICRS means of those historical positional groups, with no coordinate transformation in this recovery beyond the spherical unit-vector representation used for matching. The legacy identifier is not treated as a public key. It mixes positive public-looking values with internal hashed values; 169,611 of the 195,829 values are negative. Within the released 181-row cohort, ten legacy identifiers are negative. This behavior is preserved as audit history in `original_internal_tid`, while the recovered public `targetid` is the only archive join key recommended for users.

When a cluster contains more than one DESI anomaly member, its canonical historical member is the one with the largest original score; ties are resolved by the smaller original row number. The score and residual fields are carried forward verbatim as ranking metadata. They are not recomputed in this work, and they are not interpreted as calibrated probabilities, physical significances, or novelty measures.

Exactly one released cluster has more than one historical anomaly member: `P3-DESI-000030` (`cluster_id`=71390, `n_detections`=2). Its recovered public target is 0.979656″ from the cluster coordinate but 1.979009″ from the canonical original DESI member. The public 1″ selection is explicitly the target-to-cluster separation, not a target-to-original-member cut. The release stores both quantities as `match_separation_arcsec` and `original_member_separation_arcsec`. An explicit alternative rule, `original_member_separation_arcsec`$\leq$ 1.0, retains 180 of 181 rows and removes only `P3-DESI-000030`, whose original-member separation is 1.979009″. The r6 machine-readable sensitivity sidecar publishes this Boolean expression and result for every row; the alternative is a sensitivity analysis, not a retroactive change to the frozen release.

2.2. *Public DESI DR1 redshift catalog*

The public identity substrate is the `ZCATALOG` binary-table HDU (extension 1) of `zall-pix-iron.fits` from DESI DR1. The local file has 28,425,963 rows and 22,371,272,640 bytes. Its SHA-256 is `2d95ad99361039b556c402b49e0e7c84df5f00106dc5731d44476a58b128b49b`, equal to the value in DESI's current official checksum list. The FITS table contains repeat observations, survey and program labels, targeting bitmasks, Redrock redshift products, and flags defining primary rows. We read only the fields listed in Table 1; no 21 GB in-memory table is formed.

3. RECOVERY AND SELECTION METHOD

3.1. *Memory-bounded positional join*

All cluster coordinates are converted from (α, δ) to three-dimensional unit vectors,

$$\boldsymbol{x} = (\cos\delta\cos\alpha, \cos\delta\sin\alpha, \sin\delta), \tag{1}$$

and indexed once with a `cKDTree`. The FITS table is then traversed in blocks of 200,000 rows using `fitsio`. Numeric arrays are converted from FITS big-endian storage to native byte order before Parquet serialization; targeting masks are evaluated as unsigned 64-bit integers so bits 60 and 61 cannot be lost to signed shifts. Within each block we first apply the survey and target bit mask, convert the remaining coordinates to unit vectors, and query the nearest cluster. Chord distances are converted back to great-circle arcseconds. Only matches within the predeclared 1″ radius are retained.

Each completed block is written atomically to a small Parquet part. A JSON checkpoint records the input signature, completed block starts, cumulative row counts, and the last completed row; a fsynced JSON-lines progress file records the same counts with timestamps. Restarting with the same signature skips completed parts. A changed FITS path, size, column set, radius, chunk size, or cluster count invalidates the checkpoint rather than silently mixing runs. The full scan took 97.1 s on the audit machine and retained only 2,468 match rows. The release does not pin that machine's hardware; the streaming design bounds the resident working set to a single 200,000-row chunk rather than

Table 1. DESI fields read from each streamed FITS chunk

Group	Fields	Role in this work
Identity/position	TARGETID, TARGET_RA, TARGET_DEC	public rejoin and angular match
Observing context	SURVEY, PROGRAM	main-survey selection and description
Targeting	DESI_TARGET, BGS_TARGET, MWS_TARGET, SCND_TARGET	science-bit selection and raw metadata
Redshift fit	Z, ZWARN, SPECTYPE, DELTACHI2	quality gate and descriptive metadata
Coadd status	COADD_FIBERSTATUS	descriptive quality cross-check
Primary flags	MAIN_NSPEC, MAIN_PRIMARY, ZCAT_NSPEC, ZCAT_PRIMARY	repeat accounting and final primary gate

Note—Column names follow the public DESI zcatalog. Only SURVEY, DESI_TARGET, ZCAT_PRIMARY, and ZWARN enter the final Boolean selection; coordinates enter the positional join. SPECTYPE and Z are not selection cuts.

the full catalog, so the scan reproduces on commodity hardware such as the Apple M5 (10-core, 24 GB RAM) reference machine used to reproduce it here. Runtime is reported for reproducibility, not as a cross-platform benchmark.

The published release was generated by a fresh replay of all 28,425,963 FITS rows into new output and checkpoint directories using only the three named input files. The 143 checkpoint parts used by the independent validation were outputs of that same replay, not parts inherited from an older catalog run. This establishes a clean public-input replay at the data-product level; it is not a claim of a hermetic operating-system image.

The nearest-cluster assignment is also checked against the complete set of neighbors within $1''$. Across the 190,015 DESI-containing clusters, the minimum cluster–cluster nearest- neighbor separation is $5.02294''$; none is within $2''$ of another cluster. Direct all-neighbor queries return exactly 2,468 parent pairs and 181 strict pairs, with zero public rows matched to multiple clusters and every stored nearest cluster present in its all-neighbor set. Thus nearest-only and all-neighbor definitions are identical for these exact frozen inputs. The full nearest-neighbor quantiles and direct pair counts are emitted by the independent validator.

3.2. *Science-target parent cohort*

The parent cohort uses the exact historical bitmask definition so that the earlier 2,468 count can be tested. A row is eligible when

$$\begin{aligned}&\texttt{SURVEY} = \texttt{main},\\ &\texttt{DESI\_TARGET}\ \&\ \left(2^0\,|\,2^1\,|\,2^2\,|\,2^{60}\,|\,2^{61}\right) \neq 0,\end{aligned} \tag{2}$$

where bits 0, 1, 2, 60, and 61 denote LRG, ELG, QSO, BGS_ANY, and MWS_ANY, respectively. The exact unsigned 64-bit union mask is 0x3000000000000007. These definitions follow the authoritative DESI target-mask data model [6]. These classes are not mutually exclusive. The streamed scan finds 20,299,155 eligible rows, exactly reproducing the historical denominator. Their nearest-neighbor join yields 2,468 rows within $1''$, exactly reproducing the historical match count. This agreement tests the coordinate transform, endian-correct bitmask evaluation, row traversal, and radius together.

3.3. *Quality cohort and duplicate rules*

The public redshift catalog may contain multiple rows associated with one target. Within each reported cohort, we first apply its predicates and only then deduplicate. For the strict cohort we therefore require ZCAT_PRIMARY=true, the global primary redshift-row flag, and then require ZWARN=0. This latter gate is intentionally conservative: it removes any Redrock fit carrying a warning bit. We do not require a particular SPECTYPE, redshift range, or minimum DELTACHI2; imposing such post-hoc cuts would redefine which anomaly morphologies survive.

During the positional join, each eligible FITS row is assigned to its nearest anomaly cluster only. After applying a cohort's predicates, duplicate cluster matches are ordered by smaller angular separation, larger DELTACHI2, smaller TARGETID, and smaller FITS row. A second pass uses the same ordering to enforce one row per public TARGETID. No tie-breaking is needed in the final cohort, but publishing the rule makes reruns deterministic. The final 181 rows have unique candidate IDs, cluster IDs, public TARGETIDs, and FITS row numbers.

The executable order audit finds 2,468 parent rows, 2,448 rows satisfying ZCAT_PRIMARY, and 181 satisfying both ZCAT_PRIMARY and ZWARN=0 before deduplication. All three cohorts already have unique cluster and TARGETID keys. Consequently, a deliberately unsafe counterfactual that deduplicates the broad parent before applying the strict predicates also returns the same

181 identities and loses zero valid rows. That empirical equality checks this release; it does not replace the declared predicate-before-deduplication rule.

3.4. *Match-quality tier*

The 1″ radius was fixed before inspecting the quality cohort and is retained for exact historical count reproduction. The primary release contract therefore retains all 181 positional associations but divides them into a 170-row high-coordinate-consistency core at or below 0.1″ and 11 lower-confidence associations between 0.1″ and 1″, including eight above 0.5″ and five above 0.75″; the maximum is 0.9906″. Rather than silently deleting these valid-by-definition matches, we publish `match_quality_tier` with values `coordinate_consistent_le_0p1arcsec` and `positional_match_gt_0p1_le_1arcsec`. Users can reproduce either the declared catalog or the 170-row high-coordinate-consistency subset. The full 181-row list is intentionally not purity-controlled, and neither tier is a secure candidate-level identity claim. Unlike the 1″ join radius, the 0.1″ boundary was introduced post hoc as a descriptive quality tier; it is not a predeclared selection cut and does not alter the 181-row catalog membership.

For direct scrutiny, the 11 positional-tail identifiers and separations are listed in Table 2; the continuous values remain authoritative in the full catalog.

Table 2. Released candidates in the 0.1–1″ positional tail

Candidate	Separation (arcsec)
P3-DESI-000003	0.273888
P3-DESI-000005	0.990574
P3-DESI-000017	0.130508
P3-DESI-000022	0.785982
P3-DESI-000030	0.979656
P3-DESI-000032	0.754671
P3-DESI-000039	0.221264
P3-DESI-000045	0.621958
P3-DESI-000053	0.656096
P3-DESI-000140	0.969344
P3-DESI-000141	0.512381

Note—Separations are target-to-cluster great-circle arcseconds. All satisfy the predeclared 1″ catalog definition, but these 11 rows are lower-confidence positional associations; the full 1″ list is not purity-controlled and no row is secured by position alone.

The target-to-canonical-original-member separations have minimum, median, and 90th percentile zero, 95th percentile 0.2213″, 99th percentile 0.9740″, and maximum 1.9790″. Eleven exceed 0.1″, eight exceed 0.5″, five exceed 0.75″, one exceeds 1″, and none exceeds 2″. A second full scan of all 20,299,155 eligible public parent rows finds no competing eligible row within 1″ of either the cluster or canonical-original-member coordinate for any released candidate. One case, `P3-DESI-000039`, has a second warning-free primary eligible row at 1.0498″ from both centers, hence within 2″ but outside the declared radius; the released row is at 0.2213″. The machine-readable audit publishes every row found within both radii.

3.5. *Deterministic local-shift association control*

Catalog-key uniqueness does not estimate false association. We therefore repeated the spherical join after shifting every one of the 190,015 cluster positions along great circles by 60 and 120″ at each of eight fixed position angles. These 16 realizations retain the local sky density and footprint on arcminute scales while moving well beyond the declared radius. The 28,425,963-row FITS table was streamed once against the observed and shifted trees. For each realization we retained the nearest eligible target to each shifted cluster and counted associated cluster positions over the fixed radius grid shown in Figure 1.

For the 20,299,155-row science-bit parent, 2,468 observed cluster positions have an association within 1″, versus a shifted mean of 86.69 (sample standard deviation 14.42; range 66–109). At 0.1″, the comparison is 2,456 observed versus a shifted mean of 0.75. Applying the global-primary and `ZWARN`=0 predicates to the background targets gives 181 observed versus 76.19 (13.30; 61–103) within 1″, and 170 observed versus 0.625 within 0.1″. The observed strict 0.1–1″ annulus contains 11 rows, whereas the shifted annulus mean is 75.56 (13.01; 61–101). The sub-0.1″ comparison, however, is not an independent association test. The 170 core rows are single-member DESI clusters (Section 2.1) whose centroid is the seed member's own `TARGET_RA/TARGET_DEC`, and the public catalog matched against still contains that seed row (Section 3.1), so recovering it at $\ll 0.1''$ is the expected signature of re-identifying the seed rather than evidence of an astrophysical association. The already-reported median target-to-cluster separation of 0.00127″ (Section 4.3) and the zero minimum, median, and 90th-percentile target-to-original-member separation (Section 3.4) confirm this: for the core the recovered public object is, by construction, at essentially the member coordinate that defined the centroid. A 60–120″ shift moves the centroid off its seed member and necessarily destroys this self-recovery, which is why the shifted sub-0.1″ count collapses to 0.625. The 170-versus-0.625 gap is therefore a successful end-to-end seed self-recovery verification of the join — confirmation that the recovered public identifiers reproduce the coordinates that generated the cluster centroids — not an independent aggregate association excess. The local-shift control is informative only for the 0.1–1″ tail, where the 11 observed rows against a shifted mean of 75.56 are consistent with chance and do not acquire secure candidate-level identity from positional coincidence alone. That the observed annulus count of 11 falls below the shifted mean is itself a corollary of the core self-recovery: each core cluster's nearest-neighbor slot is consumed by its own seed member inside 0.1″, shielding those clusters from additional matches in the surrounding annulus. We retain and flag the tail for auditability rather than subtracting an estimated background or claiming a corrected purity. The shifts are correlated local controls, not an independent Poisson experiment or a selection-function model.

4. CATALOG RESULTS

4.1. *Selection waterfall*

Table 4 and Figure 2 give the full waterfall. Of the 2,468 parent matches, only 20 are removed by the global-primary requirement. The dominant reduction is `ZWARN`=0: 2,267 of the 2,448 primary rows carry at least one warning bit, leaving 181. Requiring `MAIN_PRIMARY` in addition changes nothing; requiring `COADD_FIBERSTATUS`=0 after the adopted gates also changes nothing. These equalities are reported as cross-checks, not as independent evidence that the sample is unbiased.

The excluded global-primary rows are also published as the clearly secondary v3.2.0-r5 warned-primary auxiliary table. Its 2,267 stable IDs run from `P3-DESI-WARNED-000001` through `P3-DESI-WARNED-002267`; every row carries the same source and historical-lineage fields as the primary table plus the exact integer and hexadecimal `ZWARN` mask and decoded set bits. Each row is marked SECONDARY_WARNING_BEARING_NOT_PRIMARY_NOT_PHYSICALLY_VALIDATED and `primary_catalog_member=false`. This disclosure does not weaken the warning-free primary gate, estimate its bias, or identify a physical anomaly class.

The rejected-primary warning distribution is concentrated in the three Redrock bits defined by the DESI data model [6]: LITTLE_COVERAGE (bit 1), SMALL_DELTA_CHI2 (bit 2), and POORDATA (bit 11). Their nonexclusive counts are 2,110, 1,467, and 34. The seven exact masks and counts are shown in Table 3; the 2,267 warning-bearing primaries comprise 2,194 GALAXY, 72 QSO, and one STAR Redrock classifications. These flags describe known fitting or data-quality conditions; they do not identify physical anomaly classes.

The exact auxiliary product also permits a direct descriptive comparison, rather than inference from warning counts alone. Accepted versus warning-bearing medians are respectively $z = 1.10842$ versus 0.88605, historical score 5.32442 versus 5.84182, separation 0.001269″ versus 0.001350″, and `DELTACHI2`=61.0634 versus 4.91948. Their respective interquartile ranges are [0.61636, 1.40103] versus [0.40655, 1.31527] in z; [5.15557, 5.64065] versus [5.35522, 6.66982] in historical score; [0.000833, 0.001668] versus [0.000926, 0.001685]″ in separation; and [30.9988, 235.084] versus [1.77359, 17.3439] in `DELTACHI2`. The warned table contains 2,194 GALAXY, 72 QSO, and one STAR rows, while the accepted table contains 157, 23, and one. These contrasts show that the gate is associated most directly with fit discrimination and also changes the score/redshift composition. They do not estimate selection probabilities, correct bias, or promote warning-bearing rows into the primary catalog.

The large `ZWARN` reduction is scientifically important. An anomaly-selected spectrum is more likely than an ordinary spectrum to be difficult for a standard template-fitting pipeline. Consequently, the warning-free subset is cleaner for catalog-level redshift metadata but is not a random subset of all unusual spectra. This tradeoff motivates the restricted interpretation in Section 5.

4.2. *Descriptive composition*

Redrock classifies 157 released rows as GALAXY, 23 as QSO, and one as STAR. The observing programs are 162 dark and 19 bright. Target-bit labels are overlapping: 157 rows carry ELG, 18 BGS_ANY, seven QSO, one LRG, and one MWS_ANY; a row can contribute to more than one count. These values describe the catalog after selection. They do not measure selection efficiencies by class because the relevant class-specific parent denominators and warning mechanisms have not been modeled.

The redshift metadata span $-0.00034 \le z \le 6.0878$, with quartiles 0.616, 1.108, and 1.401. Two rows have $z < 0$; 36 lie in $0 \le z < 0.5$, 36 in $0.5 \le z < 1$, 76 in $1 \le z < 1.5$, eight in $1.5 \le z < 2$, 13 in $2 \le z < 3$, and ten at $z \ge 3$. The two small negative values are retained because `ZWARN`=0 is the declared

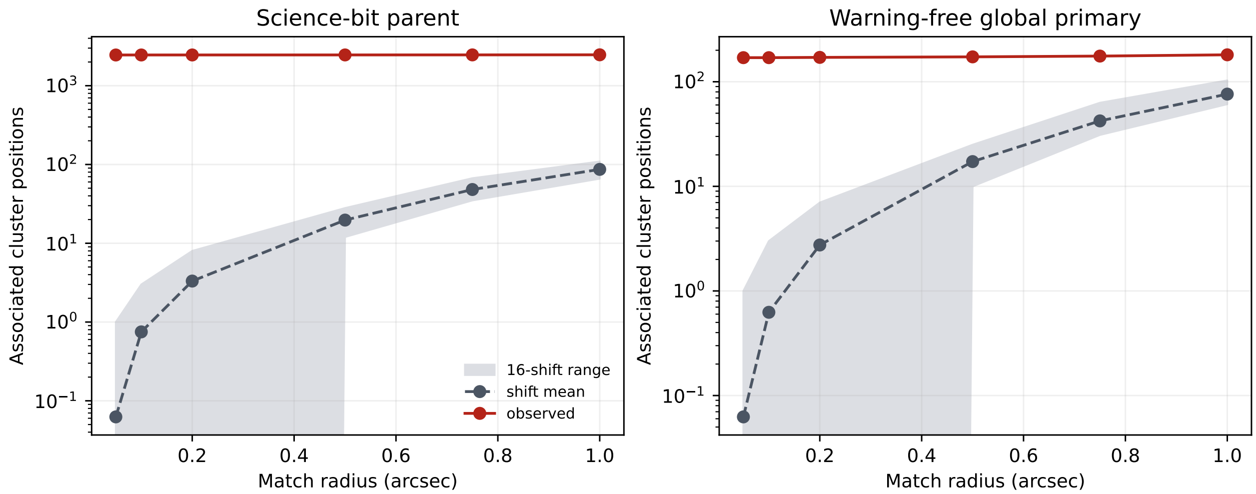

Figure 1. Observed association-radius curves and 16 deterministic local-shift controls for the science-bit parent (left) and warning-free global-primary cohort (right). Shading spans the minimum and maximum shifted count; dashed curves give shifted means. Logarithmic count axes show the sub-0.1″ core, where the large observed count is the expected seed self-recovery of the member coordinates that define the cluster centroids (the local shift verifies this recovery rather than testing an independent association), alongside the shift-controlled 0.1–1″ tail, where the observed and shifted counts are chance-compatible.

Table 3. Exact ZWARN distribution among rejected global-primary rows

Mask	Binary components	Official bit names	Rows
2	2^1	LITTLE_COVERAGE	787
4	2^2	SMALL_DELTA_CHI2	152
6	$2^1 + 2^2$	LITTLE_COVERAGE + SMALL_DELTA_CHI2	1,294
2048	2^{11}	POORDATA	3
2050	$2^{11} + 2^1$	POORDATA + LITTLE_COVERAGE	10
2052	$2^{11} + 2^2$	POORDATA + SMALL_DELTA_CHI2	2
2054	$2^{11} + 2^2 + 2^1$	POORDATA + SMALL_DELTA_CHI2 + LITTLE_COVERAGE	19

Note—Counts sum to 2,267. Bit counts are nonexclusive because a row may set multiple bits. Names and meanings follow the authoritative DESI ZWARN data model.

Table 4. Declared selection waterfall

Stage	Rows
DESI-containing anomaly clusters	190,015
Main science-bit DESI rows searched	20,299,155
$\leq 1''$ positional matches	2,468
Remove non-`ZCAT_PRIMARY`	−20
Global-primary matches	2,448
Remove primary rows with `ZWARN`$\neq 0$	$-2,267$
Released candidates	**181**

Note—The 181-row release is 7.33% of the positional science-bit cohort. This is a quality-gate survival fraction, not an anomaly rate; see Section 5.

quality criterion and no physical-redshift cut was predeclared. They are `P3-DESI-000018` ($z = -0.00033819$) and `P3-DESI-000163` ($z = -0.00033733$), both with descriptive Redrock `SPECTYPE=GALAXY`, BGS_ANY targeting, and high coordinate consistency. Their `DELTACHI2` values are 31,619.4203 and 14,278.0779, and their match separations are 0.000553″ and 0.001766″, respectively. Both belong to the 20-target public-viewer audit in Section 6.2 and pass the same target retrieval, metadata agreement, and B/R/Z array-presence checks. We retain those metadata without inferring that either is a template artifact. Users interested in a positive-redshift science subset should add that cut explicitly and report the resulting 179 rows.

The original anomaly-score range in the release is 5.0006–7.7057, with a median of 5.3244. This narrow

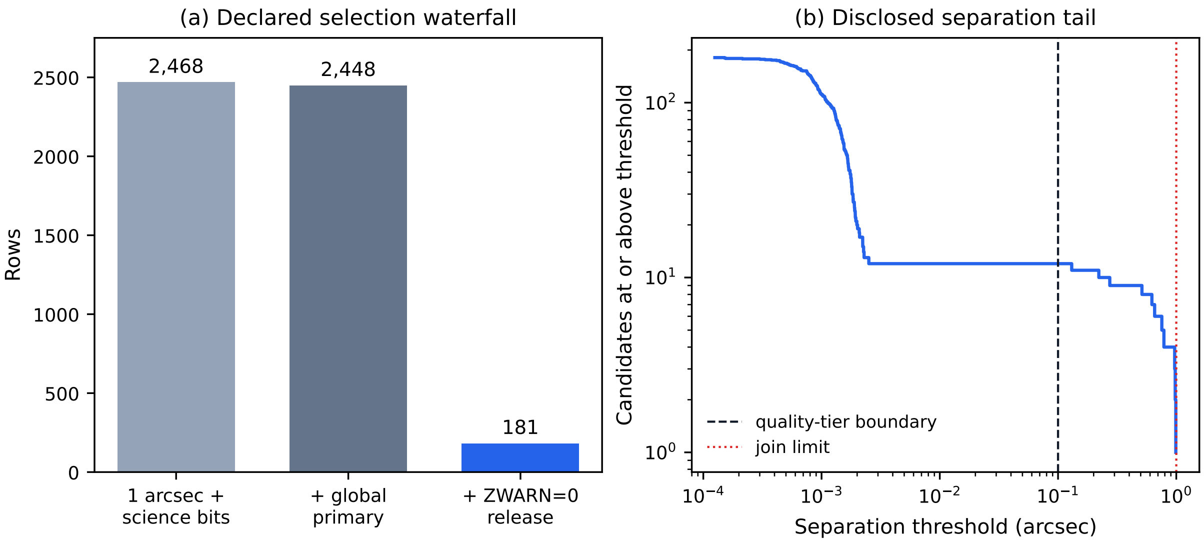

Figure 2. *Left:* Selection waterfall from the exact historical bitmask cohort to the warning-free global-primary release. *Right:* Number of released candidates at or above each positional separation. The dashed line marks the published quality-tier boundary; the dotted line is the predeclared join limit.

range is inherited from the historical anomaly threshold. We do not convert it to a Gaussian significance or compare it across models. Figure 3 shows the sky, redshift, score, and separation metadata without attaching a detection interpretation.

4.3. *Sky and positional coverage*

The released positions range from 3.73° to 352.15° in right ascension and −19.31° to 75.49° in declination. There are 134 candidates north and 47 south of the celestial equator. All twelve 30-degree right-ascension bins are occupied, while four of six coarse equal-area declination bins are occupied. These summaries demonstrate geographic coverage within the input, not isotropy. The footprint is inherited from DESI DR1 and the historical anomaly sample; no all-sky uniformity correction is attempted.

The median positional separation is $0.00127''$, with quartiles $0.00083''$ and $0.00167''$. The abrupt tail of 11 larger separations is why both the continuous separation and categorical quality tier are released. Positional coincidence inside $1''$ supports a catalog association, but it does not prove that the historical anomaly-row identity and public spectrum are physically identical in crowded or blended fields. Those 11 rows are natural priorities for image-level and spectrum-level manual verification.

4.4. *Example rows*

Table 5 uses a deterministic twelve-row illustrative set: the six highest historical scores, the highest-score QSO, the sole STAR, the maximum-redshift row, both negative-z rows, and the maximum-separation row. Because these strata overlap, any unfilled slot is assigned to the next highest-score row not already selected. Inclusion is not a recommendation that an object is novel or astrophysically important. The complete machine-readable catalog contains all 181 rows.

5. RECOVERY SCOPE AND APPROPRIATE USE

5.1. *What the catalog selects*

The release is the intersection of four operations: membership in a frozen historical anomaly-cluster list, positional coincidence with a public DESI row inside $1''$, membership in the declared main-survey science-bit parent, and the `ZCAT_PRIMARY`/`ZWARN`$=0$ quality gates. It is therefore conditional on the historical anomaly model, preprocessing, score threshold, survey footprint, target selection, public zcatalog construction, and Redrock warning behavior. This work makes the public-ID recovery and downstream gates reproducible from the frozen list. It does not reconstruct the upstream model/preprocessing selection function or make those inherited conditions ignorable.

In particular, the released fraction $181/2468 = 7.33\%$ must not be called an anomaly rate, a validation efficiency, or a purity estimate. The numerator is conditioned on warning-free template fits, while the denominator contains the very spectra for which template fitting may be difficult. A population analysis would require at least the class-, magnitude-, redshift-, signal-to-noise-, and warning-dependent inclusion probabilities for both the historical anomaly stage and the DESI rejoin stage. Those quantities are not available in this release.

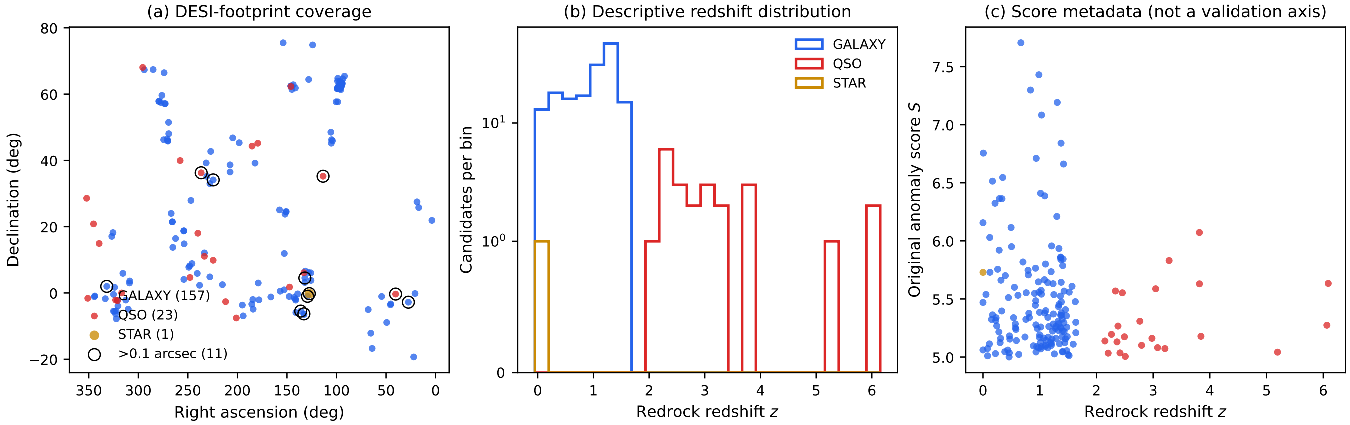

Figure 3. Released v3.2.0-r2 candidates. (a) DESI-footprint sky distribution colored by descriptive Redrock `SPECTYPE`; open circles mark the 11 rows in the 0.1–1″ separation tier. (b) Redshift metadata by spectral type. (c) Historical anomaly score versus Redrock redshift. Neither `SPECTYPE` nor redshift was used to select the release, and the score is not a calibrated physical significance.

Table 5. Deterministic example rows: six highest-score rows plus the highest-score QSO, sole STAR, maximum-z, two negative-z, and maximum-separation strata, with overlaps filled by the next highest-score unused row

Candidate	TARGETID	R.A.	Decl.	Type	z	S	Sep. (arcsec)
P3-DESI-000001	39633349869831142	272.82236	57.12256	GALAXY	0.668	7.71	0.001
P3-DESI-000002	39627706974869534	45.28046	-3.31238	GALAXY	0.986	7.43	0.002
P3-DESI-000003	39627895609492977	131.94836	4.51377	GALAXY	0.838	7.30	0.274
P3-DESI-000004	39633407088529181	97.49930	61.65664	GALAXY	1.310	7.19	0.001
P3-DESI-000005	39627636166626827	132.90847	-6.30883	GALAXY	1.033	7.08	0.991
P3-DESI-000006	39633418404759790	95.13019	62.74366	GALAXY	1.376	6.84	0.000
P3-DESI-000020	39627931743422578	132.93267	6.11482	QSO	3.816	6.07	0.001
P3-DESI-000039	39627780823979311	127.45356	-0.20596	STAR	0.000	5.73	0.221
P3-DESI-000047	39627903570285763	247.89529	4.66991	QSO	6.088	5.63	0.002
P3-DESI-000018	39628296954055370	265.65886	21.45037	GALAXY	-0.000	6.16	0.001
P3-DESI-000163	39628296954055072	265.64950	21.45594	GALAXY	-0.000	5.06	0.002
P3-DESI-000007	39636182677589446	64.05184	-16.74091	GALAXY	0.006	6.76	0.002

Note—R.A. and declination are ICRS degrees. Type and z are public Redrock metadata. S is preserved historical anomaly-score metadata. Values are rounded for display only; the full 181-row, 43-column table is supplied at submission as an AAS machine-readable digital asset. The two rows shown as $z = -0.000$ (`P3-DESI-000018`, `P3-DESI-000163`) are the small negative redshifts $z = -0.00033819$ and $z = -0.00033733$, reported to full precision in Section 4.2; they are distinct from exact zero.

5.2. *Recommended uses*

The catalog is designed for:

- deterministic retrieval of 181 public DESI spectra by `TARGETID`;
- manual or automated spectral vetting using the full public data products;
- testing alternative anomaly models on a fixed, warning-free candidate list;
- follow-up prioritization using continuous score, separation, target-bit, and redshift metadata; and
- reproducibility exercises in archive recovery and catalog selection.

It is not designed for estimating an anomaly occurrence rate, claiming new object classes, measuring cosmological parameters, or comparing raw counts with a differently selected survey. No SIMBAD-based novelty fraction, process-volume multiplier, cross-survey total, or cosmological inference is part of v3.2.0-r2. Establishing novelty requires object-by-object literature and archive review beyond a coordinate database match. Establishing a physical anomaly requires inspection of calibrated spectra, masks, observing conditions, and repeat observations.

5.3. *Relationship to prior DESI anomaly searches*

Prior DESI studies have demonstrated useful outlier-search strategies for particular target classes and data subsets [3, 4]. Direct candidate-count comparisons are not meaningful here because the parent populations, preprocessing, models, thresholds, and quality gates differ. The contribution of v3.2.0-r2 is orthogonal: it converts a historically unrejoinable coordinate-and-score product into a small public-key catalog with a fully enumerated waterfall. Model comparisons can use these 181 public spectra as a common follow-up set, but evaluation against a representative control sample would still be required.

6. REPRODUCIBILITY AND INTEGRITY AUDIT

6.1. *Exact source-row rejoin*

The independent validator reads the released `fits_row` values from the final Parquet, reopens the public FITS file, and retrieves all 18 source fields listed in Table 1. Integer, Boolean, and string fields are compared exactly; floating fields are also compared at zero tolerance because the Parquet values were copied directly. All 181 rows pass all 18 field comparisons. Independently of the released table, the validator reconstructs the strict cohort from the 143 scan checkpoints and compares the complete `(cluster_id, targetid, fits_row)` sets. The sets are exactly equal (181 expected, 181 released, zero missing, and zero unexpected), rather than merely having equal counts. The validator separately confirms unique candidate, cluster, `TARGETID`, and FITS-row keys; zero null cells; main-survey membership; at least one declared science bit; `ZCAT_PRIMARY`=true; `ZWARN`=0; and separation at or below $1''$.

6.2. *Public-viewer retrieval audit*

We queried the DESI DR1 target-specific Legacy Survey viewer documented by the official DESI data access guide for every one of the 11 rows above $0.1''$ and for the 12-row representative set in Table 5; overlap gives 20 unique targets. Every page returned the released `TARGETID`, redshift, `SPECTYPE`, `ZWARN`, and `DELTACHI2` within the declared serialization tolerances, and served B, R, and Z spectral arrays containing 2,751, 2,326, and 2,881 pixels, respectively, plus an imaging-cutout link. The governed browser capture also confirmed that each target-specific viewer page rendered a spectrum and coordinate marker.

The exact URLs, returned metadata, arm lengths, checks, and screenshots are preserved with the r3 audit artifacts; the machine-readable result is pipelines/p3_anomaly_engine/audits/p3_v320_r3_public_viewer_audit.json. This establishes public retrieval and metadata agreement. It is not a visual classification, novelty result, or proof that a historical anomaly score has an astrophysical origin. Imaging background appearance varied in the captured pages, so no crowded-field physical judgment is inferred from the screenshots.

6.3. *Provenance and checksum reconciliation*

The full local 22.37 GB FITS SHA-256 was recomputed. A no-cache retrieval of DESI's current official checksum file reports the same value. The live FITS endpoint reports the same byte size, an ETag of `6a516695-5356e87c0`, and a last-modified time of 2026-07-10 21:39:33 GMT. The machine validator compared eight predeclared remote byte ranges, including the first and last, against the local FITS and required exact HTTP range metadata, byte counts, and SHA-256 digests; all eight passed. Full transaction-level results remain in `SELECTION_AUDIT.json`. An older local provenance sidecar written in May 2026 records a different pre-replacement SHA-256. We preserve that stale value in `PROVENANCE.json` and explicitly exclude it from current validation rather than overwriting the history.

The two historical input Parquets are also hashed: the cluster table is `b14deb02...d6138c643` (12,087,824 bytes) and the anomaly table is `0a36b8d6...f103ec65` (10,514,503 bytes). Full hashes, not abbreviated display forms, are in the release manifest and provenance file.

6.4. *Release contents*

The clean v3.2.0-r2 directory contains one science table and its audit artifacts. Its primary contract retains 181 positional associations, explicitly partitioned by `match_quality_tier` into the 170-row high-coordinate-consistency core and 11-row lower-confidence tail; it makes no secure-identity or purity claim:

1. desi_dr1_science_anomaly_candidates_v3.2.0-r2.parquet;
2. `DATA_DICTIONARY.md` and `README.md`;

3. COHORT_COUNTS.json, QC_REPORT.json, SELECTION_AUDIT.json, and SELECTION_AUDIT.md;

4. exact build and independent-validation scripts;

5. `PROVENANCE.json`; and

6. `RELEASE_MANIFEST.json` with SHA-256 and byte size for every payload file.

The manifest hashes ten payload files and excludes itself to avoid a self-referential hash; the directory therefore contains eleven files total. An independent post-generation pass recomputes every listed hash and byte count. The release contains no Gaia, eROSITA, LAMOST, SDSS, Planck, ACT, or heterogeneous aggregate table. Earlier releases remain historical and are not moved or rewritten.

The separate directory desi_warned_primary_aux_v3.2.0-r5 contains the secondary warning-bearing Parquet table, dictionary, warning-forward README, provenance and license files, checkpoint and release manifests, exact-set replay assertion, and bundled builder/validator. Its validator independently reconstructs the 2,267 identities from all 143 clean-replay checkpoint parts, requires exact carried-field equality, and verifies the seven masks and decoded bits in Table 3. It is not part of the 181-row primary release. The definitive v3.2.0-r7 submission directory binds both products, the exact AAS table serialization, their builders/validators and dictionaries, the r6 controls, and the recovered coordinate-lineage evidence in one SHA-256 manifest. Component version labels remain unchanged, and the AAS digital-asset DOI remains pending.

7. LIMITATIONS

The catalog has six central limitations.

First, historical score lineage. The preserved evidence identifies a 496-to-128, five-seed-mean-MSE BigAE lineage, but the production normalization/resampling and original object-to-spectrum mapping are unavailable. The score and residual columns are therefore preserved from the anomaly stream rather than recomputed from public spectra. Neither v3.2.0-r2 nor the auxiliary product claims exact historical-score reproduction or a physical-feature sensitivity function.

Second, quality-conditioned incompleteness. Requiring `ZWARN`=0 removes 92.6% of primary positional matches. The released list is cleaner for metadata use but incomplete by construction and likely biased against the most difficult-to-fit spectra. The secondary v3.2.0-r5 table makes those exclusions inspectable but does not estimate or correct that bias and does not make them primary-catalog members.

Third, positional association. A $1''$ match is not a proof of identity in every field. The sub-$0.1''$ core recovers the seed DESI members whose own coordinates define the cluster centroids, so its local-shift comparison verifies the recovery rather than establishing an independent astrophysical association; the shift control is informative only for the 11-object 0.1–$1''$ tail, which it does not secure at the individual-row level. Those rows are flagged and all users retain the continuous separation. An original-member-centered $1''$ counterfactual retains 180 rows and removes `P3-DESI-000030`.

Fourth, no physical classification. Redrock class, redshift, and `DELTACHI2` are catalog metadata. The 20-target viewer audit verifies public retrieval but does not classify spectral features. We have not physically interpreted every spectrum, image, mask, or repeat observation, and we make no novelty or physical-anomaly claim.

Fifth, inherited footprint and targeting. The sample follows DESI DR1 coverage and the declared target bits. Target classes overlap, and no completeness correction is provided.

Sixth, version specificity. Public redshift products can be regenerated or superseded. The release pins file size and checksum so later catalog versions can be compared rather than silently substituted.

These limitations are not peripheral caveats: they define the valid use of the product. The release should be cited as a candidate catalog and reproducibility artifact.

8. CONCLUSIONS

We have replaced a heterogeneous, public-ID-limited anomaly product with a focused DESI DR1 catalog whose public-ID recovery and provenance are testable from end to end. Preserved code traces the historical DESI anomaly coordinates to FIBERMAP `TARGET_RA`/`TARGET_DEC` and the frozen cluster coordinates to member-coordinate means, while leaving the unavailable production object-to-spectrum mapping explicitly unresolved. A memory-bounded scan of 28,425,963 public zcatalog rows exactly reproduces the prior 20,299,155-row science-bit denominator and 2,468 one-arcsecond positional matches. The declared global-primary and warning-free gates yield 181 unique public `TARGETID` associations. The primary contract retains all 181 but distinguishes a 170-row high-coordinate-consistency core from 11 lower-confidence 0.1–$1''$ positional associations; neither is presented as a secure identity or purity estimate. All carried DESI fields rejoin exactly, all released keys are unique, and every payload hash and byte size passes an independent manifest audit. The 2,267 warning-bearing global-primary matches are preserved in a separately labeled auxiliary table with exact warning masks and replay assertions; they remain outside the primary catalog and are not physically classified. Deterministic local shifts confirm that the sub-0.1″ core is expected seed self-recovery of the member coordinates that define the cluster centroids — an end-to-end verification of the recovery rather than an independent association excess — while for the 11-object 0.1–1″ tail, where the shift control is genuinely informative, the rows are not secure from position alone. The exact warned-row comparison exposes the quality gate's compositional effect without correcting it, and an explicit original-member-centered 1″ counterfactual retains 180 rows and removes only `P3-DESI-000030`.

The smaller, auditable count is the catalog-integrity improvement. Relative to the frozen 195,829-row historical anomaly table, it provides 181 traceable public-ID candidates and a clear account of the 2,287 exclusions. The release retains the full 1″ cohort definition while exposing the 11-row separation tail, preserves historical score metadata without turning it into a significance, and states that the warning-free cohort is unsuitable for occurrence-rate inference. Its immediate purpose is reproducible spectral retrieval and follow-up; physical classification is not claimed by this catalog.

DATA AVAILABILITY

The v3.2.0-r2 primary release is published under the correction tag p3-v3.2.0-r2 in the public Hugging Face dataset (accessed 2026-07-14) https://huggingface.co/datasets/bamfai/bigbounce-anomaly-catalog, path releases/p3-v3.2.0-r2/. At publication the tag resolves to commit `1a9e85ee004894956665444b4f110111f1090b79`; that commit hash is the authoritative immutable identifier, and exact consumers should pin it rather than depend on a tag name alone. All 11 files in that public snapshot were downloaded and verified byte-for-byte at its freeze. The r5 closure validator, README, and audits are additive repository evidence; they do not rewrite and are not claimed to be present at the older public snapshot commit. The historical `p3-v3.2.0` and `p3-v3.2.0-r1` tags and their artifacts are preserved; p3-v3.2.0-r2 is an additive successor correction.

The historical cluster and anomaly inputs (accessed 2026-07-14) are respectively https://huggingface.co/datasets/bamfai/bigbounce-anomaly-catalog/resolve/cdaaa03a72c69d86f011be128d93f261dc5b39a8/pathc_unique_objects.parquet and https://huggingface.co/datasets/bamfai/bigbounce-anomaly-catalog/resolve/cdaaa03a72c69d86f011be128d93f261dc5b39a8/desi_dr1_anomalies.parquet. Both are pinned by `p3-v3.1.161`, which resolves to immutable commit `cdaaa03a...5b39a8`; their SHA-256 values are recorded in `PROVENANCE.json`. The public DESI input and official checksum list (accessed 2026-07-14) are available at https://data.desi.lbl.gov/public/dr1/spectro/redux/iron/zcatalog/v1/. After downloading the three named inputs, the concrete build command is:

```
BUILD=build_desi_science_catalog_v320_r2.py
VALIDATE=validate_desi_science_catalog_v320_r2.py
RELEASE=desi_science_catalog_v3.2.0-r2
FITS=zall-pix-iron.fits
FITS_SHA256=2d95ad99361039b556c402b49e0e7c84\
df5f00106dc5731d44476a58b128b49b
python3 "$BUILD" \
  --clusters pathc_unique_objects.parquet \
  --anomalies desi_dr1_anomalies.parquet \
  --fits "$FITS" \
  --output-dir "$RELEASE" \
  --chunk-rows 200000 \
  --fits-sha256 "$FITS_SHA256"
python3 "$VALIDATE" \
  --release-dir "$RELEASE" \
  --fits "$FITS" \
  --clusters pathc_unique_objects.parquet \
  --parts-dir ".$RELEASE.build/match_parts"
```

The expected FITS digest is assigned on one shell line and passed through an environment variable, avoiding any copy-paste ambiguity. The validator also derives the shown checkpoint path from `--release-dir` when `--parts-dir` is omitted; both forms pass from an unrelated fresh temporary directory. The release README gives the same download, build, validation, and explicit `sha256sum -c` commands for the two historical inputs. The source tree contains the same scripts at pipelines/p3_anomaly_engine/scripts/. No secret credentials or private data are required to validate the released catalog against the public DESI file.

The secondary v3.2.0-r5 product is available at pipelines/p3_anomaly_engine/desi_warned_primary_aux_v3.2.0-r5/ (accessed 2026-07-14). Its exact identity-set digest is `2cc02e4b839073a3d2664c57614a6b20205a3d1d9d68d71851b72ccb4794ee78`; the release manifest records the SHA-256 and byte size of each payload. This repository artifact is secondary, warning-bearing, outside the primary catalog, and not physically validated.

The definitive checksum-bound v3.2.0-r7 ApJS submission unit is pipelines/p3_anomaly_engine/apjs_submission_bundle_v3.2.0-r7/. Its `BUNDLE_MANIFEST.json` binds the primary release, warned auxiliary, AAS machine-readable table and builder, dictionaries and validators, r6 controls, and coordinate-lineage evidence by role, source path, byte size, and SHA-256. The manifest defines the 170/11 primary contract and records the AAS digital-asset DOI as pending rather than claimed; `SHA256SUMS` also binds the manifest and bundle README.

The r6 science-control script, full JSON result, 181-row original-member sensitivity sidecar, 11-row positional-tail extract, and plotted radius curve are repository artifacts at pipelines/p3_anomaly_engine/scripts/p3_apjs_r6_science_controls.py, pipelines/p3_anomaly_engine/outputs/p3_apjs_r6_science_controls.json, and the adjacent r6 output/figure paths. The JSON records all 16 realizations, input hashes, cohort descriptives, and the deterministic original-member expression; it makes no archive-DOI claim.

The printed Table 5 is only a 12-row, eight-column illustrative excerpt. The separate full 181-row, 43-column catalog table (`tab3.tsv`) is prepared for submission with this manuscript, accompanied by a plain-text column dictionary and SHA-256 manifest in pipelines/p3_anomaly_engine/aas_submission_v3.2.0-r4/. It will be submitted as an AAS contributed digital asset. AAS assigns the journal digital-asset DOI during its publication workflow; no such DOI has yet been assigned or is claimed here. The immutable Hugging Face commit above remains the present public byte-level reference. The historical dataset and catalog data/documentation are CC BY 4.0; bundled code retains the repository license. The AAS asset is the planned journal archive, with its DOI remaining an honest publication-workflow gate. This follows the AAS data guide (accessed 2026-07-14) (https://journals.aas.org/data-guide/) and digital-asset DOI policy (https://journals.aas.org/news/digital_asset_doi_landingpage/).

SOFTWARE

The build used Python 3.9.6, NumPy 1.26.4 [7], pandas 2.1.4 [8], SciPy 1.13.1 [9], fitsio 1.3.0 [11], and PyArrow 21.0.0 [12]. FITS is described by the FITS standard [10]. Exact versions and the chunk size are recorded in `PROVENANCE.json`.

ACKNOWLEDGMENTS

This research used data obtained with the Dark Energy Spectroscopic Instrument (DESI). DESI construction and operations is managed by the Lawrence Berkeley National Laboratory. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of High-Energy Physics, under Contract No. DE–AC02–05CH11231, and by the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility under the same contract. Additional support for DESI was provided by the U.S. National Science Foundation (NSF), Division of Astronomical Sciences under Contract No. AST-0950945 to the NSF's National Optical-Infrared Astronomy Research Laboratory; the Science and Technology Facilities Council of the United Kingdom; the Gordon and Betty Moore Foundation; the Heising-Simons Foundation; the French Alternative Energies and Atomic Energy Commission (CEA); the National Council of Humanities, Science and Technology of Mexico (CONAHCYT); the Ministry of Science and Innovation of Spain (MICINN), and by the DESI Member Institutions: https://www.desi.lbl.gov/collaborating-institutions. The DESI collaboration is honored to be permitted to conduct scientific research on I'oligam Du'ag (Kitt Peak), a mountain with particular significance to the Tohono O'odham Nation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the U.S. National Science Foundation, the U.S. Department of Energy, or any of the listed funding agencies.

The analysis benefited from automated code and manuscript review; all catalog assertions reported here were checked against machine-readable artifacts, and no review-system output was accepted as evidence without a local or public-source verification.

REFERENCES

[1] DESI Collaboration, "Data Release 1 of the Dark Energy Spectroscopic Instrument," Astron. J. **171**, 285 (2026), doi:10.3847/1538-3881/ae4c43, arXiv:2503.14745.

[2] D. Baron and D. Poznanski, "The weirdest SDSS galaxies: results from an outlier detection algorithm," Mon. Not. R. Astron. Soc. **465**, 4530 (2017).

[3] Y. Liang et al., "Outlier Detection in the DESI Bright Galaxy Survey," Astrophys. J. Lett. **956**, L6 (2023), arXiv:2307.07664.

[4] C. Nicolaou et al., "Identifying Anomalous DESI Galaxy Spectra with a Variational Autoencoder," Mon. Not. R. Astron. Soc. **547**(2), stag010 (2026), doi:10.1093/mnras/stag010, arXiv:2506.17376.

[5] H. Golden, "BigBounce Multi-Survey Autoencoder Anomaly Catalog (Paper 3)," version `p3-v3.1.161`, CC BY 4.0 (2026), immutable revision `cdaaa03a...5b39a8`, accessed 2026-07-14.

[6] DESI Data Systems, "Bit Masks in DESI: target masks and Redrock ZWARN," DESI Data Model 25.3 (2025), https://desidatamodel.readthedocs.io/en/25.3/bitmasks.html, accessed 2026-07-14.

[7]C. R. Harris et al., “Array programming with NumPy,” Nature **585**, 357 (2020), doi:10.1038/s41586-020-2649-2.

[8]W. McKinney, “Data Structures for Statistical Computing in Python,” in Proceedings of the 9th Python in Science Conference, 56 (2010), doi:10.25080/Majora-92bf1922-00a.

[9]P. Virtanen et al., “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python,” Nat. Methods **17**, 261 (2020).

[10]W. D. Pence, L. Chiappetti, C. G. Page, R. A. Shaw, and E. Stobie, “Definition of the Flexible Image Transport System (FITS), version 3.0,” Astron. Astrophys. **524**, A42 (2010).

[11]E. Sheldon, “fitsio: a Python library to read from and write to FITS files,” version 1.3.0, https://github.com/esheldon/fitsio.

[12]Apache Software Foundation, “Apache Arrow and PyArrow,” version 21.0.0, https://arrow.apache.org/docs/.

APPENDIX

A. EXECUTABLE SELECTION SPECIFICATION

For clarity, the released cohort can be represented by the following ordered specification:

1. Select the 190,015 historical clusters with `desi_dr1` in their survey membership.
2. For each cluster, select the highest-score DESI historical member, breaking ties by source row.
3. Stream public zcatalog rows and retain `SURVEY==main` with at least one declared `DESI_TARGET` bit.
4. Assign each eligible zcatalog row to its nearest cluster in spherical unit-vector space.
5. Retain separations $\leq 1''$.
6. Verify by an all-neighbor query that no eligible row has a second cluster within $1''$ and that parent/strict pair counts remain 2,468/181.
7. Define the parent, global-primary, and strict predicates; the strict predicate requires `ZCAT_PRIMARY`=true and `ZWARN`=0.
8. Apply the chosen cohort predicate before duplicate handling.
9. Within that predicate-filtered cohort, keep the nearest row per cluster; break ties by descending `DELTACHI2`, ascending `TARGETID`, and ascending FITS row.
10. Apply the same ordering to keep one cluster per `TARGETID`.
11. Label separation $\leq 0.1''$ as coordinate-consistent and the remaining declared $1''$ matches as the positional tail.
12. Sort by `cluster_id` and assign sequential release-local candidate IDs.

The script writes checkpoint parts before forming the final cohort. Thus the selection counts can be inspected independently of the manuscript and an interrupted full-catalog scan can resume without replaying completed chunks.

B. COLUMN GROUPS

The release contains 43 columns after addition of the separation-quality tier and the explicit target-to-original-member separation. The authoritative field-by-field descriptions and storage dtypes are in `DATA_DICTIONARY.md`. The two tables below define the public and historical field groups without duplicating them in a third summary table.

Table 6. Release and public-DESI field definitions

Column	Definition
`candidate_id`	Stable release-local ID ordered by cluster ID.
`fits_row`	Zero-based row in the public DR1 ZCATALOG extension used by the exact rejoin audit.
`cluster_table_row`	Zero-based row in the committed historical cluster table.
`match_separation_arcsec`	Great-circle target-to-cluster separation in arcseconds.
`match_quality_tier`	At most $0.1''$, or greater than $0.1''$ and at most $1''$.
`original_member_separation_arcsec`	Great-circle target-to-canonical-original-member separation; not used as a selection cut.
`targetid`	Public DESI `TARGETID`; recommended archive key.
`target_ra`, `target_dec`	Public DESI target coordinates in ICRS degrees.
`survey`, `program`	DESI survey and observing-program labels.
`desi_target`	Raw `DESI_TARGET` bitmask used for the declared science-bit gate.
`bgs_target`, `mws_target`, `scnd_target`	Raw public class-specific targeting bitmasks.
`z`	Redrock redshift estimate; descriptive, not a selection label.
`zwarn`	Redrock warning mask; required to equal zero.
`spectype`	Redrock best-fit spectral type; not used as a selection cut.
`deltachi2`	Best-versus-next-best Redrock template $\Delta\chi^2$.
`coadd_fiberstatus`	Bitwise OR of fiber-status flags contributing to the coadd.
`main_nspec`	Number of associated main-survey spectra.
`main_primary`	Primary-within-main-survey flag.
`zcat_nspec`	Number of spectra in the global zcatalog grouping.
`zcat_primary`	Global primary redshift-row flag; required to be true.

Note—Raw integer bitmasks are preserved so additional DESI software can decode more detailed classes without relying on the release's compact label.

Table 7. Historical-cluster and anomaly-metadata field definitions

Column	Definition
`cluster_id`	Stable identifier from the historical positional clustering.
`n_detections`	Number of historical anomaly members in the cluster.
`n_surveys`	Number of historical survey labels in the cluster.
`survey_list`	Comma-separated historical membership; audit provenance only.
`cluster_ra_deg`, `cluster_dec_deg`	Mean historical cluster coordinate in ICRS degrees.
`cluster_best_score`	Maximum historical score over all cluster members.
`member_ids`	Pipe-separated legacy row labels; not public archive keys.
`best_survey`	Historical survey supplying the cluster-best score.
`desi_source_row`	Row of the canonical DESI member in the historical anomaly table.
`original_internal_tid`	Legacy mixed/hash identifier; negative values are expected and it must not be used as a public key.
`original_ra_deg`, `original_dec_deg`	Coordinates of the canonical historical DESI anomaly member.
`original_score`	Preserved historical anomaly-ranking score, not a calibrated significance.
`original_worst_band`	Band with the largest historical residual summary.
`original_residual_b`	Preserved historical B-band residual summary.
`original_residual_r`	Preserved historical R-band residual summary.
`original_residual_z`	Preserved historical Z-band residual summary.
`science_target_class`	Pipe-separated decoded selected bits; labels are nonexclusive.

Note—These fields preserve the recovery lineage. They do not convert historical model output into public identity or astrophysical validation.

C. AUDIT MATRIX

Table 8. Independent v3.2.0-r2 integrity checks and v3.2.0-r10 closure audits

Check	Result	Evidence
Historical denominator	PASS: 20,299,155	streamed cumulative counter
Historical match count	PASS: 2,468	checkpoint parts and cohort JSON
Strict cohort identity	PASS: exact 181/181 set	`(cluster_id,targetid,fits_row)`; zero missing/unexpected
Public identity	PASS: 181/181	direct FITS-row `TARGETID` comparison
Carried DESI fields	PASS: 18/18 fields for 181 rows	zero-tolerance source-row comparison
Uniqueness	PASS	candidate, cluster, TARGETID, and FITS row
Null cells	PASS: none	full Parquet null scan
Selection gates	PASS	main, science bit, primary, warning-free, radius
Gate/deduplication order	PASS	predicate first; 2,468/2,448/181; unsafe counterfactual loses zero rows
Positional competitors	PASS	none within $1''$; one candidate has one competitor within $2''$
Nearest/all-neighbor equivalence	PASS	cluster minimum $5.02294''$; exact 2,468/181 pairs
Rejected ZWARN audit	PASS: 2,267	seven exact masks; bits 1/2/11 counted and defined
Warning-bearing auxiliary	PASS: exact 2,267/2,267 set	stable IDs, carried-field replay, decoded masks, manifest
Local-shift association control	PASS: 16/16 realizations	$60/120''$, eight angles, parent and strict radius curves
Accepted/warned comparison	PASS: 181/2,267 exact rows	medians, quartiles, and Redrock class counts from Parquet
Original-member sensitivity	PASS: 180 retained, one removed	explicit $\leq 1''$ expression; P3-DESI-000030
Public viewer retrieval	PASS: 20/20	all 11 tail plus representative set; matching metadata and B/R/Z arrays
FITS provenance	PASS	full local SHA equals current official checksum
Remote parity	PASS: 8/8	live 206/Content-Range/1 MiB/digest validation
Payload manifest	PASS: 10 payloads	manifest self-excluded; 11 directory files total
AAS table package	PASS: 181×43	exact typed TSV round-trip; journal DOI not yet assigned

Note—PASS describes integrity relative to the declared release contract. It is not a claim that the candidates are astrophysically anomalous or novel.