SHV Groups

Evidence-Carrying Anomaly Discovery

Error-controlled, cross-instrument discovery of unusual objects in space telescope imaging, with a pre-registered protocol and a simulation study

Apeiron, Department of Cosmology and Astrophysics (DoCA), SHV Groups

Corresponding author: Shaurya Verma, Apeiron DoCA, SHV Groups Pvt. Ltd., India

Version 1.1, November 2026. Methods paper with a simulation study. No astrophysical results are reported.

DOI: 10.5281/zenodo.23883469. Licence: CC BY 4.0.

Abstract

Wide and deep space surveys now produce more sources than experts can inspect, and machine learning can rank them all by how unusual they look. A ranking is not a discovery. Current anomaly searches do not say how many of the objects they flag are expected to be false, whether a flag reflects the sky or the instrument, or what evidence would let another group check the claim. We formalise anomaly discovery as multiple testing over sources observed in several views under changing observing conditions, and propose Evidence-Carrying Anomaly Discovery (ECAD). ECAD computes conformal p-values for each view within strata of observing conditions, applies one data-quality rule identically to reference and test views, requires replication across artifact-independent views with a partial-conjunction test, controls the false discovery rate across sources with the Benjamini-Hochberg procedure, separates rare members of known classes from unexplained objects with a class-conditional conformal test, and attaches a versioned evidence bundle and a discovery level to every candidate. We show that any search of this kind needs a reference population of at least $\kappa/(q\pi)$ views per stratum and a contamination below $q\pi/\kappa$ to find objects of prevalence π at level q; the second condition cannot be met by adding data. A simulation study on a synthetic two-instrument sky confirms that ECAD controls the false discovery rate at the replicated level (observed 0.015 at a target of 0.10), suppresses single-exposure artifacts by more than two orders of magnitude, and makes no discoveries when there is nothing to find. It also exposed three failures in the first version of the framework, which this version corrects. We specify a pre-registered evaluation on public HST, JWST, Euclid and Rubin data, with pass, fail and abandonment criteria fixed in advance and rehearsed in simulation. No astrophysical results are reported.

Keywords: anomaly detection; conformal inference; false discovery rate; partial conjunction; replication; JWST; Euclid; Rubin Observatory; Roman Space Telescope; pre-registration; human-in-the-loop discovery

1 Introduction

1.1 The scientific problem

Many of the important advances in astrophysics began with an object that did not fit. Hanny's Voorwerp was flagged by a Galaxy Zoo volunteer (Lintott et al., 2009). KIC 8462852 showed dimming that no standard model predicted (Boyajian et al., 2016). Fast radio bursts were found in archival pulsar data (Lorimer et al., 2007). In the JWST era, compact red sources at high redshift, now called "little red dots", were first noticed as unexpected photometric outliers, and their nature is still contested. Matthee et al. (2024) interpreted them as faint active galactic nuclei. Rusakov

et al. (2026) argued that in the best-observed cases their broad lines come from electron scattering in dense ionised cocoons around young black holes, implying black-hole masses of 10^5 to $10^7\,M_\odot$, two orders of magnitude below earlier estimates. Other groups dispute this interpretation for individual objects. Galaxies that appear unexpectedly massive at early times (Labbé et al., 2023) are a second example of outliers that put pressure on galaxy formation models.

The volume of imaging is now growing faster than the expert time available to inspect it. Euclid's first quick release (Q1, 19 March 2025) covered 63.1 deg^2 and about 30 million objects (Euclid Collaboration et al., 2025a). Rubin Observatory's Data Preview 1 contains 2,299,726 catalogued objects from 1,792 exposures taken over 48 nights in seven fields totalling about 15 deg^2 (Rubin Observatory Team, 2026). The Nancy Grace Roman Space Telescope launched on 30 August 2026; NASA describes its Wide Field Instrument as giving images as sharp as Hubble's over a field of view at least 100 times larger (0.28 deg^2).

1.2 What is unresolved

Machine-learning anomaly detection in astronomy is mature as a ranking technology. O'Ryan and Gómez (2025) searched 99.6 million Hubble Legacy Archive cutouts with AnomalyMatch (Gómez et al., 2026) in two to three days and, after expert review of the 5,000 highest-scoring cutouts, reported 1,176 previously undocumented anomalies, including candidate strong lenses, jellyfish galaxies and mergers. Three gaps remain.

1. **No error control.** Published searches report how many objects experts confirmed in a top-k list. They give no statement, made before inspection, of the expected fraction of false claims among the objects they flag. Anomaly scores have no common scale across instruments, so thresholds are set by hand.
2. **Artifacts are tested by eye.** Blind detectors surface unusual but uninteresting sources, including image artifacts (Storey-Fisher et al., 2021; Lochner and Rudnick, 2025). A 2026 Euclid VIS pipeline study found that image-quality failures were among the candidates its anomaly workflow returned (arXiv:2609.32753, 2026). No published framework formally requires a candidate to replicate across independent exposures or instruments.
3. **No evidence standard.** A flag is not a discovery. Candidate catalogues rarely include the provenance, calibration state, null distribution, completeness at the candidate's observing conditions and review history that another team would need to check a claim.

1.3 Contributions

1. **A formal problem statement** that treats anomaly discovery as multiple testing over sources observed in several views, with a hierarchy of null hypotheses: typical object, instrument artifact and known class (Section 3).
2. **ECAD, an inference pipeline** that combines conformal p-values stratified by observing conditions, symmetric data-quality gating, partial-conjunction replication across artifact-independent views, Benjamini-Hochberg control of the false discovery rate across sources, and a class-conditional conformal test that separates rare members of known classes from unexplained objects (Section 4).
3. **A five-level discovery scale and an evidence-bundle schema** that fix what may be called a candidate, a replicated anomaly, an unexplained object and a confirmed discovery.
4. **Two conditions that bound any search of this kind** (Section 3.5): each reference stratum must hold at least $\kappa/(q\pi)$ views to resolve objects of prevalence π, and its contamination must stay below $q\pi/\kappa$. The second cannot be fixed with more data, rules out calibrating a

survey against itself, and sets the rarest class of object that an error-controlled search can claim.

5. **A simulation study** on a synthetic two-instrument sky with realistic observing conditions, five artifact families and four anomaly classes (Section 5). It exercises every component and checks every guarantee the method claims. It exposed three failures in version 1.0 of this framework, which this version corrects: a calibration-size requirement that could not resolve realistic prevalences, a loss of most power when the reference population carries artifacts, and a rule for declaring an object unexplained that never did so.
6. **A pre-registered evaluation protocol** on public HST, JWST, Euclid and Rubin data, with null tests, injection-recovery tests, synthetic artifact controls, named baselines, ablations, blinded expert review, and pass, fail and abandonment thresholds fixed in advance (Section 7).

The simulation results describe how the method behaves on synthetic data with known ground truth. They are not astrophysical results, and they do not show how ECAD will perform on real archives. That is what the pre-registered protocol is for.

2 Background and related work

2.1 Why rare objects matter now

The standard cosmological model is under observational strain, and outliers are often where models fail first. The Planck 2018 analysis gives $H_0 = 67.4 \pm 0.5\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ under flat ΛCDM (Planck Collaboration et al., 2020), while the SH0ES Cepheid-calibrated distance ladder gives 73.04 ± 1.04 (Riess et al., 2022). The Chicago-Carnegie Hubble Program's JWST-era tip of the red giant branch analysis prefers an intermediate $70.39 \pm 1.22\,(\mathrm{stat}) \pm 1.33\,(\mathrm{sys}) \pm 0.70\,(\sigma_{\mathrm{SN}})$ and its authors argue that it is consistent with ΛCDM (Freedman et al., 2025). A 2026 review describes the disagreement as unresolved (arXiv:2606.20434, 2026). DESI DR2 baryon acoustic oscillation measurements combined with CMB data prefer evolving dark energy over a cosmological constant at 3.1σ, rising to between 2.8σ and 4.2σ depending on the supernova sample (DESI Collaboration et al., 2025); after the DES supernova recalibration, analyses report lower significances near 3.2σ to 3.3σ (arXiv:2609.10567, 2026). These are hints, not detections.

Rare objects are probes in their own right. Strong gravitational lenses test mass distributions and give time-delay distances for H_0. Jellyfish galaxies trace ram-pressure stripping. Collisional ring galaxies and mergers constrain assembly histories. Extreme emission-line objects and little red dots test black-hole seeding and early star formation (Matthee et al., 2024). The value of a rare-object search depends on completeness, meaning which rare objects are found, and purity, meaning how many claims are wrong. This paper is about making both measurable.

2.2 Instruments and their failure modes

Each telescope has failure modes that can imitate astrophysical structure. HST ACS and WFC3 imaging carries charge-transfer inefficiency trails, cosmic rays, ghosts and drizzling artifacts at mosaic edges. JWST NIRCam and MIRI (Gardner et al., 2023) have documented artifacts including snowballs, wisps, $1/f$ noise and persistence, which the COSMOS-Web reduction corrects explicitly (Franco et al., 2025). Euclid VIS and NISP (Euclid Collaboration et al., 2025b) contribute ghosts and stray light. Rubin DP1 was taken with the commissioning camera and has a median PSF FWHM of about 1.14 arcsec (Rubin Observatory Team, 2026). Artifacts are tied to an instrument, an exposure or a detector position, while astrophysical sources persist across independent observations. ECAD turns that difference into a formal test.

2.3 Anomaly detection in astronomy

Unsupervised and generative approaches. Storey-Fisher et al. (2021) trained a Wasserstein GAN on nearly one million Hyper Suprime-Cam galaxy images and found the discriminator more attuned to interesting anomalies than the generator. Pearce-Casey et al. (2023) used a conditional GAN to predict JWST NIRCam long-wavelength fluxes and treated poor predictions as anomaly indicators. Pruzhinskaya et al. (2019) applied isolation forests to supernova light curves.

Active and personalised approaches. Astronomaly (Lochner and Bassett, 2021) combines anomaly scoring with active learning so that "interesting" is learned from a few user labels, and has been run on about four million galaxies (Etsebeth et al., 2024). Astronomaly Protege (Lochner and Rudnick, 2025) replaces generic anomaly scores with a regression of user interest on self-supervised features. Active anomaly detection has also been developed for time-domain data (Ishida et al., 2021).

Semi-supervised archive-scale search. AnomalyMatch (Gómez et al., 2026) combines semi-supervised and active learning. It produced the Hubble Legacy Archive catalogue (O'Ryan and Gómez, 2025) starting from a handful of labelled examples and has since been applied to lens discovery in JWST NIRCam imaging (arXiv:2605.03442, 2026).

Zero-shot and vision-language approaches. A 2026 study applied a CLIP-based zero-shot anomaly detector to 1,342 galaxies in CEERS and NGDEEP (arXiv:2610.29868, 2026), and another used vision-language models to recover the reference sample of O'Ryan and Gómez (2025) (arXiv:2610.37747, 2026).

Shared limitation. None of these methods states a calibrated false discovery rate for the set of objects it claims, and none formally tests whether a candidate replicates across artifact-independent views. ECAD sits on top of any of them and uses their scores as inputs.

2.4 Representation learning and foundation models

Self-supervised learning produces embeddings in which nearest-neighbour and density-based anomaly scores can be computed (Hayat et al., 2021), as do general-purpose backbones such as BYOL (Grill et al., 2020) and DINOv2 (Oquab et al., 2024). Zoobot (Walmsley et al., 2023) models fine-tuned on Galaxy Zoo labels produced the Euclid Q1 morphology catalogue of about 378,000 galaxies (Euclid Collaboration et al., 2025d). AstroCLIP aligns galaxy images and spectra with a contrastive loss (Parker et al., 2024); AstroPT is an autoregressive multimodal model (Smith et al., 2024); applied to Euclid Q1, isolation forest and local outlier factor detectors on its embeddings helped isolate rare objects (Euclid Collaboration et al., 2025c). AION-1 extends this to an omnimodal model (Parker et al., 2025), and the Multimodal Universe provides about 100 TB of machine-learning-ready data (Multimodal Universe Collaboration et al., 2024).

These models are optimised to represent the bulk of the data distribution. Deep generative likelihoods can assign high likelihood to simple out-of-distribution inputs (Nalisnick et al., 2019), and embedding distances depend on instrument and depth as well as on astrophysics. For discovery, the question is whether a novelty score is calibrated with respect to observing conditions. ECAD treats the representation as interchangeable.

2.5 Conformal inference, multiple testing and replication

Conformal inference (Vovk et al., 2005; Angelopoulos and Bates, 2023) turns any score into p-values that are valid in finite samples when data are exchangeable. Bates et al. (2023) showed that conformal p-values computed against a shared calibration set of inliers are positively regression dependent on a subset (PRDS), so that the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995) applied to them controls the false discovery rate (FDR). Marandon et al. (2024) extended this to adaptive novelty detection. Weighted conformal prediction handles

Table 1. What ECAD adds to existing anomaly searches.

Component	Prior work	ECAD
Anomaly score	GAN, autoencoder, isolation forest, LOF, active and semi-supervised learning	Any of these, as an interchangeable input
Calibration	Thresholds set by hand, or top-k lists	Conformal p-values per view, stratified by observing conditions, sized to the target prevalence
Error control	Expert-confirmed counts after the fact	False discovery rate fixed in advance (BH; BY as a robustness check)
Artifacts	Visual inspection	Symmetric data-quality gating plus partial-conjunction replication across artifact-independent views
Known or unexplained	Expert classification after the fact	Class-conditional conformal test with a stated error rate
Output	Ranked list with images	Versioned evidence bundle and discovery level per candidate
Evaluation	Retrospective	Pre-registered, with abandonment thresholds, rehearsed in simulation

covariate shift when the likelihood ratio is known or estimable (Tibshirani et al., 2019); Mondrian conformal prediction gives validity within strata (Vovk et al., 2005). E-values and the e-BH procedure (Wang and Ramdas, 2022) and the Benjamini-Yekutieli procedure (Benjamini and Yekutieli, 2001) control FDR under arbitrary dependence at a cost in power. Recent work refines conformal novelty detection further (arXiv:2601.02610, 2026; Jin et al., 2026).

Partial-conjunction testing (Benjamini and Heller, 2008) tests whether at least u of n component hypotheses are false and was developed for replicability analysis. To our knowledge it has not been used to require cross-exposure or cross-instrument replication of astronomical anomaly candidates. That use is the core statistical idea of ECAD. Within a group of views, the Simes combination (Simes, 1986) is valid under the positive dependence that conformal p-values from shared calibration sets exhibit (Sarkar, 1998).

3 Problem formulation

3.1 Sources, views and conditions

Let $\mathcal{S}$ be a set of sources on the sky. Source j is observed in a set of views $\mathcal{V}_j \subseteq \{1, \ldots, K\}$, each an image or spectrum in a particular band, exposure or instrument. For view k we observe $x_j^{(k)}$ together with a condition vector $c_j^{(k)}$ recording instrument, filter, PSF FWHM, depth, exposure time, background, detector position, pipeline version and calibration context.

Views are partitioned into *artifact-independence groups* $G(k)$. Two views belong to different groups when they come from different visits, detectors or instruments, so that no single instrumental event (a cosmic ray, a satellite trail, a persistence image, a ghost or a drizzling defect) can plausibly appear in both. Views taken simultaneously through a shared optical path are kept in one group unless an artifact class is shown to be specific to one of them. Source j has n_j observed groups. The number n_j depends only on survey coverage, never on image content.

3.2 Null hypotheses

For source j and view k we define three nested nulls.

- $H_{0,j}^{\text{typ},(k)}$ (typicality): conditional on the stratum $g(c_j^{(k)})$, view $x_j^{(k)}$ is exchangeable with a reference population of views of typical sources observed in the same stratum.
- $H_{0,j}^{\text{art}}$ (artifact): fewer than u artifact-independent groups of source j contain a non-typical view. A source whose only departure from typicality is an instrumental event in one group satisfies this null.
- $H_{0,j}^{\text{known}}$ (known class): source j belongs to a class in a pre-specified catalogue $\mathcal{C}$, such as strong lens, merger, ring galaxy, jellyfish galaxy, AGN, star, or a known artifact type.

3.3 Per-view conformal p-values

Let f_k be a fixed encoder and $s_k(\cdot)$ a nonconformity score, for example the mean distance to the m nearest neighbours in a reference set. For stratum g, let $\mathcal{D}_{k,g}^{\text{cal}}$ be a disjoint calibration set of n_g reference views. Then

$$p_j^{(k)} = \frac{1 + \#\{i \in \mathcal{D}_{k,g}^{\text{cal}} : s_k(x_i) \geq s_k(x_j^{(k)})\}}{n_g + 1}. \tag{1}$$

Under $H_{0,j}^{\text{typ},(k)}$, with exchangeability inside the stratum, $\Pr\left(p_j^{(k)} \leq t\right) \leq t$ for all $t \in [0, 1]$ (Vovk et al., 2005; Bates et al., 2023).

3.4 Replication and source-level testing

Within each group, the view p-values are combined into one group p-value $\tilde{p}_j^{(1)}, \ldots, \tilde{p}_j^{(n_j)}$ (Section 4.5). The partial-conjunction p-value for "at least u of n_j groups are non-typical" is, with the Bonferroni combination of Benjamini and Heller (2008),

$$p_j^{u/n_j} = \min\left\{1,\ (n_j - u + 1)\,\tilde{p}_{j,(u)}\right\}, \tag{2}$$

where $\tilde{p}_{j,(u)}$ is the u-th smallest group p-value. This is valid under arbitrary dependence between groups. ECAD uses $u = 2$. Sources with $n_j < u$ cannot reach the replication level and are reported separately. Given a target level q, BH applied to $\{p_j^{u/n_j}\}_j$ gives the rejection set $\mathcal{R}_q$.

Scope of the guarantee. BH controls the expected fraction of false rejections among sources whose partial-conjunction null is true. For $u = 2$ that set includes every typical source with an artifact in at most one group. It does not include a typical source with independent artifacts in two groups: such a source is non-typical in two groups, so rejecting it is statistically correct and scientifically wrong. Section 4.3 reduces this case with symmetric data-quality gating, and the simulation study measures what is left.

Theoretical status. BH control for conformal p-values with a shared calibration set follows from PRDS (Bates et al., 2023). It is not established that PRDS survives the within-group combination and the partial-conjunction transformation. ECAD therefore reports Benjamini-Yekutieli (BY) results as a guaranteed but conservative alternative, and checks BH empirically, both in simulation (Section 5) and in the pre-registered null tests (Section 7).

3.5 The resolution and contamination conditions

Two conditions limit what any conformal search of this kind can find. Both follow from Equation 1 and the BH threshold, and the simulation study confirms both (Section 5).

Resolution. A conformal p-value cannot be smaller than $1/(n_g + 1)$, and the partial-conjunction transform multiplies it by up to $\kappa = n_j - u + 1$. BH can reject R hypotheses out of N only if the R smallest p-values fall below qR/N. A search can therefore make R discoveries at level q only if

$$\frac{\kappa}{n_g + 1} \leq \frac{qR}{N} \quad \Longleftrightarrow \quad n_g + 1 \geq \frac{\kappa}{q\,\pi}, \qquad \pi = R/N. \tag{3}$$

The calibration size each stratum needs is set by the prevalence π of the objects one hopes to find, not by the size of the survey. The condition is necessary, not sufficient: it assumes that every anomaly attains the smallest possible p-value. Version 1.0 of this framework required $n_g \geq 1{,}000$, which at $q = 0.1$ and $\kappa = 2$ cannot support prevalences below 2%.

Contamination. Suppose a fraction ε of the reference views in a stratum are themselves anomalies, with the same score distribution as the anomalies in the test set. The k-th most extreme of R test anomalies then has about $\varepsilon\, n_g\, k/R$ reference views scoring above it, so its p-value is about $\varepsilon k/R$, and after the partial-conjunction transform about $\kappa\varepsilon k/R$. BH rejects it only if this is below qk/N, which requires

$$\varepsilon \lesssim \frac{q\,\pi}{\kappa}. \tag{4}$$

The same argument applies to artifacts in the reference that score as high as anomalies. Unlike resolution, contamination cannot be fixed by adding reference data: a larger reference drawn from the same survey has the same ε. In particular, a survey cannot serve as its own reference, because then $\varepsilon = \pi$. Equations 3 and 4 together say that each reference stratum must be large enough to hold at least $\kappa/(q\pi)$ views and clean enough to hold fewer than one contaminant per $\kappa/(q\pi)$ views.

3.6 Discovery levels

Table 2. Discovery levels. Physical interpretation is asserted only at L4.

Level	Name	Criterion
L0	Flag	High nonconformity score in at least one view, no error control
L1	Statistical novelty	$j \in \mathcal{R}_q$ using $u = 1$
L2	Replicated anomaly	$j \in \mathcal{R}_q$ using $u \geq 2$ artifact-independent groups
L3	Unexplained after review	L2, the class-conditional test rejects every known class, and two independent reviewers agree it is neither an artifact nor a known class
L4	Physically confirmed	L3, or a known-rare L2 object, with an independent physical measurement (for example a spectroscopic redshift or line diagnostics) supporting the interpretation

ECAD outputs may be called "candidates" at L1 and L2 and "unexplained" only at L3.

3.7 Objective

The objective is to maximise the expected number of true L2 discoveries subject to FDR $= \mathbb{E}\big[\,|\mathcal{R}_q \cap \mathcal{H}_0| / \max(|\mathcal{R}_q|, 1)\,\big] \leq q$, under a fixed budget of expert review hours. A secondary objective is the artifact pass-through rate: the probability that an instrumental artifact confined to one group reaches L2.

4 The ECAD framework

4.1 Overview

ECAD has seven stages (Figure 1): (1) cutout and metadata extraction with full provenance; (2) symmetric data-quality gating; (3) representation and scoring for each view; (4) stratified conformal calibration, sized by the resolution condition; (5) within-group combination, replication testing and FDR control; (6) the known-versus-unexplained test and human review; (7) evidence-bundle generation. Stages 3 and 4 are interchangeable modules. Stages 2, 5, 6 and 7 are the methodological core.

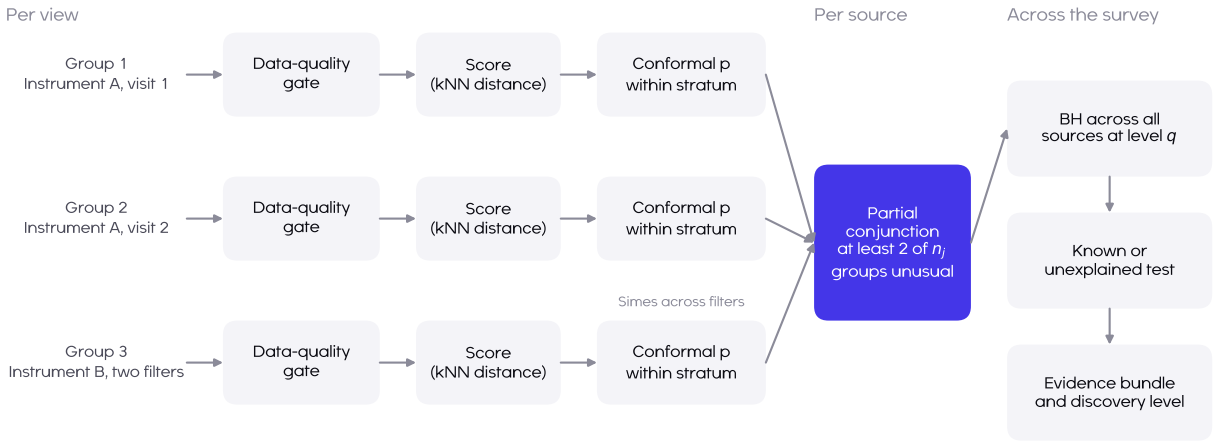


Figure 1. The ECAD pipeline for one source observed in three artifact-independent groups. Each view is gated, scored and turned into a conformal p-value against a reference population observed under the same conditions. Views inside a group are combined, groups are combined by a partial-conjunction test that requires at least two of them to be unusual, and BH controls the false discovery rate across all sources. Every candidate carries an evidence bundle.

4.2 Representations and scores

The primary configuration for real data uses three encoders: a domain foundation model (Zoobot for optical and near-infrared morphology, with AstroCLIP or AION-1 embeddings where spectra exist), a general-purpose self-supervised backbone (DINOv2), and an optional cross-instrument contrastive fine-tune in which cutouts of the same source from different instruments are positive pairs. Whether the fine-tune makes scores less dependent on observing conditions is tested by ablation, not assumed. The primary score is the mean distance to the $m = 10$ nearest reference neighbours; isolation forest (Liu et al., 2008), local outlier factor (Breunig et al., 2000), Deep SVDD (Ruff et al., 2018) and external scores such as AnomalyMatch probabilities can be calibrated in the same way.

4.3 Symmetric data-quality gating

Every reference population drawn from a real survey contains artifacts. If they stay in the calibration set they occupy its extreme tail, and the p-value of a genuine anomaly can be no smaller than the fraction of calibration artifacts that score above it. If they are removed with a rule that is not applied to test views, exchangeability fails. ECAD resolves this with one rule applied identically to both.

A data-quality rule D maps a view, and optionally the other views of the same source, to pass or fail. It may use pipeline flags, cross-exposure cosmic-ray masks, Laplacian sharpness tests for features narrower than the PSF (van Dokkum, 2001), and PSF-matched difference imaging between visits in the same filter (Alard and Lupton, 1998). It must not use the anomaly score. ECAD then (i) builds each calibration set only from views that pass D, and (ii) assigns $p_j^{(k)} = 1$ to every test view that fails D, so that a failed view carries no evidence. The group count n_j is unchanged.

Validity. Let $\mathcal{P}$ be the population of typical-source views that pass D. A test view that passes D and satisfies the typicality null is exchangeable with the calibration views, because both are draws from $\mathcal{P}$ in the same stratum, so Equation 1 remains valid. A test view that fails D has $p = 1$, which is trivially valid. Gating can only lower power, never inflate false discoveries, and it is valid whether or not D is accurate. An accurate D recovers most of the power lost to a contaminated

reference population and removes most coincident artifacts before they reach the replication test. In simulation, a gate that caught only half of all artifact views restored L2 recall from 0.063 to 0.155, against 0.185 with an artifact-free reference.

4.4 Calibration stratified by observing conditions, and its size

Exchangeability fails when calibration and test views are observed under different conditions. ECAD uses Mondrian stratification: separate calibration sets per stratum g, defined by instrument and filter, depth bins of 0.5 mag, PSF FWHM tertiles and, where relevant, detector region. Validity then holds within each stratum by construction.

Design rule for calibration size. Equation 3 replaces the fixed minimum of version 1.0. Before any confirmatory data are opened, the protocol fixes the smallest prevalence $\pi_{\min}$ the search must be able to detect, and requires in every stratum

$$n_g \;\geq\; c\,\frac{\kappa}{q\,\pi_{\min}}, \qquad \kappa = \max_j\,(n_j - u + 1), \quad c = 20. \tag{5}$$

The safety factor c comes from the simulation study, where recall first appeared at about five times the necessary condition and saturated by about twenty times it (Figure 7). Strata that cannot meet the rule are merged with neighbours. If that still fails, the search reports that it cannot reach $\pi_{\min}$ in those strata, rather than lowering the bar silently. The rule also disciplines stratification itself: every extra stratification axis divides the calibration population, and in simulation adding a brightness axis to the strata removed every L2 discovery. Strata are therefore defined by the conditions that change the score's extreme tail most, and no more.

Reference purity, and how to certify it. Equation 4 requires the contamination of each reference stratum to stay below $q\pi_{\min}/\kappa$, counting both anomalies and any artifacts that survive gating with high scores. Symmetric gating handles artifacts. Anomalies have to be kept out by building references from fields that have been inspected, for example sources classified as ordinary by Galaxy Zoo volunteers or by expert review, and the contamination has to be measured, not assumed. The protocol certifies purity with a random audit: if m randomly chosen reference views are inspected and none is anomalous, the 95% upper bound on ε is about $3/m$, so certifying $\varepsilon \leq q\pi_{\min}/\kappa$ needs $m \approx 3\kappa/(q\pi_{\min})$ inspections with none found. For $\pi_{\min} = 10^{-3}$ that is about 6,000 views, within reach of a small team or a citizen-science project; for $\pi_{\min} = 10^{-5}$ it is about 600,000, which only a large volunteer effort could provide. These two numbers bound what an error-controlled search can claim: objects rarer than the certified $\pi_{\min}$ can still be flagged at L0, but not claimed at L1 or above.

Remediation for shift. If the stratified null test fails, the pre-registered remedy is weighted conformal calibration (Tibshirani et al., 2019) with likelihood ratios $w(c)$ estimated by a probabilistic classifier on condition vectors, held to the same empirical null test. A second remedy addresses population shift between fields: calibrating each field against reference views drawn from a population matched to that field, which in simulation reduced false rejections of typical sources by a factor of four under a moderate shift.

4.5 Combining views inside a group

Views inside one group, such as simultaneous exposures in two filters, are combined with the Simes rule (Simes, 1986), $\tilde{p} = \min_i\{v\,p_{(i)}/i\}$ over the v available views. Conformal p-values computed against shared calibration sets are positively dependent, and Simes is valid under that dependence (Sarkar, 1998). Version 1.0 used a Bonferroni minimum, which is valid under any dependence but costs power when an anomaly is visible in more than one filter; in simulation the gain from Simes was small but consistent.

4.6 Known-rare versus unexplained

Version 1.0 separated known from unexplained objects with a split-conformal prediction set over $\mathcal{C} \cup \{\text{other}\}$. That rule is ill-posed: there are no training examples of "other", and a classifier forced to choose among known classes can place a genuinely new object confidently in one of them. ECAD now uses a class-conditional conformal novelty test. For each known class $c \in \mathcal{C}$, including "typical" and each known artifact type, a nonconformity score $s_c(x)$ (the mean distance to the m nearest labelled members of c) is calibrated on held-out labelled members of c, giving $p_c(x)$. A candidate is called unexplained at level α when

$$\max_{c \in \mathcal{C}} p_c(x) \leq \alpha. \tag{6}$$

For a true member of class c, $\Pr(p_c(x) \leq \alpha) \leq \alpha$, so the probability that a known-class object is wrongly called unexplained is at most α for every class, with no assumption about what unknown objects look like. The test runs on gated views: in simulation, artifacts left in the views raised the known-class error from at most 5.3% to as much as 6.9% at $\alpha = 0.05$. Labels come from the Euclid Q1 Zoobot catalogue, Galaxy Zoo, the Hubble Legacy Archive anomaly catalogue, published lens catalogues and a curated set of known artifacts.

4.7 Human review that preserves the guarantees

Active learning improves relevance (Lochner and Rudnick, 2025; Gómez et al., 2026), but if human-labelled examples enter calibration sets, exchangeability breaks. ECAD separates four pools, fixed by sky region before any labelling: a training pool for representation learning, an active-learning pool for scientific interest, a calibration pool used only for p-values and never shown to reviewers, and the confirmatory test region. The interest model can reorder the review queue within $\mathcal{R}_q$ but cannot change which sources enter it, so the FDR statement refers to novelty, not interest. Two astronomers review each L2 candidate independently; in half of the batches they are blinded to the model score so that anchoring can be measured; disagreements go to a third reviewer, and Cohen's κ is reported.

4.8 The evidence bundle

Every candidate at L1 or above receives a machine-readable bundle (JSON with linked FITS cutouts), identified by a content hash, containing: provenance (archive identifiers, pipeline and calibration versions, download times); all views with condition vectors, groups and data-quality outcomes; per-view scores and p-values with stratum and n_g; group p-values, the partial-conjunction p-value, the BH threshold at release and the BY result; injection-recovery completeness for the candidate's stratum and anomaly class; the 20 nearest reference neighbours and catalogue cross-matches; the class-conditional p-values; per-exposure replication results; the review log with blinding status; and the current discovery level with its history. Anyone with the archive data and the released code can regenerate a bundle and obtain identical statistics.

5 Simulation study

Before any real data are opened, we built a synthetic sky in which the truth is known, ran every component of ECAD on it, and checked each guarantee the method claims. The study had two purposes: to confirm that the statistical machinery behaves as the theory says when its assumptions hold, and to find the places where version 1.0 of the framework failed. It found three failures and one fundamental limit, and Sections 3 and 4 already incorporate what it taught us. Everything in this section is synthetic. None of it is evidence about real astronomical objects.

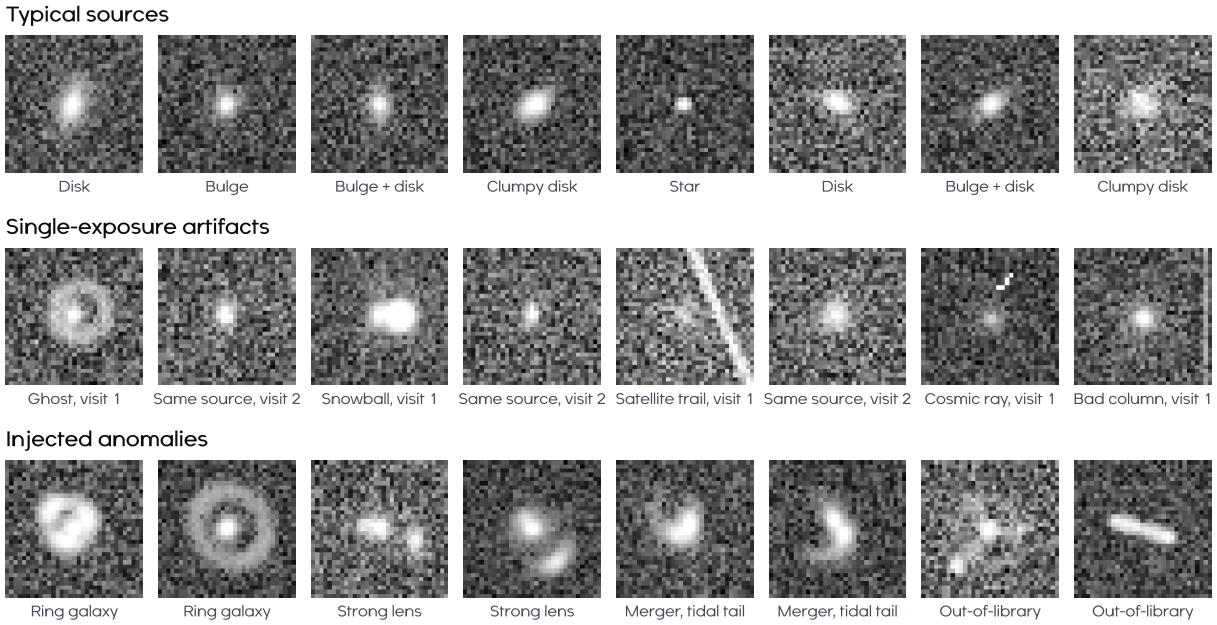


Figure 2. The synthetic sky. Top: typical sources. Middle: the five artifact families; for three of them the same source is shown in its second visit, where the artifact is absent. Bottom: the four injected anomaly classes. All cutouts are instrument-A views on an arcsinh stretch.

5.1 The synthetic sky

Sources. Each source is a 33×33 pixel cutout rendered on a $2\times$ supersampled grid (Figure 2). Typical sources are disks (Sérsic index 0.7 to 1.5), bulges (2.5 to 4.5), bulge plus disk composites, clumpy disks with one to three star-forming knots, and stars, in proportions of 38, 20, 17, 10 and 15% (Sérsic, 1963; Ciotti and Bertin, 1999). Effective radii are log-normal around 3 pixels, and 12% of fields contain a faint neighbour, so the typical population includes crowding. Bulges, disks and knots carry different colours, so a galaxy's appearance changes between instruments as real galaxies do. Typical signal-to-noise in a 50-pixel aperture is log-normal with median 40, from 10 to 1,500.

Anomalies. Four classes are injected: collisional ring galaxies (a nucleus inside a modulated, possibly offset ring); strong gravitational lenses (a singular isothermal sphere with external shear, ray-traced from a background Sérsic source, Schneider et al. 1992); mergers (two nuclei with a tidal tail along a logarithmic spiral); and an out-of-library class (spoked rings and straight chains of knots) that plays the role of a genuinely new kind of object. Anomaly signal-to-noise is log-uniform from 15 to 1,000. Each anomaly is rendered from one intrinsic model into every view, so it replicates the way a real object does.

Views and conditions. Each source has four views in three artifact-independence groups: two visits with instrument A (an optical camera; groups 1 and 2) and one visit with instrument B (a near-infrared camera with two simultaneous filters; group 3). Instrument B covers 70% of sources, as overlapping surveys do. Each visit draws its depth (deep or shallow, with noise 2.5 times higher when shallow) and its PSF (a good or a poor FWHM bin) independently, which gives 12 strata. PSFs and noise are Gaussian.

Artifacts. Every view independently has a 2% chance of one of five artifacts: a cosmic ray (a sharp track of one to five pixels), a satellite trail, a ghost (a faint offset ring that can mimic a ring galaxy), a snowball (a bright round blob with a halo) or a bad column. Because artifacts are drawn independently per view, coincidences in two groups occur at the rate they would in a real survey with these artifact rates.

Data sets. The generator produced 116,800 sources: two sets of 8,000 unlabelled survey sources

(with anomalies at 0.5% and artifacts at the natural rate), one to fit the encoder normalisation and one to train the baselines; 90,000 typical sources, of which 20,000 form the nearest-neighbour reference and the other 70,000 are split at random into a calibration population of 60,000 and a test set of 10,000 in every repetition; 6,000 anomalies (1,500 per class); and a labelled catalogue of 4,800 sources for the known-class test. Every source is drawn from its own seeded generator and can be regenerated exactly.

5.2 Encoder, score and data quality

The encoder fits a smooth elliptical model to each view (the mean surface brightness in elliptical annuli set by the image's second moments) and describes what the model fails to explain: the noise-debiased significance of the residual in three radial zones, the fraction of flux in significant residuals, the number of significant residual peaks, the significance of any outward rise in the radial profile (a ring signature), the amplitudes of the first four rotational harmonics of the residual, and the axis ratio, concentration, size, brightness and centroid offset. Every feature is either a significance or a dimensionless shape statistic, and all are invariant to rotation. The score is the mean Euclidean distance to the ten nearest of the 40,000 instrument-A reference views (or 14,000 per instrument-B filter) (Pedregosa et al., 2011). We also tried 24 principal components of the stretched cutout plus classical morphology statistics (Conselice, 2003; Abraham et al., 2003; Lotz et al., 2004); that encoder separated anomalies less well in the extreme tail and was not used. ECAD's guarantees do not depend on the encoder; its power does.

The data-quality rule D has two parts, neither of which uses the anomaly score. A Laplacian sharpness test flags features narrower than any PSF (van Dokkum, 2001). PSF-matched difference imaging between the two instrument-A visits flags a significant positive residual in either visit (Alard and Lupton, 1998). D fails 0.04% of artifact-free instrument-A views of typical sources and none of the artifact-free anomaly views. It catches 99% of cosmic rays and 100% of instrument-A snowballs, but only 21% of ghosts and between 32 and 41% of bad columns, and it cannot see satellite trails, ghosts or snowballs in instrument B, which has no second visit. Overall it catches 51% of artifact views, deliberately short of perfect.

5.3 Configurations

We compare three configurations (Table 3). In each repetition the test set holds 10,000 typical sources and 300 anomalies, 75 per class, a prevalence of 2.9%. This prevalence is far higher than real searches face; we chose it so that the guarantees could be measured with small Monte Carlo error, and we examine lower prevalences separately (Section 5.4). Main results average 200 repetitions; ablations average 50 to 100.

Table 3. Configurations compared in the simulation study.

Name	Description
v1.0-ideal	Version 1.0 as written, with a calibration population free of artifacts, which no real survey provides
v1.0-real	Version 1.0 with a realistic calibration population that carries artifacts at the natural rate
v1.1	Realistic calibration population, symmetric data-quality gating, Simes combination within groups

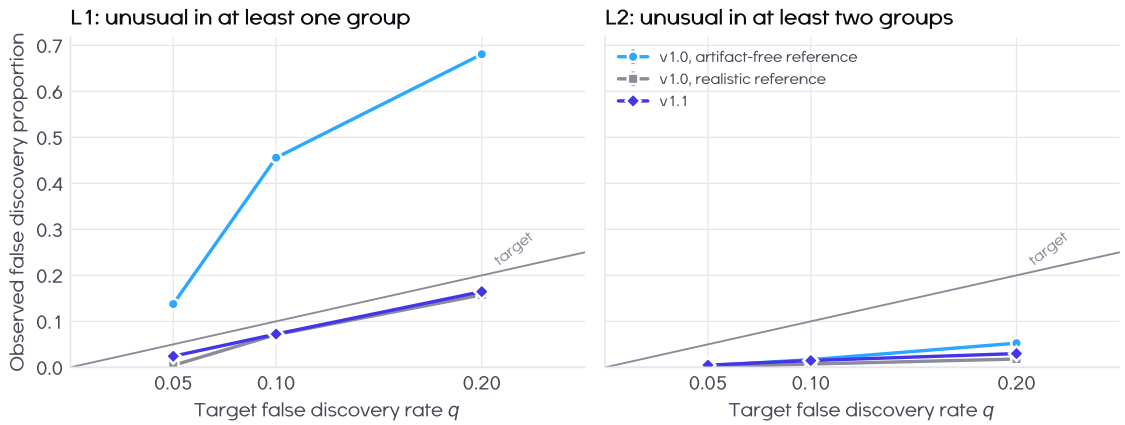


Figure 3. Observed false discovery proportion against the target, counting every rejected non-anomaly as false (200 repetitions; error bars are 95% intervals and are mostly smaller than the markers). Left: single-group evidence (L1). Right: replicated evidence (L2).

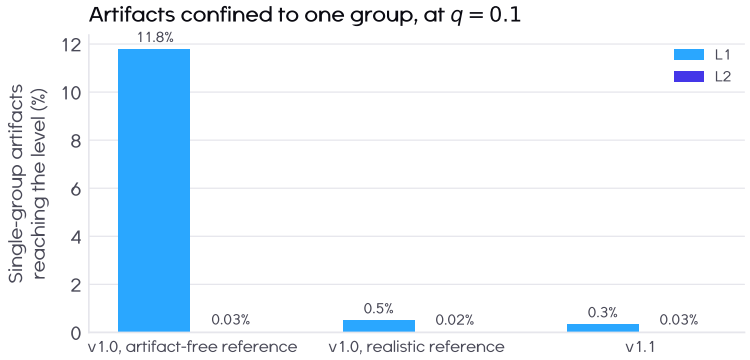


Figure 4. Fraction of sources with an artifact confined to one group that reach each level, at $q = 0.1$.

5.4 Results

The false discovery rate is controlled at L2, and replication is what makes it meaningful. Figure 3 shows the observed false discovery proportion (FDP) against the target q. Here a false discovery is any rejected source that is not an injected anomaly, whatever the reason. At L2 every configuration stays well below the target: for v1.1 the mean FDP is 0.005, 0.015 and 0.030 at $q = 0.05$, 0.10 and 0.20. At L1 without gating, v1.0-ideal reaches 0.14, 0.46 and 0.68. This is not a failure of the L1 guarantee, which concerns typicality: artifact views are genuinely atypical, and the false rejections of artifact-free typical sources at L1 stay at 2.8%, 7.0% and 13.5% of all rejections, below q each time. It shows instead that a guarantee about typicality is not a guarantee about science, which is why ECAD requires replication before it calls anything a candidate. Across 200 repetitions with no anomalies at all, L2 never made a single rejection at any q, in either configuration tested (v1.0-ideal and v1.1). The conservative Benjamini-Yekutieli procedure made no rejections in any setting, so at these calibration sizes it is a safety check, not a working alternative.

Replication removes single-group artifacts almost completely. At $q = 0.1$, 11.8% of sources whose only problem is an artifact in one group reach L1 under v1.0-ideal; 0.03% reach L2 (Figure 4). With gating, L1 pass-through falls to 0.33% and L2 stays at 0.03%. Each test set contains on average 13.7 typical sources with independent artifacts in two groups, the case that lies outside the partial-conjunction guarantee. Under v1.0-ideal, 0.01 of them per run reach L2 at $q = 0.1$ and 0.66 at $q = 0.2$; under v1.1, none did in 200 runs.

Power depends on brightness and on the class, and replication costs some of it. Overall L2

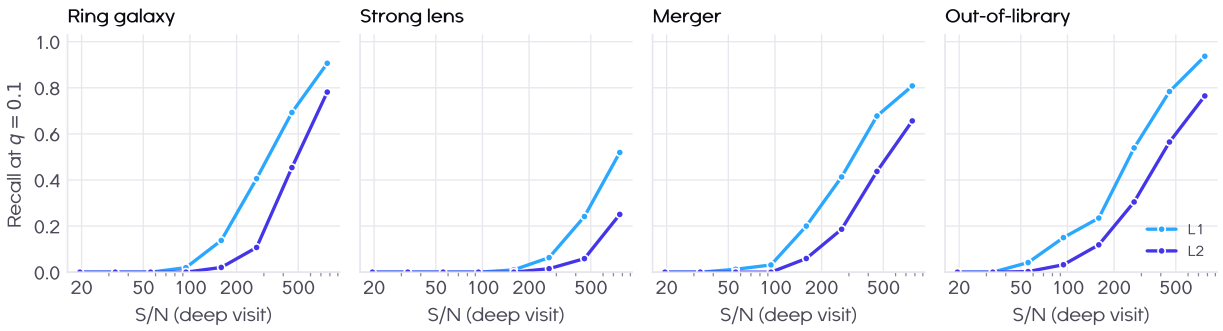


Figure 5. Recall at $q = 0.1$ as a function of signal-to-noise in a deep visit, for v1.1 (40 repetitions, 12,000 injected anomalies).

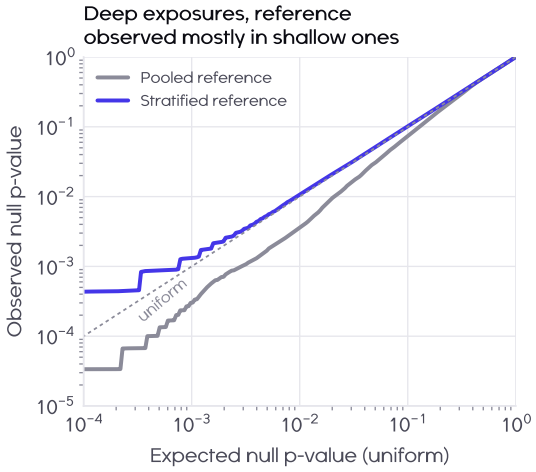


Figure 6. Null p-values of artifact-free deep views of typical sources, when the calibration population is observed mostly in shallow conditions (100 repetitions). Points below the dashed line are anti-conservative. The stratified curve flattens at small p-values because the deep strata are small, which is the resolution floor.

recall at $q = 0.1$ is 0.187 for v1.0-ideal and 0.153 for v1.1. Recall is driven by signal-to-noise (Figure 5): below a deep-visit S/N of 100 almost nothing is found, and above 250 v1.1 finds 51% of ring galaxies, 50% of mergers and 59% of out-of-library objects at L2. Strong lenses are the exception, at 12%: with Einstein radii of 4 to 8 pixels and PSFs of about 2 pixels, many arcs blend into the deflector, and the encoder does not separate them from bright, slightly irregular galaxies. Replication costs recall in every class, because an anomaly must be extreme in two groups, one of which is often shallow. That cost buys the artifact suppression above.

Stratification protects calibration when observing conditions differ. To test what happens when the calibration population is observed under different conditions from the test field, we kept 85% of shallow and 15% of deep calibration views, while the test field stayed half deep and half shallow. The score's extreme tail depends on depth: its 99.9th percentile is 4.6 for deep views and 3.5 for shallow ones, because faint structure such as neighbours and clumps becomes significant in deep data. With a pooled reference, null p-values of deep views are anti-conservative (Figure 6): 0.28% fall below 0.001 and 2.1% below 0.01, about 2.8 and 2.1 times the nominal rates. The false rejection rate of artifact-free typical sources at L1 rises to 0.112 at $q = 0.1$, above the target, and at L2 with $q = 0.2$ the FDP is 0.107, against 0.031 when stratified, with six times as many false rejections of artifact-free typical sources. With stratification, deep-view null p-values sit at or below their nominal rates (0.08% below 0.001, 0.93% below 0.01). The cost is visible too: with only 15% of the deep calibration views kept, the deep strata are small, the floor $1/(n_g + 1)$ is reached, and L2 makes almost no discoveries at $q = 0.1$, which leads directly to the next result.

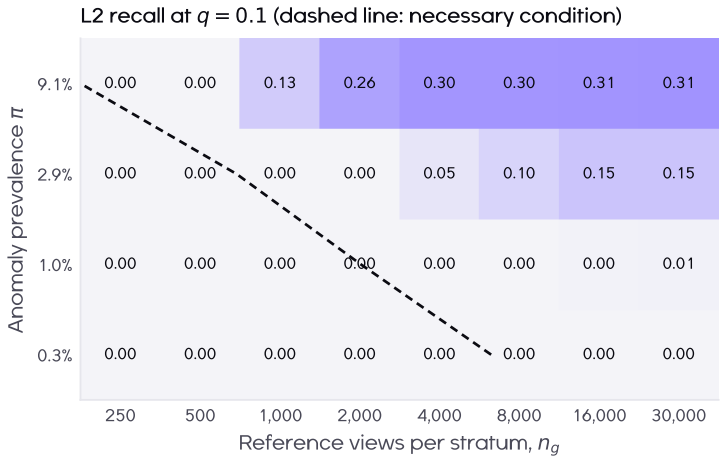


Figure 7. L2 recall at $q = 0.1$ against calibration size per stratum and anomaly prevalence (v1.1, 30 repetitions per cell). The dashed line is the necessary condition of Equation 3 with $\kappa = 2$. Calibration sizes above about 10,500 are capped by the size of the instrument-B strata.

The resolution condition is necessary, and in practice far too optimistic. We varied the calibration size per stratum from 250 to 30,000 and the prevalence from 0.3% to 9.1% (Figure 7; 30 repetitions per cell). No discoveries were made anywhere to the left of the boundary from Equation 3, as the theory requires. Recall first appears at about five times the boundary and saturates by about twenty times it: at 9.1% prevalence, recall is 0.13 at $n_g = 1{,}000$ and 0.30 from 4,000; at 2.9%, it is 0.05 at 4,000 and 0.15 from 16,000. The boundary assumes that every anomaly attains the smallest possible p-value; real anomalies spread across the tail, so only a fraction reaches it. At 1.0% prevalence even 30,000 views per stratum were not enough, because the instrument-B strata could hold at most about 10,500 views. Version 1.0's requirement of $n_g \geq 1{,}000$ would have made no discoveries at all at the prevalences real searches face. Adding a brightness axis to the strata (36 instead of 12) had the same effect: at $q = 0.1$, L2 recall fell from 0.152 to zero, because every stratum became three times smaller. The design rule in Section 4.4 therefore includes a safety factor of twenty.

What the calibration population contains matters as much as its size. With a realistic calibration population carrying artifacts at the natural 2% rate, L2 recall at $q = 0.1$ falls from 0.185 to 0.063 (Figure 8, left), because calibration artifacts occupy the extreme tail and put a floor under the p-values of genuine anomalies. Symmetric gating restores it to 0.155 while keeping FDR control, even though the gate catches only half of the artifact views. Simes combination within the instrument-B group adds a small gain over a Bonferroni minimum (0.155 against 0.149). Anomalies in the calibration population are more damaging still (Figure 8, right): recall is unchanged at a contamination of 0.01%, drops to 0.154 at 0.03%, to 0.052 at 0.1%, and to zero by 0.3%, in line with the contamination condition of Section 3.5, which puts the limit at about $q\pi/\kappa \approx 0.15\%$ here.

Self-calibration does not work. A tempting way to obtain the very large calibration sets that low prevalences require is to calibrate the survey against itself, for instance by splitting the sky in two and letting each half calibrate the other. The contamination condition rules this out: the reference then contains anomalies at exactly the prevalence being sought, so $\varepsilon = \pi > q\pi/\kappa$. In simulation, contaminating the reference at about a third of the test prevalence (1%) already removed every discovery. We considered this design for this version and removed it.

The class-conditional test says "unexplained" when it should, at a stated error rate. The prediction-set rule of version 1.0 never called anything unexplained at $\alpha = 0.05$: not a single known-class object, and not a single out-of-library object (Figure 9). Even at $\alpha = 0.1$ it called 4.5% of out-of-library objects unexplained. The class-conditional test of Section 4.6 called 5.1% of ring galaxies, 3.9% of lenses and 5.3% of mergers unexplained at $\alpha = 0.05$, consistent with the guarantee given

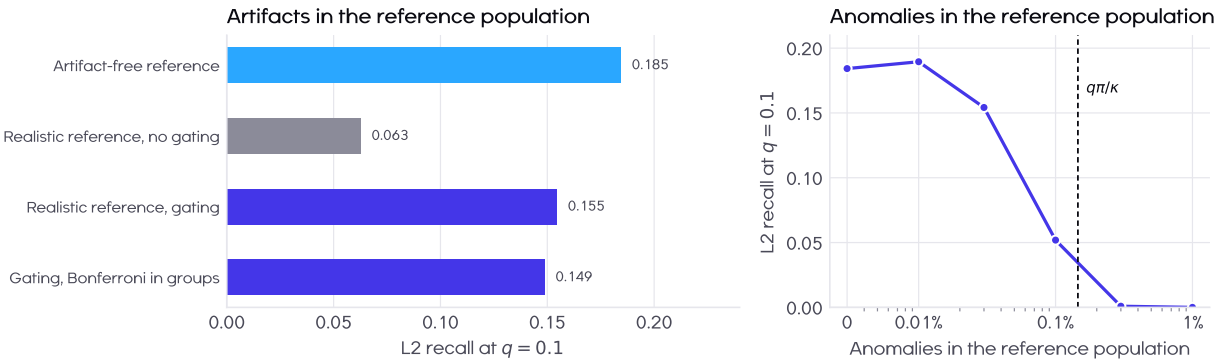


Figure 8. L2 recall at $q = 0.1$ for different calibration populations. Left: artifacts at the natural 2% rate, with and without symmetric gating (100 repetitions). Right: an artifact-free calibration population contaminated with anomalies (60 repetitions per point); the dashed line is the contamination condition $q\pi/\kappa$.

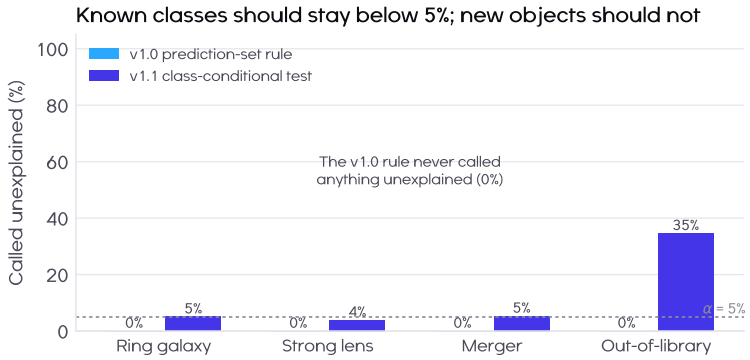


Figure 9. Fraction of each class called unexplained at $\alpha = 0.05$ (1,500 objects per class, artifact-free views). Known classes should stay near 5%; out-of-library objects should not.

the Monte Carlo error of about 0.6%, and 35% of out-of-library objects. When test views still carried artifacts, the known-class error rose to between 5.7% and 6.9%, so the test must run on gated views.

What an expert would inspect is purer. At a fixed review budget of the 100 highest-ranked sources (Figure 10), an isolation forest on a single view returns 59% real anomalies and a local outlier factor 63%. The same score as ECAD, using the best single view with gating and stratified calibration but no group combination, returns 88%; ECAD's L1 ranking 90%; and its L2 ranking 95%. The controlled L2 set at $q = 0.1$ held on average 48 sources, 98.8% of them real anomalies.

A moderate population shift between fields did not break FDR control. When the test field held three times as many clumpy disks and twice as many composites as the calibration population, L2 FDP at $q = 0.1$ was 0.015 and the false rejection rate of artifact-free typical sources at L1 was 0.042, both below target. Calibrating against reference views drawn from a population matched to the field cut L2 false rejections of artifact-free typical sources from 0.62 to 0.16 per run, at a cost in recall from a smaller reference (0.159 to 0.107). A larger shift could break control, so field-matched calibration stays in the protocol as the first remedy.

5.5 Dress rehearsal of the pre-registered criteria

Table 4 applies the criteria of Section 7 to the simulation. Two criteria fail, and we report them as failures.

P3 fails for strong lenses because the encoder does not represent arcs well at this resolution; the protocol keeps the criterion, and a lens-aware encoder is among the encoders the confirmatory

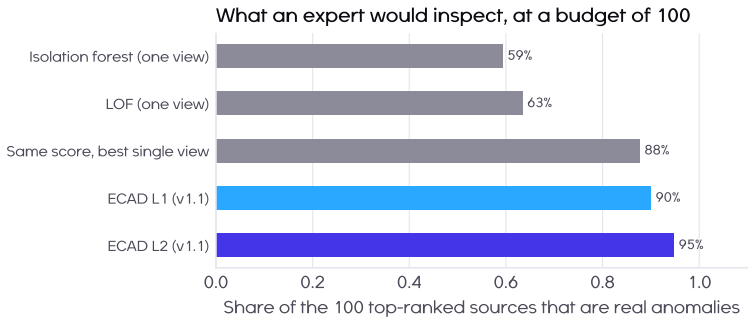


Figure 10. Share of the 100 highest-ranked sources that are real anomalies (40 repetitions).

Table 4. The pre-registered criteria applied to the simulation (v1.1). Simulation analogues are noted where the real criterion cannot apply.

ID	Simulation outcome	Result
P1	Mean L2 FDP 0.005, 0.015, 0.030 at q = 0.05, 0.10, 0.20 (limits 0.07, 0.12, 0.22); assessed over all strata together	Pass
P2	Single-group artifacts reaching L2: 0.03%, against 0.33% at L1 (91% lower)	Pass
P3	L2 recall above a deep-visit S/N of 250 (stand-in for the 5σ limit): rings 0.51, mergers 0.50, out-of-library 0.59, lenses 0.12	Fail (lenses)
P4	External recovery has no simulation analogue	n/a
P5	Purity at a budget of 100: 94.8% for L2 against 90.0% for ECAD-L1 (+4.8 points) and 59 to 63% for single-view baselines (+31 to +35)	Fail (vs ECAD-L1)
P6	No anti-conservative stratum in 50 runs; a two-sided test flagged 10% of runs, all for conservative p-values caused by calibration contamination	Pass (one-sided)
P7	Known classes called unexplained at $\alpha = 0.05$: at most 5.3% (6.9% with artifacted views)	Pass

study will compare. P5 fails against ECAD-L1, not against external methods: in simulation, gating already removes enough artifacts that replication adds about 5 points of purity at this budget, not 10. Real artifacts are more varied than simulated ones and gating will be less complete, so the confirmatory test remains informative either way, and we keep the threshold. P6 was refined from a two-sided to a one-sided test, because contamination makes calibration conservative, which is acceptable, while the criterion exists to catch anti-conservative calibration.

5.6 What the simulation does and does not show

It shows that, when its assumptions hold, ECAD controls the false discovery rate at the replicated level, suppresses single-group artifacts by more than two orders of magnitude, makes no discoveries when there is nothing to find, and gives the known-class test its stated error rate. It shows that three parts of version 1.0 failed: the calibration size requirement, the treatment of a realistic calibration population, and the rule for declaring an object unexplained. And it exposes a limit that no amount of calibration data removes: the reference must be cleaner than $q\pi/\kappa$. It does not show how ECAD will perform on real data. The simulated PSFs, noise, artifacts and galaxies are simpler than real ones, the encoder is deliberately basic, and the prevalences are far higher than those of genuinely new objects. Those questions belong to the pre-registered confirmatory study.

6 Data for the confirmatory study

All primary data are public. Access conditions are noted where they apply.

Hubble Space Telescope. The Hubble Legacy Archive source cutouts used by O'Ryan and Gómez (2025) (99.6 million ACS F814W cutouts) and their catalogue of 1,176 anomalies, published on Zenodo, serve as an external positive reference set and a single-view benchmark. Hubble Legacy Fields GOODS-S mosaics (Whitaker et al., 2019) and COSMOS ACS F814W imaging (Koekemoer et al., 2007), from MAST, provide the HST views in the anchor fields.

JWST. COSMOS-Web (Casey et al., 2023) covers 0.54 deg^2 with NIRCam in F115W, F150W, F277W and F444W, plus 0.19 deg^2 with MIRI F770W; reduced mosaics are public (Franco et al., 2025). JADES provides NIRCam and MIRI mosaics in GOODS-S and GOODS-N, and its NIRSpec spectroscopy (D'Eugenio et al., 2025) provides the spectroscopic views needed for physical confirmation. CEERS (Finkelstein et al., 2023), in the Extended Groth Strip, is an additional held-out field. Per-exposure calibrated files from MAST are used for exposure-level replication and for data-quality gating.

Euclid. The Q1 release covers 63.1 deg^2 over the three Euclid Deep Fields with VIS and NISP imaging, NISP spectroscopy, catalogues and masks (Euclid Collaboration et al., 2025a). The Q1 Zoobot morphology catalogue (Euclid Collaboration et al., 2025d) supplies known-class labels. Whether the Euclid Deep Field Fornax overlaps GOODS-S, and by how much, will be checked and recorded before the protocol is frozen.

Vera C. Rubin Observatory. Data Preview 1 (Rubin Observatory Team, 2026) provides 1,792 exposures from 48 nights of LSSTComCam commissioning over about 15 deg^2 in seven fields in $ugrizy$, including the Extended Chandra Deep Field South. DP1 is available to Rubin data-rights holders through the Rubin Science Platform; analyses using it will be run by or with rights-holding collaborators, and only derived products permitted by Rubin policy will be released.

Nancy Grace Roman Space Telescope. Roman launched on 30 August 2026. No Roman science data are used in this protocol. ECAD will be checked on Roman pre-launch simulated imaging, and Roman will be added as a new view under a pre-registered amendment.

Labelled and multimodal collections. The Multimodal Universe (Multimodal Universe Collaboration et al., 2024) supports pretraining and spectroscopic views. Galaxy Zoo labels (Lintott et al., 2008; Walmsley et al., 2022) support the known-class test.

Simulations for injection. IllustrisTNG public data (Nelson et al., 2019), forward-modelled to each instrument's PSF, pixel scale and noise, provide realistic mergers, tidal features and ring-like systems. CAMELS (Villaescusa-Navarro et al., 2021) subgrid-physics variations test whether recall depends on the simulation model. `lenstronomy` (Birrer and Amara, 2018) produces synthetic strong lenses.

Anchor fields. The confirmatory analysis focuses on COSMOS (HST ACS, JWST COSMOS-Web, ground-based data) and GOODS-S/ECDFS (Hubble Legacy Fields, JADES, Rubin DP1, Euclid Q1 where it overlaps, NIRSpec spectra), where most sources have two or more artifact-independent groups. HLA-only and Euclid-only regions serve as single-view comparison sets.

7 Pre-registered evaluation protocol

This section will be deposited with a timestamp in a public pre-registration repository before any confirmatory data are opened, together with hashes of the code and the data manifest (Nosek et al., 2018). Any later change will be logged as a dated amendment with its reasons. The simulation study in Section 5 is a rehearsal of this protocol on synthetic data; none of its numbers are confirmatory.

7.1 Splits and blinding

Sky regions are split before analysis. The development set is half of the COSMOS-Web tiles (a seeded random draw), the Hubble Legacy Archive outside the anchor fields, and Euclid Q1 North. The confirmatory set is the remaining COSMOS-Web tiles, GOODS-S/ECDFS, CEERS and Euclid Q1 South. All hyperparameters, strata definitions, data-quality rules, $\pi_{\min}$ and reviewer instructions are frozen on the development set. Each reference stratum is sized by Equation 5, and its purity is certified by a random audit of $3\kappa/(q\pi_{\min})$ views before the confirmatory set is opened (Section 4.4).

7.2 Metrics

- Empirical false discovery proportion (FDP) on null and injection experiments, averaged over 200 resampled calibration and test partitions.
- Recall at fixed FDR, per injected anomaly class and stratum.
- Artifact pass-through: the fraction of injected single-group artifacts that reach L2, and the number of coincident-artifact sources that reach L2.
- Recall of the external Hubble Legacy Archive reference set within $K = 5{,}000$ inspected cutouts, matching O'Ryan and Gómez (2025).
- Expert-confirmed precision from blinded reviewers, with Wilson 95% intervals.
- Calibration diagnostics: one-sided Kolmogorov-Smirnov tests for an excess of small null p-values per stratum.
- Known-class error: the rate at which labelled members of known classes are called unexplained.
- Reviewer agreement (Cohen's κ).

7.3 Baselines

All baselines use identical cutouts and review budgets: isolation forest and LOF on foundation-model embeddings, as in Euclid Collaboration et al. (2025c); WGAN discriminator scores (Storey-Fisher et al., 2021); Astronomaly and Astronomaly Protege (Lochner and Bassett, 2021; Lochner and Rudnick, 2025); AnomalyMatch rerun under the same label budget (Gómez et al., 2026); kNN on DINOv2, AstroCLIP and AION-1 embeddings; Deep SVDD (Ruff et al., 2018); zero-shot AdaCLIP (arXiv:2610.29868, 2026); and ECAD-L1, which is ECAD without replication. Each baseline is evaluated both at its own top-k and after its score passes through ECAD's conformal, gating and BH layers, which separates the value of the calibration layer from the value of the score.

7.4 Injection-recovery tests

Positive injections, at matched surface brightness and signal-to-noise per stratum: synthetic strong lenses (`lenstronomy`); IllustrisTNG mergers, tidal features and ring-like systems; jellyfish-like stripped tails; compact very red point-like sources resembling little red dots; and out-of-library morphologies made by procedurally distorting real galaxies. Each injection is placed consistently in every view with correct per-band fluxes.

Negative-control injections: cosmic rays, satellite trails, persistence ghosts, snowball-like and wisp-like features, and drizzling edge defects, injected into one artifact-independent group only, plus a separate set injected independently into two groups to measure coincidences.

Null injections: typical reference galaxies, held out from calibration, re-inserted to measure FDP directly.

7.5 Pass and fail criteria

Table 5. Pre-registered criteria. The simulation rehearsal of each is reported in Table 4.

ID	Criterion	Pass threshold
P1	FDR validity (BH, L2)	Mean FDP $\leq q + 0.02$ for $q \in \{0.05, 0.10, 0.20\}$, in every stratum meeting the design rule
P2	Artifact suppression	Single-group artifact pass-through to L2 $\leq 2\%$, and at least 80% lower than at L1
P3	Power	For each positive injection class above the stratum's 5σ surface-brightness limit, L2 recall at $q = 0.10$ is ≥ 0.50, and no more than 0.05 below the best baseline evaluated through the same conformal layer
P4	External recovery	Recall of the Hubble Legacy Archive reference set within 5,000 inspections is within 10 percentage points of AnomalyMatch rerun under the same label budget
P5	Expert precision	On 300 blinded L2 candidates against 300 from the best baseline at a matched budget, ECAD's expert-confirmed precision is higher by at least 10 percentage points, with the 95% interval of the difference excluding zero
P6	Calibration	No stratum shows a one-sided Kolmogorov-Smirnov $p < 0.01$, after Holm correction, for an excess of small null p-values (conservative calibration is acceptable)
P7	Known-class error	At $\alpha = 0.05$, no known class is called unexplained more than 7% of the time

7.6 Ablations

Remove replication ($u = 1$); remove gating; vary the reference contamination; pooled against stratified calibration; weighted against Mondrian calibration; Simes against Bonferroni within groups; each encoder alone, with and without the cross-instrument fine-tune; BH against BY against e-BH with conformal e-values (Wang and Ramdas, 2022); $m \in \{5, 10, 20\}$; exposure-level against mosaic-level replication; calibration sizes on both sides of the design rule; and including human-labelled examples in calibration, which deliberately breaks the guarantee to measure how much it matters.

7.7 Abandonment thresholds

- **A1.** If P1 fails, one pre-registered remediation is applied (weighted calibration, then field-matched calibration, then coarser strata). If P1 still fails, all FDR claims are withdrawn and the failure analysis is published.
- **A2.** If P2 and P5 both fail, the hypothesis that cross-view replication improves discovery quality is abandoned and published as a negative result.
- **A3.** If P3 fails for every positive injection class, the discovery campaign stops, because the power cost of replication would be too high for practical use.
- **A4.** If P4 fails but P1, P2 and P5 pass, ECAD is reported as a precision-oriented complement to recall-oriented tools, not a replacement.
- **A5.** If P7 fails, the word "unexplained" is not used for any candidate, and L3 is suspended until the known-class test is revised and re-registered.

All outcomes will be published with full tables, whether or not they are favourable.

7.8 Discovery campaign rules

Only after P1 to P3 pass will ECAD run in discovery mode on the confirmatory fields. All L2 candidates are released with evidence bundles. Candidates are described as unexplained only at L3. Follow-up proposals, for example for JWST NIRSpec time, target L3 objects, and their outcome determines L4.

8 Risks and limitations

Exchangeability violations not captured by strata, such as field-to-field differences in the galaxy population, crowding or Galactic latitude. In simulation, a moderate population shift did not break FDR control, and field-matched calibration cut false rejections of typical sources by a factor of four (Section 5); a larger shift could break control. The protocol uses field-matched calibration as its first remedy, with P6 diagnostics per field.

Rare objects beyond reach. The contamination condition sets a floor on the prevalence an error-controlled search can claim, and certifying purity for prevalences near 10^{-5} needs hundreds of thousands of inspected reference views. Genuinely new kinds of object may be rarer than that. For them ECAD can still rank and flag at L0, with the evidence bundle, but it cannot attach an error rate, and it will say so.

Anomalies that are physically confined to one view. Little red dots, for example, stand out in NIRCam long-wavelength bands. ECAD defines replication over artifact-independent exposures, preferably in the same or adjacent bands, not across wavelengths. Cross-band consistency is checked against spectral-energy-distribution expectations, not used as a gate.

Time-variable sources fail replication by nature and need a separate time-domain track (Ishida et al., 2021).

Faint anomalies. Replication costs recall where an anomaly is detectable in only one group, typically when the other groups are shallow. The simulation quantifies this cost as a function of signal-to-noise (Figure 5); it is the price of artifact suppression and is reported per stratum.

Coincident artifacts. Independent artifacts in two groups of the same source fall outside the partial-conjunction guarantee. Their expected number grows with the square of the per-view artifact rate and linearly with survey size, while true discoveries grow with prevalence. In simulation each test set held about 14 such sources; without gating, 0.66 per run reached L2 at $q = 0.2$, and with gating none did in 200 runs. The protocol counts them directly with two-group negative-control injections.

Representation bias. Foundation models trained mostly on low-redshift optical data may compress high-redshift JWST morphology. The protocol uses several encoders and an ablation of the cross-instrument fine-tune.

Theoretical gap. BH validity after the within-group combination and the partial-conjunction transform is checked empirically, not proven. BY results are always reported alongside.

Simulation dependence. The simulation study uses analytic galaxy profiles, Gaussian PSFs, Gaussian noise and parametric artifacts. Real PSFs have extended wings and diffraction spikes, real noise is correlated after drizzling, and real artifacts are more varied. Its results show that the statistical machinery behaves as claimed when its assumptions hold and identify where it breaks; they do not predict recall or purity on real archives.

Astrophysical systematics. Chance projections can mimic interactions without redshift information (Euclid Collaboration et al., 2025c); lensing magnification changes apparent morphology; dust can make normal galaxies look like extreme red objects; photometric redshift outliers can mislabel populations. The known-class test and L4 spectroscopic confirmation handle these effects; ECAD does not claim to remove them.

Scope. ECAD addresses imaging anomalies with optional spectroscopic views. It does not test cosmological models directly, does not address time-domain anomalies, and makes no claim about scientific interest, which remains a human judgement.

9 Broader impact and reproducibility

Open science. The paper is openly available under CC BY 4.0 (doi:10.5281/zenodo.23883469). The simulation code, generator seeds, results files and simulated data for Section 5 are available on request to researchers through a restricted Zenodo record linked from that DOI, and will be released openly with the confirmatory study. Every result in Section 5 regenerates exactly from the code and seeds. All code for the confirmatory study will be released under an OSI-approved licence, with a frozen, hashed pre-registration and data manifests listing archive identifiers. Evidence bundles will be released with persistent identifiers under CC BY 4.0, following the FAIR principles (Wilkinson et al., 2016). Negative and null results will be published in full. For restricted data (Rubin DP1), bundles reference rather than redistribute restricted pixels.

People stay in charge. The statistical layer decides only which objects deserve attention, at a stated error rate. Astronomers decide what an object is, whether it matters and what is said publicly. Every L3 decision needs two independent expert judgements recorded in the bundle, and the interest model only reorders the review queue.

Responsible communication. The discovery levels give a public vocabulary: "candidate" (L1 or L2), "unexplained after expert review" (L3) and "confirmed" (L4). SHV Groups commits to using only this vocabulary in its own communications. All outputs will acknowledge NASA, ESA, CSA, the Rubin Observatory, the survey teams and Galaxy Zoo volunteers as each archive requires.

Relation to SHV's work. This framework is the methodological basis for the effort at Apeiron, SHV Groups' Department of Cosmology and Astrophysics, to find unusual objects across Webb, Hubble and Roman imaging, whose guiding rule is that every candidate must carry reproducible evidence. The method stands on its own and does not depend on any proprietary model.

10 Conclusion

Astronomy can rank hundreds of millions of sources by how unusual they look, but it has no shared standard for when such a ranking supports a discovery claim. ECAD proposes one: conformal p-values calibrated within observing conditions and sized to the prevalence being sought, symmetric data-quality gating, false discovery rate control across sources, formal replication across artifact-independent views, a known-class test with a stated error rate, and a hashed evidence bundle for every claim, sorted into five discovery levels. A simulation study with known ground truth shows that these pieces deliver the guarantees they claim when their assumptions hold and measures the price of each. It corrected three failures in the first version of the framework before any real data were touched, and it exposed a limit that more data cannot remove: a reference population must be cleaner than $q\pi/\kappa$, so the rarest objects can be flagged but not claimed with an error rate. The paper reports no astrophysical results. It commits in advance to the tests that would show the framework fails on real archives, and to publishing those outcomes. If the pre-registered criteria pass, ECAD can be applied unchanged to Euclid and Rubin releases and to Roman's first public data.

References

R. G. Abraham, S. van den Bergh, and P. Nair. A New Approach to Galaxy Morphology. I. Analysis of the Sloan Digital Sky Survey Early Data Release. *ApJ*, 588:218–229, 2003.

C. Alard and R. H. Lupton. A Method for Optimal Image Subtraction. *ApJ*, 503:325–331, 1998.

A. N. Angelopoulos and S. Bates. Conformal Prediction: A Gentle Introduction. *Foundations and Trends in Machine Learning*, 16(4), 2023. arXiv:2107.07511.

arXiv:2601.02610. Conformal novelty detection with false discovery rate control at the boundary, 2026. arXiv:2601.02610.

arXiv:2605.03442. High-Redshift Gravitational Lens Discoveries in JWST NIRCam using Anomaly-Match, 2026. arXiv:2605.03442.

arXiv:2606.20434. The Hubble Tension: A Decade Review, 2026. arXiv:2606.20434.

arXiv:2609.10567. Study of dynamical dark energy using DESI DR2 BAO, CMB, and Type Ia Supernovae data, 2026. arXiv:2609.10567.

arXiv:2609.32753. Making Euclid VIS Imaging AI-Ready: A Scalable Pipeline for Morphology, Anomaly Detection, and Similarity Search, 2026. arXiv:2609.32753.

arXiv:2610.29868. Evaluating Zero-Shot Anomaly Detection for Identifying Unusual Galaxy Morphologies in CEERS and NGDEEP, 2026. arXiv:2610.29868.

arXiv:2610.37747. Identifying Scientifically Interesting Galaxies with Vision-Language Models, 2026. arXiv:2610.37747.

S. Bates, E. Candès, L. Lei, Y. Romano, and M. Sesia. Testing for outliers with conformal p-values. *Annals of Statistics*, 51(1):149–178, 2023.

Y. Benjamini and R. Heller. Screening for partial conjunction hypotheses. *Biometrics*, 64(4):1215–1222, 2008.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57(1):289–300, 1995.

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, 2001.

S. Birrer and A. Amara. lenstronomy: multi-purpose gravitational lens modelling software package. *Physics of the Dark Universe*, 22:189–201, 2018.

T. S. Boyajian et al. Planet Hunters IX. KIC 8462852: where's the flux? *MNRAS*, 457(4):3988–4004, 2016.

M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. LOF: identifying density-based local outliers. In *Proc. ACM SIGMOD*, pages 93–104, 2000.

C. M. Casey et al. COSMOS-Web: An Overview of the JWST Cosmic Origins Survey. *ApJ*, 954:31, 2023.

L. Ciotti and G. Bertin. Analytical properties of the $R^{1/m}$ law. *A&A*, 352:447–451, 1999.

C. J. Conselice. The Relationship between Stellar Light Distributions of Galaxies and Their Formation Histories. *ApJS*, 147:1–28, 2003.

DESI Collaboration, M. Abdul-Karim, et al. DESI DR2 Results II: Measurements of Baryon Acoustic Oscillations and Cosmological Constraints. *Physical Review D*, 2025. arXiv:2503.14738.

F. D'Eugenio et al. JADES Data Release 3: NIRSpec/Microshutter Assembly Spectroscopy for 4000 Galaxies in the GOODS Fields. *ApJS*, 277:4, 2025.

V. Etsebeth, M. Lochner, M. Walmsley, and M. Grespan. Astronomaly at scale: searching for anomalies amongst 4 million galaxies. *MNRAS*, 2024.

Euclid Collaboration, H. Aussel, et al. Euclid Quick Data Release (Q1): Data release overview, 2025a. arXiv:2503.15302.

Euclid Collaboration, Y. Mellier, et al. Euclid. I. Overview of the Euclid mission. *A&A*, 697:A1, 2025b.

Euclid Collaboration, M. Siudek, et al. Euclid Quick Data Release (Q1): Exploring galaxy properties with a multi-modal foundation model, 2025c. arXiv:2503.15312.

Euclid Collaboration, M. Walmsley, et al. Euclid Quick Data Release (Q1): First visual morphology catalogue. *A&A*, 2025d. doi:10.1051/0004-6361/202555140.

S. L. Finkelstein et al. CEERS Key Paper I: An Early Look into the First 500 Myr of Galaxy Formation with JWST. *ApJL*, 946:L13, 2023.

M. Franco et al. COSMOS-Web: Comprehensive Data Reduction for Wide-Area JWST NIRCam Imaging, 2025. arXiv:2506.03256.

W. L. Freedman, B. F. Madore, T. J. Hoyt, I. S. Jang, A. J. Lee, and K. A. Owens. Status Report on the Chicago-Carnegie Hubble Program (CCHP): Measurement of the Hubble Constant Using the Hubble and James Webb Space Telescopes. *ApJ*, 985:203, 2025.

J. P. Gardner et al. The James Webb Space Telescope Mission. *PASP*, 135:068001, 2023.

J.-B. Grill et al. Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. In *NeurIPS*, volume 33, pages 21271–21284, 2020.

P. Gómez, L. E. Ruhberg, M. T. Nardone, and D. O'Ryan. AnomalyMatch: Discovering Rare Objects of Interest with Semi-supervised and Active Learning. *RAS Techniques and Instruments*, 2026. doi:10.1093/rasti/rzag045.

M. A. Hayat, G. Stein, P. Harrington, Z. Lukić, and M. Mustafa. Self-supervised Representation Learning for Astronomical Images. *ApJL*, 911:L33, 2021.

E. E. O. Ishida et al. Active anomaly detection for time-domain discoveries. *A&A*, 650:A195, 2021.

Y. Jin et al. Everywhere Valid Bounds on False Discovery Proportions in Conformal Inference, 2026. arXiv:2605.20726.

A. M. Koekemoer et al. The COSMOS Survey: Hubble Space Telescope Advanced Camera for Surveys Observations and Data Processing. *ApJS*, 172:196, 2007.

I. Labbé et al. A population of red candidate massive galaxies ~600 Myr after the Big Bang. *Nature*, 616:266–269, 2023.

C. J. Lintott et al. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *MNRAS*, 389(3):1179–1189, 2008.

C. J. Lintott et al. Galaxy Zoo: 'Hanny's Voorwerp', a quasar light echo? *MNRAS*, 399:129–140, 2009.

F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation Forest. In *Proc. IEEE ICDM*, pages 413–422, 2008.

M. Lochner and B. A. Bassett. ASTRONOMALY: Personalised active anomaly detection in astronomical data. *Astronomy and Computing*, 36:100481, 2021.

M. Lochner and L. Rudnick. Astronomaly Protege: Discovery through Human-Machine Collaboration. *AJ*, 2025. arXiv:2411.04188.

D. R. Lorimer, M. Bailes, M. A. McLaughlin, D. J. Narkevic, and F. Crawford. A Bright Millisecond Radio Burst of Extragalactic Origin. *Science*, 318:777–780, 2007.

J. M. Lotz, J. Primack, and P. Madau. A New Nonparametric Approach to Galaxy Morphological Classification. *AJ*, 128:163–182, 2004.

A. Marandon, L. Lei, D. Mary, and E. Roquain. Adaptive novelty detection with false discovery rate guarantee. *Annals of Statistics*, 52(1):157–183, 2024.

J. Matthee et al. Little Red Dots: An Abundant Population of Faint Active Galactic Nuclei at z~5 Revealed by the EIGER and FRESCO JWST Surveys. *ApJ*, 963(2):129, 2024.

Multimodal Universe Collaboration, E. Angeloudi, et al. The Multimodal Universe: Enabling Large-Scale Machine Learning with 100 TB of Astronomical Scientific Data. In *NeurIPS Datasets and Benchmarks Track*, volume 37, 2024.

E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan. Do Deep Generative Models Know What They Don't Know? In *ICLR*, 2019.

D. Nelson et al. The IllustrisTNG simulations: public data release. *Computational Astrophysics and Cosmology*, 6:2, 2019.

B. A. Nosek, C. R. Ebersole, A. C. DeHaven, and D. T. Mellor. The preregistration revolution. *PNAS*, 115(11):2600–2606, 2018.

M. Oquab et al. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, 2024.

D. O'Ryan and P. Gómez. Identifying astrophysical anomalies in 99.6 million source cutouts from the Hubble legacy archive using AnomalyMatch. *A&A*, 704:A227, 2025.

L. Parker et al. AstroCLIP: a cross-modal foundation model for galaxies. *MNRAS*, 531(4):4990–5011, 2024.

L. Parker et al. AION-1: Omnimodal Foundation Model for Astronomical Sciences, 2025. arXiv:2510.17960.

R. Pearce-Casey et al. Using cGANs for Anomaly Detection: Identifying Astronomical Anomalies in JWST NIRCam Imaging, 2023. arXiv:2310.09073.

F. Pedregosa et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Planck Collaboration, N. Aghanim, et al. Planck 2018 results. VI. Cosmological parameters. *A&A*, 641:A6, 2020.

M. V. Pruzhinskaya, K. L. Malanchev, M. V. Kornilov, E. E. O. Ishida, F. Mondon, A. A. Volnova, and V. S. Korolev. Anomaly detection in the Open Supernova Catalog. *MNRAS*, 489(3):3591–3608, 2019.

A. G. Riess et al. A Comprehensive Measurement of the Local Value of the Hubble Constant with 1 km/s/Mpc Uncertainty from the Hubble Space Telescope and the SH0ES Team. *ApJL*, 934:L7, 2022.

Rubin Observatory Team. The Vera C. Rubin Observatory Data Preview 1. *AJ*, 2026. doi:10.3847/1538-3881/ae521f.

L. Ruff et al. Deep One-Class Classification. In *ICML*, pages 4393–4402, 2018.

V. Rusakov et al. Little red dots as young supermassive black holes in dense ionized cocoons. *Nature*, 649:574–579, 2026.

S. K. Sarkar. Some probability inequalities for ordered MTP2 random variables: a proof of the Simes conjecture. *Annals of Statistics*, 26(2):494–504, 1998.

P. Schneider, J. Ehlers, and E. E. Falco. *Gravitational Lenses*. Springer, Berlin, 1992.

R. J. Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73 (3):751–754, 1986.

M. J. Smith, R. J. Roberts, E. Angeloudi, and M. Huertas-Company. AstroPT: Scaling Large Observation Models for Astronomy, 2024. arXiv:2405.14930.

K. Storey-Fisher et al. Anomaly detection in Hyper Suprime-Cam galaxy images with generative adversarial networks. *MNRAS*, 508(2):2946–2963, 2021.

J. L. Sérsic. Influence of the atmospheric and instrumental dispersion on the brightness distribution in a galaxy. *Boletin de la Asociacion Argentina de Astronomia*, 6:41–43, 1963.

R. J. Tibshirani, R. F. Barber, E. J. Candès, and A. Ramdas. Conformal Prediction Under Covariate Shift. In *NeurIPS*, volume 32, 2019.

P. G. van Dokkum. Cosmic-Ray Rejection by Laplacian Edge Detection. *PASP*, 113:1420–1427, 2001.

F. Villaescusa-Navarro et al. The CAMELS Project: Cosmology and Astrophysics with Machine-learning Simulations. *ApJ*, 915:71, 2021.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

M. Walmsley et al. Galaxy Zoo DECaLS: Detailed visual morphology measurements from volunteers and deep learning for 314,000 galaxies. *MNRAS*, 509(3):3966–3988, 2022.

M. Walmsley et al. Zoobot: Adaptable Deep Learning Models for Galaxy Morphology. *Journal of Open Source Software*, 8(85):5312, 2023.

R. Wang and A. Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society B*, 84(3):822–852, 2022.

K. E. Whitaker et al. The Hubble Legacy Field GOODS-S Photometric Catalog. *ApJS*, 244:16, 2019.

M. D. Wilkinson et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018, 2016.

A Simulation parameters

Table 6. Generator settings. The full code, seeds and simulated data are available on request (Section 9).

Item	Setting
Cutout	33×33 pixels, rendered on a $2\times$ supersampled grid
Typical mix	Disk 38%, bulge 20%, bulge plus disk 17%, clumpy disk 10%, star 15%; faint neighbour in 12% of fields
Sizes	Effective radius log-normal, median 3.0 px, 0.35 in natural log, clipped to 1.2 to 7 px
Brightness	Typical S/N (50-pixel aperture, deep visit) log-normal, median 40, clipped to 10 to 1,500; anomalies log-uniform 15 to 1,000
Colours	Instrument B to A flux ratios: disk 0.9 and 0.8, bulge 1.3 and 1.6, knots 0.7 and 0.5 (short and long filter), with 8% scatter
Instrument A	Two visits; PSF FWHM 1.5 to 1.9 px (good) or 2.1 to 2.6 px (poor); noise 1.0 (deep) or 2.5 (shallow), ± 10%
Instrument B	One visit, two filters; short-filter FWHM 1.8 to 2.2 or 2.4 to 2.9 px, long filter 1.45 times wider; noise 0.8 or 2.0; coverage 70%
Artifacts	2% per view, independent; cosmic ray 8 to 60σ, satellite trail 3 to 10σ, ghost ring 1.5 to 4σ, snowball 10 to 40σ, bad column 2 to 6σ
Ring galaxies	Ring radius 4.5 to 9 px, width 0.8 to 1.6 px, azimuthal modulation up to 60%, nucleus 15 to 35% of flux
Strong lenses	Singular isothermal sphere, Einstein radius 4 to 8 px, shear up to 0.1, source offset up to 0.6 Einstein radii, deflector 50 to 75% of flux
Mergers	Separation 3 to 7 px, flux ratio 1 to 4, tidal tail 8 to 14 px holding 8 to 20% of flux
Out-of-library	Spoked ring (three to five spokes) or straight chain of four to six knots, equally often
Encoder	Residuals from a smooth elliptical model: three zonal significances, residual flux fraction, residual peak count, profile-rise significance, four rotational harmonics, axis ratio, concentration, size, brightness, centroid offset (15 features, standardised)
Score	Mean distance to the 10 nearest of 40,000 (instrument A) or about 14,000 (each instrument-B filter) reference views
Splits	Per repetition: 60,000 typical sources for calibration, 10,000 typical sources plus 300 anomalies for test
Data quality	Laplacian sharpness: excess $> 7\sigma$ and > 0.8 of the pixel value; difference imaging: smoothed residual > 6 robust σ

B Changes from version 1.0

1. The fixed minimum calibration size ($n_g \geq 1{,}000$) is replaced by the design rule of Equation 5, derived from the resolution condition, with a safety factor set by the simulation study.
2. The contamination condition (Equation 4) is stated, with a random audit to certify the purity of each reference stratum, and an explicit floor on the prevalence the search can claim.
3. Symmetric data-quality gating is added, so that realistic reference populations can be used without losing validity.
4. Views inside a group are combined with Simes instead of a Bonferroni minimum.
5. The prediction-set rule for separating known from unexplained objects is replaced by a class-conditional conformal novelty test with a per-class error guarantee, run on gated views.
6. The scope of the FDR guarantee is stated explicitly, including the coincident-artifact case.

7. Field-matched calibration is added as the first remedy for population shift between fields.
8. Criterion P6 is made one-sided; criterion P7 and abandonment rule A5 are added for the known-class test.
9. A simulation study rehearses every criterion before any real data are used. Two criteria (P3 for strong lenses, P5 against ECAD-L1) fail in the rehearsal and are kept unchanged.