

A quantitative benchmark for explainable artificial intelligence in astronomical transient classification

A. Pagliaro^{1,2,3,*} 

¹ INAF – Istituto di Astrofisica Spaziale e Fisica Cosmica, Via Ugo La Malfa 153, 90146 Palermo, Italy

² ICSC – Centro Nazionale di Ricerca in HPC, Big Data e Quantum Computing, 40121 Bologna, Italy

³ Istituto Nazionale di Fisica Nucleare, Sezione di Catania, Via Santa Sofia 64, 95123 Catania, Italy

Received 22 April 2026 / Accepted 19 August 2026

ABSTRACT

Explainable artificial intelligence (xAI) methods are increasingly used to interpret machine learning classifiers of astronomical transients, yet their reliability is rarely assessed quantitatively. We present *astro-xai-bench*, an open benchmark that evaluates explanation quality along five complementary axes – perturbation-based faithfulness, deletion and insertion AUC, complexity, cross-representation plausibility, and the standard model- and data-randomisation sanity checks – applied to nine classifiers spanning tree-based, deep-learning, and physics-informed architectures on the PLAsTiCC photometric transient dataset (14 classes, six pass-bands), with all metrics averaged over five random seeds. A detailed case study of XGBoost (tree-based) and CNN2D (convolutional on Gramian angular field images) illustrates how each xAI axis interacts with model architecture: faithfulness is representation-dependent, complexity rankings invert under basis change, and projected plausibility is sensitive to the projection rather than the model. Three general findings emerge: (i) deep-learning models on raw light curves align better with an astrophysically informed reference mask than tree-based models trained directly on the hand-crafted features that define the mask; (ii) the sanity checks reveal that LSTM, CNN1D, and Random Forest attributions are dominated by the input distribution, while XGBoost, CNN2D, ViT, and both physics-informed baselines (Avocado, ParSNIP) pass all applicable checks; (iii) no single model dominates all five axes, making multi-axis evaluation essential. All code and evaluation pipelines are released as a public benchmark.

Key words. methods: data analysis – methods: statistical – techniques: photometric

1. Introduction

The next generation of time-domain astronomical surveys – including the *Vera C. Rubin* Observatory’s Legacy Survey of Space and Time (LSST; Ivezić et al. 2019) and the planned Wide Field Survey Telescope – will detect millions of transient and variable sources per night. The same challenge is already pressing for current-generation facilities: the *Zwicky* Transient Facility (ZTF; Bellm et al. 2019), in routine operation since 2018, issues of the order of 10^5 transient alerts per night, motivating the automated classifiers studied here well before LSST reaches full survey operations. Automated classification pipelines based on machine learning (ML) are essential to triage this data volume (Narayan et al. 2018; Muthukrishna et al. 2019; Boone 2019).

While deep learning models have achieved impressive classification accuracy on photometric light curves (Möller & de Boissière 2020; Allam & McEwen 2021; Qu et al. 2021), their ‘black-box’ nature poses challenges for scientific trust and discovery. Explainable AI (xAI) methods have emerged as a means to understand why a classifier assigns a particular label, with SHAP (Lundberg & Lee 2017), Grad-CAM (Selvaraju et al. 2017), and attention-based visualisations being the most widely adopted in astronomy (Jamal & Bloom 2020; Villar et al. 2021; Carrasco-Davis et al. 2021).

However, most applications of xAI in astronomy are qualitative: authors generate saliency maps or feature importance plots and interpret them by visual inspection. Quantitative evaluation

of explanation quality – whether the highlighted features are genuinely driving the prediction and whether the explanations are consistent with astrophysical knowledge – is largely absent from the literature. Recent efforts such as the explainable-by-design subsequence features of uSAST (Mbouopda et al. 2026) and the FIX framework for expert-alignment scoring (Jin et al. 2024) have begun to address this gap, but a systematic benchmark spanning multiple model families, data representations, and complementary evaluation axes has been lacking. Similarly, Kumaran et al. (2023) classified *Chandra* X-ray point sources with LightGBM, and their follow-up (Kumaran et al. 2026) extended the approach with TreeSHAP for feature-level physical validation, while large-scale survey pipelines (e.g. Carrasco-Davis et al. 2021) increasingly rely on post hoc explanations without standardised quality checks.

This gap is significant. An xAI method that highlights physically irrelevant features (e.g. survey cadence artefacts) while appearing plausible could mislead scientific interpretation. Conversely, a faithful explanation that concentrates on astrophysically meaningful features provides genuine insights into what the model has learned.

In this work, we address this gap by presenting *astro-xai-bench*, an open benchmark for quantitative xAI evaluation in transient classification. Our contributions are:

1. A systematic comparison of seven model architectures spanning tabular, sequential, and image-based representations of photometric light curves, complemented by two state-of-the-art physics-informed PLAsTiCC classifiers – Avocado (Boone 2019) and ParSNIP (Boone 2021) – so that the xAI

* Corresponding author: antonio.pagliaro@inaf.it

Table 1. PLAsTiCC transient classes used in this work, with the number of objects per class in the unblinded training set ($N = 7848$).

Index	Class	N
0	μ -Lens-Single	151
1	TDE	495
2	EBE	924
3	SNII	1193
4	SNIax	183
5	Mira	30
6	SNIbc	484
7	KN	102
8	M-dwarf	981
9	SNIa-91bg	208
10	AGN	370
11	SNIa	2313
12	RRL	239
13	SLSN-I	175

Notes. The benchmark uses a stratified split of these objects into 70% training, 10% validation, and 20% test. The strong imbalance – spanning two orders of magnitude, from SNIa (2313) down to Mira (30), KN (102), and SLSN-I (175) – is intrinsic to PLAsTiCC and is reflected in the per-class F1 scores. We stress that these class proportions were chosen by the challenge designers for classification difficulty, not to reproduce astrophysical volumetric rates (see Hložek et al. 2023, for a discussion of realistic rate assumptions); the benchmark therefore evaluates xAI methods under the class distribution of the dataset, not under a physically motivated prior.

metrics are calibrated against domain-specific models and not only against generic architectures.

- Five complementary xAI evaluation axes – perturbation-based faithfulness, deletion and insertion AUC, complexity, cross-representation plausibility, and the model- and data-randomisation sanity checks – adapted and validated for the astronomical context.
- An astrophysically motivated reference feature mask, grounded in domain knowledge, for plausibility evaluation.
- A fully reproducible, open-source evaluation pipeline.

The remainder of this paper is organised as follows. Section 2 describes the dataset and data representations. Section 3 presents the model architectures. Section 4 defines the xAI methods and quantitative metrics. Section 5 reports our experimental results. Section 6 discusses implications, Sect. 7 shows how to evaluate a new classifier with the benchmark, and Sect. 8 summarises our conclusions.

2. Data

2.1. PLAsTiCC dataset

We used the Photometric LSST Astronomical Time-series Classification Challenge (PLAsTiCC; Kessler et al. 2019; Hložek et al. 2023) training set, which contains 7,848 light curves observed in six passbands (u, g, r, i, z, Y) spanning 14 astrophysical classes (Table 1). Each light curve consists of irregularly sampled flux measurements with associated uncertainties.

We applied a stratified split: 70% training, 10% validation, 20% test. All preprocessing and evaluation used the test set exclusively. To assess statistical robustness, we repeated every experiment with five random seeds (42, 123, 456, 789, and 1024) and report all metrics as cross-seed mean $\pm$ standard deviation

in Tables 4–5. Redshift was deliberately not used by the seven generic architectures: this isolated the comparison to flux-only representations and the effect of representation on attribution quality. The two physics-informed baselines of Sect. 3.5 did use the host photometric redshift, as described there, and served as the complementary domain-informed anchor.

2.2. Data representations

We considered three input representations to enable a comparison across model families:

Tabular features. Following Boone (2019) and Muthukrishna et al. (2019), we extracted ten statistical features per passband (60 total): the mean, standard deviation, amplitude, median, number of observations, skewness, kurtosis, mean signal-to-noise ratio, linear slope, and fraction of flux above the mean. These served as inputs for tree-based models.

Time series. Raw flux values were arranged as tensors of shape (6256), whereby each passband was zero-padded or truncated to 256 time steps and normalised per-object per-band. This representation is used by the LSTM and 1D-CNN.

Gramian angular field (GAF) images. Following Wang & Oates (2015), each per-band light curve was converted to a Gramian angular summation field image of size 224×224 via piecewise aggregate approximation, angular encoding, and outer-product computation. Concretely, the series was min–max scaled to $[-1, 1]$, each value was mapped to an angle $\phi_i = \arccos(\tilde{x}_i)$, and the image was formed as $G_{ij} = \cos(\phi_i + \phi_j)$ (Fig. A.1). Intuitively, the GAF re-encodes the light curve so that local temporal correlations become spatial structure: each pixel (i, j) measures the angular similarity between the values at times i and j , so the main diagonal traces the curve’s overall shape while off-diagonal structure captures how strongly distant epochs co-vary – rapid rise and decline produces sharp off-diagonal contrast, slowly varying segments produce smooth blocks. This makes temporal evolution accessible to the spatial inductive biases of 2-D convolutional and vision-transformer architectures. The six passband images are stacked as channels, yielding tensors of shape (6, 224, 224). This representation is used by the 2D-CNN and Vision Transformer.

Redshift-aware raw light curves. The two physics-informed baselines (Sect. 3.5) do not consume any of the three fixed representations above; instead they ingest the irregularly sampled (time, flux, flux_error, passband) measurements together with the host-galaxy photometric redshift. Because our preprocessed tensors discard both the absolute MJD sampling and the redshift, we recover them by joining our PLAsTiCC object identifiers against the original survey metadata; the empirical specz distribution of the full PLAsTiCC test-set metadata is used solely as the redshift prior for Avocado’s augmentation (Sect. 3.5), while no test-set light curve or label enters training at any stage.

Redshift and amplitude treatment. Table 2 summarises how each model handles redshift and amplitude – two factors that are critical for transient classification because observed brightness scales with luminosity distance. The seven generic architectures deliberately exclude redshift: the tabular features are raw flux statistics (not rest-frame), while the time-series and GAF representations apply a per-object, per-band normalisation that removes the absolute flux scale but does not correct for

Table 2. Model architectures, input representations, and treatment of redshift and amplitude.

Model	Input	Params	z	Norm.
RF	Tabular (60)	–	–	–
XGB	Tabular (60)	–	–	–
LSTM	Series (6×256)	~400K	–	band
CNN1D	Series (6×256)	~200K	–	band
CNN2D	GAF (6×224 ²)	~300K	–	band
ViT	GAF (6×224 ²)	~22M	–	band
Ens.	Meta (84)	–	–	–
Avocado	LC+ z (41 GP)	–	✓	GP
ParSNIP	LC+ z (VAE)	~1.2M	✓	A

Notes. ‘ z ’ indicates whether the host photometric redshift is used; ‘Norm.’ indicates per-object normalisation (Sect. 2.2).

cosmological effects (time dilation, K-correction). This design choice isolates the comparison to the flux-only representation; redshift-dependent performance differences are absorbed into the reported classification scores. The two physics-informed baselines take the complementary approach: Avocado conditions its Gaussian-process augmentation on the host photo- z and extracts rest-frame colour features, while ParSNIP’s VAE marginalises over an explicit amplitude parameter A in the decoder and conditions the latent space on redshift. This dichotomy – flux-only versus redshift-aware – is a deliberate experimental axis of the benchmark.

3. Models

We evaluated seven architectures of increasing complexity, spanning three representation families (Table 2).

The first seven models are generic architectures benchmarked on our three fixed representations; the last two (Avocado, ParSNIP) are physics-informed baselines described in Sect. 3.5. The stacking ensemble combines only the six generic base models and does not include the physics-informed baselines.

3.1. Tree-based models

Random Forest. A scikit-learn `RandomForestClassifier` with 500 estimators, trained on the 60 tabular features.

XGBoost. An `XGBClassifier` with 500 estimators, a max depth of 8, and a learning rate of 0.1.

3.2. Sequence models

LSTM. A two-layer bidirectional LSTM (hidden size 128) with masked mean pooling over valid time steps, followed by dropout (0.3) and a linear classification head.

CNN1D. Three convolutional blocks (64 → 128 → 256 filters, kernel sizes 7, 5, 3) with batch normalisation, ReLU, and max and adaptive pooling, followed by dropout and a linear head.

3.3. Image-based models

CNN2D. Four residual-style convolutional blocks (32 → 64 → 128 → 256 filters) operating on six-channel GAF images, with adaptive average pooling and a linear head.

Vision Transformer (ViT). A `vit_small_patch16_224` backbone from the `timm` library (Wightman 2021), pretrained on ImageNet and fine-tuned with six input channels, dropout 0.1, and a linear head.

3.4. Stacking ensemble

A logistic regression meta-learner trained on the concatenated `predict_proba` outputs of all six base models (84 meta-features). To avoid target leakage – which would otherwise occur because the same validation set is used both to early-stop the deep learning base models and to fit the meta-learner – we generated the meta-features via five-fold stratified cross-validation on the training set: in each fold the six base models were retrained on 4/5 of the training samples and asked to predict the held-out 1/5, and the resulting out-of-fold predictions were concatenated to form a leak-free training set for the meta-learner. At inference time the base models trained on the full training (with validation set used for early stopping, as described above) were used to compute meta-features on the test set, and the meta-learner produced the final classification.

3.5. Physics-informed baselines

To calibrate the xAI metrics against models that encode astrophysical structure by design – rather than only against the seven generic architectures above – we additionally evaluated the two classifiers that have defined the state of the art on PLAsTiCC. Both were retrained from scratch on our 14-class split, once per random seed, using the same training, validation, and test partition as all other models.

Avocado. The Gaussian-process feature-extraction and gradient-boosting pipeline of Boone (2019), winner of the PLAsTiCC Kaggle challenge. Each light curve was fitted with a 2D GP across wavelength and time; ~41 features (peak magnitude, colours, rise and decline timescales, GP length scale, percentile fluxes, host photo- z) are extracted and classified with LightGBM (Ke et al. 2017). We applied Avocado’s redshift augmentation with a factor of 10× rather than the 100× used in Boone (2019). Avocado discards augmented light curves that fail its detection-quality cuts, so the nominal 10× yields 7.7× in practice: the resulting training set holds ~42 300 objects (42 293 at seed 42, and between 42 293 and 42 461 across the five seeds), substantially exceeding the unaugmented training set (5493) while retaining the full redshift augmentation of the published protocol (Boone 2019, Sect. 3.6); the 100× pipeline produces a serialised augmented set (~5 GB) that exceeds available workstation memory. Classification performance may therefore represent a lower bound relative to the original 100× configuration. We further note that Boone’s published results were obtained on the full 3.5 M-object PLAsTiCC test set as the unlabelled augmentation source, whereas our benchmark is restricted to the 7 848 labelled training objects, which independently depresses absolute performance relative to the original work. Augmentation is reseeded independently for each of the five cross-validation seeds, so that the reported standard deviations reflect both augmentation stochasticity and classifier-training stochasticity; this contrasts with shared-augmentation protocols common in ML benchmarks, where only the classifier seed varies.

ParSNIP. The generative variational auto-encoder of Boone (2021), which learns a low-dimensional intrinsic latent

representation of supernova light curves (three intrinsic latent dimensions plus reference time, colour and amplitude). We train the VAE per seed on our split via the `lcdatal/sncosmo` interface and classify the latent predictions with a downstream LightGBM head (Ke et al. 2017) retrained for the 14 PLAsTiCC classes (the VAE itself is class-agnostic). As with Avocado, the host photometric redshift is supplied to the decoder.

3.6. Training protocol

All PyTorch models were trained with AdamW, cosine annealing learning rate schedule, gradient clipping (1.0), and early stopping (patience 10) on validation loss. Sequence models used $\text{lr} = 10^{-3}$; the ViT used $\text{lr} = 10^{-4}$ with weight decay 0.01. All experiments were conducted on a single NVIDIA RTX 4090 GPU.

We did not perform an exhaustive hyperparameter search; the choices reflect a mix of established defaults and light validation-set guidance, as follows. Random Forest and XGBoost use 500 estimators (above scikit-learn’s default of 100, for ranking stability) with otherwise conventional settings for tabular multiclass problems (XGBoost: max depth 8, learning rate 0.1); neither was grid-searched, and each is fully determined by a fixed `random_state` per seed. The LSTM (hidden size 128, two layers) and CNN channel widths were selected from a small set of candidates by validation macro-F1; dropout (0.3 for the LSTM and CNN1D, 0.1 for the ViT), the AdamW optimiser, the cosine schedule, and the patience-10 early stopping are standard and were held fixed across architectures. The ViT is the ImageNet-pretrained `vit_small_patch16_224` from `tim` with its default configuration, fine-tuned at the lower $\text{lr} = 10^{-4}$ appropriate for a pretrained backbone. The two physics-informed baselines follow their published protocols rather than any tuning of ours: Avocado uses the Gaussian-process and LightGBM settings of Boone (2019) (the only deliberate deviation being the $10\times$ augmentation factor; Sect. 3.5), and ParSNIP uses the default VAE configuration of Boone (2021) (three intrinsic latent dimensions) with a LightGBM head at package defaults. This design favours reproducibility and a level playing field over per-model optimisation; absolute accuracies could likely be improved by tuning, but the relative xAI comparison – our primary object of study – is unaffected.

4. Explainability methods and metrics

4.1. Attribution methods

We applied different xAI methods according to model family:

TreeSHAP. (Lundberg et al. 2020) for Random Forest and XGBoost, providing exact Shapley-value attributions per feature per class.

Integrated Gradients. (Sundararajan et al. 2017) for LSTM, CNN1D, CNN2D, and ViT: the path integral of the gradient of the predicted-class logit along the straight line from a zero baseline $\mathbf{x}' = \mathbf{0}$ to the input $\mathbf{x}$,

$$\text{IG}_i(\mathbf{x}) = (x_i - x'_i) \int_0^1 \frac{\partial f_c(\mathbf{x}' + \alpha(\mathbf{x} - \mathbf{x}'))}{\partial x_i} d\alpha, \quad (1)$$

approximated with $n_{\text{steps}} = 50$ Riemann sum points using the `captum` library. We used IG rather than vanilla gradient saliency because the latter does not satisfy the completeness or sensitivity

axioms and is known to produce noisy attributions (Sundararajan et al. 2017).

Grad-CAM. (Selvaraju et al. 2017) for the CNN architectures: class-discriminative localisation maps from the last convolutional layer.

Attention rollout. (Abnar & Zuidema 2020) for the ViT: recursive multiplication of attention matrices across transformer blocks, yielding a spatial attention map (Fig. C.1). Attention rollout and Grad-CAM (Fig. 4) enter the benchmark as qualitative diagnostics only: every quantitative xAI metric of Sect. 4 was computed from TreeSHAP attributions for the tabular models and Integrated Gradients for the sequence and image models. What they show is nonetheless informative. The CNN1D highlights temporal regions around peak brightness and the CNN2D diagonal GAF structure corresponding to rapid flux evolution, while the ViT attends preferentially to off-diagonal regions encoding temporal correlations at different timescales (Fig. C.1). We caution against treating attention maps as stand-alone explanations, since the literature reports divergent findings on whether attention weights can be interpreted causally (Jain & Wallace 2019; Wiegrefe & Pinter 2019). For the stacking ensemble, we used the logistic regression coefficients as attribution proxies over the meta-feature space.

For the two physics-informed baselines we applied TreeSHAP (Lundberg et al. 2020) to their LightGBM classification heads, yielding exact Shapley attributions over Avocado’s 41 GP-derived features and over ParSNIP’s latent-prediction features, respectively. These native attribution spaces differ from the 60-dimensional tabular space of the expert mask; the projection used for plausibility is described in Sect. 4.2.4.

4.2. Quantitative xAI metrics

We evaluated explanation quality along five complementary axes. We describe the first three – faithfulness, complexity, and plausibility – in this section and the remaining two – the model- and data-randomisation sanity checks – in Sect. 4.3.

4.2.1. Faithfulness

Faithfulness measures whether the features identified as important by the attribution genuinely drive the model’s prediction. We computed it via two complementary protocols.

Perturbation correlation. As a continuous summary statistic, we adopted the perturbation-based approach in which features are progressively masked in descending order of attribution magnitude, replaced by random Gaussian noise, and the Pearson correlation between cumulative removed attribution and prediction probability change is computed. Higher correlation indicates greater faithfulness. Formally, for a sample, $\mathbf{x}$, with an attribution, $\mathbf{a}$, and predicted class, c :

$$\text{Faith. (corr)} = \text{corr} \left(\sum_{j=1}^k |a_{\pi(j)}|, p_c(\mathbf{x}) - p_c(\mathbf{x}_{\text{masked}}^{(k)}) \right)_{k=1}^K, \quad (2)$$

where π orders features by $|a_j|$ descending.

Insertion and deletion AUC. Gaussian-noise masking is known to push the perturbed input off the data manifold, which may bias the correlation metric (Hooker et al. 2019). We therefore complement it with the deletion and insertion AUC of

Petsiuk et al. (2018), which uses zero-baseline masking instead. Let $\mathbf{x}_{[k]}^{\text{del}}$ be $\mathbf{x}$ with the top- k features (in attribution order) replaced by zero, and $\mathbf{x}_{[k]}^{\text{ins}}$ the all-zero baseline with the top- k features filled in from $\mathbf{x}$:

$$\text{Del. AUC} = \int_0^1 p_c(\mathbf{x}_{[\lfloor \kappa K \rfloor]}^{\text{del}}) d\kappa, \quad (3)$$

$$\text{Ins. AUC} = \int_0^1 p_c(\mathbf{x}_{[\lfloor \kappa K \rfloor]}^{\text{ins}}) d\kappa. \quad (4)$$

A faithful attribution causes the deletion curve to drop quickly (low Del. AUC) and the insertion curve to rise quickly (high Ins. AUC). The atomic units used for ranking are: the 60 features for tabular models, the six photometric passbands for time-series DL models, and a 16×16 patch grid (matching the ViT patch size) for the GAF-image models. These atomic units serve solely as the ranking granularity for the deletion and insertion protocol; the underlying photometric data – including all six passbands – are never physically masked or removed from any model’s input, only progressively replaced by the zero baseline in attribution order when computing the curves. Curve points for the patch-based variant are sub-sampled to 50 to keep the evaluation tractable on a single consumer GPU.

4.2.2. Complexity

Complexity quantifies how concentrated or diffuse the explanation is, using the normalised Shannon entropy of the attribution distribution:

$$\text{Compl. (raw)} = \frac{H(\hat{\mathbf{a}})}{\log N_{\text{eff}}}, \quad \text{where} \quad \hat{a}_j = \frac{|a_j|}{\sum_i |a_i|}, \quad (5)$$

and N_{eff} is the number of features above a relative threshold (5% of max attribution). Values near 0 indicate simple explanations dominated by few features; values near 1 indicate uniform spread.

Band-aggregated complexity. The raw definition above was computed on the native attribution space, which is 60 dimensions for the tabular models but $\sim 300,000$ dimensions for the GAF-image models. Comparing entropies across spaces of such vastly different cardinality structurally favours the lower-dimensional representations, since the same fraction of dominant signal yields a lower normalised entropy when the support is smaller. To enable a fair cross-representation comparison we additionally report a band-aggregated complexity computed on a fixed six-dimensional basis (one entry per photometric passband):

$$\text{Compl. (band)} = \frac{H(\hat{\mathbf{b}})}{\log 6}, \quad \text{where} \quad \hat{b}_p = \frac{\sum_{j \in \mathcal{B}_p} |a_j|}{\sum_i |a_i|}, \quad (6)$$

and $\mathcal{B}_p$ is the set of attribution components belonging to band p (the ten hand-crafted features for tabular models, the 256 timesteps for time-series models, and the 224×224 pixels for image models). The band-aggregated complexity is the metric we use for cross-representation comparisons; the raw complexity is retained as an architecture-specific diagnostic.

4.2.3. Plausibility

Plausibility measures the alignment between model attributions and astrophysical domain knowledge. We stress that what we call the ‘expert mask’ is not a set of fundamental physical parameters

Table 3. Expert feature mask: 21 astrophysically motivated features from the 60-dimensional tabular representation (Kessler et al. 2019; Muthukrishna et al. 2019; Boone 2019).

Feature	Bands	Physical motivation
Amplitude	u, g, r, i, z, Y	Peak brightness
Std	u, g, r, i, z, Y	Total variability proxy
Slope	g, r, i	Evolutionary timescale
Skewness	g, r, i	Rise and decline asymmetry
Mean S/N	g, r, i	Per-band detectability

Excluded: mean and median (raw statistical moments, not domain knowledge in their own right), `n_obs` (survey artefact), kurtosis (outlier-sensitive), `frac_above_mean` (redundant with skewness and amplitude)

Notes. Mean and median are excluded because raw statistical moments are not domain knowledge in their own right. The flux-only representation does not provide direct access to timestamp-aware features (time-above-half-peak) or band-difference features ($r - i$ at peak); we treat this as a deliberate methodological scope choice (Sect. 4.2.4). The mask contains 21 features.

(e.g. peak luminosity, ejecta velocity, or temperature); rather, it is a set of 21 statistical features of the observed flux – amplitude, variability, signal-to-noise ratio, timescales – that the transient-classification literature consistently identifies as diagnostically relevant (Richards et al. 2011; Narayan et al. 2018; Boone 2019). The mask therefore encodes a community consensus on which observable proxies carry astrophysical information, and the plausibility IoU measures whether a model’s attributions concentrate on those proxies relative to the full 60-feature basis. Its value lies in the relative comparison across models, not in the absolute IoU. We defined an expert feature mask $\mathbf{m} \in \{0, 1\}^{60}$ encoding 21 features expected to be physically diagnostic for transient classification (Table 3), and computed the intersection-over-union (IoU) between thresholded attributions and the mask:

$$\text{Plausibility} = \text{IoU}(\mathbf{1}[\mathbf{a}] \geq \text{median}(|\mathbf{a}|_{>0}), \mathbf{m}). \quad (7)$$

The median was taken per sample and over the strictly positive entries of $|\mathbf{a}|$, and the IoU was then averaged over test samples. The restriction to positive entries matters: it makes the size of the selected set depend on how sparse a model’s attribution is, and hence the chance level of the IoU model-dependent (Sect. 6.4). For tabular models the mask was applied directly to the SHAP attribution; for sequence and image models the IG attributions were first projected into the tabular feature space (Sect. 4.2.4).

4.2.4. Cross-representation plausibility

The 60-dimensional flux-only representation captures statistical moments of multi-band light curves but does not provide direct access to timestamp-aware or band-differencing features such as time-above-half-peak, peak magnitude, or $r - i$ colour at peak. We therefore exclude mean and median from the expert mask – they are raw statistical moments, not domain-motivated features in their own right – along with `n_obs`, kurtosis and `frac_above_mean`; retaining amplitude and std in all six bands, and slope, skewness, and mean signal-to-noise ratio (S/N) in g, r, i only, leaves a 21-feature mask (Table 3). The remaining features (amplitude, slope, std, skewness, mean S/N) are defensible proxies for peak brightness, evolutionary timescale, rise

and decline asymmetry, variability, and per-band detectability, respectively. The flux-only restriction is a deliberate methodological choice for the seven generic architectures: it isolates the effect of representation on attribution quality from the richness of upstream feature engineering. The physics-informed Avocado baseline (Sect. 3.5) provides the complementary anchor – its 41 GP-derived features include explicit colour, peak time, and rise and decline timescale, and the partial-mapping discussion of Sect. 6.4 quantifies what is gained by domain-specific feature engineering. The reader can therefore read Avocado’s classification advantage as a direct measure of what the flux-only mask misses.

A naive application of the IoU mask to a deep learning model’s attribution would compare incompatible spaces: a CNN1D produces attributions over 6×256 timesteps, while the expert mask lives in the 60-dimensional tabular feature space. To make plausibility comparable across model families, we introduced a representation-agnostic projection that maps any per-sample attribution into the same 60-dimensional vector before applying the mask.

For time-series models (LSTM, CNN1D), the IG attribution $\mathbf{A} \in \mathbb{R}^{6 \times 256}$ is reduced to ten band-aggregated statistics per passband, computed on the attribution itself: mean, standard deviation, amplitude, median, fraction of active timesteps, [skewness], [kurtosis], peak magnitude, slope proxy, and fraction above mean. The slope proxy is the absolute Pearson correlation between the attribution magnitude and the original flux profile, which captures whether IG follows the temporal evolution of the band signal – precisely the astrophysical content of the tabular ‘slope’ feature. The ten statistics are concatenated across the six passbands, producing a 60-vector aligned with the expert mask.

For image models (CNN2D, ViT), the IG attribution $\mathbf{A} \in \mathbb{R}^{6 \times 224 \times 224}$ on the GAF representation is reduced via a similar lookup table. Five core geometric statistics were computed per band – mean, peak, fraction of active pixels, attribution concentration ($1 - H/H_{\max}$, where H is the Shannon entropy of the heatmap), and a diagonality score (mean attribution along the main diagonal minus mean off-diagonal attribution). The diagonality score serves as the slope proxy because GAF images encode local temporal evolution along the main diagonal, so high diagonal energy in the attribution indicates that the model is attending to rapidly evolving flux. These statistics are mapped to the ten tabular slots per band; the ‘skewness’ and ‘kurtosis’ slots both receive the concentration value, since 2D heatmaps do not separate the two moments cleanly.

For the stacking ensemble, plausibility is reported as the meta-coefficient-weighted average of the six base-model plausibilities, where each base model is weighted by the sum of $|\mathbf{w}_{\text{meta}}|$ over the 14 meta-features it contributes to the logistic head. The reported standard deviation propagates the per-model variances assuming independence.

For the physics-informed baselines, the native attribution spaces are again incommensurate with the 60-dimensional mask. Avocado’s 41 GP-derived features differ structurally from the 60 per-band statistics: we define a partial mapping from 21 Avocado features to 20 of the 60 expert-mask slots, grouped by physical analogy (e.g. flux ratios $\rightarrow$ amplitude, GP length scale $\rightarrow$ slope). The remaining 20 Avocado features fall into two categories – 11 astrophysically meaningful features with no direct analogue in our per-band statistics (multi-peak ratios, host photo- z) and nine observing-cadence artefacts that carry no physical content – and are zero-padded. Avocado plausibility should therefore be read as a lower bound relative to the tabular models, and we show in Sect. 6.4 that the partial mapping induces a

structural degeneracy in its score. For ParSNIP, the latent-prediction features are mapped heuristically onto the colour, amplitude and timescale slots of the mask; this mapping is inherently lossy and is reported with the same lower-bound caveat.

This projection is approximate by construction: it relies on a hand-designed mapping rather than a learned alignment, and the GAF skewness and kurtosis slots share information. Nonetheless, it allows plausibility to be reported on a common scale across all nine models, which is, to our knowledge, the first such comparison in the time-domain astronomy literature.

4.2.5. Validation of the projection

To check that the projection-and-scoring pipeline does not itself distort the plausibility score, we used XGBoost as an identity control. Because XGBoost’s SHAP attribution already lives natively in the 60-dimensional tabular space, routing it through the cross-representation dispatcher with an identity map must return exactly the plausibility obtained by applying the IoU directly to the SHAP attribution. The two agree to numerical precision ($|\Delta| < 10^{-6}$; Table B.5), confirming that the binarisation-and-IoU stage shared by the tabular, time-series, and image entry points is implemented consistently. This control validates the scoring stage; the per-band aggregation that the time-series and image projections add on top of it is an approximation whose limitations we discuss in Sect. 6.4.

4.3. Saliency sanity checks

Following Adebayo et al. (2018), we asked whether our explanation maps actually depend on the trained model and not on architectural priors alone. We ran two complementary sanity checks for every deep learning model:

Model randomisation test. Starting from the trained network, we progressively replaced each layer’s parameters with random Gaussian noise in top-down order (output side first). At each cascading step, we recomputed the IG attributions for a fixed set of test samples and report their Spearman rank correlation with the trained-model attributions. A faithful saliency method should produce attributions that decorrelate from the trained baseline as randomisation propagates inwards; persistent correlation would indicate that the explanations are largely architecture-driven and weight-insensitive.

Data randomisation test. We retrained each model from scratch on a label-permuted copy of the training set (single seed, $\sigma_y = 42$), keeping the architecture, optimiser, and early-stopping schedule identical to the trained baseline. The Spearman correlation between the trained-model and random-label-model attributions on the same test samples measures the extent to which the explanations depend on the supervision signal: a value close to zero indicates that label information is required to produce the observed attribution structure.

For the tree-based models (Random Forest, XGBoost), we ran only the data randomisation test, since SHAP attributions on a single tree have no meaningful model-randomisation analogue. The same applies to Avocado, whose classifier is a LightGBM ensemble: we ran only the data randomisation test, retraining on permuted labels without class-balanced sample weights so that the protocol matches the one applied to Random Forest and XGBoost (see Sect. 5.3). For ParSNIP we randomised and retrained only the downstream LightGBM head while holding the pre-trained VAE encoder fixed, which tests the sensitivity of

Table 4. Classification performance (mean $\pm$ std over five seeds).

Model	Acc.	F1 _m	AUC
RF	.646 $\pm$.004	.430 $\pm$.006	.907 $\pm$.001
XGB	.668 $\pm$.000	.529 $\pm$.000	.909 $\pm$.000
LSTM	.605 $\pm$.007	.376 $\pm$.009	.886 $\pm$.003
CNN1D	.529 $\pm$.036	.289 $\pm$.040	.834 $\pm$.029
CNN2D	.560 $\pm$.009	.435 $\pm$.006	.880 $\pm$.004
ViT	.559 $\pm$.012	.447 $\pm$.014	.866 $\pm$.006
Ens.	.709 $\pm$.006	.568 $\pm$.015	.922 $\pm$.003
Avocado	.801$\pm$.005	.777$\pm$.007	.979$\pm$.001
ParSNIP	.647 $\pm$.024	.581 $\pm$.034	.952 $\pm$.008

Notes. The ensemble uses out-of-fold stacking (Sect. 3.4). The two rows below the lower rule are the physics-informed baselines (Sect. 3.5); Avocado is the strongest classifier overall (bold).

the classification stage given a fixed learned representation; this is a deliberately narrower scope than the full-network randomisation applied to the deep learning models and is stated as such. The results of both tests are summarised in Table 6.

5. Results

5.1. Classification performance

Table 4 reports the classification metrics, averaged across five random seeds. The out-of-fold stacking ensemble (Sect. 3.4) achieves both the highest accuracy (0.709 ± 0.006) and the highest macro-F1 (0.568 ± 0.015) among the seven generic architectures, surpassing XGBoost (0.668 ± 0.000 accuracy, 0.529 ± 0.000 macro-F1) and Random Forest (0.646 ± 0.004 accuracy, 0.430 ± 0.006 macro-F1). The XGBoost cross-seed standard deviation collapses to zero because, with a fixed `random_state` and identical training data, the algorithm is fully deterministic – the only meaningful seed dependence in this work occurs for the stochastic models. We note that the ensemble out-of-fold accuracy (0.709) is higher than the value we obtained when fitting the meta-learner naively on validation-set predictions of the same base models (0.688). The improvement is not due to leakage – the OOF protocol explicitly removes it – but to the larger number of training samples available to the meta-learner under cross-validation (5,493 OOF training points vs ~ 785 validation samples in the naive fit), which allows the logistic head to better calibrate the relative weights of the six base models.

Among the non-ensemble architectures, the Vision Transformer (0.447 ± 0.014) and CNN2D (0.435 ± 0.006) outperform Random Forest (0.430 ± 0.006) in macro-F1 despite trailing it in accuracy, indicating that the deep learning models on GAF images, while less accurate overall, are more balanced across the 14 classes. The CNN1D is the weakest classifier on every metric ($\text{acc} = 0.529 \pm 0.036$, $\text{F1}_{\text{macro}} = 0.289 \pm 0.040$) and also the most variable: its cross-seed standard deviation is roughly six times larger than that of the other architectures, reflecting the well-known instability of small 1D convolutional networks under aggressive early stopping.

All models except CNN1D achieve $\text{AUC} > 0.86$, indicating reasonable ranking ability even when hard classification accuracy is moderate – consistent with the strong class imbalance in PLAsTiCC. The normalised confusion matrices (Fig. 1) reveal that tree-based models produce the cleanest diagonal structure, while deep learning models show systematic confusion among

the supernova subtypes (SNI, SNIax, SNIbc, SNIa-91bg and SNIa; indices 3, 4, 6, 9 and 11). All panels are drawn from seed 42; the CNN1D panel is a low realisation for that model, with a support-weighted accuracy of 0.46 against its five-seed mean of 0.529 ± 0.036 , whereas the other five panels fall within 0.013 of their own means.

The two physics-informed baselines bracket the generic architectures from opposite directions. Avocado is the strongest classifier in the entire benchmark by a wide margin (0.801 ± 0.005 accuracy, 0.777 ± 0.007 macro-F1, 0.979 ± 0.001 AUC), exceeding the out-of-fold ensemble by $+0.09$ in accuracy and by more than $+0.20$ in macro-F1. The macro-F1 gain is the more telling figure: Avocado’s GP-based redshift augmentation produces usable per-class recall even for the rare supernova subtypes on which every generic model collapses (e.g. SNIax, KN), which is precisely where balanced performance is hardest on PLAsTiCC. ParSNIP is more modest (0.647 ± 0.024 accuracy, 0.581 ± 0.034 macro-F1), comparable to the tree-based models in accuracy but with a markedly higher macro-F1, consistent with its generative latent representation being tuned for supernova-like classes; its larger cross-seed spread reflects the per-seed VAE retraining. Both substantially outperform the generic models in macro-F1, confirming that architectures that encode the redshift-dependent physics of transient light curves retain a real advantage on this dataset, and that the xAI metrics below are therefore being measured on genuinely competitive classifiers rather than on weak baselines.

5.2. Explanation quality

We report the full metric summary across all models below. Sect. 6.1 then walks through XGBoost and CNN2D as a detailed case study illustrating how the five xAI axes interact with model architecture.

Faithfulness (perturbation correlation). Random Forest achieves the highest perturbation-correlation faithfulness (0.871 ± 0.018), substantially above all other models (Table 5). This is expected: TreeSHAP provides exact Shapley values for tree ensembles, and the compact 60-feature space allows perturbations to meaningfully affect predictions. Among deep learning models the GAF-image CNN achieves the highest faithfulness (0.645 ± 0.068), narrowly ahead of the Vision Transformer (0.606 ± 0.065); LSTM and CNN1D follow at 0.542 ± 0.059 and 0.281 ± 0.137 , respectively. The very low and highly variable CNN1D faithfulness mirrors its classification instability and indicates that the model’s attributions, while computed via the axiomatic integrated-gradients method, do not consistently track the features that drive the prediction. The stacking ensemble has low faithfulness (0.478 ± 0.060) because perturbations on the meta-feature space (84 base-model probability outputs; Sect. 3.4) only weakly map back to physically meaningful feature changes.

Insertion and deletion AUC. The zero-baseline insertion and deletion AUC tells a partially aligned but more nuanced story (Table 5, Fig. 5). Random Forest still wins on the deletion side (Del. AUC = 0.170 ± 0.004 , the lowest in the benchmark): masking the most-attributed features in attribution order collapses the predicted-class probability faster than for any other model. On the insertion side, however, XGBoost takes the lead (Ins. AUC = 0.623 ± 0.031), closely followed by ViT (0.600 ± 0.083) and CNN1D (0.528 ± 0.104): inserting only the top patches or features in attribution order recovers most of the prediction more

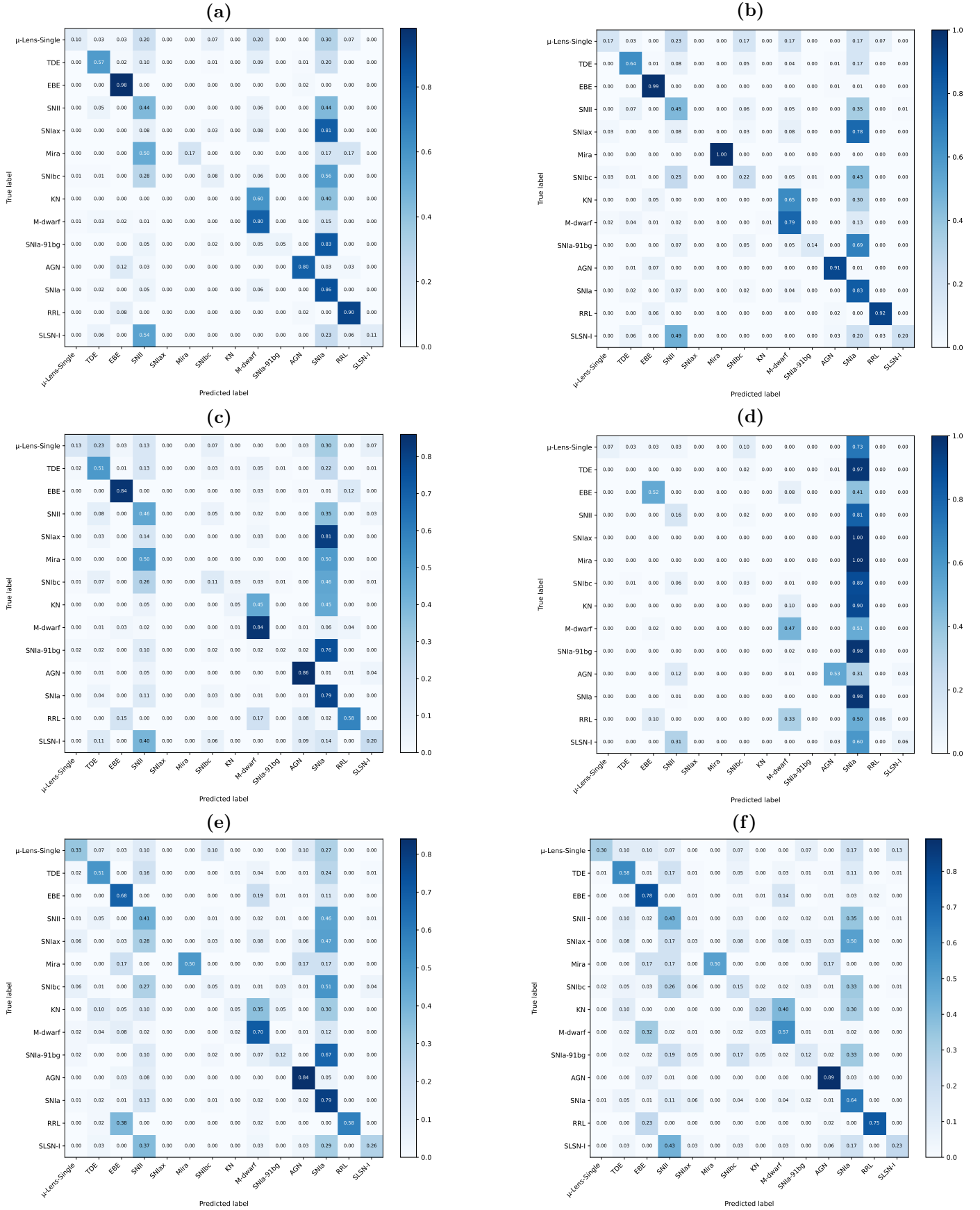


Fig. 1. Normalised confusion matrices (recall) for the six base models on the PLAsTiCC test set: (a) Random Forest and (b) XGBoost (top row, tree-based); (c) LSTM and (d) CNN1D (middle row, time-series); (e) CNN2D and (f) ViT (bottom row, GAF-image), with the model name printed above each panel.

Table 5. Explanation-quality metrics on the PLAsTiCC test set, averaged over five random seeds.

Model	Faith.	Del.↓	Ins.↑	C_{raw}	C_{band}	Plaus.	Chance	Ceil.	Resc.
RF	.871±.018	.170±.004	.385±.017	.911±.003	.955±.003	.229±.002	.262	.700	−.08
XGB	.671±.034	.217±.020	.623±.031	.889±.006	.990±.000	.224±.000	.262	.700	−.09
LSTM	.542±.059	.316±.050	.460±.043	.913±.012	.941±.009	.381±.008	.250	.767	.25
CNN1D	.281±.137	.393±.109	.528±.104	.923±.003	.951±.005	.398±.005	.249	.771	.29
CNN2D	.645±.068	.209±.022	.405±.024	.987±.000	.965±.003	.149±.005	.255	.736	−.22
ViT	.606±.065	.330±.042	.600±.083	.981±.002	.986±.002	.257±.002	.262	.700	−.01
Ens.	.478±.060	—	—	.834±.003	—	.250±.002	.259	.715	−.02
Avocado	.631±.054	—	—	.903±.003	—	.399±.014	.140	.524	.67
ParSNIP	.706±.033	—	—	.946±.007	—	.215±.012	.133	.493	.23

Notes. Faith.: perturbation-based Pearson correlation; Del. and Ins.: deletion and insertion AUC (Petsiuk et al. 2018), for which lower deletion and higher insertion values are better; $C_{\text{raw}}/C_{\text{band}}$: entropy-based complexity on native and six-band basis (Sect. 4.2.2); Plaus.: IoU against the 21-feature expert mask (Sect. 4.2.4). The two rows below the lower rule are the physics-informed baselines (Sect. 3.5); their deletion and insertion AUC and band complexity are undefined because they operate on derived feature spaces with no shared per-band atomic basis, and are marked ‘—’. The expert mask excludes the per-band mean and median (raw statistical moments rather than domain features; Sect. 4.2.4). Chance: permutation baseline for the plausibility IoU (Sect. 6.4), obtained by applying the same estimator to a random permutation of each sample’s attribution vector; averaged over 200 permutations per sample and over the five seeds. It is model-dependent because the binarisation threshold is the median of the positive attribution entries, so a sparse attribution selects fewer features: the two physics-informed baselines select ~11 of 60 against ~30 for the dense attributions. Ceil.: the largest IoU reachable at that selected-set size, $\min(|A|, 21)/\max(|A|, 21)$; it too varies, from 0.700 where $|A| = 30$ to 0.493 for ParSNIP. Resc.: the position within the attainable range, $(\text{Plaus.} - \text{Chance})/(\text{Ceil.} - \text{Chance})$, which is the only one of the three columns that may be compared across model families – negative values denote alignment poorer than a random selection. The ensemble has no attribution vector of its own to permute, its plausibility being a coefficient-weighted mean of the base-model scores; its chance level and ceiling are the same weighted means of theirs.

efficiently than for Random Forest (0.385 ± 0.017) or CNN2D (0.405 ± 0.024). The CNN2D in particular has a low Ins. AUC paired with a low Del. AUC, suggesting that its attributions correctly identify which patches matter for removal but those patches are not on their own sufficient to reconstruct the prediction – a signature of distributed encoding across the GAF image. The two faithfulness protocols therefore agree that CNN1D is the least faithful DL model and that Random Forest is the most faithful tabular model, but they disagree on the ranking of the GAF-image architectures, which we attribute to the off-manifold concerns of correlation-based perturbation (Hooker et al. 2019). The curves themselves (Fig. 5) show the mechanism directly: a faithful attribution makes the deletion curve fall quickly, since removing the most-attributed features collapses the predicted-class probability, and the insertion curve rise quickly, since restoring those same features recovers most of the prediction. The Random Forest curves separate cleanly, consistent with its high faithfulness, whereas the ViT curves are flatter, reflecting the diffuse nature of pixel-level attribution on GAF images.

Calibrated faithfulness baselines. Table B.1 compares attribution-ordered AUC against random and oracle baselines. All models outperform the random baseline on at least one protocol, confirming that the explanations carry genuine signal; the one exception is CNN2D on the deletion side, where the attribution-ordered AUC matches the random baseline (0.172) – the signal concentrates entirely on the insertion protocol. The largest gaps are those of the ViT on deletion (0.186) and of the CNN2D on insertion (0.154), the smallest that of the LSTM on insertion (0.057); the ordering therefore does not track the perturbation-correlation ranking, a further indication that the two faithfulness protocols probe different properties of an attribution. RF’s attribution is already close to the oracle on deletion (0.172 vs 0.159).

Complexity (raw vs band-aggregated). The raw complexity, computed on each model’s native attribution space, ranks

the ensemble meta-learner first (0.834 ± 0.003), followed by XGBoost (0.889 ± 0.006) and Random Forest (0.911 ± 0.003); the GAF-image models are pushed to near-uniform attribution distributions (ViT = 0.981 ± 0.002 , CNN2D = 0.987 ± 0.000). When we recompute complexity on the fair six-band basis (Sect. 4.2.2), the ranking changes substantially: LSTM becomes the simplest model (0.941 ± 0.009), followed by CNN1D (0.951 ± 0.005), Random Forest (0.955 ± 0.003), CNN2D (0.965 ± 0.003), ViT (0.986 ± 0.002), and XGBoost last at 0.990 ± 0.000 . The reversal between XGBoost (best on raw complexity among the base models, worst on band complexity) and the time-series DL models illustrates the importance of basis choice when comparing attribution entropies across heterogeneous representations: the raw metric structurally rewards the lower-dimensional input spaces, while the band metric reveals that XGBoost actually distributes its SHAP importance most uniformly across the six photometric passbands.

A Gini-index robustness check (0 = perfectly uniform, 1 = maximally sparse) sharpens the same conclusion. On the common 60-feature basis it ranks XGBoost as the most evenly spread (0.41), then Random Forest (0.56), with the projected time-series attributions much the sparsest (LSTM and CNN1D 0.89) – precisely the ordering of the band-aggregated complexity, and the reverse of the raw one. An independent sparsity measure therefore sides with C_{band} against C_{raw} . That C_{raw} dissents even for the two tabular models, whose native attribution space is the 60-feature basis and for which no change of basis is involved, points to its normalisation rather than to the basis: the $\log N_{\text{eff}}$ denominator of Sect. 4.2.2 is itself model-dependent, so C_{raw} is not a monotone function of sparsity and should not be read as one.

Plausibility. The cross-representation projection (Sect. 4.2.4) yields the most striking finding of the benchmark and inverts the naive expectation that tree-based models on hand-crafted features should align best with the expert mask. The two time-series deep learning models score higher plausibility than the tabular tree models: CNN1D = 0.398 ± 0.005 and

LSTM = 0.381 ± 0.008 versus RF = 0.229 ± 0.002 and XGBoost = 0.224 ± 0.000 . The Vision Transformer (0.257 ± 0.002) and especially the CNN2D (0.149 ± 0.005), both operating on GAF images, sit at the opposite end. The stacking ensemble lands in the middle (0.250 ± 0.002), as expected from the coefficient-weighted average over six heterogeneous base models. The cross-seed variances for the seven generic architectures are small (≤ 0.008); the two physics-informed baselines show somewhat larger variances (up to 0.014 for Avocado), still within an acceptable range, indicating that the projection is robust to seed-driven training fluctuations. All values use the 21-feature expert mask (Sect. 4.2.4). We discuss the astrophysical interpretation of this ordering in Sect. 6.4.

Physics-informed baselines. Across the xAI axes the two domain-specific models behave very differently from their accuracy ranking. On TreeSHAP attributions of their LightGBM heads, ParSNIP attains the second-highest faithfulness in the benchmark on the perturbation-correlation protocol (0.706 ± 0.033 , behind only Random Forest), while Avocado is intermediate (0.631 ± 0.054); their raw complexities (0.946 ± 0.007 and 0.903 ± 0.003) are comparable to the tree-based models. Because both classifiers operate on their own derived feature spaces rather than on our three fixed representations, the deletion and insertion AUC and the band-aggregated complexity – which require a shared per-band atomic basis – are not defined for them and are left blank in Table 5. Their projected plausibility scores diverge: Avocado (0.399 ± 0.014) attains the highest value in the benchmark, numerically level with the best time-series model, CNN1D (0.398 ± 0.005), while ParSNIP (0.215 ± 0.012) falls below both the tree-based and the time-series baselines. Both coincidences are artefacts of scale: because the two physics-informed baselines produce far sparser projected attributions, they are measured against lower chance levels and lower ceilings than the generic models (Table 5). Rescaled onto the attainable range as $(\text{IoU} - \text{chance})/(\text{ceiling} - \text{chance})$ – the ‘Resc.’ column of Table 5 – Avocado reaches 0.67 against 0.29 for CNN1D, a lead of a factor 2.4 rather than a tie, and ParSNIP reaches 0.23, level with the LSTM (0.25) and above both tree-based models, which are negative. We use the raw IoU for continuity with Table 5 but read all cross-family comparisons on the rescaled quantity. The Avocado advantage reflects its explicit colour, peak-time and rise and decline features mapping into the expert mask; the ParSNIP drop reflects a sensitivity of its heuristic latent-to-tabular projection to the exclusion of the mean and median slots – a limitation of the cross-representation projection rather than of the model itself, consistent with the caveats in Sect. 6.4.

Plausibility threshold sensitivity. A sensitivity analysis across four IoU percentile thresholds (Table B.2) confirms that the ordering of the tree-based and time-series models is stable in the p25–p50 range; at p75 and above, the diffuse DL attributions are thresholded away while tree models degrade more gracefully.

5.3. Sanity check results

The model- and data-randomisation tests of Sect. 4.3 are summarised in Table 6 and yield a clear partition of the benchmark into models whose attributions are weight- and label-sensitive and models whose attributions are not. CNN2D and ViT pass all applicable checks cleanly: cascading randomisation drives the Spearman correlation to 0.000 ± 0.000 , and label permutation yields $\rho_{\text{data}} = 0.255$ and $\rho_{\text{data}} = 0.341$, respectively. XGBoost

Table 6. Saliency sanity checks following Adebayo et al. (2018).

Model	Model rand. ρ	Data rand. ρ
Random Forest	–	0.862
XGBoost	–	0.348
LSTM	0.995 ± 0.000	0.996
CNN1D	0.000 ± 0.000	0.995
CNN2D	0.000 ± 0.000	0.255
ViT	0.000 ± 0.000	0.341
Avocado	–	0.265
ParSNIP	–	–0.300

Notes. Lower Spearman correlation, ρ , between trained-model and perturbed-model attributions is better, indicating that the explanations actually depend on (a) the learned weights and (b) the supervision signal. Tabulated values are absolute-value Spearman correlations, with the single exception of ParSNIP, for which the signed value is quoted in order to expose the sign inversion discussed below; Table B.3 decomposes the two variants for the deep-learning models, where they differ substantially for the LSTM. Tree-based models do not have a meaningful model-randomisation analogue and are therefore reported only for the data randomisation test. The LSTM and CNN1D values are reported with the per-band mean of the training light curves as the IG baseline; we verified that switching to a zero baseline changes each ρ by less than 0.003, consistent with the analytical observation that no choice of baseline can rescue the degeneracy (Sect. 6.5). For CNN1D, the data-randomisation ρ was computed on a random-label model trained for 50 forced epochs without early stopping, ruling out an init-proximity artefact. The two physics-informed baselines (below the rule) are tree-based classifiers and are reported for the data-randomisation test only: Avocado without class-balanced sample weights (matching RF and XGBoost), and ParSNIP by randomising only its downstream LightGBM head while holding the VAE encoder fixed. ParSNIP’s negative ρ indicates that random-label attributions are systematically inverted relative to the trained-model attributions rather than merely decorrelated; in absolute terms $|\rho| = 0.300$, comparable to CNN2D (0.255) and XGBoost (0.348).

passes the data randomisation test ($\rho_{\text{data}} = 0.348$); the model-randomisation test does not apply to TreeSHAP and is left blank.

Three models fail one or both checks. CNN1D passes the model-randomisation test ($\rho_{\text{model}} = 0.000$) but fails the data-randomisation test ($\rho_{\text{data}} = 0.995$): the random-label retraining early-stops within ~ 15 epochs at chance accuracy, so the resulting weights produce attributions that are essentially a fixed function of the input rather than of the weights, which is in turn nearly identical to the input-dominated trained-model attributions. Forcing 50 training epochs without early stopping does not change the result ($\rho_{\text{data}} = 0.995$), ruling out an initialisation-proximity artefact. Random Forest also fails the data-randomisation test ($\rho_{\text{data}} = 0.862$), indicating that TreeSHAP is dominated by the marginal feature distribution rather than by the supervised signal at the precision used here. We verified that both tree models achieve 100% training accuracy on the shuffled labels (vs a chance level of $1/14 \approx 7.1\%$), confirming that the random-label models fully memorise the permuted targets and that the high Spearman correlations reflect genuine attribution insensitivity rather than model underfitting. Most strikingly, the LSTM fails both checks ($\rho_{\text{model}} = 0.995 \pm 0.000$, $\rho_{\text{data}} = 0.996$): even fully randomising the LSTM weights barely changes its IG attributions. We attribute this to the combination of a fixed reference baseline and the very small per-sample gradients produced by the bidirectional pooled architecture: the IG

approximation degenerates to $IG_i \approx x_i \cdot c$ for some near-constant c , which makes the attribution a function of the input alone.

Signed versus absolute correlations. Following Adebayo et al. (2018), Table B.3 decomposes the Spearman correlation into signed and absolute-value variants. CNN1D, CNN2D, and ViT drop to zero under both variants of the model-randomisation test. The LSTM reveals a striking dissociation: the signed correlation drops to $\rho_{\text{signed}} = 0.148$ after full randomisation, while the absolute correlation remains at $\rho_{|\cdot|} = 0.997$ – randomisation scrambles the sign but preserves the magnitude profile, precisely the pattern expected when $|IG_i| \propto |x_i|$ regardless of the weights.

We verified that these failures are not artefacts of the choice of IG baseline: repeating with per-band mean and Gaussian-noise baselines yields essentially identical attribution profiles (Table B.4), and therefore identical sanity-check correlations. This invariance is mathematically expected: Spearman correlation is invariant to monotone rescalings, and the IG expression degenerates to a monotone function of $\mathbf{x}$ whenever the integrated gradient is approximately uniform across input dimensions.

Both physics-informed baselines pass the data-randomisation test. Avocado’s TreeSHAP attributions on a permuted-label retraining correlate with the trained-model attributions at only $\rho_{\text{data}} = 0.265^1$ – markedly better than the $\rho_{\text{data}} = 0.862$ of the otherwise-similar Random Forest, indicating that Avocado’s GP-derived features carry genuinely label-dependent attribution structure. ParSNIP’s downstream-head randomisation yields a negative signed correlation, $\rho_{\text{data}} = -0.300$, indicating that the random-label classifier’s SHAP rankings are systematically inverted relative to the trained-model attributions rather than merely decorrelated from them. In absolute terms ($|\rho_{\text{data}}| = 0.300$) ParSNIP joins the passing cluster alongside CNN2D (0.255) and XGBoost (0.348); the negative sign is an interesting methodological signature – not a better score than a near-zero $|\rho|$ such as that of CNN2D. We reiterate that ParSNIP’s test randomises only the LightGBM head and not the VAE encoder, so this result certifies the classification stage rather than the learned representation.

The implication for users of the benchmark is that LSTM integrated-gradient attributions and TreeSHAP on Random Forest, while numerically well defined, should not be over-interpreted on this dataset; XGBoost SHAP and the GAF-image deep models produce attributions that demonstrably depend on both the learned weights and the supervision signal. We discuss the methodological implications in Sect. 6.5.

6. Discussion

6.1. Case study: XGBoost versus CNN2D

To illustrate how the five xAI axes interact with architecture, we walk through two representative models in detail: XGBoost (tree-based, tabular) and CNN2D (convolutional, GAF images). Both pass all applicable sanity checks, removing a confound that would complicate interpretation.

¹ An initial implementation of the Avocado data-randomisation test applied class-balanced sample weights to the random-label retraining. This artefactually depressed the correlation (from 0.265 to 0.060) by giving the random-label model a calibration advantage that made its attributions more similar to those of the genuinely trained model. The reported value was computed without sample weighting, matching the protocol applied to Random Forest and XGBoost.

Classification. XGBoost (0.668 accuracy, 0.529 macro-F1) outperforms CNN2D (0.560 accuracy, 0.435 macro-F1). This gap is expected: the 60-feature tabular space is compact relative to $\sim 5\,500$ training samples, whereas the 6×224^2 GAF image is vastly over-parameterised for the same sample size (Sect. 6.6). The xAI metrics below should therefore be read in the context of a stronger versus a moderate classifier.

Faithfulness. XGBoost’s perturbation-correlation faithfulness (0.671 ± 0.034) exceeds CNN2D’s (0.645 ± 0.068), but the gap is smaller than the accuracy gap. TreeSHAP provides exact Shapley values in a 60-dimensional space where individual feature perturbations are well defined; Integrated Gradients on a 6×224^2 image distributes attribution across $\sim 300\text{K}$ pixels, so that masking individual pixels has a weaker per-unit effect. On the deletion side, CNN2D (0.209) actually outperforms XGBoost (0.217): when the most-attributed patches are removed in order, CNN2D’s prediction collapses faster, indicating that the IG attribution correctly identifies the driving spatial regions even though it is diffuse at the pixel level. The lesson is that faithfulness is representation-dependent: a ‘less faithful’ attribution on a high-dimensional input under the perturbation-correlation protocol may simply reflect a less localised prediction mechanism, not a less trustworthy explanation.

Complexity. The raw entropy-based complexity of CNN2D (0.987) is near the theoretical maximum, while XGBoost’s (0.889) is substantially lower. This ranking inverts under band aggregation (Sect. 4.2.2): XGBoost becomes the most complex model (0.990), distributing its SHAP importance nearly uniformly across the six passbands, while CNN2D drops to 0.965. The inversion illustrates that ‘complexity’ is not a property of the model alone but of the $\langle \text{model}, \text{basis} \rangle$ pair: XGBoost concentrates attribution on a few features within each band (low raw entropy) but spreads it evenly across bands (high band entropy), whereas CNN2D does the opposite.

Plausibility. Both models score low in absolute plausibility, with XGBoost (0.224) above CNN2D (0.149). For XGBoost, TreeSHAP concentrates on per-band S/N and skewness features that intersect only partially with the 21-feature expert mask, yielding a modest IoU. For CNN2D, the cross-representation projection (Sect. 4.2.4) must map 224×224 spatial attributions onto 60 tabular slots via a hand-designed lookup, which penalises off-diagonal GAF structure that does not correspond to any single tabular statistic. This is a limitation of the projection, not of the model: CNN2D’s attributions are spatially concentrated on regions encoding rapid flux evolution (Fig. 4), which is astrophysically meaningful but not captured by the IoU with a tabular mask.

Sanity checks. Both pass the data-randomisation test: XGBoost $\rho_{\text{data}} = 0.348$, CNN2D $\rho_{\text{data}} = 0.255$. CNN2D additionally passes the model-randomisation test ($\rho_{\text{model}} = 0.000$). The passing scores confirm that the attributions of both models depend on the learned function and the supervision signal, and that the metrics above – faithfulness, complexity, and plausibility – are measuring genuine explanation properties rather than input-distribution artefacts.

Summary. XGBoost provides more faithful and simpler explanations in the raw feature space; CNN2D provides more targeted spatial attributions that are harder to project onto a tabular basis but pass stricter sanity checks. Neither model dominates the other across all five axes, which is precisely the motivation for a

multi-axis benchmark: a single metric would mask the trade-off between architectural transparency and representational richness.

6.2. The accuracy–interpretability trade-off

Our results quantify a tension that is often discussed qualitatively: models with higher accuracy do not necessarily produce better explanations. The out-of-fold stacking ensemble dominates the classification metrics (Table 4; 0.709 ± 0.006 accuracy and 0.568 ± 0.015 macro-F1; Fig. C.2 shows the confusion matrix of the non-OOF construction) but ranks second-to-last in perturbation–correlation faithfulness (0.478 ± 0.060 , only above CNN1D), and the meta-feature space on which it operates does not admit the insertion and deletion AUC by construction. Random Forest sits at the opposite end: the lowest accuracy among the tabular models but the highest faithfulness on both the perturbation correlation (0.871 ± 0.018) and the deletion AUC (0.170 ± 0.004). XGBoost is the most balanced compromise: highest macro-F1 among the generic non-ensemble models (0.529), highest insertion AUC (0.623), and the second-best perturbation faithfulness among the generic architectures (0.671; ParSNIP reaches 0.706). When the practitioner must commit to a single model that is both accurate and interpretable on this dataset, gradient-boosted trees with TreeSHAP currently sit at the sweet spot.

6.3. Deep learning explanations are diffuse

The high complexity scores for the deep learning models, in particular for the GAF-image architectures (≥ 0.98), reflect a fundamental challenge: even Integrated Gradients, which satisfies the completeness and sensitivity axioms unlike vanilla gradient saliency (Sundararajan et al. 2017), distributes attribution across thousands of input dimensions, making it difficult to identify a concise set of ‘reasons’ for a prediction at the raw-input level. The time-series models score noticeably lower complexity (~ 0.91 – 0.92), consistent with the smaller input dimensionality. Importantly, our cross-representation plausibility (Sect. 4.2.4) shows that high pixel-level complexity is not equivalent to physical irrelevance: when the attributions are aggregated to physically meaningful units (per-band statistics), the time-series deep learning models in fact exceed the tree-based models in plausibility. Methods that perform such aggregation to a domain-grounded basis are therefore a productive direction for bridging the apparent gap between feature-level and pixel-level xAI.

For the CNN2D, a brief comparison of IG versus Grad-CAM on the deletion and insertion protocol shows largely overlapping important regions, though IG’s pixel-level granularity gives it an edge on deletion AUC (0.172 vs 0.243; both values are seed-42 on 15 test samples, bracketing the five-seed CNN2D mean of 0.209 in Table 5). Per-sample attribution times range from ~ 2 ms (TreeSHAP on XGBoost) to ~ 700 ms (IG on LSTM), all tractable for real-time broker scenarios.

6.4. Cross-representation plausibility reveals representation effects

The expert feature mask approach introduces subjectivity but also domain grounding that generic xAI metrics lack. Our mask encodes consensus knowledge from the transient classification literature, but we emphasise that it represents a necessary set of features, not a sufficient one. Models that attend to features outside the mask may be exploiting genuine but previously

unrecognised patterns – or survey artefacts. Disentangling these cases remains an open challenge.

The cross-representation projection (Sect. 4.2.4) produces the most surprising result of the benchmark: time-series DL models score higher plausibility than the tree-based models operating on the same hand-crafted features. Two complementary factors explain this. First, the projection rewards attributions spread across multiple bands that correlate with the flux profile; time-series models naturally produce such patterns, while tree-based SHAP concentrates on a selective feature subset that intersects the expert mask less generously under the IoU threshold. Second, the DL models are not as ‘physically blind’ as their high complexity suggests: the global geometry of their attributions is consistent with astrophysically relevant structure, a meaningful nuance for the xAI literature on transient classification.

Such a comparison is only interpretable against a null model, which the IoU of Sect. 4.2.3 does not supply by construction. We therefore computed a permutation baseline: the same estimator applied to a random permutation of each sample’s attribution vector, which leaves the sparsity pattern – and hence the number of selected features – untouched while randomising which features are selected. The baseline is model-dependent, because the binarisation threshold is the median of the positive attribution entries: models with dense attributions select 30 of the 60 features and sit at a chance IoU of 0.26, whereas Avocado and ParSNIP, whose projected attributions are 65–73% zero, select ~ 11 and sit at 0.13–0.14 (Table 5). Read against their own baselines, four models exceed chance – Avocado (2.9 $\times$), ParSNIP (1.6 $\times$), CNN1D (1.6 $\times$) and LSTM (1.5 $\times$); the ViT and the ensemble sit at chance (0.98 $\times$ and 0.97 $\times$); and three fall below it: Random Forest (0.87 $\times$), XGBoost (0.85 $\times$) and CNN2D (0.58 $\times$). A single analytic chance level would therefore misread the two physics-informed baselines, whose raw IoU is depressed by the sparsity of their projected attributions rather than by any misalignment with the mask.

The tree-based result is the sharpest form of the finding above: the two models trained directly on the hand-crafted features that define the expert mask align with it slightly worse than a random reordering of their own attributions. Figures 2 and 3 show why. Their TreeSHAP rankings are dominated by `mean_snr` and `skewness`; no std feature reaches the top 20 and `amplitude` appears only at rank 18–19, yet `amplitude` and `std` together occupy 12 of the 21 mask slots. Even the statistic they do share is only half-shared: `mean_snr` enters the mask for g , r , i alone, while several of the leading tree features are its u , z and Y counterparts. Counted over the whole selected set rather than the top 20, 0.32 of the features Random Forest selects fall inside the mask and 0.31 of XGBoost’s, against the $21/60 = 0.35$ expected of a random selection, whereas the time-series models reach 0.48–0.50 and Avocado 0.85. The margin below chance is thus small, and the substantive point is not that the tree models are perverse but that the mask and the tree models encode two genuinely different views of which measurements carry class information.

This reading is threshold-dependent, and we state where it holds. At the p25 and p50 binarisations the pattern above is stable, but at p75 it inverts: XGBoost rises above its own chance level (0.187 against ≈ 0.17) while CNN1D and LSTM fall below theirs (0.161 and 0.156), because at stringent thresholds nearly all of the diffuse deep-learning attribution mass is discarded while the sparse tree attributions survive (Table B.2). Chance-normalised plausibility should therefore be read in the p25–p50 range, the same range in which the raw ranking is stable, and a

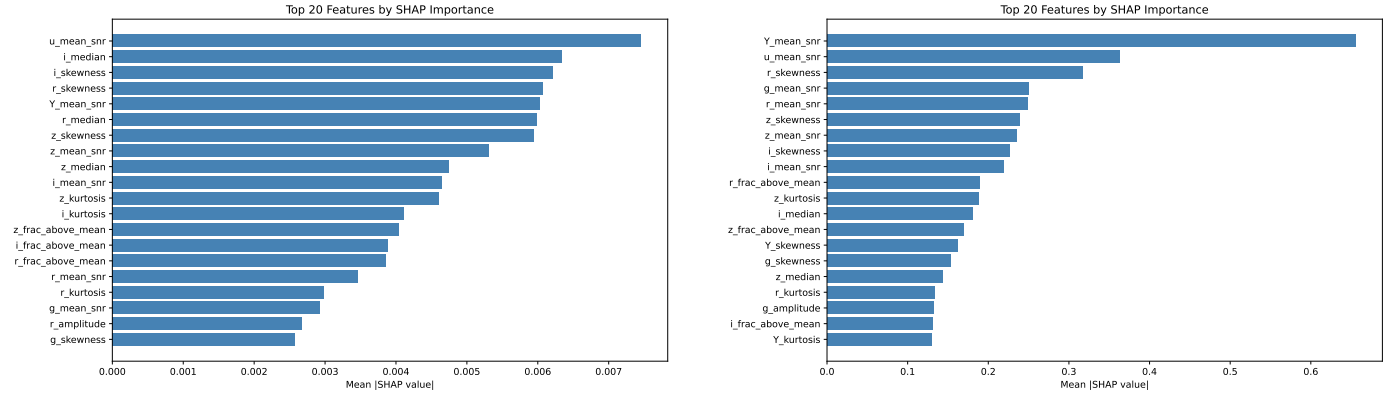


Fig. 2. Top 20 features by mean |SHAP| for Random Forest (*left*) and XGBoost (*right*).

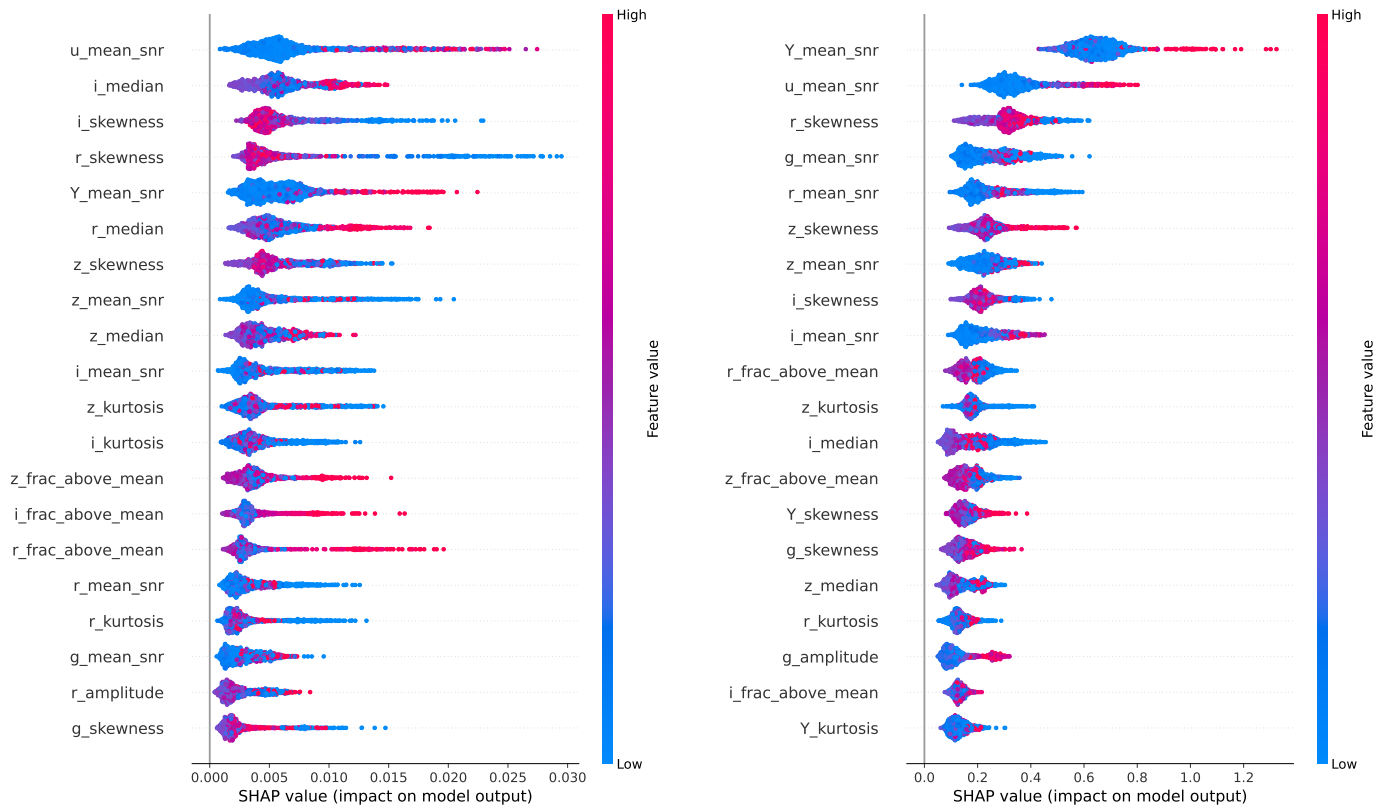


Fig. 3. SHAP beeswarm summary plots for Random Forest (*left*) and XGBoost (*right*), showing the distribution of feature attributions across test samples. Colour encodes feature value (red = high, blue = low).

practical corollary for users of the benchmark is that a plausibility score is interpretable only alongside its own permutation baseline at a declared threshold.

Conversely, the GAF-image models score the lowest plausibility (CNN2D = 0.149, ViT = 0.257). The GAF transformation encodes per-band temporal information into 2D spatial structure that is not separable into the same statistical moments that define the tabular features; our projection, with its hand-designed lookup table, can only approximate the mapping. The CNN2D in particular concentrates its attribution off the diagonal, which under our diagonality-based slope proxy reduces alignment with the expert mask. We caution against reading the low CNN2D plausibility as evidence that the GAF representation is physically uninformative – it more likely reflects a limitation of the cross-representation projection and points to a productive direction for

future work: a learned alignment between heterogeneous attribution spaces, in the spirit of the FIX framework (Jin et al. 2024). Intrinsically interpretable approaches such as uSAST (Mboupda et al. 2026), which extract semantically aligned subsequence features by design, could also serve as calibration baselines for plausibility in future extensions of this benchmark.

The physics-informed baselines illustrate how sensitive a fixed-mask IoU plausibility is to the construction of the expert mask itself. The 21 Avocado features that map onto our 60-dimensional space concentrate on amplitude-, colour-, and timescale-related slots; were the mask enlarged to contain all of them, the projected attribution would be a strict subset of the mask and the IoU would degenerate to a seed- and magnitude-independent constant $n_{above}/|m|$ – a structural saturation that conveys nothing about attribution quality. The 21-feature mask

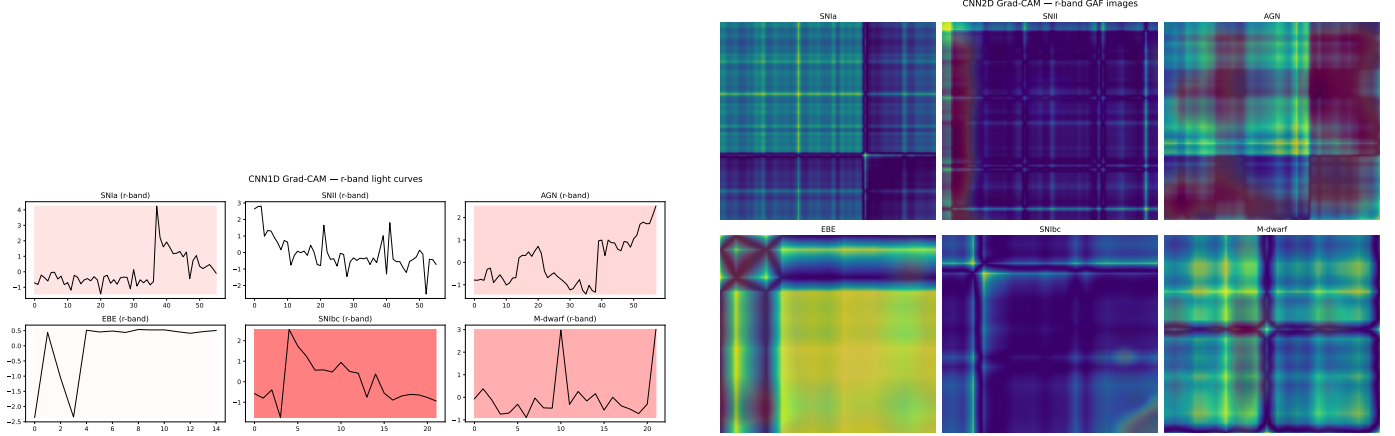


Fig. 4. Grad-CAM visualisations for CNN1D on *r*-band light curves (*left*) and CNN2D on *r*-band GAF images (*right*). Red shading indicates regions of high attribution.



Fig. 5. Insertion and deletion curves (Petsiuk et al. 2018) for Random Forest (*left*) and the Vision Transformer (*right*), averaged across the 50 test samples used to compute the AUC (seed 42).

avoids this degeneracy precisely because six of Avocado’s mapped targets (the per-band mean and median colour proxies) fall outside it, so its plausibility is a genuine, magnitude- and seed-dependent IoU (0.399 ± 0.014 ; Table 5). ParSNIP illustrates the complementary effect: its heuristic latent-to-tabular projection concentrates on the same colour-related slots, so excluding them depresses its plausibility to 0.215 ± 0.012 – a low absolute value that nonetheless sits at $1.6\times$ its own permutation baseline, on a par with the time-series models. Both cases show that the absolute plausibility of a model whose attributions are projected from a foreign feature space is as much a property of the mask–projection pair as of the model, and reinforce the case for learned-alignment plausibility approaches (FIX, Jin et al. 2024) and perturbation-based alternatives (ROAR and KAR, Hooker et al. 2019) that do not depend on a hand-designed projection or a fixed shared-feature mask.

6.5. Sanity checks discriminate trustworthy from untrustworthy attributions

The Adebayo sanity checks (Sect. 5.3 and Table 6) are arguably the single most consequential addition of this benchmark. Of the eight model–attribution combinations tested (the ensemble is excluded because its meta-feature space does not admit a meaningful randomisation analogue), the four tree-based combinations – XGBoost, Random Forest, Avocado, and ParSNIP, all using TreeSHAP on a gradient-boosted or random-forest classifier – have no meaningful model-randomisation analogue

and are therefore assessed on data randomisation alone. Three of them pass it (XGBoost $\rho_{\text{data}} = 0.348$, Avocado 0.265, and ParSNIP $|-0.300| = 0.300$, the latter via randomisation of its downstream LightGBM head only), while Random Forest fails ($\rho_{\text{data}} = 0.862$). Among the four deep-learning combinations, which admit both tests, CNN2D IG and ViT IG pass both cleanly, CNN1D IG passes model randomisation ($\rho_{\text{model}} = 0.000$) but fails data randomisation ($\rho_{\text{data}} = 0.995$), and LSTM IG fails both. In total, five of the eight combinations pass all of their applicable checks (XGBoost, Avocado, ParSNIP, CNN2D, and ViT), one passes one of its two checks (CNN1D), and two fail (Random Forest on its single applicable check, LSTM on both).

The LSTM failure is a structural property of gradient-based attribution on this architecture, not a matter of hyper-parameter tuning. As shown in Sect. 5.3, the signed versus absolute decomposition reveals that model randomisation scrambles the sign of the IG but preserves the magnitude profile – the hallmark of an attribution whose absolute value depends only on the input. The bidirectional pooled LSTM falls into exactly this regime because masked-mean pooling and the bidirectional sum produce a structurally smooth integrated gradient, causing IG to encode only the input geometry. The CNN1D failure is subtler: its attributions depend on having weights but not on which labels those weights were trained against, consistent with the model fitting a label-independent input geometry rather than a discriminative representation.

The lesson for the time-domain xAI community is concrete and actionable: any attribution method should be subjected to

the Adebayo tests before its outputs are used to draw scientific conclusions, because in our benchmark the failures are neither rare nor confined to ‘known bad’ methods. Moreover, when a model fails one of the tests, switching the IG baseline is not a viable workaround – the failure is invariant to monotone rescalings of the attribution (Table B.4). Practitioners should instead replace gradient-based methods with perturbation-based alternatives that do not share the $(\mathbf{x} - \mathbf{x}') \cdot \int \nabla F$ form, such as feature occlusion or KernelSHAP (but not GradientShap, which is itself an expectation of $(\mathbf{x} - \mathbf{x}') \cdot \nabla F$ over sampled baselines and would inherit the same degeneracy). We leave a systematic comparison of these alternatives on the LSTM and CNN1D for future work.

The Random Forest result is subtler: TreeSHAP attributions on a model trained with shuffled labels remain correlated with the trained-model attributions at $\rho = 0.862$, indicating that the marginal feature distribution dominates the SHAP signal. This does not contradict the fact that Random Forest attains the best perturbation faithfulness in our benchmark – faithfulness measures whether the top features drive the prediction, while data randomisation measures whether the top features depend on the labels. A model can achieve both high faithfulness and high data-randomisation correlation if its predictions are dominated by the same features the labels happen to correlate with. The combination, however, suggests that the Random Forest SHAP ranking should be interpreted as ‘which features the model leans on’ rather than ‘which features the label-conditional structure of the data points to’, and that complementary methods such as conditional permutation importance may be needed when the latter question is of scientific interest. We note that Kumaran et al. (2026) found TreeSHAP to be physically valid on tabular X-ray data with well-separated classes; our result suggests that this trustworthiness may be context-dependent and that sanity checks should be routine in any SHAP-based scientific analysis.

6.6. Limitations

Several limitations should be noted. First, we use only the PLAsTiCC training set (~7,800 objects), which limits deep learning performance; larger datasets may shift the accuracy balance. We do not apply class-weighting or focal loss to address the strong class imbalance in PLAsTiCC; this design choice ensures that all models face identical training conditions, isolating the effect of architecture on explanation quality rather than conflating it with loss-function differences. Class-balanced training may affect both accuracy and attribution structure, and is a natural direction for future work.

Second, our cross-representation plausibility relies on a hand-designed projection from per-timestep or per-pixel attributions to the tabular feature space (Sect. 4.2.4); a learned alignment between representations – in the spirit of representation-removal benchmarks such as ROAR and KAR (Hooker et al. 2019) or the FIX framework of Jin et al. (2024) – could yield more accurate plausibility estimates for the deep learning models. The IoU threshold is fixed at the median of $|a_j|$; a sensitivity analysis across percentiles (Table B.2) confirms that the ordering of the tree-based and time-series models is stable in the p25–p50 range but collapses for deep learning models at more aggressive thresholds. Moreover, the expert feature mask itself was designed by a single author based on the transient classification literature; independent validation by multiple domain experts would strengthen the plausibility metric, and we encourage future users of the benchmark to contribute alternative masks.

Third, faithfulness via feature perturbation assumes feature independence, which may not hold for correlated light curve

statistics; we mitigate this by reporting the zero-baseline insertion and deletion AUC of Petsiuk et al. (2018) alongside the perturbation correlation, but a perturbation strategy that respects the joint distribution of multi-band light curves remains an open methodological question. The faithfulness metrics use different atomic units across representations (60 features for tabular models, six bands for time series, 16×16 patches for images), which is inevitable given the representations but limits direct cross-family comparisons of absolute AUC values. The ensemble is excluded from insertion and deletion because its meta-features are probability outputs of base models, not physical inputs; a coarse faithfulness proxy via per-model meta-feature occlusion is left for future work.

Fourth, the CNN1D exhibits ~6× higher cross-seed variance than all other models ($\sigma_{\text{acc}} = 0.036$), and the LSTM and CNN1D sanity-check failures imply that IG is unsuitable for these architectures; non-gradient alternatives (feature occlusion, KernelSHAP) are a natural next step. Fifth, the benchmark uses fixed-length representations of the PLAsTiCC light curves – zero-padded or truncated series, PAA-reduced GAF images (Sect. 2.2) – rather than their native irregular cadence; extending to irregularly cadenced real survey data (ZTF, ELAsTiCC-2) and multimodal inputs would test generalisability.

Sixth, there is an intrinsic asymmetry in sample efficiency between the model families: tree-based classifiers operating on 60 hand-crafted features are in a comparatively data-rich regime at ~5 500 training samples, whereas deep-learning models consuming raw time series (6×256) or GAF images (6×224^2) face a much higher-dimensional input space relative to the same sample size. This imbalance is inherent to the dataset and to the choice of comparing heterogeneous representations on a common task; it favours tree-based accuracy but does not invalidate the xAI comparison, which is our primary object of study. Users should bear in mind that the classification gap between tree-based and neural models would likely narrow on larger datasets.

Seventh, the two physics-informed baselines are evaluated under conditions that make their reported figures conservative. Avocado uses a 10× rather than 100× redshift augmentation and, like ParSNIP, is restricted to the 7 848 labelled PLAsTiCC objects rather than the full 3.5 M-object test set used as an unlabelled augmentation source in the original work (Sect. 3.5); their absolute accuracies are therefore lower bounds on what the published pipelines achieve. Their plausibility scores rest on a partial, hand-designed feature mapping (only 21 of Avocado’s 41 features map into the mask, and ParSNIP’s latent dimensions are projected heuristically), and remain sensitive to the choice of projection and expert mask (Sect. 6.4); and ParSNIP’s sanity check randomises only its classification head, not its VAE encoder. These baselines should thus be read as competitive reference points for the xAI metrics, not as a re-benchmarking of the underlying classifiers.

7. Using the benchmark

7.1. Evaluating a new model

A central goal of *astro-xai-bench* is to let a practitioner evaluate their own classifier on the same five xAI axes. The framework is released as an installable Python package; any model exposing a scikit-learn-style `predict_proba` interface (and, optionally, `get_embeddings`) can be dropped into the pipeline without modifying the metric code. Tabular models are scored directly on their SHAP attributions; sequence- and image-based

models additionally supply a per-sample attribution map (e.g. Integrated Gradients), which the cross-representation projection of Sect. 4.2.4 maps onto the 60-dimensional tabular basis, against which the 21-feature expert mask is then applied, before the plausibility IoU is computed. The minimal working example in Fig. C.3 evaluates a user-supplied tabular model, returning faithfulness, complexity, plausibility, and the data-randomisation sanity correlation. For deep-learning models the same four calls apply after replacing the SHAP step with an integrated-gradient attribution and passing `model_type="timeseries"` or `"image"` to `compute_plausibility`, which selects the appropriate projection (Sect. 4.2.4).

We intend to host a public comparison of submitted models. To keep entries comparable, contributors supply (i) a configuration file specifying the data split and seeds and (ii) the per-sample attribution arrays; the metrics are then recomputed by the benchmark code rather than taken from the submission, so that the expert mask, binarisation thresholds, and evaluation sample sets are identical across all entries. The repository, its documentation and the submission instructions are given in the Data availability section.

7.2. Porting the benchmark to another survey

The five axes are properties of a model and its attributions rather than of PLAsTiCC: the metric definitions of Sect. 4, the five-seed averaging, and the randomisation protocol of Adebayo et al. (2018) carry over unchanged to any photometric transient dataset. What does not carry over is the domain anchor against which explanations are scored. We set out here the four components that must be re-instantiated for a new survey, both to make such an extension a matter of configuration rather than reimplementation, and to be explicit that the effort it requires is one of domain design rather than computation.

(i) The tabular feature basis. The 60 features of Sect. 2.2 are ten per-band statistics over six passbands. A survey with a different filter set – ZTF’s g, r pair, say – yields a basis of different dimension and different physical content; the statistics must be recomputed from that survey’s photometry rather than remapped from ours. (ii) The expert mask. Plausibility is the IoU between a binarised attribution and an astrophysically motivated subset of that basis (Sect. 4.2.3, Table 3); because the mask indexes into the feature basis, changing the basis invalidates it. Redesigning it – deciding which per-band statistics an expert would regard as discriminative for the new target classes – is a domain-knowledge task, and is the one step of the extension that cannot be automated. It is also the step that most affects the outcome: as Sect. 6.4 and Table B.2 show, fixed-mask IoU plausibility is sensitive both to the mask’s construction and to the binarisation threshold, which is why we report a sensitivity analysis rather than a single number and encourage contributed alternative masks.

(iii) The cross-representation projection. The mapping from per-timestep or per-pixel attributions onto the feature basis (Sect. 4.2.4) is hand-designed and inherits the survey’s band count and sequence length; the validation of Sect. 4.2.5 should be repeated for each new instantiation. (iv) The atomic units and the class taxonomy. Faithfulness and complexity are computed over representation-specific atomic units, and the target-class taxonomy fixes both the number of output units and the per-class support that drives the imbalance discussed in Sect. 6.6.

Two consequences follow. First, absolute metric values are comparable only within a single instantiation: a plausibility of

0.4 obtained under one feature basis and expert mask carries no information about a 0.4 obtained under another, so cross-survey statements should concern model rankings and which sanity checks are passed, never raw magnitudes. Second, because the metrics are recomputed by the benchmark code from submitted attributions rather than taken from reported values, a new survey defines a new leaderboard rather than extending the existing one. Applying this protocol to irregularly cadenced data – ZTF or ELAsTiCC-2, whose two- and six-band photometry would each require their own expert mask – is the natural next step, and we regard PLAsTiCC here as the reference instantiation of the framework rather than as its scope.

8. Conclusions

We have presented *astro-xai-bench*, a benchmark for quantitative evaluation of xAI in astronomical transient classification, evaluated across five random seeds for statistical robustness. By comparing seven model architectures plus two physics-informed baselines (Avocado, ParSNIP) across five xAI axes – perturbation faithfulness, insertion and deletion AUC, raw and band-aggregated complexity, cross-representation plausibility, and the Adebayo sanity checks – we find that:

1. The out-of-fold stacking ensemble dominates classification among the seven generic architectures, achieving both the highest accuracy (0.709 ± 0.006) and the highest macro-F1 (0.568 ± 0.015) in that group. The OOF construction (Sect. 3.4) not only removes target leakage but also outperforms the leak-prone validation-set fit (0.688 accuracy), because cross-validation gives the meta-learner $\sim 7\times$ more training samples;
2. The physics-informed baselines bracket the benchmark from opposite ends on the xAI axes. Avocado leads classification overall (0.801 ± 0.005 accuracy, 0.777 ± 0.007 macro-F1) and also tops the projected-plausibility ranking (0.399 ± 0.014 , and by a factor of 2.4 over the best time-series model once corrected for the chance level and ceiling implied by its attribution sparsity), confirming that GP-based redshift augmentation retains a decisive advantage on PLAsTiCC even under our conservative $10\times$, labelled-set-only configuration. ParSNIP, by contrast, passes the data-randomisation sanity check (signed $\rho_{\text{data}} = -0.300$, $|\rho| = 0.300$, a pass comparable to CNN2D and XGBoost) but exhibits low projected plausibility (0.215 ± 0.012) under the expert mask – a sensitivity of its heuristic latent-to-tabular projection rather than a property of the model. Classification strength, sanity-check behaviour, and projection-based plausibility are therefore largely orthogonal, which is the central methodological message of the benchmark;
3. Random Forest produces the most faithful explanations on both protocols (perturbation correlation 0.871 ± 0.018 , deletion AUC 0.170 ± 0.004), while XGBoost wins the insertion AUC (0.623 ± 0.031) and is the most balanced choice when both classification and interpretability matter. The CNN2D on GAF images is the most faithful deep learning model under the perturbation correlation; CNN1D is the weakest and most variable classifier, with cross-seed standard deviations roughly six times larger than the other architectures;
4. The raw attribution complexity is biased by input dimensionality and ranks the GAF-image models last; the band-aggregated complexity reverses the ordering and reveals that XGBoost actually distributes its SHAP importance most uniformly across the six photometric passbands.

Cross-representation comparisons of attribution entropy must use a fair common basis;

5. Our cross-representation plausibility metric (Sect. 4.2.4) reveals a counter-intuitive ordering: the time-series deep learning models (CNN1D 0.398 ± 0.005 , LSTM 0.381 ± 0.008) align better with the astrophysical expert mask than the tree-based models that operate directly on the same hand-crafted features (RF 0.229 ± 0.002 , XGBoost 0.224 ± 0.000). The GAF-image models sit at the opposite end, suggesting that attributions over GAF representations do not project cleanly onto the tabular feature basis;
6. The Adebayo sanity checks (Sect. 5.3) reveal that LSTM integrated-gradient attributions are dominated by the input rather than by the learned weights ($\rho_{\text{model}} \approx 0.995$, $\rho_{\text{data}} \approx 0.996$), that CNN1D IG attributions depend on having weights but not on which labels those weights were trained against ($\rho_{\text{model}} = 0$, $\rho_{\text{data}} = 0.995$), and that TreeSHAP on Random Forest is dominated by the marginal feature distribution rather than the supervised signal ($\rho_{\text{data}} = 0.862$). All three results are robust to baseline choice. CNN2D IG and ViT IG pass all applicable checks cleanly (both model- and data-randomisation), as does XGBoost SHAP on its single applicable check (data-randomisation; the model-randomisation analogue does not apply to tree-based attributions). We strongly recommend that future astronomical xAI studies routinely run sanity checks before drawing scientific conclusions from attribution maps, and that gradient-based attribution be replaced with perturbation-based alternatives on architectures that fail. The two physics-informed baselines both pass the data-randomisation check (Avocado $\rho_{\text{data}} = 0.265$ and ParSNIP signed $\rho_{\text{data}} = -0.300$, i.e. $|\rho| = 0.300$); we report only data randomisation for both because their classifiers are tree-based and TreeSHAP has no meaningful model-randomisation analogue;
7. All cross-seed standard deviations are small (≤ 0.04 for accuracy, ≤ 0.08 for the xAI metrics, the single exception being the ViT insertion AUC at 0.083) for all architectures except CNN1D, indicating that the relative ordering of models is robust to random seed variation.

Data availability

All code, trained models, and evaluation pipelines are publicly available, to support reproducible xAI benchmarking in time-domain astronomy. The repository, together with its documentation and the instructions for submitting a model to the public comparison of Sect. 7, is at <https://github.com/antopag/astro-xai-bench> and is released under the MIT licence; the version used for this paper is archived on Zenodo (DOI: 10.5281/zenodo.21901215). The benchmark is built on the PLAsTiCC training set (Kessler et al. 2019), which is itself public.

Acknowledgements. We thank the anonymous referee for two exceptionally constructive reports. Their guidance reshaped this paper well beyond matters of presentation: the request to restrict the expert feature mask to strictly domain-motivated features changed the plausibility results reported here, the call for a worked demonstration of how the five axes are used in practice produced Sect. 7, and the suggestion to consider further training sets prompted the extension protocol of Sect. 7.2. This work made use of the PLAsTiCC dataset (Kessler et al. 2019). AP acknowledges support from INAF and from ICSC – Centro Nazionale di Ricerca in HPC, Big Data e Quantum Computing, funded by the European Union – NextGenerationEU. AP is also supported by INFN, Sezione di Catania. This research made use of Python packages including PyTorch, scikit-learn, XGBoost, SHAP, captum, timm, and matplotlib.

References

- Abnar, S., & Zuidema, W. 2020, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4190
- Adebayo, J., Gilmer, J., Muelly, M., et al. 2018, in *Advances in Neural Information Processing Systems*, 31
- Allam, T., & McEwen, J. D. 2021, arXiv e-prints [arXiv:2105.06178]
- Bellm, E. C. et al. 2019, *PASP*, 131, 018002
- Boone, K. 2019, *AJ*, 158, 257
- Boone, K. 2021, *AJ*, 162, 275
- Carrasco-Davis, R., et al. 2021, *AJ*, 162, 231
- Hložek, R., et al. 2023, *ApJS*, 267, 25
- Hooker, S., Erhan, D., Kindermans, P.-J., & Kim, B. 2019, in *Advances in Neural Information Processing Systems*, 32 (Curran Associates)
- Ivezić, Ž., et al. 2019, *ApJ*, 873, 111
- Jain, S., & Wallace, B. C. 2019, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 3543
- Jamal, S., & Bloom, J. S. 2020, *ApJS*, 250, 30
- Jin, H., Havaladar, S., Kim, C., et al. 2024, *Journal of Data-centric Machine Learning Research*, 2, 17
- Ke, G., Meng, Q., Finley, T., et al. 2017, in *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 3146
- Kessler, R., et al. 2019, *PASP*, 131, 094501
- Kumaran, S., Mandal, S., Bhattacharyya, S., & Mishra, D. 2023, *MNRAS*, 520, 5065
- Kumaran, S., Mandal, S., & Bhattacharyya, S. 2026, arXiv e-prints [arXiv:2601.04650]
- Lundberg, S. M., & Lee, S.-I. 2017, in *Advances in Neural Information Processing Systems*, 30
- Lundberg, S. M., et al. 2020, *Nat. Mach. Intell.*, 2, 56
- Mbouopda, M. F., Ishida, E. E. O., Mephu Nguifo, E., & Gangler, E. 2026, *Big Data Res.*, 43, 100591
- Möller, A., & de Boissière, T. 2020, *MNRAS*, 491, 4277
- Muthukrishna, D., Narayan, G., Mandel, K. S., Biswas, R., & Hložek, R. 2019, *PASP*, 131, 118002
- Narayan, G., et al. 2018, *ApJS*, 236, 9
- Petsiuk, V., Das, A., & Saenko, K. 2018, in *British Machine Vision Conference*
- Qu, H., et al. 2021, *AJ*, 162, 67
- Richards, J. W., Starr, D. L., Butler, N. R., et al. 2011, *ApJ*, 733, 10
- Selvaraju, R. R. et al. 2017, in *IEEE International Conference on Computer Vision*, 618
- Sundararajan, M., Taly, A., & Yan, Q. 2017, in *Proceedings of the 34th International Conference on Machine Learning*, 3319
- Villar, V. A. et al. 2021, *ApJS*, 255, 24
- Wang, Z., & Oates, T. 2015, in *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, 3939
- Wiegrefe, S., & Pinter, Y. 2019, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 11
- Wightman, R. 2021, *PyTorch Image Models*, <https://github.com/huggingface/pytorch-image-models>

Appendix A: The Gramian angular field transformation

The GAF is less widely used in the time-domain astronomy community than the tabular or raw light-curve representations, so we illustrate the full encoding pipeline in Fig. A.1 on a representative single-band transient.

Appendix B: Supplementary tables

Tables B.1–B.5 provide supplementary robustness checks and ablation results referenced in the main text.

Table B.1. Deletion and insertion AUC under three attribution rankings.

Model	Del. AUC ↓			Ins. AUC ↑		
	attr	rand	oracle	attr	rand	oracle
RF	.172	.278	.159	.395	.271	.450
XGB	.324	.480	.244	.588	.453	.722
LSTM	.301	.434	.290	.488	.431	.524
CNN1D	.525	.637	.429	.694	.560	.728
CNN2D	.172	.172	–	.316	.162	–
ViT	.277	.463	–	.545	.424	–

Notes. The three rankings are the model’s own attribution (‘attr’), a random permutation (‘rand’), and an oracle ranking based on leave-one-out prediction drop (‘oracle’; omitted for image models due to cost). These AUC values were computed on a smaller sample than those of Table 5 and for a single seed, which is the source of the numerical differences between the two tables; only the attribution-to-random gaps within a row are intended to be read here. The gap between attribution and random rankings quantifies the effect size of the explanation; the oracle provides an upper bound. All values are seed-42, $N = 30$ (tabular and time-series) or $N = 15$ (image) test samples.

Table B.2. Plausibility IoU sensitivity to the attribution binarisation threshold, expressed as a percentile of the positive $|a_j|$ distribution (seed 42, $N = 50$ test samples).

Model	p25	p50	p75	p90
Random Forest	0.272	0.233	0.133	0.064
XGBoost	0.257	0.223	0.187	0.109
LSTM	0.364	0.375	0.156	0.000
CNN1D	0.399	0.397	0.161	0.001
CNN2D	0.328	0.150	0.022	0.000
ViT	0.398	0.255	0.110	0.001
Avocado	0.544	0.412	0.232	0.119
ParSNIP	0.336	0.210	0.132	0.057

Notes. The median (p50) is our default; p25 is more permissive, p75–p90 more stringent. The ordering of the tree-based and time-series models is stable at p25–p50 but collapses for deep learning models at p90, where nearly all attribution mass falls below the threshold. All values use the 21-feature expert mask (Sect. 4.2.4). Avocado’s row is an ordinary, magnitude-dependent IoU rather than a fixed $n/|\mathbf{m}|$ sequence: its mapped features would form a strict subset of any mask that also contained the per-band mean and median slots – which would force a seed-independent constant – but the 21-feature mask excludes those slots (Sect. 6.4).

Table B.3. Signed vs absolute-value Spearman correlations for the model-randomisation and data-randomisation sanity checks (seed 42).

Model	Model rand.		Data rand.	
	signed	.	signed	.
LSTM	0.148	0.997	−0.014	0.994
CNN1D	0.000	0.000	–	–
CNN2D	0.000	0.000	–	–
ViT	0.000	0.000	–	–

Notes. Values were computed on the seed-42 attribution sample and may differ from those of Table 6 in the third decimal; what these rows are meant to separate is the signed from the absolute variant, not those residual differences. Following Adebayo et al. (2018), we report both variants to distinguish whether the randomisation changes only the sign of attributions (signed ρ drops, absolute ρ stays high) or also their magnitude (both drop). The LSTM exhibits exactly the former pattern: randomisation scrambles the direction of IG but preserves its magnitude profile, confirming that the magnitude is input-dominated.

Table B.4. IG baseline ablation for LSTM and CNN1D (seed 42, $N = 15$ test samples).

Model	IG baseline	$ \text{attr} $	Complexity	Time (s)
LSTM	zero	0.0105	0.906	9.1
	mean	0.0090	0.890	9.3
	noise	0.0038	0.878	8.9
CNN1D	zero	0.0144	0.919	0.2
	mean	0.0147	0.920	0.2
	noise	0.0403	0.941	0.1

Notes. All configurations use 50 integration steps. The attribution ranking is stable across baselines, while the overall magnitude is not: $|\text{attr}|$ varies by up to a factor of 2.8. Since the sanity-check correlations are rank-based, and rank is invariant under the monotone rescaling that the baseline change induces, the failures are structural rather than artefacts of the baseline choice.

Table B.5. Validation of the cross-representation projection using XGBoost as an identity control (Sect. 4.2.5).

Model	Direct IoU	Via projection (identity)	$ \Delta $
XGBoost	0.224	0.224	$< 10^{-6}$

Notes. Because XGBoost lives natively in the 60-dimensional tabular space, the plausibility computed directly on its SHAP attribution and the plausibility obtained by routing the same attribution through the projection dispatcher with an identity map are identical to numerical precision. Values use the 21-feature mask, all test samples, cross-seed mean.

Appendix C: Supplementary figures

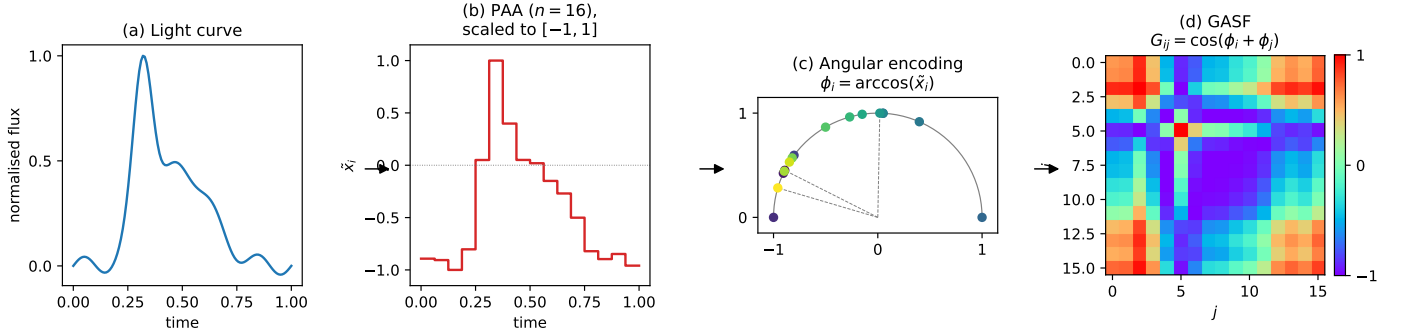


Fig. A.1. Gramian angular summation field (GASF) pipeline (Sect. 2.2). (a) A single-band light curve is (b) reduced by Piecewise Aggregate Approximation (PAA) to n segments and min–max scaled to $[-1, 1]$; (c) each scaled value is mapped to an angle $\phi_i = \arccos(\tilde{x}_i)$ on the unit semicircle (colour encodes time order); (d) the GASF image is the angular Gram matrix $G_{ij} = \cos(\phi_i + \phi_j)$, whose off-diagonal structure encodes temporal correlations between epochs. For visual clarity the series is reduced to $n = 16$ points here; the benchmark uses 224 points per band.

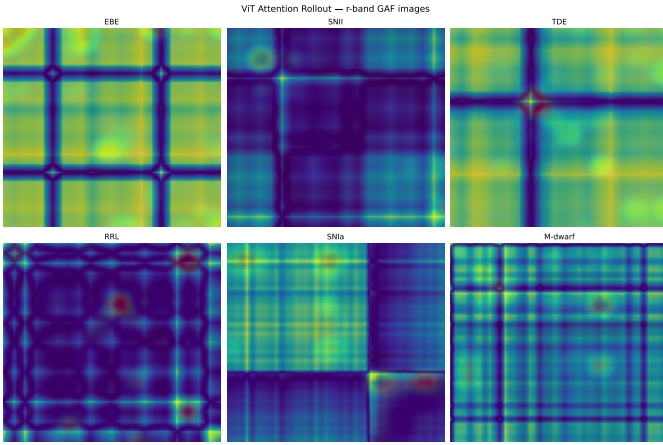


Fig. C.1. Attention rollout maps for the Vision Transformer on r -band GAF images, showing spatial attention patterns for six representative transient classes.

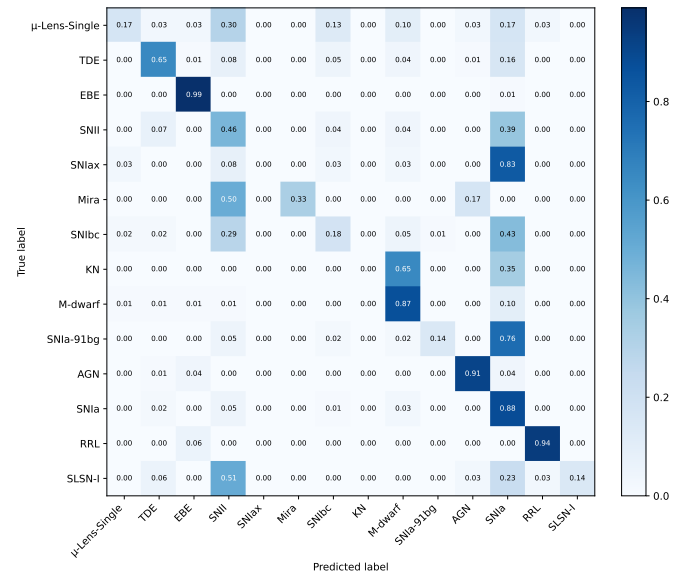


Fig. C.2. Confusion matrix for the stacking ensemble fitted naively on validation-set predictions of the base models (seed 42, 0.688 accuracy on the test set). The out-of-fold construction adopted in the paper (Sect. 3.4) improves this to 0.709 ± 0.006 (Table 4) without qualitatively changing the confusion structure.

```

from astro_xai_bench.xai import (compute_shap_values, compute_plausibility,
                                faithfulness_correlation, explanation_complexity)
from astro_xai_bench.xai.shap_utils import normalize_shap_values
from astro_xai_bench.xai.metrics import (build_expert_mask_tabular,
                                         get_feature_names)
from astro_xai_bench.xai.sanity_checks import data_randomization_test
import numpy as np

model.fit(X_train, y_train) # any classifier with predict_proba
shap = compute_shap_values(model, X_test, method="tree")
attr = normalize_shap_values(shap) # mean |SHAP| over classes -> (N, 60)
mask = build_expert_mask_tabular(get_feature_names()) # 21-feature mask
faith = np.mean([faithfulness_correlation(model, X_test[i], attr[i])
                 for i in range(len(X_test))])
compl = np.mean([explanation_complexity(attr[i]) for i in range(len(X_test))])
plaus = compute_plausibility(shap, mask, model_type="tabular")
rho = data_randomization_test(model, X_train, y_train, X_test)
print(faith, compl, plaus["plausibility_mean"], rho)

```

Fig. C.3. Minimal working example: evaluating a user-supplied tabular classifier on four of the benchmark's five xAI axes (faithfulness, complexity, plausibility, and the data-randomisation correlation ρ). The model only needs a scikit-learn-style `predict_proba`.