Draft version February 17, 2026
Typeset using LaTeX **twocolumn** style in AASTeX7.0.1

J-PAS: Semi-Supervised Sim-to-Obs Transfer for Robust Star–Galaxy–Quasar Classification

Daniel López-Cano [1,2] L. Raul Abramo [1] L. Nakazono [3] I. Pérez-Ràfols [4] G. Martínez-Solaeche [5] J. Chaves-Montero [6] Matthew M. Pieri,[7] Jailson Alcaniz,[3] Narciso Benitez,[8] Silvia Bonoli,[9] Saulo Carneiro,[3] Javier Cenarro,[10,11] David Cristóbal-Hornillos,[10] Simone Daflon,[3] Renato Dupke,[3] Alessandro Ederoclite,[10] Rosa González Delgado,[12] Antonio Hernán-Caballero,[10,11] Carlos Hernández–Monteagudo,[13,14] Jifeng Liu,[15] Carlos López-Sanjuan,[10,11] Antonio Marín-Franch,[10,11] Claudia Mendes de Oliveira,[16] Mariano Moles,[10] Fernando Roig,[3] Laerte Sodré Jr.,[16] Keith Taylor,[17] Jesús Varela,[10] Héctor Vázquez Ramió,[10,11] Jose Vilchez,[12] and Javier Zaragoza-Cardiel[10,11]

[1] *Department of Mathematical Physics, Institute of Physics, University of São Paulo, R. do Matão 1371, 05508-090, São Paulo, SP, Brazil*
[2] *Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), R. Pio XI, 1500, Alto da Lapa, 05468-901, São Paulo, SP, Brasil*
[3] *Observatório Nacional, Rua General José Cristino, 77, São Cristóvão, 20921-400 Rio de Janeiro, RJ, Brazil*
[4] *Departament de Física, EEBE, Universitat Politècnica de Catalunya, c/Eduard Maristany 10, 08930 Barcelona, Spain*
[5] *Instituto de Astrofísica de Andalucía (IAA-CSIC), P.O. Box 3004, 18080 Granada, Spain*
[6] *Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, 08193 Bellaterra (Barcelona), Spain*
[7] *Aix Marseille Univ, CNRS, CNES, LAM, Marseille, France*
[8] *Independent Researcher*
[9] *Donostia International Physics Center (DIPC), Manuel Lardizabal Ibilbidea, 4, San Sebastián, Spain*
[10] *Centro de Estudios de Física del Cosmos de Aragón (CEFCA), Plaza San Juan, 1, E-44001, Teruel, Spain*
[11] *Unidad Asociada CEFCA-IAA, CEFCA, Unidad Asociada al CSIC por el IAA y el IFCA, Plaza San Juan 1, 44001 Teruel, Spain*
[12] *Instituto de Astrofísica de Andalucía - CSIC, Apdo 3004, E-18080, Granada, Spain*
[13] *Instituto de Astrofísica de Canarias, C/ Vía Láctea, s/n, E-38205, La Laguna, Tenerife, Spain*
[14] *Universidad de La Laguna, Avda Francisco Sánchez, E-38206, San Cristóbal de La Laguna, Tenerife, Spain*
[15] *National Astronomical Observatory of China, Chinese Academy of Sciences, Beijing, China*
[16] *Departamento de Astronomia, Instituto de Astronomia, Geofísica e Ciências Atmosféricas, Universidade de São Paulo, Brazil*
[17] *Instruments4, 4121 Pembury Place, La Canada Flintridge, CA 91011, U.S.A.*

ABSTRACT

Modern studies in astrophysics and cosmology increasingly rely on simulations and cross-survey analyses, yet differences in data generation, instrumentation, calibration, and unmodeled physics introduce distribution mismatches between datasets ("domain shift"). In machine-learning pipelines, this occurs when the joint distribution of inputs and labels differs between the training (source) and application (target) domains, causing source-trained models to underperform on the target. Transfer learning and domain adaptation provide principled ways to mitigate this effect. We study a concrete simulation-to-observation case: semi-supervised domain adaptation (SSDA) to transfer a four-class spectral classifier—high-redshift quasars, low-redshift quasars, galaxies, and stars—from J-PAS mock catalogs based on DESI spectra to real J-PAS observations. Our pipeline pretrains on abundant labeled DESI→J-PAS mocks and adapts to the target domain using a small labeled J-PAS subset. We benchmark SSDA against two baselines: a J-PAS–only supervised model trained with the same target-label budget, and a mocks-only model evaluated on held-out J-PAS data. On this held-out J-PAS data, SSDA achieves a macro-F1 score (balancing precision and recall) of 0.82 and an overall true positive rate of 0.89, compared to 0.79/0.85 for the J-PAS–only baseline and 0.73/0.87 for the mocks-only model. The gains are driven primarily by improved quasar classification, especially in the high-redshift subclass (F1 = 0.66 vs. 0.55/0.37), yielding better-calibrated candidate lists for spectroscopic targeting (e.g., WEAVE-QSO) and AGN searches. This study shows how modest target supervision enables robust, data-efficient simulation-to-observation transfer when simulations are plentiful but target labels are scarce.

Email: daniellopezcano13@gmail.com

Keywords: Astronomy data analysis (1858) — Astrostatistics (1882) — Photometry (1234) — Surveys (1671) — Catalogs (205)

1. INTRODUCTION

Astronomical analyses increasingly aim to combine datasets that share a common goal but differ in project-specific characteristics. On the survey side, many current/forthcoming missions such as the Dark Energy Spectroscopic Instrument (DESI, A. G. Adame et al. 2025), Euclid (Euclid Collaboration et al. 2025), the Legacy Survey of Space and Time (LSST, Ž. Ivezić et al. 2019), WEAVE (G. Dalton et al. 2016; S. Jin et al. 2024), or J-PAS (N. Benitez et al. 2014), will observe overlapping sky regions with different passbands, depths, footprints, and calibration strategies. Analysing them together to perform cross-survey analyses can effectively enlarge survey area and redshift reach, improve completeness, and enable stronger systematic testing. At the same time, heterogeneity in selection, observing conditions, and processing pipelines makes cross-survey work nontrivial.

Simulations constitute another pillar in modern cosmological analyses. For example, N-body runs model the formation and evolution of haloes and of the large-scale structure of the Universe (e.g., Y. Feng et al. 2016; D. Potter et al. 2017; L. H. Garrison et al. 2021; V. Springel et al. 2021; C. Rampf et al. 2025), hydrodynamical simulations describe galaxy formation and feedback processes (e.g., M. Vogelsberger et al. 2014; R. A. Crain et al. 2015; J. Schaye et al. 2023), and strong lensing pipelines can generate mock images resembling direct observations (e.g., P. Bergamini et al. 2023; N. Anau Montel et al. 2023). These tools can be employed to conduct survey validation checks, produce end-to-end forecasts, and train machine-learning (ML) pipelines to perform different tasks. However, rendering any of these simulation outputs into realistic instrument-level observables requires making several assumptions (selection, noise, blending, reduction, etc.), and small differences at any level can accumulate, so simulated mocks and real data typically exhibit some degree of mismatch in properties and characteristics.

Taken together, survey heterogeneity and the approximations required to translate simulations into realistic observations mean that any model trained in one setting will face a different data–label relationship in another (e.g. S. Pierre et al. 2025). We refer to this mismatch as *domain shift*: the joint distribution over inputs and labels differs between the *source* (training) and *target* (application) domains. As a result, the performance of a given model, when evaluated over the target domain, will typically show reduced accuracy, worse calibration, and biased downstream results.

A practical set of techniques has recently been developed by the ML community to mitigate these effects (see X. Liu et al. 2022, for a recent review). *Domain adaptation (DA)* methods transfer knowledge from a labelled source to a related target domain under the same task but different input distributions.

The practical regimes are largely set by target-label availability: with many target labels available, it's common to employ supervised transfer techniques (e.g. Z. Han et al. 2024); with none, *unsupervised domain adaptation (UDA)* aligns source/target representations (e.g., Y. Ganin et al. 2015; E. Tzeng et al. 2017; C.-L. Li et al. 2017); and with a small labelled target subset, *semi-supervised DA (SSDA)* uses those few labels to improve model calibration (e.g., D. Berthelot et al. 2021; Y.-C. Yu & H.-T. Lin 2023).

Related approaches include *domain generalization* – for robustness across multiple sources without seeing the target (see K. Zhou et al. 2021; J. Wang et al. 2021, for recent reviews) and *contrastive (CL)/self-supervised learning (SSL)* – for determining domain-agnostic embeddings from unlabeled data, both of which can improve invariance to nuisance variations (e.g., T. Chen et al. 2020; A. Bardes et al. 2021; S. Alshammari et al. 2025).

These approaches have already proved useful in astronomy and cosmology (see M. Huertas-Company et al. 2023 for a review of CL/SSL): feature-alignment improves robustness to changing depth/PSF in image-based strong lensing tasks (e.g., S. Alexander et al. 2023; P. Swierc et al. 2023, 2024; S. Agarwal et al. 2024; H. Parul et al. 2025; S. Pandya et al. 2025); DA/CL reduce bias and improve coverage in simulation-based inference, and enhance the performance of forward model frameworks (e.g., A. Roncoli et al. 2023; A. Akhmetzhanova et al. 2024; J.-Y. Lee et al. 2024; D. López-Cano et al. 2024; S. Andrianomena & S. Hassan 2025; A. Akhmetzhanova et al. 2025); and DA/DG/SSL improve label efficiency and robustness in cross-survey classification and photometric-redshift pipelines (e.g., M. A. Hayat et al. 2021; A. Ćiprijanović et al. 2021, 2023; J. Vega-Ferrero et al. 2024; S. Gilda et al. 2024).

Overall, these results support a simple message: domain shift is widespread in astrophysical pipelines, and domain-aware learning can improve robustness, cali-

bration, and reproducibility when transferring models across surveys or from simulations to observations.

In this work, we apply one of these domain-shift-aware techniques—in particular, a Semi-Supervised Domain Adaptation (SSDA) approach—to improve the performance of an ML model for classifying astronomical sources based on their J-PAS narrow-band flux vectors (commonly referred to as *J-spectra*). The Javalambre Physics of the Accelerating Universe Astrophysical Survey (J-PAS, N. Benitez et al. 2014) is an ongoing wide-field multiband survey providing quasi-spectroscopic photometry over several thousand square degrees (see https://www.j-pas.org). J-PAS employs 54 contiguous narrow-band filters supplemented by two medium-band and one broad-band filter to deliver precise photometric redshifts for galaxies and quasars (QSOs), map stellar populations, and probe large-scale structure. In this work we use the J-PAS Early Data Release (EDR; https://www.j-pas.org/datareleases/jpas_early_data_release); J-PAS software and data products are maintained within the CEFCA infrastructure, including https://gitlab.cefca.es.

We focus on a concrete, relevant case: training a four-way classifier for *J-spectra* data distinguishing high-redshift quasars, low-redshift quasars, galaxies, and stars. This problem is not merely methodological; accurate quasar probabilities directly enable operational candidate selection for spectroscopic follow-up, where false positives translate into wasted fibres and reduced science yield. In practice, target-selection programs require ranking sources at very low false-positive rates while preserving completeness for rare classes; this places a premium on well-calibrated probabilities rather than just hard labels. In this sense, improved quasar classification feeds directly into survey strategy: it increases the purity of high-z candidate lists, stabilizes forecasts, and helps make the resulting selection function explicit for downstream LSS analyses.

We use the public J-PAS Early Data Release (EDR) as a testing ground for domain-shift–aware classification, leveraging public DESI EDR spectroscopy (DESI Collaboration et al. 2024) together with DESI→*J-spectra* mocks built by projecting DESI spectra into the J-PAS narrow-band system (following C. Queiroz et al. 2023). Source classification in J-PAS and J-PLUS has been widely explored using ML techniques (e.g., P. O. Baqui et al. 2021; R. von Marttens et al. 2024; A. del Pino et al. 2024; A. P. Jeakel et al. 2025), including the mini-JPAS quasar-selection efforts for creating combined catalogs (I. Pérez-Ràfols et al. 2025, 2023; G. Martínez-Solaeche et al. 2023; N. V. N. Rodrigues et al. 2023; C. Queiroz et al. 2023). A recurring outcome of these pipelines is the presence of a *simulation-to-observation* gap, whereby models trained primarily on mocks degrade when applied to real data. Beyond methodology, accurate, calibrated class probabilities are operationally relevant for targeting in upcoming spectroscopic efforts such as WEAVE (M. M. Pieri et al. 2016; S. Jin et al. 2024), where low false-positive rates at fixed completeness for rare high-z quasars directly translate into improved fibre usage and clearer selection functions for downstream large-scale structure analyses.

We adopt a SSDA scheme: pretrain on DESI→J-PAS mocks and adapt on J-PAS using a small labelled subset from the DESI×J-PAS cross-match, so that mocks supply coverage while target labels guide class-conditional alignment. We compare three settings on J-PAS: (i) *SSDA* (this work), (ii) a *target-only* supervised baseline using the same labelled J-PAS budget, and (iii) a *mocks-only* model evaluated zero-shot on J-PAS (i.e., trained solely on DESI→J-PAS mocks and applied directly to J-PAS observations with no fine-tuning, reweighting, or calibration using J-PAS sample). Performance is quantified with class-resolved diagnostics, emphasising the quasar subclasses most consequential for follow-up targeting under domain shift. We release the full pipeline at https://github.com/daniellopezcano/JPAS_Domain_Adaptation—data curation, training and SSDA adaptation code, sweep configurations, and plotting scripts—including YAML files for the `wandb` sweeps and scripts to regenerate all tables and figures.

The remainder of the paper is organised as follows: Sect. 2 describes the J-PAS data, the DESI→J-PAS mocks, the DESI×J-PAS cross-match, preprocessing, and the train/validation/test splits. Sect. 3 details the model, training regimes, and losses. Sect. 4 presents the main results with class-resolved diagnostics (confusion matrices, F1, etc.). Sect. 5 summarizes our findings and discusses the implications for cross-survey and sim-to-obs pipelines.

2. DATA

Our work focuses on classifying J-PAS sources into four distinct classes denoted as: high-redshift quasars (`QSO_high`), low-redshift quasars (`QSO_low`), galaxies (`GALAXY`), and stars (`STAR`).

We employ two different datasets to train and evaluate the performance of our models. First, we build a large labelled *DESI→J-PAS mocks* sample by projecting DESI spectra into the J-PAS photometric system. Full details on the procedure to generate simulated J-PAS fluxes from spectra were presented in C. Queiroz et al. (2023), with the key difference being that for the

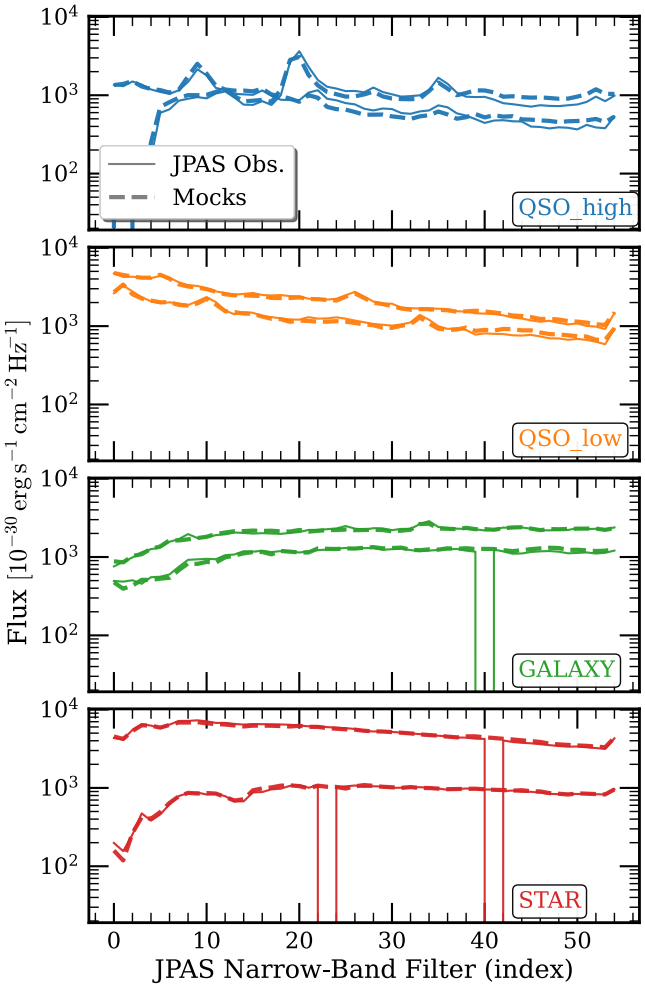

Figure 1. Representative photometric SEDs, comparing J-PAS (solid) with DESI→J-PAS *Mocks* (dashed). Each solid/dashed pair corresponds to a distinct individual object with a one-to-one mock counterpart (i.e., the dashed curve is the mock constructed for that same source); the multiple curves shown in each panel are therefore not averaged templates but separate objects plotted together for illustration. Panels correspond to different classes. Curves span the effective 55-band set after reliability masking (Section 2.2). For visual clarity, we display examples with i-band signal-to-noise ratio > 200; this high-S/N cut is used only for visualization and is not applied elsewhere in the analysis. Within each class, the plotted objects are selected from the subset passing this visualization-only S/N cut (randomly chosen from the eligible set). For legibility, we also omit error bars. We intentionally retain bands with missing flux measurements in J-PAS (visible as gaps) to reflect real observational coverage.

present paper, instead of using the SDSS spectra we now employ DESI spectra.

2.1. *DESI→J-PAS mock generation*

Each DESI spectrum is convolved with the J-PAS transmission curves and passed through an observational model (zeropoints and band-dependent noise) to produce 54 narrow-band J-PAS-like flux estimates. We intentionally exclude the two medium bands at the spectral edges of J-PAS, as the DESI spectroscopic wavelength coverage does not reliably span these passbands; generating mock fluxes there would require extrapolation and could introduce extra domain-dependent systematics. All experiments, therefore, use only the 54 narrow bands plus the i band.

We then apply an *object-by-object multiplicative normalization* to place the synthetic *J-spectra* on the same absolute scale as imaging photometry: we compute synthetic broad-band fluxes from the DESI spectrum in an anchor band and rescale by the ratio of observed-to-synthetic flux. In our pipeline the anchors are the J-PAS i band and the Legacy Surveys r band (used to tie the DESI spectrophotometry to photometry). This step is purely multiplicative (no wavelength-dependent warping) and therefore does not, by construction, introduce color-dependent distortions. The per-band noise model for the mocks is anchored to the J-PAS flux uncertainties measured in the same photometric system (see below), so the resulting synthetic *J-spectra* reproduce the signal-to-noise levels of the chosen J-PAS photometry.

This provides a synthetic data set of labeled observations with approximately 1.5×10^6 unique sources spanning the four classes with good coverage of rare classes such as `QSO_high`.

Second, we assemble a *J-PAS×DESI cross-match* – hereafter *J-PAS Observations* – that attaches DESI spectroscopically generated labels to J-PAS photometric sources (from the Early Data Release catalogue) in the *target* measurement space. This set contains 52,020 objects before quality cuts, with a strongly imbalanced composition (e.g., $\lesssim 2\%$ in `QSO_high`, similar to the mocks). While too small to train high-capacity models by itself, this dataset is ideal for calibration and for assessing performance where it ultimately matters: on real J-PAS observations.

For each J-PAS source, we extract per-band fluxes using the 3″ aperture photometry, including the catalogue PSF aperture correction computed with the help of non-saturated high signal-to-noise ratio (SNR) point-like sources. This choice is the most appropriate for the predominantly unresolved sources, but it has proven the most robust overall in terms of providing the best estimate of the intrinsic SED in a consistent aperture across bands; the corresponding flux uncertainties from this same photometry are propagated into the mock-

generation step to set the per-band noise and reproduce realistic J-PAS-like S/N. We apply catalogue-level quality filtering to remove problematic photometry before constructing the final working sets; the remaining curation steps are listed in Sect. 2.2.

Figure 1 shows representative real *J-spectra* (solid) with their corresponding DESI→J-PAS *mock* counterparts (dashed) for each class. Each data vector spans the effective 55-band set after reliability masking (see subsection 2.2) and is displayed using the same per-source normalization to facilitate visual comparison. As expected, stars show relatively smooth continua; galaxies exhibit color changes and break-like features; and the quasar subclasses (`QSO_low`, `QSO_high`) present localized modulations associated with strong emission features sampled by the narrow bands. The mocks track the overall shapes and relative fluxes well, while small, localized residuals in color and amplitude appear at specific filters and near the edges of the wavelength coverage. Occasional gaps in the J-PAS curves reflect bands removed by coverage/quality cuts, which the mocks do not inherit. This object-by-object view clarifies what the simulator reproduces faithfully (global SED structure and class-separating cues) and where residual mismatches concentrate (localized bands and missing-filter patterns). Even after careful emulation, the simulations show slight differences from the actual J-PAS observations, introducing a *domain shift*.

Unless noted otherwise, each object in our datasets is further equipped with a one-hot *The Tractor* morphology indicator (D. Lang et al. 2016). *The Tractor* is a model-based photometry framework present in the DESI Legacy Surveys, which fits a compact set of parametric light-profile families to imaging data and records the best-fit class (e.g., PSF, GPSF, REX, EXP, DEV, SER, GGAL). These classifications are available for all objects in both J-PAS observations and DESI→J-PAS mocks. The resulting discrete label provides a coarse size/shape cue—separating point-like from extended sources—and is particularly useful for disambiguating stars and QSOs from galaxies at low S/N in J-PAS. We encode this categorical field as a one-hot vector and concatenate it to the photometric features so that the classifiers can condition on morphology without overriding the spectral information.

In the remainder of this section we describe how the raw datasets are curated and organized for the classification experiments.

2.2. *Dataset curation and splits*

As introduced in Sect. 2, our working datasets – *DESI→J-PAS mocks* (source) and *J-PAS obs.* (target) – occupy the same photometric space after reliability masking: 54 J-PAS-like narrow-band filters plus the i-band measurements. Upon these two raw datasets, we apply the following curation steps:

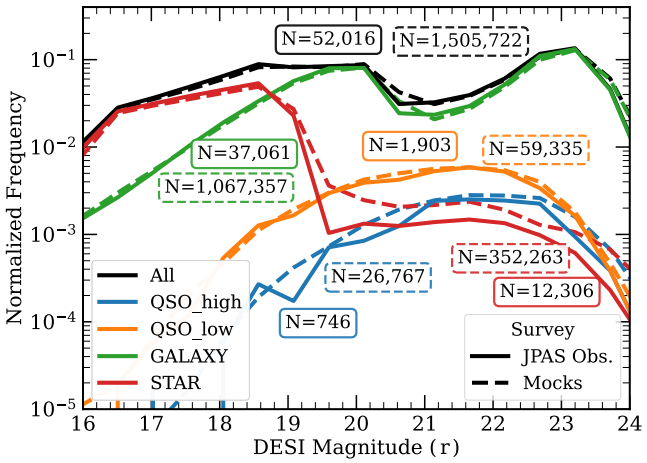

Figure 2. Class balance in our two working datasets: full DESI→J-PAS *Mocks* (source; dashed) versus the *labeled* J-PAS×DESI subset (target; solid). The y-axis shows *relative frequency per magnitude bin* normalized by the total size of the mock catalog ($\simeq 1.5 \times 10^6$ unique sources); consequently, curves appear on the same percentage scale and trace each other in shape even though absolute counts differ widely ($N_{\text{Mocks}} \gg N_{\text{J-PAS}}$; see the total per-class counts in the annotated boxes). The figure includes *all objects after our data curation* (Sect. 2.2); no train/validation/test splits or magnitude cuts are applied here.

1. *Missing and sentinel values.* We drop SED rows that are entirely missing (all entries NaN) or contain sentinel placeholders (−99 or 99). No such rows are present in the *J-PAS Obs.* set as constructed here, whereas the synthetic data occasionally include these placeholders from the mock-generation process; we remove 1,041 fully-NaN SEDs and 609 rows containing sentinel values in the mocks. For partially missing SEDs, we retain the object and linearly interpolate along the wavelength-ordered grid using the closest valid neighbors (affecting $\sim 1.18\%$ of mock SEDs). This interpolation is performed *independently for each object* (per-SED): only the valid flux values of the same SED enter the interpolation. This preserves coverage without introducing target-driven imputation.

2. *Classes and class imbalance.* We classify sources into four classes: `QSO_high` ($z \geq 2.1$), `QSO_low` ($z < 2.1$), `GALAXY`, and `STAR`. We adopt $z = 2.1$ as an operational (survey-driven) threshold commonly used in large optical spectroscopic surveys

to distinguish quasars whose spectra provide a usable Lyα forest from those that do not. For $z \gtrsim 2.1$, the Lyα forest region blueward of 1216 $\AA$ is shifted into the ground-based optical window with sufficient path length for cosmological analyses. This value is therefore widely used in target selection and pipeline definitions of "high-z" (Lyα-forest) quasars. We emphasize that this boundary does not mark a sharp physical transition in quasar properties, but rather reflects observational and analysis requirements. Figure 2 shows per-class counts as a function of magnitude in both domains. The relative abundances are closely matched overall but strongly imbalanced, especially for `QSO_high`.

Three takeaways follow from Fig. 2: (i) the dependence of counts on magnitude reflects DESI selection inherited by both domains; (ii) relative abundances track closely between domains globally and per class; and (iii) the problem is strongly imbalanced, with `QSO_high` comprising only ≈ 1.4–1.8% overall and becoming rarer at bright magnitudes. This motivates our use of a class-aware objective (balanced cross-entropy; see Sect. 3) and macro-averaged/class-resolved evaluation metrics (Sect. 4).

3. *Cross-match and de-duplication across domains (leakage prevention).* We construct a linking table between *mocks* and *J-PAS Obs.* using the `TARGETID` variable contained in both datasets[1]. Conceptually, this enables 'leakage-safe' splits (see point four below): Any mock object linked to a J-PAS source assigned to a target validation/test split is excluded from the source training pool; conversely, J-PAS objects used for training/validation (e.g., in the JPAS-supervised baseline or the small labelled set for SSDA) are excluded from the J-PAS test set. Thus, the same astrophysical source never influences training and evaluation in a different disguise (synthetic vs. observed), and target metrics reflect genuine sim→obs generalization rather than memorization via duplicates.

4. *Train/validation/test splits (and evaluation mask).* For *DESI→J-PAS mocks* (source) we adopt a 70/15/15 split with stable class fractions, removing any overlaps with the J-PAS cross-match before splitting to prevent leakage. For *J-PAS Obs.* (target) we split 30/30/40 into train/validation/test to preserve statistical power during evaluation for sparse regimes, especially `QSO_high`. The target *J-PAS Obs.* training set is therefore limited to $\sim 1.5\times10^4$ objects, which motivates the use of SSDA rather than target-only training. After defining these splits, we apply an evaluation-only magnitude mask: in the test sets of both domains, we retain only sources with $r \leq 22.5$ (our scientific target operating range), while training and validation remain untrimmed to leverage additional faint examples and stabilize rare-class learning.

5. *Normalization.* We compute per-band standardization statistics only on the source training split (mocks) and then apply the resulting fixed transform to both domains. For band b with training mock mean μ_b and standard deviation σ_b, each flux x_b is transformed to $(x_b - \mu_b)/\sigma_b$. A single, source-derived normalization enforces a shared scale while preventing the target distribution from influencing feature scaling.

In summary, after performing basic dataset curation and leak-safe splits, both the source *mocks* and the labelled *J-PAS observations*, live in the same input space: 55 JPAS-like bands plus a one-hot Tractor morphology flag. The training and validation splits will be employed for the different experiments described in Section 3, and the test samples will be used for a fair evaluation comparison in Section 4.

3. METHODS

The data setup in Sect. 2 leads naturally to our training plan: we have a large, label-rich *source* (DESI→J-PAS mocks) and a smaller but measurement-faithful *target* (J-PAS Observations), both curated to share the same feature space. In what follows we define the loss used to handle class imbalance, and compare three complementary training regimes for four-way classification: (i) a mocks-only model evaluated zero-shot on J-PAS, (ii) a semi-supervised domain adaptation (SSDA) variant that adapts source-pretrained features with a small labeled J-PAS subset, and (iii) a J-PAS-supervised baseline trained only on target labels. We also summarize the hyperparameter sweeps that establish model convergence.

3.1. *Loss function: balanced cross-entropy*

All models are trained with a multi-class *balanced cross-entropy* (BaCE) loss. Let $\mathbf{Y} \in \{0,1\}^{N\times C}$ be the

[1] We construct the DESI×J-PAS association with a nearest-neighbor sky match within a $1''$ radius, using great-circle separations on the sphere. When multiple DESI targets fall within $1''$, we keep the closest match and discard ambiguous cases.

true-label targets for a batch of size N over C=4 classes and $\hat{\mathbf{Y}} \in (0,1]^{N\times C}$ the model probabilities (softmax of the logits). The loss is

$$\mathcal{L}_{\rm BaCE}(\mathbf{Y},\hat{\mathbf{Y}}) \;=\; -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\beta_c\, Y_{ic}\,\log \hat{Y}_{ic}\,, \qquad (1)$$

where $\beta_c > 0$ are class-balance coefficients. Intuitively, Eq. 1 increases the penalty for underrepresented classes (e.g., `QSO_high`) so the loss is not dominated by more abundant types such as `STAR`/`GALAXY`. We compute the weights β_c from the empirical frequencies $\hat{\pi}_c$ of the training splits via inverse-frequency normalization:

$$\beta_c \;=\; \frac{\hat{\pi}_c^{-1}}{\frac{1}{C}\sum_{k=1}^{C}\hat{\pi}_k^{-1}}. \qquad (2)$$

During training, we also include standard weight regularization (ℓ_2 term in the loss) to improve generalization and stabilize optimization.

3.2. *Models and training regimes*

All experiments share the input representation and preprocessing from Sect. 2.2: 55 J-PAS bands after reliability masking concatenated with *Tractor* morphology flags. The different classifiers we are going to train are comprised of an encoder f_θ and a downstream head g_ϕ (both multi-layer perceptrons) producing a four-way softmax over `QSO_high`, `QSO_low`, `GALAXY`, `STAR`. We compare three regimes:

(i) Mocks-only (no-DA) zero-shot transfer.—Train (f_θ, g_ϕ) on DESI→J-PAS *mocks* using Eq. 1; select the checkpoint that minimizes the *mock-validation* BaCE (In practice, ranking by mock-validation BaCE or by mock-validation macro-F1 leads to the same top-performing models). This baseline measures the raw sim→obs gap once we test the model on real *J-PAS observations*.

(ii) Semi-supervised domain adaptation (SSDA; DA).—Start from several top-performing no-DA checkpoints, *freeze the downstream head* g_ϕ to keep the decision boundaries (learned from abundant source labels) fixed, and *re-train only the encoder* f_θ on the labeled J-PAS *training* split so that target-domain features align with the head's fixed class regions. Concretely, during SSDA we keep all head downstream model parameters fixed, while the encoder remains trainable. This step warps features so that J-PAS examples fall into the correct class-conditioned regions while keeping the decision surfaces fixed, improving alignment and calibration with modest target supervision.

(iii) J-PAS-supervised (target-only) baseline.—Train (f_θ, g_ϕ) from scratch on the labeled J-PAS *training* split and select by the minimum *J-PAS-validation* BaCE. This baseline quantifies performance when relying exclusively on target labels.

3.3. *Hyperparameter search and convergence*

To ensure convergence and a fair comparison between models, we ran a hyperparameter sweep using the `Weights & Biases` package (L. Biewald 2020) for each regime over network depth/width, dropout, learning rate and weight decay, batch size, and gradient clipping. Model selection uses the validation loss appropriate to each regime: mock-validation BaCE for the no-DA pretraining; J-PAS-validation BaCE for SSDA and J-PAS-supervised models. The top configurations cluster tightly in validation metrics, and re-runs with different seeds reproduce the same training plateaus, indicating that results are not driven by lucky initializations or idiosyncratic settings.

All configuration files (`wandb` sweep specs, seeds, and exact hyperparameters) and the best trained checkpoints for the three regimes are publicly available at: https://github.com/daniellopezcano/JPAS_Domain_Adaptation

4. RESULTS

All results in this section use held-out test splits that were never employed during the training/validation process (see Sect. 3). On the *target* side, we evaluate on the reserved J-PAS Observations test split with an evaluation-only cut of $r \leq 22.5$ (see Sect. 2). On the *source* side, we evaluate the results on the DESI→J-PAS *mock* test split to establish in-domain performance of the *mocks-only* model and quantify the sim→obs gap. After our selection cuts, the J-PAS test split contains $N^{\rm test}_{\rm J\text{-}PAS} = 14{,}777$ objects, and the corresponding mocks test split contains $\approx 10^6$ individual sources.

We report the *best* model in each regime – J-PAS-supervised, Mocks no-DA, and DA – selected using the validation criteria defined in Sect. 3.2.

Figure 3 compiles four confusion matrices[2], one per training/evaluation regime. (Top-left) *J-PAS supervised*: target-only training, evaluated on J-PAS observations test split, (Top-right) *Mocks no-DA (in-domain)*: trained and evaluated on the DESI→J-PAS mock splits, (Bottom-left) *J-PAS no-DA (zero-shot)*:

[2] A confusion matrix is a cross-tabulation of *true* (rows) versus *predicted* (columns) class counts; diagonal cells are correct classifications, and off-diagonals quantify specific misclassifications. Row-normalization makes the color scale comparable across rows/classes.

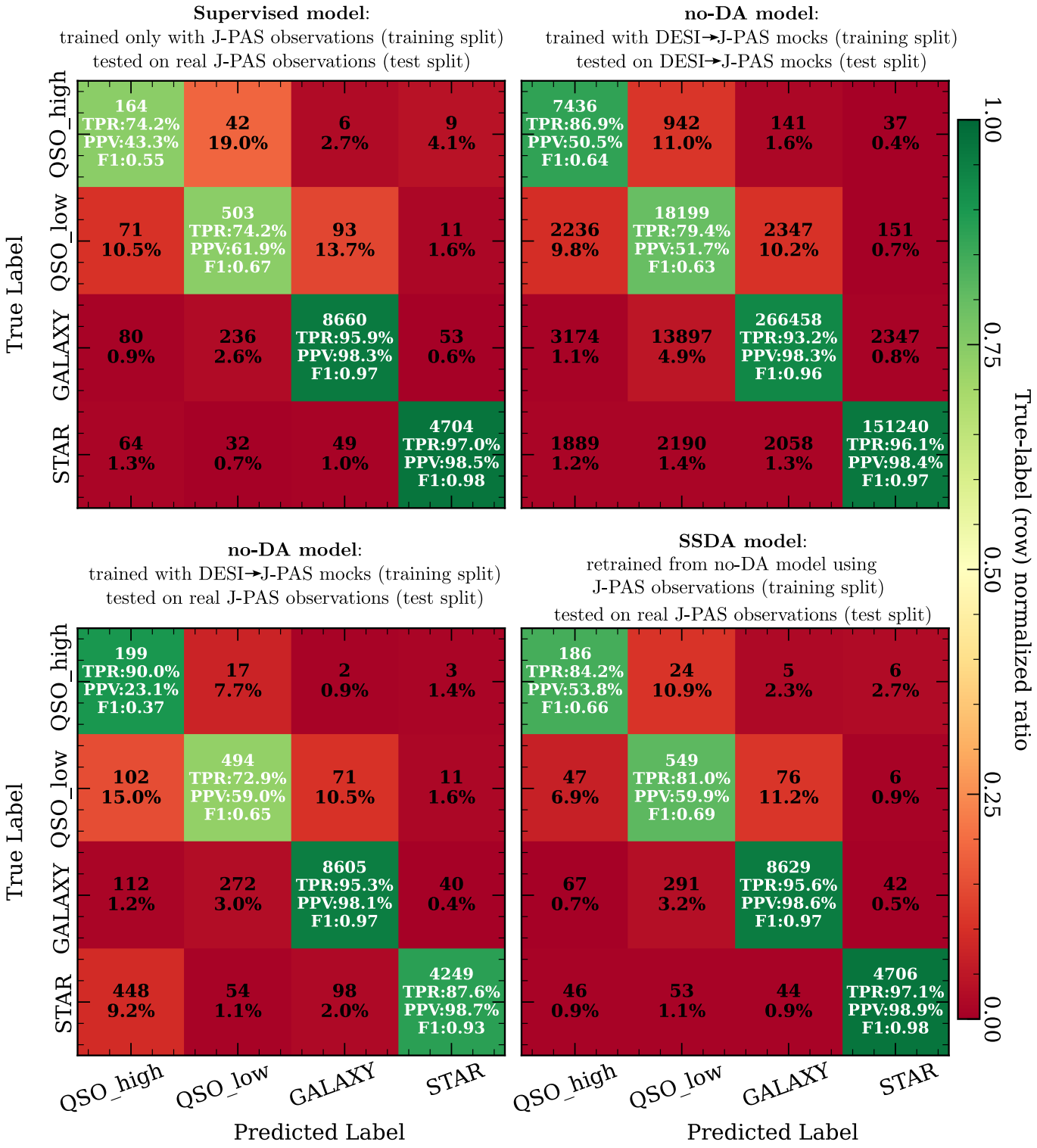

Figure 3. Confusion matrices for four complementary evaluations: (top left) *J-PAS supervised*: target-only baseline, (top right) *Mocks no-DA*: source-domain evaluation on the mock test split, (bottom left) *J-PAS no-DA*: zero-shot transfer of the mocks-trained model, and (bottom right) *J-PAS DA*: the same model after semi-supervised adaptation. Cells show *counts*; the colormap is *row-normalized* to aid visual comparison. Diagonals also list TPR (recall), PPV (precision), and F1 (see Eq. 3) for each class.

trained on mocks, evaluated on J-PAS test observations, and (Bottom-right) *J-PAS DA (SSDA)*: mocks-pretrained, encoder adapted with a small labeled J-PAS set while keeping the head fixed and then evaluated on the J-PAS observations test split. Each cell shows the absolute number of objects in that predicted/true bin, while the colormap is row-normalized to the total number of sources in the corresponding true class to aid cross-class comparison. For the diagonal entries we additionally report three per-class diagnostics: the true positive rate (TPR/recall), the positive predictive value (PPV/precision), and the F1 score, defined from the usual confusion-matrix counts (true positives TP, false positives FP, false negatives FN) as

$$\begin{aligned} \mathrm{TPR} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \\ \mathrm{PPV} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \\ \mathrm{F1} &= \frac{2\cdot\mathrm{TPR}\cdot\mathrm{PPV}}{\mathrm{TPR}+\mathrm{PPV}}. \end{aligned} \tag{3}$$

Across all panels, the residual off–diagonal counts follow physically expected effects (see Appendix A, Fig. 5). First, at low redshift ($z \lesssim 0.3$) host-galaxy light dilutes AGN features in the J-PAS narrow bands, shifting probability from `QSO_low` toward `GALAXY`. Second, near $z \simeq 2.1$, the separation between `QSO_low` and `QSO_high` is set by an explicit redshift threshold, so any method that effectively infers z from *J-spectra* (either explicitly or implicitly, as in our case) will show unavoidable uncertainty-driven mixing across the boundary. Finite filter width and measurement noise limit the redshift precision, and quasars with true redshifts near 2.1 naturally receive similar support for both subclasses.

Third, for *true* `QSO_high` at higher redshift, narrow redshift intervals show additional leakage into `QSO_low` consistent with line-identification aliases—e.g., configurations where Lyα or C IV falling in specific bands can mimic the lower-z template dominated by Mg II in the J-PAS passbands. Appendix A quantifies these behaviours with probability–vs–z medians and dispersions, including a zoom that confirms the structural ambiguity at $z \simeq 2.1$.

Differences between panels reflect the interplay of dataset sizes and domain shift. The *Mocks no-DA* matrix (tested employing in-domain mock observations; top-right) shows very tight behaviour, consistent with training and evaluation sharing the same synthetic distribution. However, when the same model is applied zero-shot to J-PAS (bottom-left), the off-diagonal leakage increases, exposing the sim→obs shift discussed in Sect. 1. A clear effect in the *J-PAS no-DA (zero-shot)* panel appears as extra confusion between `STAR` and QSO classes: in mocks about 1.2% of `STAR` are labelled as `QSO_high`, while on real J-PAS observations this rises to 9.2%. After SSDA, the confusion drops to $\simeq 0.9\%$, comparable to the in-domain mock value, indicating that the $\sim$8% excess in the zero-shot case is predominantly a sim→obs domain-shift effect rather than a physical degeneracy.

The *J-PAS supervised* baseline (top-left) is competitive overall but constrained by the limited training dataset and imbalanced target-label budget. The rare-class boundaries are less stable, and higher levels of `QSO_high`/`QSO_low` confusion and `QSO_low`/`GALAXY` mixing remains. By contrast, the *J-PAS DA* panel (bottom-right) shows a visible re-concentration of counts along the diagonal for both quasar classes, while `GALAXY`/`STAR` remain essentially saturated. In other words, the adaptation step warps the encoder so that J-PAS examples fall into the correct class-conditional regions constructed from abundant source labels (Sect. 3.2), reducing the ambiguities expected from spectral physics and measurement noise. Visual inspection already suggests that SSDA yields the best on-target performance; to avoid qualitative comparisons, we now turn to compact, class-resolved metrics. A per-class breakdown of accuracy, F1, recall, precision, 'area-under-curve' (AUC), and calibration is provided in Appendix B.

Figure 4 summarizes per-class behavior in two complementary ways. The top panel (radar plot) shows the F1-scores by class; this makes the rare-class trade-offs easy to read. The global performance of the models can be interpreted as the enclosed area defined by the different closed-curves. The bottom panel displays per-class ROC curves on test-split J-PAS observations with AUC values. The low-FPR region (left side) is particularly relevant for constructing clean candidate lists for follow-up surveys. Both views tell a consistent story: SSDA lifts the performance for the two QSO subclasses simultaneously: F1 increases and the Receiver Operating Characteristic (ROC) curves shift up/left—while leaving `GALAXY`/`STAR` essentially unchanged. Minor dips for a majority class can occur (the training objective is the balanced cross-entropy; Sect. 3.1), but the net effect is positive: macro metrics improve and the label-scarce, shift-sensitive QSO classes benefit the most. SSDA is optimized to improve target-domain performance under our training objective loss (Eq. 1), but it does not guarantee strictly monotonic gains in every derived local metric; for `QSO_high`, for instance, the improvement of the F_1-score and the PPV value can come with a reduction ($\sim 6\%$, see Figure 3) in the TPR value.

To guide operational thresholds for future spectroscopic surveys focused on QSO detection such as

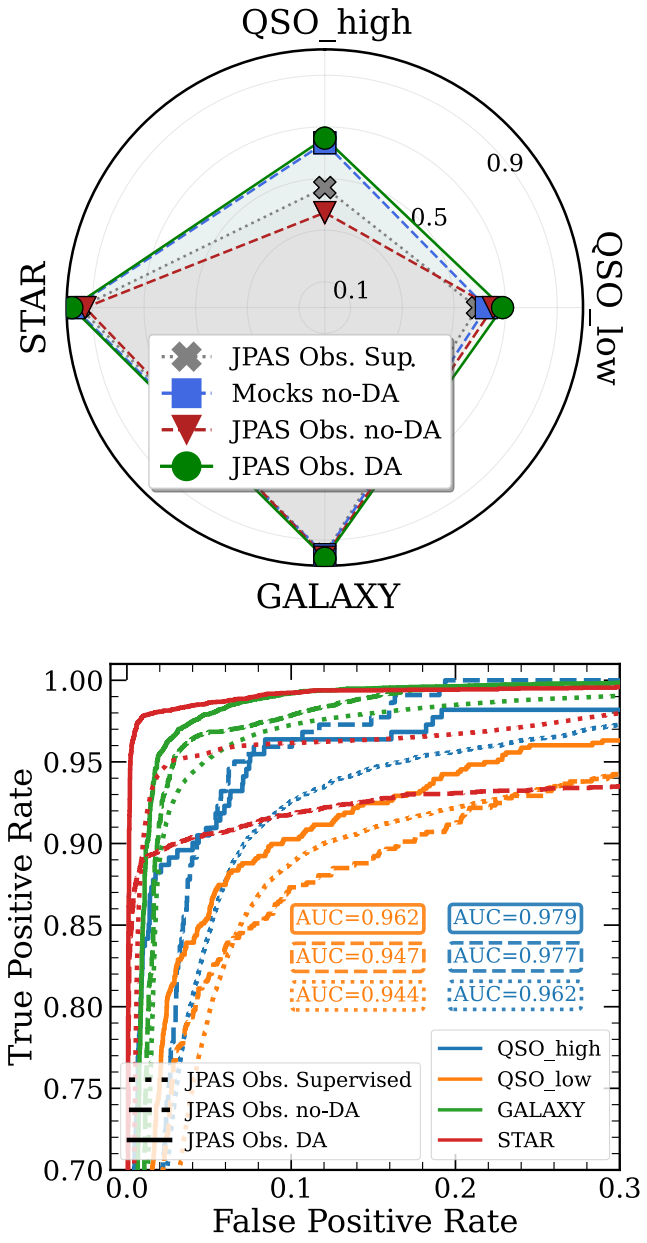

Figure 4. Per-class summary diagnostics. *Top:* F1 scores by class (radar plot) shown for all four regimes (different markers with distinct colors and linestyles). *Bottom:* Per-class ROC curves and AUC values; the low-FPR region (left) is most relevant for building clean follow-up target lists. Both panels reflect the same trend seen in Fig. 3: SSDA boosts the quasar subclasses while leaving `GALAXY`/`STAR` essentially saturated.

WEAVE-QSO, Appendix B examines how class-wise F1 varies with simple r-band magnitude cuts, see Fig. 7.

Table 1 consolidates global statistics on the corresponding test-splits for the four different experiments. We report overall accuracy (Acc), macro-averaged F1, TPR/recall, PPV/precision, AUC, and the Expected

Table 1. Global metrics on the test splits for each experiment. Column keys correspond to: J-sup (J-PAS supervised), M-noDA (mocks in-domain), J-noDA (zero-shot from mocks to J-PAS), and J-SSDA (domain-adapted). We summarize overall accuracy (Acc), macro-F1, macro recall (TPR), macro precision (PPV), macro area under the ROC curve (AUC), and expected calibration error (ECE; lower is better).

Metric	J-sup	M-noDA	J-noDA	J-SSDA
Acc	0.950	0.934	0.917	**0.952**
F1	0.792	0.798	0.729	**0.824**
TPR	0.853	0.889	0.865	**0.894**
PPV	0.755	0.747	0.697	**0.778**
AUC	0.957	**0.977**	0.959	0.975
ECE	0.054	0.047	**0.045**	0.048

Calibration Error (ECE), a standard summary statistic of probabilistic calibration computed by binning predictions by confidence. With M bins B_m of sizes n_m, empirical accuracies $\mathrm{acc}(B_m)$, and mean confidences $\mathrm{conf}(B_m)$,

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{n_m}{N} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|. \qquad (4)$$

Higher values indicate better performance except for ECE (where lower means better). "Macro" scores are unweighted means of the per-class values. The AUC summarizes global threshold-free ranking quality.

As anticipated from the confusion-matrix trends, the SSDA model delivers the strongest *on-target* performance on J-PAS across most metrics: it attains the highest accuracy (0.952), macro-F1 (0.824), macro-TPR/recall (0.894), and macro-PPV/precision (0.778). For the AUC score, the *Mocks no-DA (in-domain)* model reaches 0.977—unsurprising, since this evaluates ranking on the same synthetic distribution used for training. On real J-PAS test-set observations, SSDA's AUC (0.975) is the best among the J-PAS evaluations and consistent with its gains elsewhere. Regarding calibration, the lowest ECE is observed for *J-PAS no-DA* model (0.045) while SSDA is close (0.048); these close values indicate similarly good calibration in practice. The *J-PAS supervised* baseline remains competitive but is ultimately limited by the scarcity and class imbalance of labelled `QSO_high`, which SSDA mitigates by leveraging abundant source labels plus modest target supervision. These aggregate numbers are consistent with the class-resolved trends summarized in Appendix B.

5. CONCLUSIONS

Domain shift is transversal across many studies in astronomy and cosmology: simulators and overlapping

surveys do not share identical characteristics, so the performance of most models drops when deployed out of domain. Here we focused on a concrete case, J-PAS four-way source classification, and asked how far a minimal Semi-Supervised Domain Adaptation step can bridge the simulations-to-observations gap.

The approach is simple. We pretrain on DESI→J-PAS *mocks* (DESI spectra convolved with the J-PAS narrow-band transmission curves to yield *J-spectra*-like with DESI labels) and then nudge the representation to real J-PAS by adapting only the encoder on a small, leak-safe labeled subset while keeping the classification head fixed. On held-out J-PAS data, our adapted model improves accuracy to 0.952 (from 0.917), macro-F1 to 0.824 (from 0.729), and AUC to 0.975 (from 0.959), with well-behaved calibration (ECE$\approx$0.05). The quasar subclasses benefit most: the distinction between high and low redshift quasars (split at $z = 2.1$) is cleaner and confusion with galactic sources is reduced; for example, F1(`QSO_high`) reaches 0.66 versus 0.55 (J-PAS–supervised) and 0.37 (no-DA zero-shot).

For WEAVE-QSO survey (M. M. Pieri et al. 2016; S. Jin et al. 2024), calibrated class probabilities, especially for quasars, support clean candidate lists built at low false-positive rates; PPV-based yield forecasts allow field-by-field budgeting. The adaptation is data-efficient, so periodic refreshes can track slow changes in seeing or zeropoints. Keeping per-object selection probabilities makes the selection function explicit for $n(z)$ and clustering analyses.

Looking ahead, the same techniques applied in this work can be extended to related studies involving cross-survey classification, photo-z estimation, and forward-modeling pipelines. Contrastive/self-supervised pretraining on unlabeled data can learn domain-agnostic embeddings that transfer well across different simulation-specific characteristics (sub-grid physics, resolution effects, etc.) and instrument specifications (passbands, depths, and PSFs); coupled with domain adaptation techniques, this helps bridge survey-to-survey and sim-to-obs gaps by reducing bias and calibration drift, preserving class boundaries, and improving coverage and likelihood calibration. We encourage treating such domain-aware methods as standard practice so that large simulated or auxiliary datasets translate into reliable, reproducible results on the target observations.

ACKNOWLEDGMENTS

We are grateful to Rodrigo Voivodic and Natalia Villa Nova Rodrigues for insightful discussions, constructive suggestions, and many helpful comments throughout this work.

DLC is supported by the São Paulo Research Foundation (FAPESP) via a Postdoctoral Fellowship (Process No. 2024/05768-9), linked to the FAPESP Thematic Project (Process No. 2019/26492-3).

RA acknowledges support from the São Paulo Research Foundation (FAPESP) through the same Thematic Project (Process No. 2019/26492-3) and from the Brazil–France FAPESP–ANR program (FAPESP Process No. 2022/03426-8).

IPR was supported by funding from the grant PID2023-151122NA-I00 by MICIU/AEI/10.13039/501100011033 and by ERDF/EU.

G.M.S. acknowledges financial support from Severo Ochoa grant CEX2021-001131S funded by MICIU/AEI/10.13039/501100011033 and project PID2022-141755NB-I00. G.M.S. also acknowledges support from AST22_00001_Subp 26 and 11, funded by the European Union–NextGenerationEU; the Ministerio de Ciencia, Innovación y Universidades; the Plan de Recuperación, Transformación y Resiliencia; the Consejería de Universidad, Investigación e Innovación (Junta de Andalucía); and the Consejo Superior de Investigaciones Científicas

RGD acknowledges financial support from the project PID2022-141755NB-I00 and the Severo Ochoa grant CEX2021-001131-S, funded by MICIU/AEI/10.13039/501100011033.

This research used data obtained with the Dark Energy Spectroscopic Instrument (DESI). DESI construction and operations is managed by the Lawrence Berkeley National Laboratory. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of High-Energy Physics, under Contract No. DE–AC02–05CH11231, and by the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility under the same contract. Additional support for DESI was provided by the U.S. National Science Foundation (NSF), Division of Astronomical Sciences under Contract No. AST-0950945 to the NSF's National Optical-Infrared Astronomy Research Laboratory; the Science and Technology Facilities Council of the United Kingdom; the Gordon and Betty Moore Foundation; the Heising-Simons Foundation; the French Alternative Energies and Atomic Energy Commission (CEA); the National Council of Humanities, Science and Technology of Mexico (CONAHCYT); the Ministry of Science and Innovation of Spain (MICINN), and by the DESI Member Institutions: https://www.desi.lbl.gov/collaborating-institutions. The DESI collaboration is honored to be permitted to conduct scien-

tific research on I'oligam Du'ag (Kitt Peak), a mountain with particular significance to the Tohono O'odham Nation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. National Science Foundation, the U.S. Department of Energy, or any of the listed funding agencies.

This paper has gone through an internal review by the J-PAS collaboration. Based on observations made with the JST/T250 telescope and JPCam at the Observatorio Astrofísico de Javalambre (OAJ), in Teruel, owned, managed, and operated by the Centro de Estudios de Física del Cosmos de Aragón (CEFCA). We acknowledge the OAJ Data Processing and Archiving Unit (UPAD) for reducing and calibrating the OAJ data used in this work. Funding for the J-PAS Project has been provided by the Governments of Spain and Aragón through the Fondo de Inversiones de Teruel; the Aragonese Government through the Research Groups E96, E103, E16_17R, E16_20R, and E16_23R; the Spanish Ministry of Science and Innovation (MCIN/AEI/10.13039/501100011033 y FEDER, Una manera de hacer Europa) with grants PID2021-124918NB-C41, PID2021-124918NB-C42, PID2021-124918NA-C43, and PID2021-124918NB-C44; the Spanish Ministry of Science, Innovation and Universities (MCIU/AEI/FEDER, UE) with grants PGC2018-097585-B-C21 and PGC2018-097585-B-C22; the Spanish Ministry of Economy and Competitiveness (MINECO) under AYA2015-66211-C2-1-P, AYA2015-66211-C2-2, and AYA2012-30789; and European FEDER funding (FCDD10-4E-867, FCDD13-4E-2685). The Brazilian agencies FINEP, FAPESP, FAPERJ and the National Observatory of Brazil have also contributed to this project. Additional funding was provided by the Tartu Observatory and by the J-PAS Chinese Astronomical Consortium.

AUTHOR CONTRIBUTIONS

Contributions are described using the CRediT (Contributor Roles Taxonomy; ANSI/NISO) framework (https://credit.niso.org).

DLC: *Conceptualization*: defined the sim→obs transfer problem and the domain adaptation pipeline. *Methodology*: designed the SSDA recipe, loss function, and evaluation metrics. *Software*: implemented data pipeline, training/evaluation code, and `wandb` sweeps. *Data curation*: created leak-safe splits; applied reliability masking; standardized bands. *Formal analysis*: Trained the different classification models; provided physical interpretation of the results. *Validation*: computed evaluation metrics (macro-F1, TPR/PPV, AUC, ECE). *Visualization*: produced all figures. *Investigation*: carried out a literature review of previous works employing DA/CL techniques in astrophysics and cosmology. *Resources*: packaged the public repository; built the machine where the analysis was performed. *Writing—original draft*: wrote the manuscript core sections and appendix materials.

RA: *Conceptualization*: helped frame the sim→obs strategy and target metrics. Contributed to scientific discussions within the QSO identification working group. *Resources*: generated DESI→J–PAS mock catalogs and the instrument/noise model used to render *J-spectra*. *Writing—review & editing*: provided iterative feedback on methods, results, and interpretation. *Funding acquisition*: secured support via the FAPESP Thematic Project (Process No. 2019/26492–3) and the Brazil–France FAPESP–ANR program (FAPESP Process No. 2022/03426–8), enabling data preparation, compute resources, and personnel time.

LN: *Data curation*: cleaned J–PAS IDR inputs, applied reliability/coverage masks. *Resources*: produced the cross-matched catalogs and raw data products. *Writing—review & editing*: verified data descriptions and clarified pipeline details. *Conceptualization*: contributed to scientific discussions within the QSO identification working group.

IPR: *Data curation*: performed the J–PAS×DESI cross-matching and prepared the final match tables. *Resources*: constructed the raw datasets delivered to the analysis pipeline. *Writing—review & editing*: reviewed the data construction and results sections. *Conceptualization*: contributed to scientific discussions within the QSO identification working group.

GMS: *Writing—review & editing*: provided detailed feedback on multiple drafts of the manuscript, particularly regarding quasar classification strategy. *Conceptualization*: contributed to scientific discussions within the QSO identification working group.

JCM: *Writing—review & editing*: reviewed the final manuscript and provided useful suggestions to this work.

All authors reviewed and approved the final manuscript.

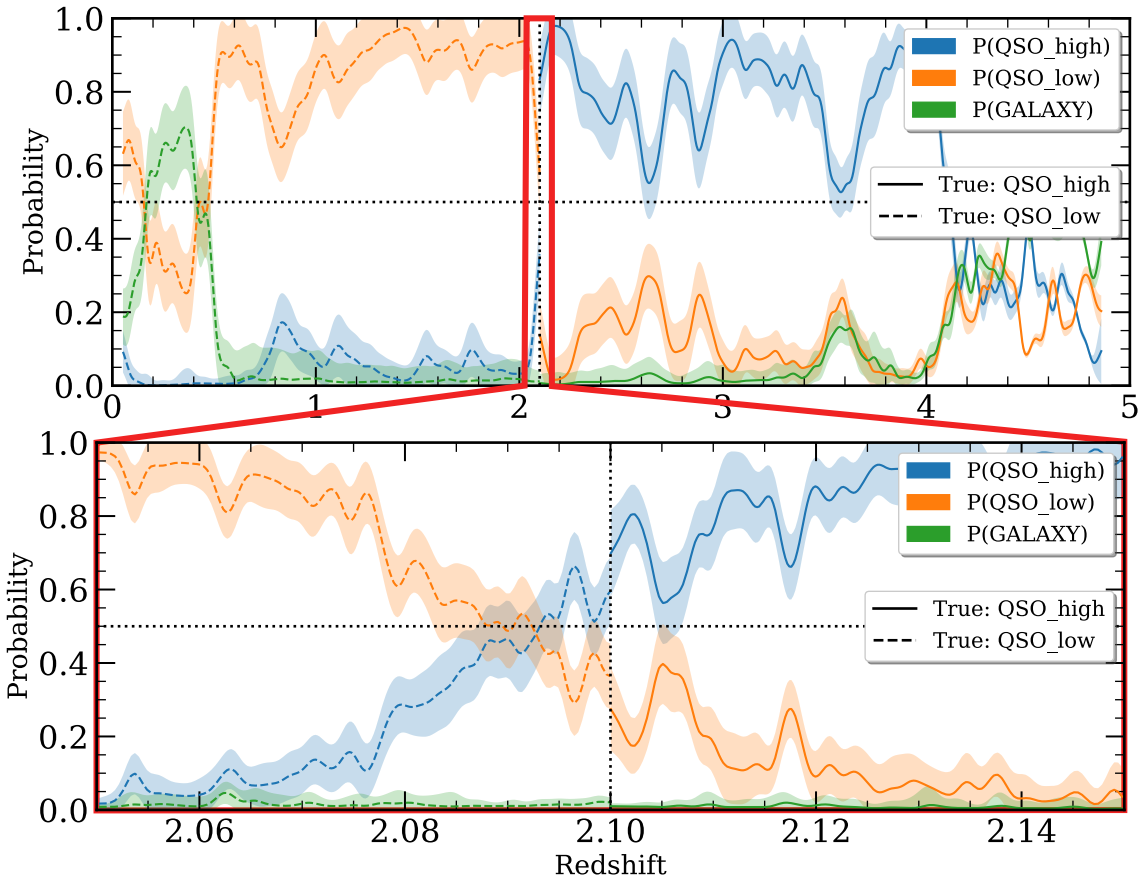

Figure 5. Quasar class probabilities as a function of redshift (two-panel view). *Top:* range $0 < z < 5$. *Bottom:* zoom on the transition near $z \simeq 2.1$ (red box in the top panel marks the window magnified below). For objects whose *true* labels are `QSO_high` (solid curves) or `QSO_low` (dashed curves), we plot the predicted class probabilities as a function of spectroscopic redshift: blue = P(`QSO_high`), orange = P(`QSO_low`), green = P(`GALAXY`) (the `STAR` channel is omitted for clarity). Curves show the *binned median* probability in redshift, smoothed for readability; shaded bands indicate a conservative dispersion of $\pm\frac{1}{3}\sigma$ (one third of the per-bin sample standard deviation) around each median. The vertical dotted line denotes the subclass split at $z = 2.1$; the horizontal dotted line marks the reference level $p = 0.5$. Together, the panels visualize the two main sources of ambiguity discussed in the text: low-z host dilution (`QSO_low` vs. `GALAXY`) and mutual `QSO_high`/`QSO_low` confusion at the $z \simeq 2.1$ boundary.

Facilities: We use observations from the J-PAS survey obtained with JPCam at the Observatorio Astrofísico de Javalambre (OAJ).

Software: Python, Matplotlib (J. D. Hunter 2007), PyTorch (A. Paszke et al. 2019), Numpy (S. van der Walt et al. 2011), Weights & Biases (L. Biewald 2020), Source Extractor (E. Bertin & S. Arnouts 1996)

APPENDIX

A. QUASAR CLASS PROBABILITIES AS A FUNCTION OF REDSHIFT

All diagnostics in this Appendix are computed using the *no-DA* model evaluated on the DESI→J-PAS *mocks*. We use the mock sample here because it provides much larger statistics, making broad trends and fine structure in the probability–vs–z behaviour easier to visualize.

To interpret the off–diagonal structure of the confusion matrices presented in section 4, we examine how the predicted class probabilities vary with redshift for objects whose *true* labels are `QSO_high` or `QSO_low`. Figure 5 shows two views of the same diagnostics: the **top panel** covers the redshift range ($0 < z < 5$), and the **bottom panel** zooms into the transition near $z \simeq 2.1$.

Each curve is the binned *median* probability as a function of redshift; shaded bands indicate a conservative dispersion of $\pm\frac{1}{3}\sigma$ around the median.

The vertical dotted line marks $z = 2.1$ (our `QSO_high`/`QSO_low` split) and the horizontal line marks $p = 0.5$. Solid and dashed linestyles indicate the *true* class label (`QSO_high` vs. `QSO_low`).

Top panel: global trends ($0 < z < 5$).—Two regimes explain most of the `QSO_low` confusion:

1. At low redshift ($z \sim 0.0$–0.5), the median p(`GALAXY`) rises while p(`QSO_low`) softens. This matches the expectation that host-galaxy light increasingly dilutes the AGN signal in the J-PAS narrow bands, the *J-spectra* of low-z AGN can be dominated by resolved host-galaxy emission when the nucleus is comparatively faint, and the classification may be driven by host-galaxy properties rather than by nuclear emission alone.

2. Near the subclass boundary ($z \simeq 2.1$), p(`QSO_low`) and p(`QSO_high`) converge, as expected for a hard redshift threshold combined with finite *J-spectra* resolution: finite filter widths and measurement noise limit the effective redshift precision, so objects with true redshifts near 2.1 naturally receive comparable support for both subclasses.

For `QSO_high`, the same boundary effect appears around $z \simeq 2.1$. At higher redshift the curves show additional, narrower bumps where p(`QSO_high`) dips and p(`QSO_low`) momentarily increases. These features are consistent with occasional line-identification aliases—effectively, the classifier infers redshift from spectral features, and certain filter/line coincidences can transiently favour the wrong subclass. At the highest redshifts there are very few detected sources, and the smaller S/N for these objects results in uncertain class probabilities.

Bottom panel: zoom on the transition ($2.05 \lesssim z \lesssim 2.15$).—The crossover is clearer in the zoom: the median p(`QSO_high`) climbs through ~ 0.5 as z crosses 2.1 while p(`QSO_low`) symmetrically falls, for both true-label subsets. The overlap of the dispersion bands confirms that the ambiguity at the split is *structural*—set by J-PAS filter sampling of QSO features—rather than a mere modelling artifact.

In the smoothed-binned curves, the apparent p(`QSO_high`) $= p$(`QSO_low`) crossing can occur slightly offset from the nominal threshold (e.g., by $\Delta z \sim 0.01$). We do not assign a physical meaning to such a small displacement: it can arise from finite sample size and the binning/smoothing procedure used to summarize the point cloud in redshift, and it can also reflect a purely *model-side* effect—i.e., a particular converged solution that yields essentially the same global objective value but induces a small local shift in the median curves near the boundary, without implying an underlying feature in the data. Because this diagnostic is meant to be illustrative (and because resolving the detailed origin of percent-level offsets would require a dedicated stability/ablation analysis beyond the scope of this Appendix), we caution against over-interpreting the exact crossing location.

For completeness, we reiterate that Fig. 5 uses the no-DA model on mocks to maximize statistical power; the same qualitative behaviours are present in the J-PAS observations, but the smaller sample size makes the trends noisier.

B. ADDITIONAL PER-CLASS DIAGNOSTICS AND MAGNITUDE-CUT TRADE-OFFS

Figure 6 collects per-class metrics across the four evaluation regimes discussed in Sect. 3 and 4. Each panel reports one diagnostic (accuracy, F1, TPR/recall, precision/PPV, AUC, and ECE), with bars grouped by class (`QSO_high`, `QSO_low`, `GALAXY`, `STAR`). The plot makes explicit where adaptation helps most and where performance is already saturated. For the quasar subclasses, SSDA tightens recall and precision together, lifting F1 while keeping calibration at the $\sim$few-percent level; the `STAR` and `GALAXY` metrics remain high and flat across regimes, consistent with the confusion matrices in Fig. 3. Dashed horizontal guides indicate reference levels within each panel and help compare the regimes class-by-class without relying on a single macro average. Read together, these panels support the main-text claim that adaptation reduces star→QSO leakage and sharpens the `QSO_high`/`QSO_low` boundary around $z \simeq 2.1$ while preserving the already-strong `GALAXY`/`STAR` performance.

Figure 7 explores a simple operational knob for spectroscopic targeting: imposing r-band magnitude cuts and measuring how per-class F1 varies on the J-PAS×DESI test split. Due to the size of the cross-match test set, several magnitude bins contain few objects per class—e.g., only 11 `QSO_high` for $17 \leq r < 19$, and 54/28 `STAR` for $21 \leq r < 22$ and $22 < r \leq 22.5$—so apparent swings in F1 can be dominated by noise. With that caveat, the qualitative pattern is as expected: brighter cuts (higher S/N) tend to help both quasar subclasses. Across cuts, the SSDA model generally matches or improves upon the target-only baseline; in sparsely populated bins, uncertainties are large and differences should not be over-interpreted. Practically, we recommend selecting thresholds in the low–false-positive region of the

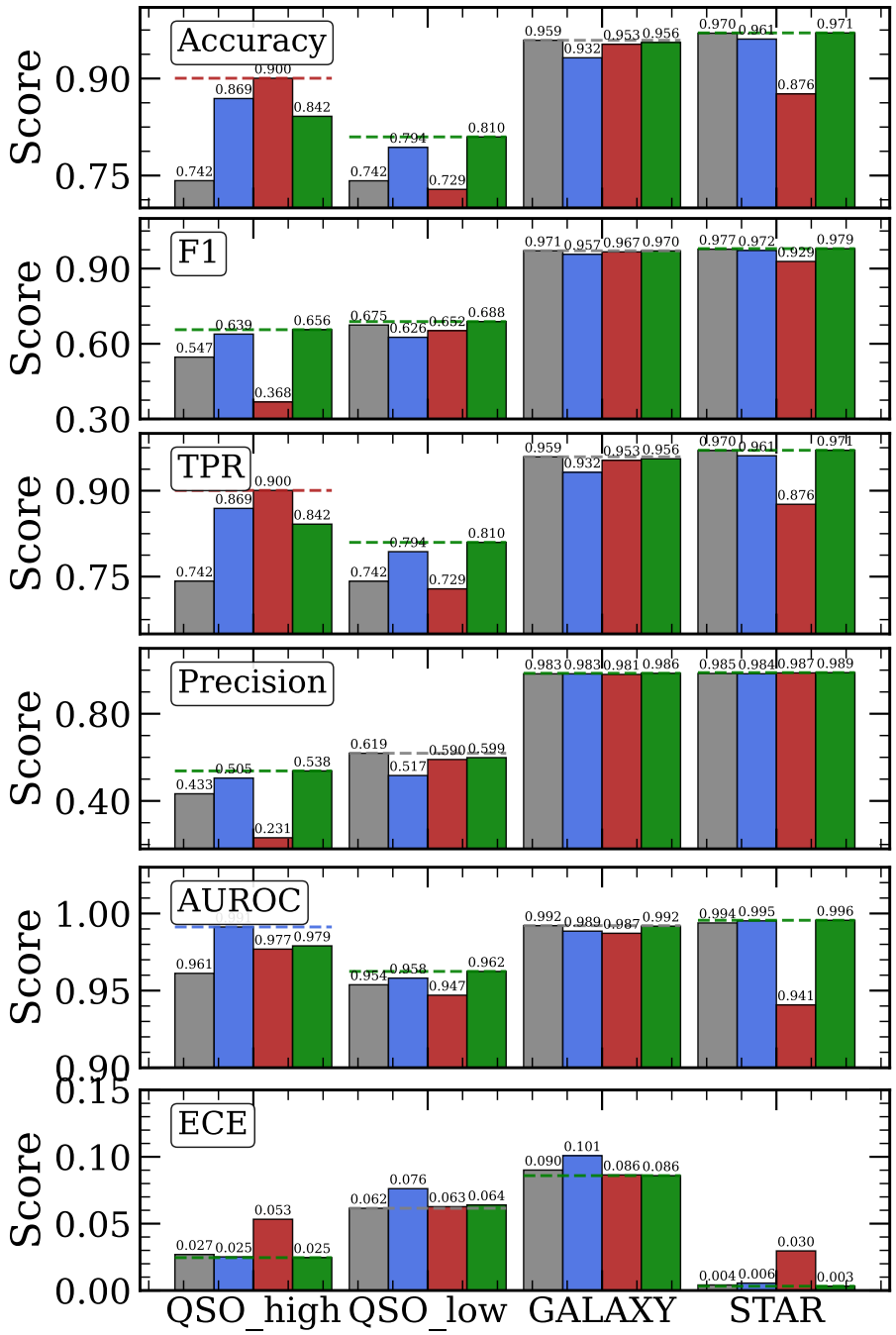

Figure 6. Per–class diagnostics across regimes. Panels show, from top to bottom: *Accuracy*, *F1*, *TPR/Recall*, *Precision/PPV*, *AUC*, and *ECE*. Within each panel, bars are grouped by true class (`QSO_high`, `QSO_low`, `GALAXY`, `STAR`); values are printed above the bars and dashed horizontal lines mark the per–class top-performance model. Colors are fixed across all panels and figures: **gray** = *J-PAS supervised* (evaluated on the J-PAS Observations test split), **blue** = *Mocks no-DA in-domain* (trained and evaluated on DESI→J-PAS mocks), **red** = *J-PAS no-DA* (trained on mocks, evaluated zero-shot on the J-PAS Observations test split), and **green** = *J-PAS SSDA* (mocks-pretrained, encoder adapted on a small labeled J-PAS subset; evaluated on the J-PAS Observations test split).

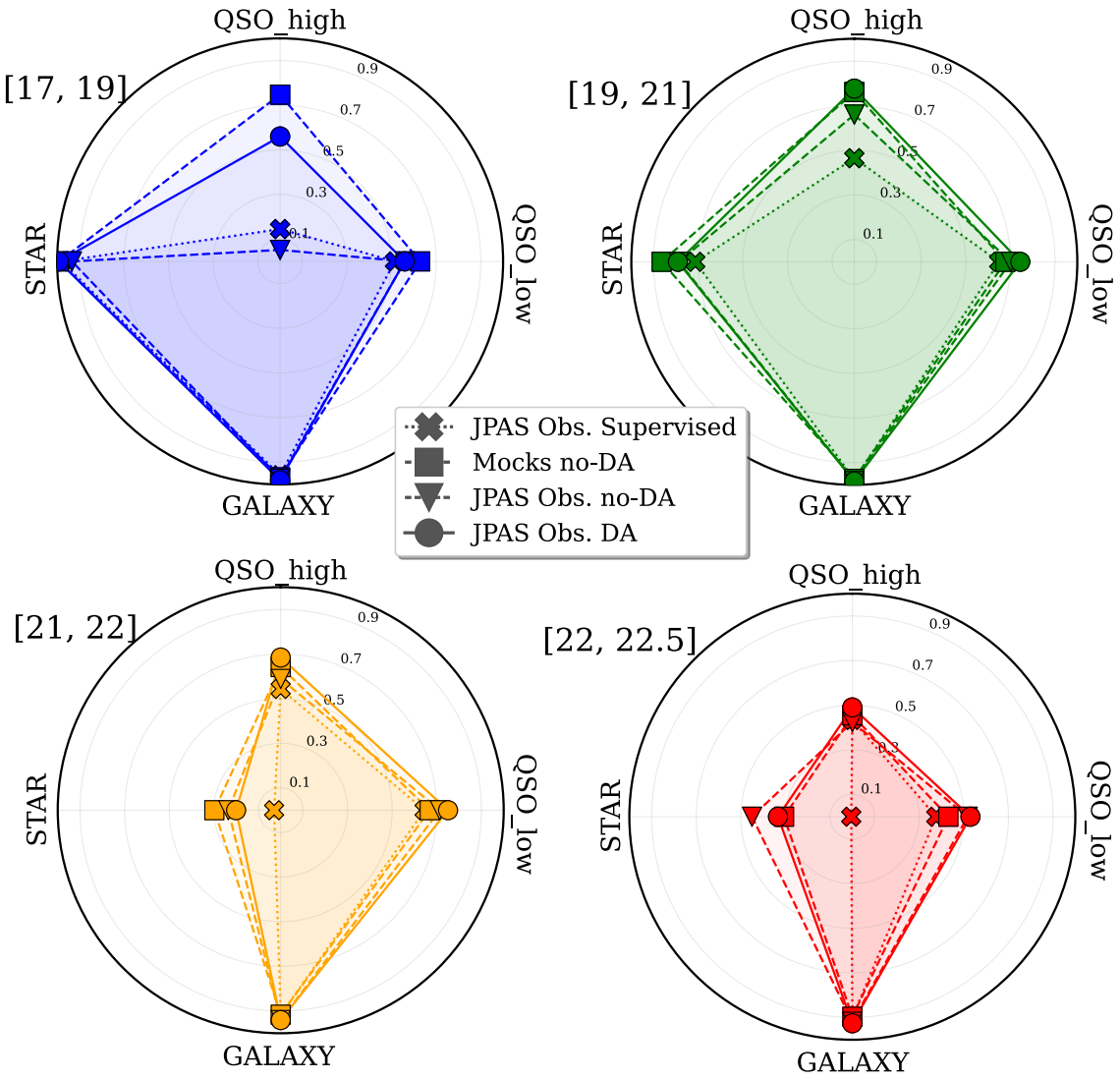

Figure 7. Class-wise F1 vs. r-band magnitude cuts. Each polygon shows per–class F1 for a specific r-band selection (cuts indicated in the legend). Axes correspond to classes (`QSO_high`, `QSO_low`, `GALAXY`, `STAR`); the radius encodes F1 in $[0, 1]$ with concentric gridlines for reference. Markers and line-styles match Fig. 4. For each magnitude cut, the J-PAS–based regimes are tested on the *J-PAS Observations test split*, while the "*Mocks no-DA*" pipeline is evaluated employing the *mocks test split*. Because the J-PAS×DESI test set is comparatively small, some cuts leave few objects per class (e.g., only 11 `QSO_high` for $17 \leq r < 19$ and 54/28 `STAR` for $21 \leq r < 22$ and $22 < r \leq 22.5$), so apparent swings can be dominated by counting noise and should not be over-interpreted.

ROC curves (Sect. 4) and aggregating magnitude bins—or enforcing a minimum per-class count—when using F1 to tune r cuts. Because SSDA uses only a small number of J-PAS labels, the same procedure can be refreshed field-by-field as depth and seeing vary; storing selection probabilities then makes the targeting function explicit for downstream $n(z)$ and clustering analyses.

REFERENCES

Adame, A. G., Aguilar, J., Ahlen, S., et al. 2025, JCAP, 2025, 021, doi: 10.1088/1475-7516/2025/02/021

Agarwal, S., Ćiprijanović, A., & Nord, B. D. 2024, arXiv e-prints, arXiv:2411.03334, doi: 10.48550/arXiv.2411.03334

Akhmetzhanova, A., Cuesta-Lazaro, C., & Mishra-Sharma, S. 2025, arXiv e-prints, arXiv:2508.05744, doi: 10.48550/arXiv.2508.05744

Akhmetzhanova, A., Mishra-Sharma, S., & Dvorkin, C. 2024, MNRAS, 527, 7459, doi: 10.1093/mnras/stad3646

Alexander, S., Gleyzer, S., Parul, H., et al. 2023, ApJ, 954, 28, doi: 10.3847/1538-4357/acdfc7

Alshammari, S., Hershey, J., Feldmann, A., Freeman, W. T., & Hamilton, M. 2025, arXiv e-prints, arXiv:2504.16929, doi: 10.48550/arXiv.2504.16929

Anau Montel, N., Coogan, A., Correa, C., Karchev, K., & Weniger, C. 2023, MNRAS, 518, 2746, doi: 10.1093/mnras/stac3215

Andrianomena, S., & Hassan, S. 2025, Ap&SS, 370, 14, doi: 10.1007/s10509-025-04405-y

Baqui, P. O., Marra, V., Casarini, L., et al. 2021, A&A, 645, A87, doi: 10.1051/0004-6361/202038986

Bardes, A., Ponce, J., & LeCun, Y. 2021, arXiv e-prints, arXiv:2105.04906, doi: 10.48550/arXiv.2105.04906

Benitez, N., Dupke, R., Moles, M., et al. 2014, arXiv e-prints, arXiv:1403.5237, doi: 10.48550/arXiv.1403.5237

Bergamini, P., Grillo, C., Rosati, P., et al. 2023, A&A, 674, A79, doi: 10.1051/0004-6361/202244834

Berthelot, D., Roelofs, R., Sohn, K., Carlini, N., & Kurakin, A. 2021, arXiv e-prints, arXiv:2106.04732, doi: 10.48550/arXiv.2106.04732

Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393, doi: 10.1051/aas:1996164

Biewald, L. 2020, Experiment Tracking with Weights and Biases, https://www.wandb.com/

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. 2020, arXiv e-prints, arXiv:2002.05709, doi: 10.48550/arXiv.2002.05709

Ćiprijanović, A., Lewis, A., Pedro, K., et al. 2023, Machine Learning: Science and Technology, 4, 025013, doi: 10.1088/2632-2153/acca5f

Ćiprijanović, A., Kafkes, D., Downey, K., et al. 2021, MNRAS, 506, 677, doi: 10.1093/mnras/stab1677

Crain, R. A., Schaye, J., Bower, R. G., et al. 2015, MNRAS, 450, 1937, doi: 10.1093/mnras/stv725

Dalton, G., Trager, S., Abrams, D. C., et al. 2016, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9908, Ground-based and Airborne Instrumentation for Astronomy VI, ed. C. J. Evans, L. Simard, & H. Takami, 99081G, doi: 10.1117/12.2231078

del Pino, A., López-Sanjuan, C., Hernán-Caballero, A., et al. 2024, A&A, 691, A221, doi: 10.1051/0004-6361/202450503

DESI Collaboration, Adame, A. G., Aguilar, J., et al. 2024, AJ, 168, 58, doi: 10.3847/1538-3881/ad3217

Euclid Collaboration, Mellier, Y., Abdurro'uf, et al. 2025, A&A, 697, A1, doi: 10.1051/0004-6361/202450810

Feng, Y., Chu, M.-Y., Seljak, U., & McDonald, P. 2016, MNRAS, 463, 2273, doi: 10.1093/mnras/stw2123

Ganin, Y., Ustinova, E., Ajakan, H., et al. 2015, arXiv e-prints, arXiv:1505.07818, doi: 10.48550/arXiv.1505.07818

Garrison, L. H., Eisenstein, D. J., Ferrer, D., Maksimova, N. A., & Pinto, P. A. 2021, MNRAS, 508, 575, doi: 10.1093/mnras/stab2482

Gilda, S., de Mathelin, A., Bellstedt, S., & Richard, G. 2024, Astronomy, 3, 189, doi: 10.3390/astronomy3030012

Han, Z., Gao, C., Liu, J., Zhang, J., & Qian Zhang, S. 2024, arXiv e-prints, arXiv:2403.14608, doi: 10.48550/arXiv.2403.14608

Hayat, M. A., Stein, G., Harrington, P., Lukić, Z., & Mustafa, M. 2021, ApJL, 911, L33, doi: 10.3847/2041-8213/abf2c7

Huertas-Company, M., Sarmiento, R., & Knapen, J. H. 2023, RAS Techniques and Instruments, 2, 441, doi: 10.1093/rasti/rzad028

Hunter, J. D. 2007, Computing in Science & Engineering, 9, 90, doi: 10.1109/MCSE.2007.55

Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111, doi: 10.3847/1538-4357/ab042c

Jeakel, A. P., Vieira dos Santos, G., Marra, V., et al. 2025, arXiv e-prints, arXiv:2511.20524, doi: 10.48550/arXiv.2511.20524

Jin, S., Trager, S. C., Dalton, G. B., et al. 2024, MNRAS, 530, 2688, doi: 10.1093/mnras/stad557

Lang, D., Hogg, D. W., & Mykytyn, D. 2016, The Tractor: Probabilistic astronomical source detection and measurement,, Astrophysics Source Code Library, record ascl:1604.008 http://ascl.net/1604.008

Lee, J.-Y., Kim, J.-h., Jung, M., et al. 2024, ApJ, 975, 38, doi: 10.3847/1538-4357/ad73d4

Li, C.-L., Chang, W.-C., Cheng, Y., Yang, Y., & Póczos, B. 2017, arXiv e-prints, arXiv:1705.08584, doi: 10.48550/arXiv.1705.08584

Liu, X., Yoo, C., Xing, F., et al. 2022, arXiv e-prints, arXiv:2208.07422, doi: 10.48550/arXiv.2208.07422

López-Cano, D., Stücker, J., Pellejero Ibañez, M., Angulo, R. E., & Franco-Barranco, D. 2024, A&A, 685, A37, doi: 10.1051/0004-6361/202348965

Martínez-Solaeche, G., Queiroz, C., González Delgado, R. M., et al. 2023, A&A, 673, A103, doi: 10.1051/0004-6361/202245750

Pandya, S., Patel, P., Nord, B. D., Walmsley, M., & Ćiprijanović, A. 2025, Machine Learning: Science and Technology, 6, 035032, doi: 10.1088/2632-2153/adf701

Parul, H., Gleyzer, S., Reddy, P., & Toomey, M. W. 2025, ApJ, 990, 47, doi: 10.3847/1538-4357/adee16

Paszke, A., Gross, S., Massa, F., et al. 2019, arXiv e-prints, arXiv:1912.01703, doi: 10.48550/arXiv.1912.01703

Pérez-Ràfols, I., Abramo, L. R., Martínez-Solaeche, G., et al. 2023, A&A, 678, A144, doi: 10.1051/0004-6361/202347488

Pérez-Ràfols, I., Abramo, L. R., Martínez-Solaeche, G., et al. 2025, arXiv e-prints, arXiv:2507.11380, doi: 10.48550/arXiv.2507.11380

Pieri, M. M., Bonoli, S., Chaves-Montero, J., et al. 2016, in SF2A-2016: Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics, ed. C. Reylé, J. Richard, L. Cambrésy, M. Deleuil, E. Pécontal, L. Tresse, & I. Vauglin, 259–266, doi: 10.48550/arXiv.1611.09388

Pierre, S., Régaldo-Saint Blancard, B., Hahn, C., & Eickenberg, M. 2025, arXiv e-prints, arXiv:2507.03086, doi: 10.48550/arXiv.2507.03086

Potter, D., Stadel, J., & Teyssier, R. 2017, Computational Astrophysics and Cosmology, 4, 2, doi: 10.1186/s40668-017-0021-1

Queiroz, C., Abramo, L. R., Rodrigues, N. V. N., et al. 2023, MNRAS, 520, 3476, doi: 10.1093/mnras/stac2962

Rampf, C., List, F., & Hahn, O. 2025, JCAP, 2025, 020, doi: 10.1088/1475-7516/2025/02/020

Rodrigues, N. V. N., Raul Abramo, L., Queiroz, C., et al. 2023, MNRAS, 520, 3494, doi: 10.1093/mnras/stac2836

Roncoli, A., Ćiprijanović, A., Voetberg, M., Villaescusa-Navarro, F., & Nord, B. 2023, arXiv e-prints, arXiv:2311.01588, doi: 10.48550/arXiv.2311.01588

Schaye, J., Kugel, R., Schaller, M., et al. 2023, MNRAS, 526, 4978, doi: 10.1093/mnras/stad2419

Springel, V., Pakmor, R., Zier, O., & Reinecke, M. 2021, MNRAS, 506, 2871, doi: 10.1093/mnras/stab1855

Swierc, P., Tamargo-Arizmendi, M., Ćiprijanović, A., & Nord, B. D. 2024, arXiv e-prints, arXiv:2410.16347, doi: 10.48550/arXiv.2410.16347

Swierc, P., Zhao, M., Ćiprijanović, A., & Nord, B. 2023, arXiv e-prints, arXiv:2311.17238, doi: 10.48550/arXiv.2311.17238

Tzeng, E., Hoffman, J., Saenko, K., & Darrell, T. 2017, arXiv e-prints, arXiv:1702.05464, doi: 10.48550/arXiv.1702.05464

van der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, Computing in Science Engineering, 13, 22, doi: 10.1109/MCSE.2011.37

Vega-Ferrero, J., Huertas-Company, M., Costantin, L., et al. 2024, ApJ, 961, 51, doi: 10.3847/1538-4357/ad05bb

Vogelsberger, M., Genel, S., Springel, V., et al. 2014, MNRAS, 444, 1518, doi: 10.1093/mnras/stu1536

von Marttens, R., Marra, V., Quartin, M., et al. 2024, MNRAS, 527, 3347, doi: 10.1093/mnras/stad3373

Wang, J., Lan, C., Liu, C., et al. 2021, arXiv e-prints, arXiv:2103.03097, doi: 10.48550/arXiv.2103.03097

Yu, Y.-C., & Lin, H.-T. 2023, arXiv e-prints, arXiv:2302.02335, doi: 10.48550/arXiv.2302.02335

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., & Change Loy, C. 2021, arXiv e-prints, arXiv:2103.02503, doi: 10.48550/arXiv.2103.02503