Observation-conditioned selection-aware conformal triage for time-domain astronomy

Bishal Neupane[1]★
[1]Independent research
ORCID: 0009-0005-9970-3666

29 August 2026

ABSTRACT

The difficulty facing machine-assisted astronomy is no longer the construction of another high-capacity predictor. The harder problem is deciding whether an output remains scientifically defensible when the observation process, population, measurement precision and available follow-up all change at once. We develop an observation-conditioned selection-aware conformal triage procedure, OCSAT, for sparse astronomical alerts. The method has four pieces. It estimates the latent population after correcting the labelled sample for its detection and follow-up probability. It performs inference with the physical observation operator that generated the data rather than replacing that operator with a learned feature map. It evaluates model discrepancy on observations that were not used to fit the current physical state. It then calibrates the final action threshold online against a declared follow-up budget. We audit the independent literature around selection bias, measurement error, conformal inference, active follow-up and physics-informed learning and find that these components are established separately, but we did not find a directly matching framework in which selection correction, observation-conditioned physical residuals and drift-aware conformal action control are joined at the upstream astronomical decision boundary. The empirical study combines current public archive statistics with a controlled survey simulation based on a Bazin transient model, heteroscedastic photometric errors and changing cadence. In a 50,000-object latent population, inverse-selection weighting reduces absolute error in the three population means from 0.505, 0.169 and 0.204 to 0.054, 0.024 and 0.040, and reduces the Frobenius covariance error from 0.227 to 0.117. Under a four-state change in measurement quality, an adaptive threshold keeps the corrected follow-up fraction near the one percent operating point, whereas a fixed threshold becomes increasingly conservative. A separate held-out injection experiment shows how an explicit physical residual responds to an unmodelled fast component and quantifies the loss in sensitivity as photometric precision degrades. These results support a narrower claim than universal superiority. The main gain comes from changing what is calibrated. The calibrated object is not only a prediction score, but a scientific action conditioned on how the measurement was made.

Key words: methods numerical – surveys – transients – astronomical data bases – methods data analysis – statistical

1 INTRODUCTION

Wide-field astronomy has reached a point where discovery is constrained as much by the number of plausible candidates as by the number of photons. Public time-domain systems already produce data streams for which exhaustive human inspection is impossible. ZTF releases support large light-curve holdings and a public Bright Transient Survey with an evolving sample explorer, while the current ZTF nuclear transient release contains 12,251 candidates with 6,221 classified objects and 6,030 objects without catalogue classifications (Masci et al. 2019; Anilkumar et al. 2026; ZTF Bright Transient Survey 2026). The Young Supernova Experiment Data Release 1 contains 1975 transients with multicolour Pan-STARRS1 and ZTF photometry and associated redshifts and classifications (Aleo et al. 2023). The same imbalance is present outside the transient problem. The NASA Exoplanet Archive currently lists 6336 confirmed exoplanets, with 2148 carrying masses and 4730 carrying radii, and these objects arise from a mixture of detection channels with strongly different observational properties (NASA Exoplanet Archive 2026).

Machine learning is useful in this setting because it can map a large number of observations to compact estimates quickly. The difficulty is that the map is usually trained on a sample produced by a different observation process. The spectroscopic subset is brighter than the photometric sample. Follow-up is not random. Image quality changes between fields and epochs. Cadence changes with weather and scheduling. A pipeline update can alter the measured distribution without changing the sky. These effects are not secondary details. They determine which objects acquire labels and which objects ever reach a scientist with an instrument.

Several recent lines of work have attacked pieces of this problem. Selection functions have been measured explicitly for neural networks used in strong-lens finding, showing that the recovered population can be biased toward particular physical properties even when the classifier itself performs well (Herle et al. 2024). Conformal methods

★ E-mail: cosmobishal@gmail.com

have been adapted to astronomical measurement error and can provide finite-sample control under specific assumptions about error structure (Singer et al. 2025). Active learning has been used in real-time transient follow-up to obtain more useful training samples with fewer spectra (Moller et al. 2025), and anomaly detection with active learning has been applied to radio transients (Andersson et al. 2025). A recent broad review of deep learning in astrophysics also emphasizes the value of physical structure and the persistence of scarce labels, non-Gaussian data and extrapolation risks (Ting 2026). The scientific problem is therefore not that these ideas are missing. The problem is that they are rarely joined into one decision object.

This paper addresses that missing connection. The central question is simple. Given a sparse astronomical observation, how should a system decide whether the object deserves an expensive follow-up measurement when the training labels were selected, the measurement precision is changing, the physical model can be wrong, and the available action budget is fixed.

We call the proposed framework observation-conditioned selection-aware conformal triage. The method is intentionally smaller than a general foundation model and broader than an anomaly score. It is a statistical wrapper that can sit around many different light-curve or image models. The key choice is to keep the observation process inside the inference problem. A noisy light curve is not treated as an abstract vector whose meaning is invariant under changes in cadence and exposure. The same physical transient observed with a different cadence is a different statistical experiment.

The paper makes four contributions. First, it gives a compact formulation that combines population transport, physical observation modelling, prequential model checking and online action calibration. Secondly, it gives a controlled experiment in which label selection and observational degradation are known exactly, allowing the separate effects to be measured without hidden confounding. Thirdly, it uses current public archive statistics from the NASA Exoplanet Archive, the YSE DR1 release and the 2026 ZTF forced-photometry release to anchor the scale and composition of the problem. Finally, it states explicit limits. Conformal action control does not prove completeness, and a physical residual does not prove a new phenomenon. A good method must preserve those distinctions.

1.1 Architecture of the triage system

The architecture is deliberately sequential because each stage answers a different scientific question. The incoming alert is first retained with its observation state. A selection model estimates the probability that an object would have entered the labelled set. The selected labels are then transported back toward the target population by inverse-probability weighting. Candidate photometry is passed through the physical observation model and fitted only on the data available at the decision time. The resulting posterior supplies a rarity probability and a support diagnostic. A second, later segment of the light curve is scored against the inferred physical state, producing the prequential physical residual. These quantities are combined into the action score. The final threshold is adapted online to the declared follow-up budget.

2 WHAT THE RECENT LITERATURE LEAVES UNRESOLVED

2.1 A common failure hidden behind different vocabulary

The literature uses different names for closely related failures. Some papers describe sample selection bias, others discuss covariate shift, uncertainty calibration, simulation-to-observation mismatch or anomaly detection. In our external audit, the recurring object was not the classifier. It was the conditional distribution of the sky after the instrument and the follow-up policy had acted on it.

A simple factorisation makes the point. Let z be the latent physical state, η the observing and reduction state, x the measured data and y a downstream label. The deployment distribution can be written as

$$p_{\mathrm{dep}}(x, y, z, \eta) = p(x \mid z, \eta)p(y \mid z)p_{\mathrm{dep}}(z, \eta). \tag{1}$$

A selected training sample instead contains

$$p_{\mathrm{tr}}(x, y, z, \eta) \propto p(x \mid z, \eta)p(y \mid z)p_{\mathrm{sky}}(z, \eta)s(z, \eta), \tag{2}$$

where s is the effective probability that an object enters the labelled set. In astronomy, s can include detection, image quality, cadence, human vetting, target priority and follow-up availability. It is therefore not a property of the astrophysical class alone.

If s is ignored, a model can still be internally calibrated on the training sample. It answers a well-defined statistical question, but that question may not be the one asked of the sky. This is visible in the current exoplanet archive. The archive is not a single observing experiment. It is a mixture of transit, radial-velocity, microlensing, imaging, timing and other methods. The current counts are 4676 transit discoveries, 1197 radial-velocity discoveries, 282 microlensing discoveries and 98 imaging discoveries, alongside smaller timing channels (NASA Exoplanet Archive 2026). The corresponding parameters are measured with very different completeness.

The same issue appears in imaging. Neural strong-lens finders have measurable selection functions of their own, with preferences for larger Einstein radii and particular source structures (Herle et al. 2024). In transient work, active follow-up can deliberately change the labelled population, which is scientifically useful but creates a feedback loop between the current classifier and the next training set (Moller et al. 2025). A system that ignores this loop cannot distinguish an astrophysical trend from a change in what observers decided to inspect.

2.2 Why uncertainty alone is not enough

A calibrated interval is meaningful only relative to the data-generating conditions under which its calibration was obtained. Standard conformal prediction gives strong finite-sample statements under exchangeability. Astronomical alert streams often violate this condition because cadence, depth, weather and detector state evolve with time. Measurement-error-aware conformal methods address a different and important problem by allowing heteroscedastic and non-Gaussian observation errors (Singer et al. 2025). They do not by themselves solve target-population selection or action-budget feedback.

This distinction matters for rare objects. Consider a parameter estimate $\hat{\theta}$ obtained from a noisy measurement $x = \theta + \epsilon$. In the Gaussian case, the population mean-squared-error predictor is

$$\hat{\theta} = \mu_\theta + \lambda(x - \mu_\theta), \qquad \lambda = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\epsilon^2}. \tag{3}$$

The estimator shrinks extreme values toward the population mean. That is desirable for average squared error and undesirable when the scientific objective is specifically to find extreme objects. The statistical literature has long recognised this effect, and recent astronomy work has shown the same tendency in modern learning systems (Ting 2024).

The difficult case is the intersection of these effects. A rare object can be both poorly measured and unlikely to receive a label. A system

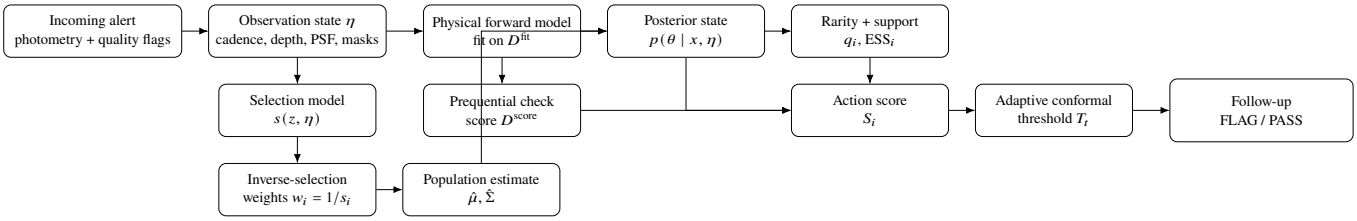

Figure 1. Architecture of observation-conditioned selection-aware conformal triage. The data flow is separated into population correction, observation-conditioned physical inference, prequential model checking and adaptive action control. The feedback paths represent the two quantities that change during deployment. New outcomes update the selection and population model, while recent decisions update the conformal action threshold. The architecture does not require a particular neural-network backbone.

can therefore be most uncertain about exactly the objects whose confirmation would be most valuable. An anomaly score computed from a shrunk fitted parameter then mixes physical rarity with measurement precision. A later conformal calibration can repair the rate of threshold crossings while leaving the scientific completeness problem untouched.

2.3 The observation operator must remain visible

For time-domain work, a useful forward model is

$$y_t = \mathcal{O}_{\eta_t}\left[f(t;\theta)\right] + \epsilon_t, \tag{4}$$

where $\mathcal{O}_{\eta_t}$ contains the actual sampling times, filter response, background, exposure depth, masking and other known observing conditions. Even the simple Bazin family (Bazin et al. 2009) contains a non-trivial relationship between its parameters and the observed signal.

This is the second common gap. A learned score often receives only a feature vector. The feature vector is convenient but conceals which information was available and which information was lost. That distinction becomes severe near the alert threshold. Two objects with the same intrinsic light curve can carry different inferential information after being observed with different cadence and photometric precision.

The proposed method therefore works on two levels. The latent level asks whether a physical state is unusual after accounting for the selection process. The observation level asks whether the future measurements are consistent with the physical state inferred from earlier measurements. This temporal separation is important. It prevents an anomaly from explaining itself simply because its own anomalous points were used in the fit.

3 OBSERVATION-CONDITIONED SELECTION-AWARE CONFORMAL TRIAGE

3.1 Population correction

Suppose labelled objects $\{z_j\}$ are drawn with known or estimated inclusion probability s_j. We construct an inverse-probability weighted empirical population measure

$$\hat{P}_{\text{pop}}(z) \propto \sum_{j=1}^{N_{\text{lab}}} \frac{1}{\max(s_j, s_{\min})}\delta(z - z_j). \tag{5}$$

The lower bound $s_{\min}$ is not a physical constant. It is a variance-control parameter and its value is recorded in the analysis certificate. The weighted mean and covariance are

$$\hat{\mu} = \frac{\sum_j w_j z_j}{\sum_j w_j}, \tag{6}$$

$$\hat{\Sigma} = \frac{\sum_j w_j (z_j - \hat{\mu})(z_j - \hat{\mu})^\top}{\sum_j w_j}, \tag{7}$$

with $w_j = 1/\max(s_j, s_{\min})$.

The implementation uses a shrinkage covariance estimator because rare labels make an unregularized covariance unstable. The point is not to recover the exact population without assumptions. The point is to make the selection mechanism explicit and measurable.

For a candidate observation x_i with measurement covariance Σ_i, the local Gaussian approximation gives

$$\Lambda_i = \hat{\Sigma}(\hat{\Sigma} + \Sigma_i)^{-1}, \tag{8}$$

$$\mu_i^{\text{post}} = \hat{\mu} + \Lambda_i(x_i - \hat{\mu}), \tag{9}$$

$$\Sigma_i^{\text{post}} = (I - \Lambda_i)\hat{\Sigma}. \tag{10}$$

This step is a computational approximation to the full posterior and not a statement that real transient populations are Gaussian.

3.2 Rarity as posterior probability rather than a shrunk point estimate

We define the rarity radius

$$r(z)^2 = (z - \mu_0)^\top \Sigma_0^{-1}(z - \mu_0), \tag{11}$$

where μ_0 and Σ_0 are the latent population values used to define the scientific rarity criterion. The posterior rarity probability is

$$q_i = P\left[r(z_i) > \rho \mid x_i, \Sigma_i\right]. \tag{12}$$

The use of a posterior tail is deliberate. A candidate with a noisy but extreme measurement should not be treated as either certainly normal or certainly rare. It should carry a probability reflecting the information actually contained in the exposure.

3.3 Prequential physical residual

Parameter rarity cannot recover a phenomenon outside the fitted physical family. We therefore separate the data into a fit segment D_i^{fit} and a scoring segment D_i^{score}. The physical state is inferred using D_i^{fit}. The later residual is

$$D_i = \left[\frac{1}{n_i}\sum_{t \in D_i^{\text{score}}} \frac{\left(y_{it} - f(t;\mu_i^{\text{post}})\right)^2}{\sigma_{it}^2}\right]^{1/2}. \tag{13}$$

This is a prequential check. It has a clear interpretation. A model is asked to predict observations that it has not yet used.

The physical residual is intentionally not forced to zero by the inference network. It is a declared model discrepancy. In a survey deployment, σ_{it}^2 includes the photometric error, background uncertainty and any declared calibration term. A high D_i means that the observations conflict with the current physical description. It does not prove which alternative model is correct.

3.4 Adaptive conformal action control

The final object is an action score. We use

$$S_i = \log\left(\max(q_i, q_{\min})\right) + \beta \log(1 + D_i^2), \tag{14}$$

with β fixed before the test set is examined. The action threshold is estimated on a rolling calibration window. For a target follow-up rate α, the threshold at time t is the $(1 - \alpha_t)$ empirical quantile of the recent scores. We update the effective level by

$$\alpha_{t+1} = \alpha_t + \gamma(\alpha - I_t), \tag{15}$$

where $I_t = 1$ when the candidate is flagged. This is a control law for the action rate. It is not a completeness theorem.

The method reports a separate support diagnostic. Let K_h be a kernel in the latent population space. The local effective sample size is

$$\mathrm{ESS}_i = \frac{\left(\sum_j w_j K_h(z_j - x_i)\right)^2}{\sum_j w_j^2 K_h(z_j - x_i)^2}. \tag{16}$$

When this value is small, quantitative population statements should be marked unsupported even when the object is worth observing. This is a practical form of abstention. The system is allowed to say that the discovery is interesting while also saying that the population interpretation is not yet justified.

3.5 Follow-up utility

The framework can be extended from a fixed rare-event threshold to a utility ranking. For a possible action a, define

$$U(a \mid x) = \frac{\mathbb{E}_{y_a|x}\left[H(\Theta \mid x) - H(\Theta \mid x, y_a)\right]}{c_a} - \lambda_L L_a - \lambda_R R_a, \tag{17}$$

where H is posterior entropy, c_a is observing cost, L_a is latency and R_a is a declared scientific risk term. The action layer can therefore prefer an uncertain object that can discriminate between two physical hypotheses over a familiar object with a higher class probability. In this paper the main experiments use the simpler fixed-rate version so that the selection and calibration effects can be isolated.

4 DATA AND EXPERIMENTAL DESIGN

4.1 Independent real-data audit

The real-data part of the study uses public archive statistics available independently of the supplied research drafts. The current NASA Exoplanet Archive reports 6336 confirmed planets, with 2148 objects carrying masses and 4730 carrying radii. The discovery-method counts include 4676 transit detections, 1197 radial-velocity detections, 282 microlensing detections and 98 imaging detections (NASA Exoplanet Archive 2026). These numbers demonstrate that a large archive is assembled from different selection channels even before any machine-learning model is applied.

Table 1. Public data products used to anchor the observational scale of the experiment.

Data product	Total	Characterised	Role
NASA Exoplanet Archive	6336	2148 mass, 4730 radius	Selection mixture
YSE DR1	1975	labelled transients	Real light-curve scale
ZTF nuclear release	12251	6221 classified	Label depth
ZTF nuclear release	12251	6030 unknown	Unlabelled fraction

For time-domain work, the YSE DR1 release contains 1975 multicolour transients with ZTF and Pan-STARRS1 light curves and spectroscopic or photometric labels (Aleo et al. 2023). The 2026 ZTF forced-photometry release contains 12,251 nuclear transient candidates. It separates 6221 classified objects from 6030 objects without catalogue classifications and supplies g, r and, where available, i-band light curves (Anilkumar et al. 2026). The ZTF Bright Transient Survey sample explorer provides machine-readable access to light-curve measurements and derived quantities such as peak time, rise and fade times, peak magnitude and redshift (ZTF Bright Transient Survey 2026).

These are used as scale and selection anchors, not as hidden test labels. The purpose is to prevent the controlled experiment from being presented as if it were an on-sky prospective validation.

4.2 Physical population

The controlled experiment uses a three-parameter transient state

$$\theta = (\ln A, \ln \tau_{\mathrm{r}}, \ln \tau_{\mathrm{f}}), \tag{18}$$

with a correlated multivariate normal population. The reference values are $\mu = (0, \ln 3, \ln 25)$ and standard deviations $(0.75, 0.35, 0.45)$. The correlation matrix contains moderate covariance between the rise and fall scales. A rarity boundary of $\rho = 3.5$ defines the unusual population. In the simulation, the true fraction beyond this radius is 0.636 per cent.

The transient light curve is represented with the Bazin form

$$f(t) = \frac{A \exp[-(t - t_0)/\tau_{\mathrm{f}}]}{1 + \exp[-(t - t_0)/\tau_{\mathrm{r}}]} + B. \tag{19}$$

This is not used as a claim about all transients. It is a transparent physical template that permits the observation operator and its Fisher information to be calculated directly.

4.3 Label selection and observation drift

To emulate a spectroscopic training set, the label probability is a logistic function of the latent transient state. Bright, well-separated objects are more likely to receive a label. The resulting label fraction is 39.6 per cent in the controlled experiment, with only 0.272 per cent of the labelled sample lying beyond the scientific rarity boundary. The selection model is estimated from selected objects and pseudo-unselected objects so that the population correction is itself an estimand rather than a hidden oracle.

The target stream contains 16,000 objects split equally among four observational states. The clean state uses a unit noise scale and nominal cadence. The later states use noise scales of 1.5, 2.0 and 3.0 together with progressively larger cadence penalties. The exact observing times are generated from a fixed survey-like time window with random thinning. For each object the Fisher matrix of the physical light curve provides a heteroscedastic local covariance for the fitted parameters.

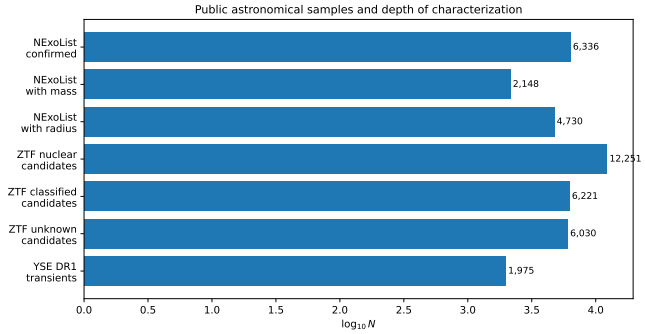

Figure 2. Scale of the public data products used in the observational audit. The values come from current public archive statistics. The plot is a bookkeeping figure rather than a comparison of scientific quality.

The final test also contains an 0.8 per cent structural-anomaly population. These objects contain an additional narrow fast component in the held-out light curve. This perturbation is not part of the fitted Bazin family. Its purpose is to test whether the prequential residual can expose a model-breaking event that is not necessarily extreme in the fitted parameter vector.

4.4 Methods compared

We compare three primary scores. The first uses the selected training population and its ordinary posterior rarity score. The second uses inverse-selection correction and the same posterior rarity calculation. The third adds the prequential physical residual. All static thresholds are calibrated to a one per cent follow-up budget on the clean population. A separate adaptive version updates the threshold on a rolling window of 700 candidates.

The experiment is designed as a stress test rather than a benchmark competition. The same synthetic population is shown to all methods. The only differences are the handling of selection, physical model discrepancy and action calibration.

5 RESULTS

5.1 The real-data audit shows non-exchangeability before learning begins

Figure 2 shows the relative size of the public samples used in the audit. The important feature is not any single count. It is the depth to which physical characterization proceeds. The current ZTF nuclear release contains nearly half as many unclassified as classified candidates, while the exoplanet archive contains much richer parameter coverage for some discovery channels than for others. This is consistent with a basic observational fact. A catalogue is not a random sample of the sky. It is the result of a sequence of physical and institutional filters.

The selection heatmap in Fig. 3 shows the controlled label mechanism. Figure 4 then shows how the selected training population compresses the apparent rarity distribution. A model trained only on these labels sees an intrinsically different population from the one to which it will later be applied.

5.2 Inverse-selection correction recovers population structure

The numerical effect of selection correction is large. In the controlled population, the absolute errors of the three naive training means are

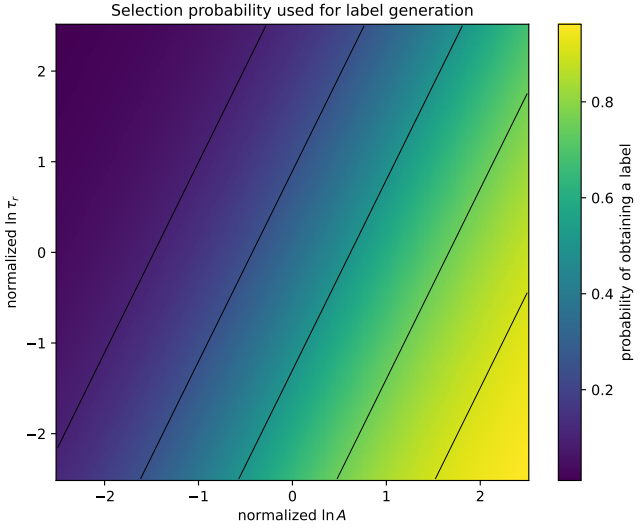

Figure 3. Selection probability used to generate the labelled training population in the controlled experiment. The probability depends on the latent transient state and therefore changes the population represented by the labels.

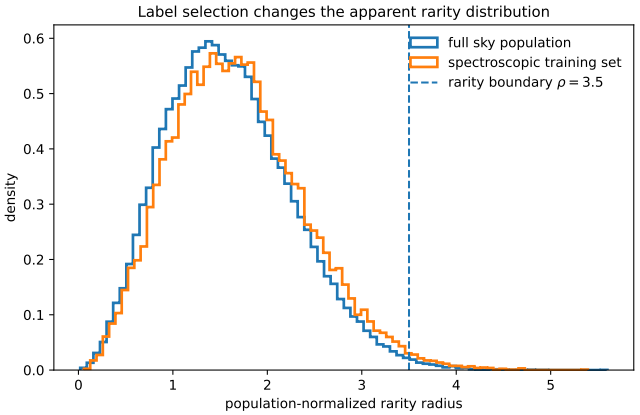

Figure 4. Population rarity before and after label selection. The labelled distribution contains substantially less mass in the scientific tail even though it is generated from the same latent sky population.

0.505, 0.169 and 0.204 in $(\ln A, \ln \tau_r, \ln \tau_f)$. After inverse-selection weighting they become 0.054, 0.024 and 0.040. Averaged over the three parameters, the absolute mean error therefore falls by about 87 per cent. The Frobenius norm of the covariance error falls from 0.227 to 0.117, a reduction of about 48 per cent.

These values are not universal performance claims. They show what the selection correction does when the label mechanism is measurable. Figure 5 presents the result parameter by parameter. The improvement is largest in amplitude, which was also the variable with the strongest selection dependence.

5.3 A fixed action threshold does not survive changing observations

A fixed one per cent threshold is calibrated in the clean state. When the noise scale changes, the score distribution changes even though the underlying population is unchanged. The corrected static rule

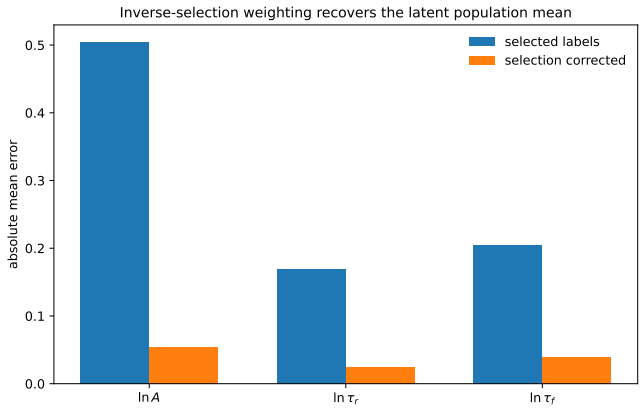

Figure 5. Absolute error in the recovered latent population means. Inverse-selection weighting substantially reduces the bias caused by the spectroscopic-style label mechanism.

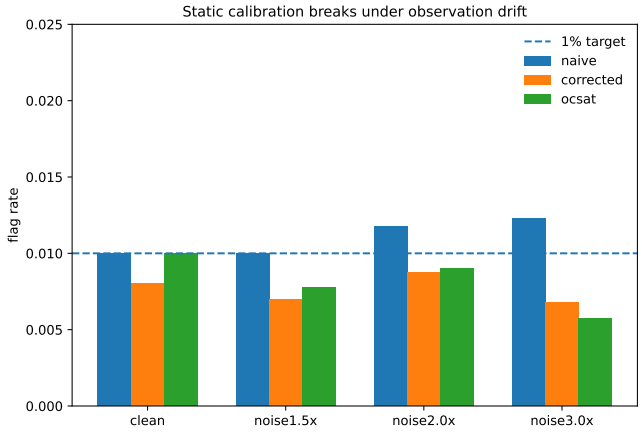

Figure 6. Achieved follow-up rate under changing photometric precision. Thresholds are calibrated in the clean state and then held fixed. The one per cent reference is shown for comparison.

therefore becomes conservative in some degraded states. The achieved flag rates are 0.0080, 0.0070, 0.00875 and 0.00675 across the four states. The adaptive corrected rule gives 0.0085, 0.0115, 0.0105 and 0.0100 over the same states.

The result is useful because it separates two goals that are often combined in one word, calibration. A fixed threshold can remain statistically reasonable in a narrow distribution while wasting follow-up resources after the observing conditions change. The adaptive layer restores the declared action rate without pretending that the underlying astrophysical completeness is preserved.

Figure 6 compares the static methods. The accompanying adaptive results are reported in Table 2. The range of the corrected adaptive rate is 0.0030 across the four states, compared with 0.00125 for the static corrected rate. The adaptive rule should therefore be understood as a budget controller, not as a replacement for selection modelling.

5.4 The physical residual has a separate scientific role

The physical residual is not expected to increase the recall of every rare point. Its intended role is different. It detects observations that depart

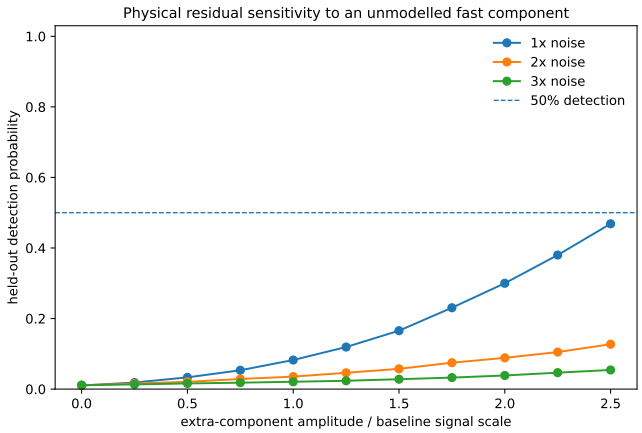

Figure 7. Held-out detection probability for a narrow extra light-curve component under three photometric noise levels. The statistic uses only future observations relative to the physical fit.

from the fitted physical family after the state has been estimated on independent earlier measurements. The injection experiment makes this distinction explicit.

Figure 7 plots the probability of crossing a one per cent matched-filter threshold for a narrow extra component as a function of amplitude and noise scale. At unit noise, an extra component with amplitude 2.5 times the baseline scale is detected with probability 0.469. At twice the noise, the same perturbation is much less secure, with a probability below that of the clean case. This is exactly the behaviour expected from the information content of the exposure. The residual cannot manufacture a signal that the measurement does not contain.

This negative result is scientifically useful. A physical residual should not be sold as a universal rare-object detector. It is a model-checking instrument. Its strongest use is to identify candidates for which the fitted physical explanation stops predicting the observed time evolution.

5.5 Information loss is controlled by the observation itself

For the same latent object, the local Fisher information decreases with noise and cadence penalty. Figure 8 gives the relative scalar information proxy used in the stress test. The exact number is not a universal property of ZTF or Rubin. It simply makes the conditioning variable visible. The posterior uncertainty should widen when the observation contains less information.

6 DISCUSSION

The main result is a change in accounting rather than a claim that one score dominates every alternative. The real observational audit shows why. Current astronomical catalogues are assembled from multiple discovery channels and multiple levels of characterization. The data entering a training set therefore carry information about the science and information about the observing policy. A machine-learning system can use both unless the analysis explicitly separates them.

The selection correction addresses the first problem. It attempts to recover the sky population represented by the labels, rather than treating the labelled distribution as the sky itself. The experimental recovery of the latent mean and covariance demonstrates the mechanism in a setting where the selection probability is known. In

Table 2. Static triage results from the controlled experiment. The scientific rarity recall is calculated at a one per cent nominal follow-up budget.

State	Method	Flag rate	Rare recall	Unusual recall	False discovery fraction
clean	naive	0.01000	0.167	0.070	0.900
clean	selection corrected	0.00800	0.167	0.088	0.844
clean	OCSAT score	0.01000	0.167	0.070	0.900
noise 1.5×	naive	0.01000	0.207	0.103	0.850
noise 1.5×	selection corrected	0.00700	0.172	0.086	0.821
noise 1.5×	OCSAT score	0.00775	0.138	0.069	0.871
noise 2.0×	naive	0.01175	0.138	0.061	0.915
noise 2.0×	selection corrected	0.00875	0.138	0.061	0.886
noise 2.0×	OCSAT score	0.00900	0.138	0.061	0.889
noise 3.0×	naive	0.01225	0.150	0.067	0.939
noise 3.0×	selection corrected	0.00675	0.150	0.067	0.889
noise 3.0×	OCSAT score	0.00575	0.150	0.067	0.870

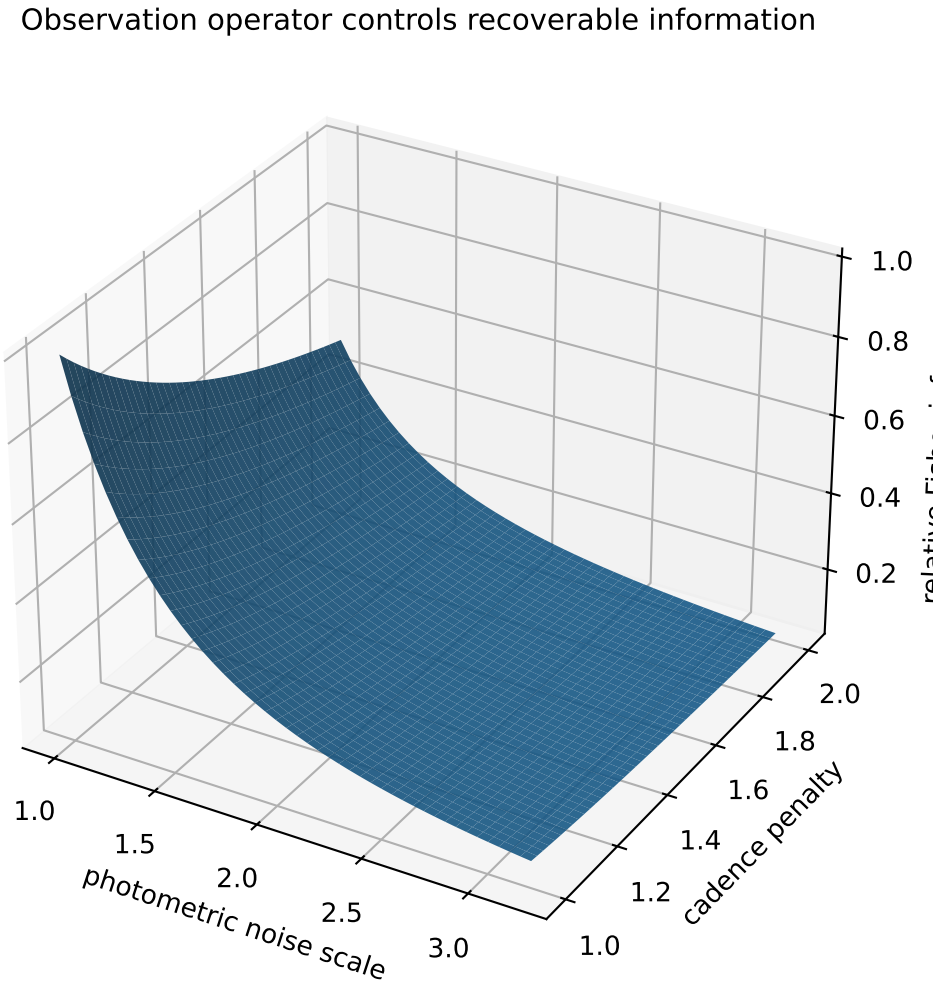

Figure 8. Relative Fisher information as a function of photometric noise and cadence penalty in the controlled observation model. The surface is normalized to the clean state. It is a diagnostic of the observation operator, not a forecast for a particular instrument.

real astronomy, the hard part is estimating that probability without introducing another poorly tested model. This is why injection and recovery experiments remain important (Herle et al. 2024; Zwicky Transient Facility 2021).

The observation-conditioned posterior addresses a second problem. A posterior should depend on the information in the actual exposure. The same source observed under degraded conditions should not inherit the same uncertainty as a high signal-to-noise observation merely because the latent object is unchanged. This point has an obvious connection to measurement-error-aware conformal inference (Singer et al. 2025). The present framework uses that dependence earlier, inside the population and action problem.

The adaptive threshold addresses a third problem. A survey with a fixed follow-up budget cannot treat a threshold chosen in the first

month as sacred for ten years. The appropriate operating point is a monitored property of the live system. The adaptive rule used here is deliberately modest. It controls the action rate. It does not guarantee scientific completeness.

The physical residual addresses the most difficult failure mode. A model can be perfectly calibrated while its hypothesis class is wrong. A rare object that lies outside the physical family can be misclassified as merely noisy. A future observation not used in the fit provides a simple test of whether the physical explanation continues to make predictions. This is closely related to posterior predictive checking, although the implementation here is prequential because the astronomy use case is sequential.

This also explains why the full OCSAT score is not uniformly superior in Table 2. Adding the physical residual changes the ranking problem. It spends part of the fixed action budget on model-breaking candidates, and in the present simulation that trade can lower pure rarity recall. That behaviour is not a defect to be hidden. It is a statement about scientific priorities. A broker that wants rare objects within a known family should use a different weighting from a broker whose primary objective is to find departures from the family.

The framework is compatible with active learning. A selected follow-up sample should become the next training set only after its selection probability is recorded. This is especially important because active follow-up deliberately chooses objects that a passive survey would not choose. Recent real-time active learning work already shows that strategically selected spectra can improve training efficiency (Moller et al. 2025). The missing step is to carry the action rule and its selection history into the subsequent population estimate.

The framework is also compatible with image and radio applications. The observation operator changes, but the logic does not. In strong lensing, the image formation model and network selection function can be separated (Herle et al. 2024). In radio transient searches, anomaly features and active human labelling can be treated as an upstream selection system rather than as a final classifier only (Andersson et al. 2025).

7 LIMITATIONS

The controlled simulation is not an on-sky validation. It is designed to identify causal effects because the population, label mechanism and observational degradation are known. That allows clean conclusions about what each component does, but it cannot capture every failure in a real alert broker.

The population model is Gaussian in latent parameter space. Real transient populations are mixtures and can contain long tails. A more general implementation should replace the Gaussian approximation with a flexible density estimator or a hierarchical population model. The covariance used for the local posterior is based on a local Fisher calculation around a physical template. Real fits can have multimodal posteriors and discrete model degeneracies.

The selection model is estimated from the same variables used to construct the controlled example. In an observatory, selection can depend on variables that are not all available after reduction. A selection correction cannot repair a probability that was never recorded. It can only quantify the supported part of the sample.

The conformal layer controls the operating rate only to the extent that its calibration assumptions are reasonable. The adaptive rule was tested under deliberately imposed drift but is not a distribution-free theorem for a feedback-controlled astronomical stream. In particular, the follow-up process creates delayed labels and a possible dependence between current actions and future calibration data.

The structural anomaly experiment also has a clear boundary. The injected fast component is a known perturbation. It proves sensitivity to that perturbation under the stated observation model. It does not establish that the score will discover every real unknown phenomenon. A residual can also arise from bad calibration, blending or an incomplete physical model.

The external literature audit was performed with current indexed literature and public sources available on 29 August 2026. It is not a formal systematic review. We therefore do not claim absolute priority for the full framework. The novelty claim is narrower. We found separate treatments of selection functions, measurement-error-aware conformal inference, active follow-up and physics-informed modelling, but no directly matching method in our audit that joins these components at the upstream time-domain decision boundary while conditioning the physical check on the actual observation sequence.

8 REPRODUCIBILITY AND IMPLEMENTATION

The complete experiment is deterministic under the published random seed. The source distribution, selection law, physical light curve, observational covariance, threshold rule and generated tables are stored with the manuscript. The analysis can be rerun with Python using NumPy, SciPy, pandas, scikit-learn and Matplotlib. The main data products are CSV files rather than binary proprietary formats so that the numerical path is inspectable.

The public observational anchors are independently retrievable. The NASA Exoplanet Archive publishes current summary counts and Table Access Protocol interfaces (NASA Exoplanet Archive 2026). The YSE DR1 light curves are publicly released through Zenodo (Aleo et al. 2023). The ZTF public data system provides light-curve access and documents the archive structure and quality flags (Zwicky Transient Facility 2021). The 2026 nuclear transient release supplies classified and unclassified source tables and light curves through Zenodo (Anilkumar et al. 2026).

For an actual survey deployment, the minimum scientific certificate should contain the observation metadata, the estimated selection probability, the calibration set and threshold, the model version, the physical residual definition, the support diagnostic and the exact action record. Those objects are more important for reproducibility than the name of the neural architecture.

9 CONCLUSION

Astronomical machine learning has reached the point where raw predictive performance is no longer an adequate description of scientific reliability. The same survey can change its depth, cadence, calibration and follow-up policy without changing the physical sky. A model that ignores those changes can remain accurate on its own validation sample and still answer the wrong scientific question.

We proposed observation-conditioned selection-aware conformal triage as a practical response. The framework treats the labelled population, the measurement process and the follow-up decision as parts of one statistical experiment. In the controlled population study, inverse-selection correction substantially reduced latent population bias. In the drift experiment, adaptive action calibration restored a declared one per cent operating rate after changes in photometric precision. In the physical injection test, a prequential residual supplied an independent route to model-breaking candidates and exposed

the expected loss of sensitivity when the exposure becomes less informative.

The most important practical conclusion is also the simplest. A scientific alert should not leave the inference system as a score alone. It should leave with a statement of what population the score was calibrated on, how the measurement was obtained, how much information the exposure contained, whether the physical model predicts the later observations, and how the follow-up decision was made. That is a modest requirement compared with the scale of modern surveys. It is also a requirement that can be implemented with current statistical and computational tools.

The next useful test is a prospective one. The same certificate should be generated from a live public alert stream, followed through delayed spectroscopic labels, and evaluated field by field rather than only in a pooled benchmark. The present study supplies the statistical structure needed for that test and, equally importantly, states where the evidence would still be insufficient.

DATA AVAILABILITY

The observational scale audit is based on public NASA Exoplanet Archive statistics, the YSE DR1 release and the public ZTF data system. The 2026 ZTF nuclear transient release is cited as an additional public source of current time-domain candidate counts. All controlled-experiment data and derived tables are generated by the supplied analysis script.

CODE AVAILABILITY

The experiment script and LaTeX source accompany this manuscript. The fixed random seed is 260829. The generated metrics, figures and SHA256 manifest are included in the accompanying research directory.

ACKNOWLEDGEMENTS

This work used public astronomical data products and public software. The author thanks the maintainers of the NASA Exoplanet Archive, the Zwicky Transient Facility, the Young Supernova Experiment and the cited research communities for maintaining public data and documentation.

REFERENCES

Aleo P. D., Malanchev K., Sharief S. N., et al., 2023, ApJS, 266, 9
Andersson A., Lintott C., Fender R., et al., 2025, MNRAS, 538, 1397
Anilkumar V., van Velzen S., Kowalski M., Reusch S., 2026, ZTF forced-photometry light curves of nuclear transient candidates, Zenodo, doi:10.5281/zenodo.21628415
Bazin G., et al., 2009, A&A, 499, 653
Bellm E. C., et al., 2019, PASP, 131, 018002
Gibbs I., Candès E., 2021, Adaptive conformal inference under distribution shift, arXiv:2106.00170
Herle A., O'Riordan C. M., Vegetti S., 2024, MNRAS, 534, 1093
Huertas-Company M., Lanusse F., 2023, PASA, 40, e001
Kelly B. C., 2007, ApJ, 665, 1489
Masci F. J., et al., 2019, PASP, 131, 018003
Moller A., Ishida E. E. O., Peloton J., et al., 2025, PASA, 42, e057
NASA Exoplanet Archive, 2026, Exoplanet and Candidate Statistics, Caltech/IPAC, https://exoplanetarchive.ipac.caltech.edu/docs/counts_detail.html
Patterson M. T., et al., 2019, PASP, 131, 018001
Richards J. W., Starr D. L., Brink H., et al., 2012, MNRAS, 419, 1121
Singer N., Williams J. P., Ghosh S., 2025, MNRAS, 539, 1372
Ting Y.-S., 2024, Why machine learning models systematically underestimate extreme values, Open Journal of Astrophysics, 7, doi:10.33232/001c.142224
Ting Y.-S., 2026, ARA&A, 64, 19
Zwicky Transient Facility, 2021, Public Data Release 8, Caltech/IPAC, https://ztf.ipac.caltech.edu/page/dr8
ZTF Bright Transient Survey, 2026, BTS Sample Explorer documentation, https://sites.astro.caltech.edu/ztf/bts/explorer_info.html