



An Unsupervised Machine Learning Approach to Identify Spectral Energy Distribution Outliers: Application to the S-PLUS DR4 Data

F. Quispe-Huaynasi¹ , F. Roig¹ , N. Holanda¹ , V. Loaiza-Tacuri^{1,2} , Romualdo Eleutério¹ , C. B. Pereira¹, S. Daflon¹ , V. M. Placco³ , R. Lopes de Oliveira² , F. Sestito⁴ , P. K. Humire⁵ , M. Borges Fernandes¹ , A. Kanaan⁶ , C. Mendes de Oliveira⁵ , T. Ribeiro⁷ , and W. Schoenell⁸

¹ Observatório Nacional, Rio de Janeiro, RJ 20921-400, Brazil

² Departamento de Física, Universidade Federal de Sergipe, São Cristóvão, SE 49100-000, Brazil

³ NSF NOIRLab, 950 N. Cherry Ave., Tucson, AZ 85719, USA

⁴ Centre for Astrophysics Research, University of Hertfordshire, Hatfield, AL10 9AB, UK

⁵ Instituto de Astronomia, Geofísica e Ciências Atmosféricas, Universidade de São Paulo, São Paulo, SP 05508-900, Brazil

⁶ Departamento de Física, Universidade Federal de Santa Catarina, Florianópolis, SC 88040-900, Brazil

⁷ Rubin Observatory Project Office, Tucson, AZ 85719, USA

⁸ GMTO Corporation, Pasadena, CA 91107, USA

Received 2025 January 28; revised 2025 April 11; accepted 2025 April 17; published 2025 May 23

Abstract

Identification of specific stellar populations using photometry for spectroscopic follow-up is a first step to confirm and better understand their nature. In this context, we present an unsupervised machine learning approach to identify candidates for spectroscopic follow-up using data from the Southern Photometric Local Universe Survey (S-PLUS). First, using an anomaly detection technique based on an autoencoder model, we select a large sample of objects ($\sim 19,000$) whose Spectral Energy Distribution is not well reconstructed by the model after training it on a well-behaved star sample. Then, we apply the t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm to the 66 color measurements from S-PLUS, complemented by information from the SIMBAD database, to identify stellar populations. Our analysis reveals 69 carbon-rich star candidates that, based on their spatial and kinematic characteristics, may belong to the CH or carbon-enhanced metal-poor categories. Among these chemically peculiar candidates, we identify four as likely carbon dwarf stars. We show that it is feasible to identify three primary white-dwarf (WD) populations: WDs with hydrogen-dominated atmospheres, WDs with neutral helium-dominated atmospheres, and the WDs main sequence binaries (WD + MS). Furthermore, by using eROSITA X-ray data, we also highlight the identification of candidates for very active low-mass stars. Finally, we identified a large number of binary systems using the autoencoder model, but did not observe a clear association between the overdensities in the t-SNE map and their orbital properties.

Unified Astronomy Thesaurus concepts: Chemically peculiar stars (226); Carbon stars (199); White dwarf stars (1799); Low mass stars (2050); Binary stars (154)

Materials only available in the online version of record: machine-readable table

1. Introduction

Over the past two decades, spectroscopic surveys at low, medium, and high resolution, have contributed to the characterization of specific stellar populations and the discovery of peculiar objects. Examples are the Large Sky Area Multi-Object Fiber Spectroscopic Telescope survey (LAMOST; X.-Q. Cui et al. 2012), the RADial Velocity Experiment (M. Steinmetz et al. 2006), the Apache Point Observatory Galactic Evolution Experiment (APOGEE; S. R. Majewski et al. 2017), the GALactic Archaeology with HERMES survey (G. M. De Silva et al. 2015), among others. More recently, studies using Gaia DR3 low-resolution spectra ($R \approx 50$) from the BP and RP photometers (Gaia BP/RP), along with machine learning techniques, have reported a large sample of candidate stars, including carbon-enhanced metal-poor stars, metal-poor stars, and white dwarfs (M. Lucey et al. 2023; M. L. Kao et al. 2024; Y. Yao et al. 2024). However, both currently and in the past, the identification of distinct

stellar populations and candidates to peculiar objects for later spectroscopic follow-up was primarily facilitated by multiband photometric surveys such as the Sloan Digital Sky Survey (SDSS; D. G. York et al. 2000)⁹ and the Panoramic Survey Telescope and Rapid Response System (K. C. Chambers et al. 2016), among others. In addition to enabling the identification and classification of stellar populations using color–magnitude or color–color diagrams (K. Finlator et al. 2000; P. Szkody et al. 2002; A. Helmi et al. 2003; N. Mauron et al. 2004), standard photometric systems can provide information on stellar parameters (D. D. Lenz et al. 1998; E. Masana et al. 2006). Furthermore, combined with infrared and/or ultraviolet data from photometric surveys like the Two-Million All Sky Survey (M. F. Skrutskie et al. 2006), WISE (E. L. Wright et al. 2010), and GALEX (D. C. Martin et al. 2005), the precision of this information can be improved (S. G. Sichevskij et al. 2014; C. Allende Prieto 2016).

On the other hand, photometric surveys like Pristine (E. Starkenburg et al. 2017), SkyMapper (G. S. Da Costa et al. 2019), the Javalambre Photometric Local Universe Survey (J-PLUS; A. J. Cenarro et al. 2019), the Southern



Original content from this work may be used under the terms of the [Creative Commons Attribution 4.0 licence](https://creativecommons.org/licenses/by/4.0/). Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

⁹ Also with low-resolution spectroscopy.

Photometric Local Universe Survey (S-PLUS; C. Mendes de Oliveira et al. 2019),¹⁰ and the Javalambre Physics of the Accelerating Universe Astrophysical Survey (N. Benítez et al. 2014), which combine broadband and narrowband filters, have shown greater potential for the identification of stellar populations and the determination of stellar parameters with very good accuracy. For instance, the *H* filter (J0660) and the Ca II triplet (J0861) in the S-PLUS, J-PLUS, and JPAS surveys can capture features that indicate stellar activity and chemical composition. Moreover, the J0660 filter facilitates the identification of H-emitting low-mass stars, which are often undetectable with broadband filters alone.

Recently, using machine learning techniques and photometric data from these surveys, it has become possible not only to determine stellar parameters but also to obtain chemical information for millions of stars (D. D. Whitten et al. 2021; Y. Huang et al. 2024; C. E. Ferreira Lopes et al. 2025). Leveraging the potential of combining broadband and narrowband filters, this study aims to identify outlier spectral energy distributions (SEDs) derived from the S-PLUS photometric data, using an anomaly detection technique with an autoencoder architecture. Subsequently, by applying a t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm to the colors derived from the 12 magnitudes provided by S-PLUS, alongside data from the SIMBAD catalog, we identify candidate M-type stars, carbon-rich stars, binary stars, white dwarfs, and other stellar populations within the anomalous sample.

This article is structured as follows. Section 2 provides a brief overview of the S-PLUS data used in this study. Section 3 outlines the process of selecting a large sample of stars that deviate from the expected behavior of well-behaved stars, employing anomaly detection technique with autoencoder. Section 4 details the use of the t-SNE dimensionality reduction to identify stellar subpopulations within the sample selected by the autoencoder. The characteristics of some of the identified populations are discussed in Section 5. Finally, Section 6 presents the discussion and conclusions.

2. S-PLUS Data

The Southern Photometric Local Universe Survey is a photometric survey that observes the southern hemisphere with a dedicated 0.8 m robotic telescope hosted at the Cerro Tololo Inter-American Observatory, Chile. The telescope is equipped with a camera (T80Cam-S) of $9.2 \text{ k} \times 9.2 \text{ k}$ pixels, that provides a 2 deg^2 field of view with a pixel scale of $0.55 \text{ arcsec pix}^{-1}$. The observational strategy, image reduction, and main scientific goals of S-PLUS are described in C. Mendes de Oliveira et al. (2019). S-PLUS combines seven narrowband filters (J0378, J0395, J0410, J0430, J0515, J0660, and J0861) and five broadband filters (*u*, *g*, *r*, *i*, and *z*) in the optical range ($\sim 3500\text{--}10000 \text{ \AA}$). The *g*-, *r*-, *i*-, and *z*-band filters are similar to the corresponding SDSS filters, while the *u*-band filter is the Javalambre *u* filter, which is an improved version of the SDSS *u* filter in terms of transmission. The characteristics and information about the filters are shown in Table 1 (C. Mendes de Oliveira et al. 2019).

The S-PLUS Data Release Four includes 1629 fields, covering about 3000 deg^2 of the southern sky, all reduced and calibrated in all the survey bands. Since this study

Table 1
S-PLUS Photometric Bands

Filter	Central Wavelength (\AA)	$\Delta\lambda$ (\AA)	Spectral Features
uJAVA	3563	352	Balmer-discontinuity region
J0378	3770	151	[O II]
J0395	3940	103	Ca H + K
J0410	4094	201	H_δ
J0430	4292	201	<i>G</i> band
gSDSS	4751	1545	SDSS-like <i>g</i>
J0515	5133	207	Mg <i>b</i> triplet
rSDSS	6258	1465	SDSS-like <i>r</i>
J0660	6614	147	H_α
iSDSS	7690	1506	SDSS-like <i>i</i>
J0861	8611	408	Ca triplet
zSDSS	8831	1182	SDSS-like <i>z</i>

exclusively uses data from the S-PLUS DR4 (F. R. Herpich et al. 2024) and from the Gaia Data Release 3 (DR3; Gaia Collaboration et al. 2023a); hereafter, we will refer to them simply as S-PLUS and Gaia, respectively.

3. Anomaly Detection and Peculiar Candidates Selection

We used an autoencoder architecture (D. E. Rumelhart et al. 1986; G. E. Hinton & R. R. Salakhutdinov 2006) to identify stars that exhibit unique SED within the S-PLUS data set. There are different techniques for performing this task; however, here we use the simplest form of autoencoder (the so-called vanilla autoencoder) because it is widely used in the literature, it is easy to implement, and it is more intuitive. Classified as an unsupervised neural network, this architecture operates by first compressing the input data into a lower-dimensional representation¹¹ (the encoder), followed by the approximate reconstruction of the original data based on the lower representation (the decoder). The goal of training an autoencoder is to minimize the reconstruction error or loss function, which is the measure of the difference between the original input and its reconstructed output. The encoder and decoder both consist of fully connected traditional neural networks. Due to their characteristics, autoencoder architectures and extensions have been widely used for dimensionality reduction tasks (G. E. Hinton & R. R. Salakhutdinov 2006; T. Yang & X. Li 2015; S. K. N. Portillo et al. 2020) and as generative models (A. Oussidi & A. Elhassouny 2018; A. Nazábal et al. 2020). Recently, a novel architecture called the scatter variational autoencoder was developed, which learns to generate a Gaia BP/RP spectra and estimate intrinsic scatter (A. Laroche & J. S. Speagle 2025). While autoencoders have multiple applications across various domains (D. Bank et al. 2023; K. Berahmand et al. 2024; I. D. Mienye & T. G. Swart 2025), our focus in this study is on their ability for the detection of anomalies in the SED. In this context, a similar approach using autoencoder has been applied to SDSS spectra by K. Sharma et al. (2019), to identify potentially anomalous objects.

In general, the goal of anomaly detection is to find data instances that do not follow the expected pattern, or that deviate significantly from the majority of the data instances,

¹⁰ <https://www.splus.iag.usp.br/>

¹¹ This low-dimensional representation is referred to as code or latent space representation in the literature.

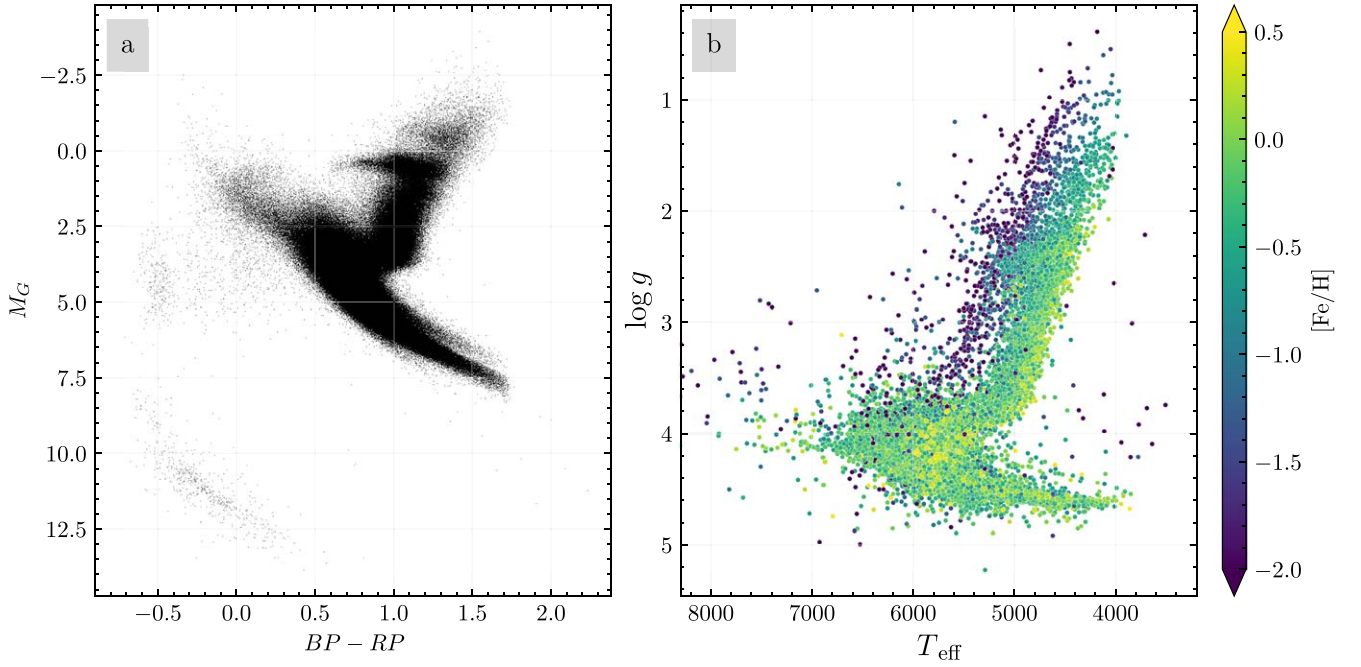


Figure 1. (a) Color–magnitude diagram for the training sample (~ 1.5 million stars). (b) Kiel diagram for objects in common between the training sample and SIMBAD, colored by metallicity ($\sim 20,000$ stars). The stellar parameters are those given in SIMBAD, and come from a diversity of independent works.

representing a small set of the total sample (G. Pang et al. 2021). In particular, our aim here is apply an autoencoder architecture to identify stars with atypical SEDs. The idea is to train the autoencoder with a sample of normal SEDs (majority class) or nonanomalous SEDs to learn to recognize the typical patterns of this normal data sample. After the training, it is expected that the reconstruction error values (anomaly score) will be lower for the normal data compared to the values for the anomalous data.

3.1. Data Training

The training sample was obtained of an S-PLUS Value Added Catalog, where the sources were classified as stars (CLASS = 1) according to the QSO/star/galaxy classification described in L. Nakazono et al. (2021). Since the goal is to select well-behaved sources, with reliable photometric and astrometric parameters, several quality cuts have been applied. Considering the photometry from S-PLUS, only objects with magnitudes lower than 19 and magnitude errors below 0.15 in all the 12 filters have been selected. Furthermore, we performed a crossmatch between S-PLUS and Gaia and applied the following quality cuts: (i) `phot_bp_rp_excess_factor` < 1.3 (for normal stars, which are neither close binaries, nor situated in crowded regions, this factor should be close to 1, D. W. Evans et al. 2018); (ii) sources with parallax greater than zero ($\varpi > 0$); (iii) `|astrometric_gof_al|` < 3 and `ruwe` < 1.2 (where `ruwe` is the Re-normalized Unit Weight Error, to ensure a good astrometric solution and avoid binary systems); (iv) `number_visibility_periods` > 15 (as recommended by Gaia, sources with visibility periods > 9 should be preferable); (v) `in_qso_candidates` = 0 and `in_galaxy_candidates` = 0 (to remove probably quasar and galaxy candidates, respectively); (vi) `phot_variable_flag` \neq VARIABLE (which excludes sources classified as variable); (vii) `non_single_star` = 0 (that filters out sources identified as

nonsingle stars, like astrometric binaries, spectroscopic binaries, or eclipsing binaries); and (viii) `duplicated_source` = 0 (to eliminate duplicated sources).

In addition to the above criteria, we consider sources with `fidelity_v2` > 0.5 and `norm_dg` < -3 flags,¹² as described in J. Rybizki et al. (2022) catalog. These cuts are used to select sources with reliable astrometric solutions and to exclude sources with potential color contamination from nearby objects, respectively.

After applying this sequence of quality cuts, the initial sample of approximately 3.5 million sources in S-PLUS is reduced to about 1.5 million objects, which will constitute our training data set. To characterize the training sample, we crossmatched it with the SIMBAD database, identifying $\sim 33,000$ objects, of which $\sim 20,000$ have information on stellar parameters: effective temperature (T_{eff}), surface gravity ($\log g$), and metallicity ($[\text{Fe}/\text{H}]$). Figure 1 presents the color–magnitude diagram of the entire training data set, and the Kiel diagram ($\log g$ versus T_{eff}) of the subset with known parameters. This indicates that the dominant population of stars in the training sample has T_{eff} ranging from 4000 to 7000 K, $\log g$ between 1.2 and 4.8 dex, and $[\text{Fe}/\text{H}]$ spanning from -2 to +0.5 dex.

3.1.1. Autoencoder Architecture

Once we have obtained our training sample, the next step is to define the autoencoder architecture that will be trained to select objects that do not follow the expected pattern for most S-PLUS stars. The autoencoder architecture used in this work is shown in Figure 2 and was implemented through `keras` (F. Chollet et al. 2015). Hyperparameters, including the number of hidden layers

¹² These flags were obtained by training a neural network, and were originally included in the external table `gedr3spur.main`, hosted at the German Astrophysical Virtual Observatory, which is currently available in the table `external.gaiaedr3_spurious` at <https://gea.esac.esa.int/archive/>.

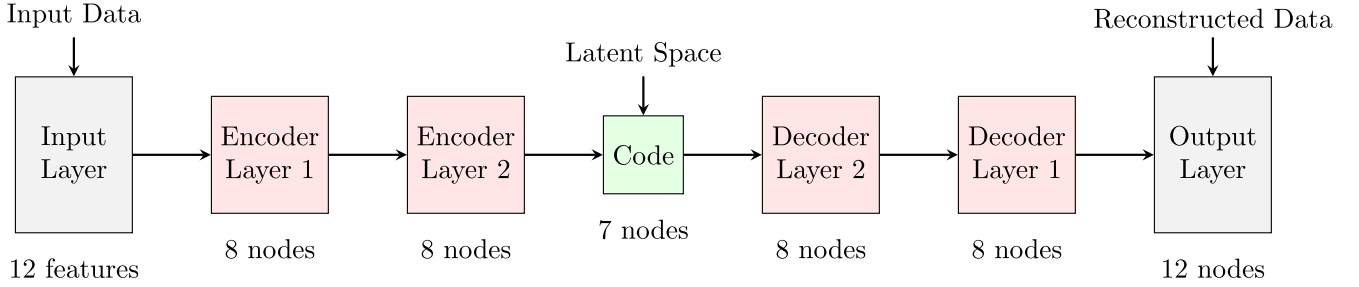


Figure 2. Schematic representation of the autoencoder architecture after hyperparameter tuning.

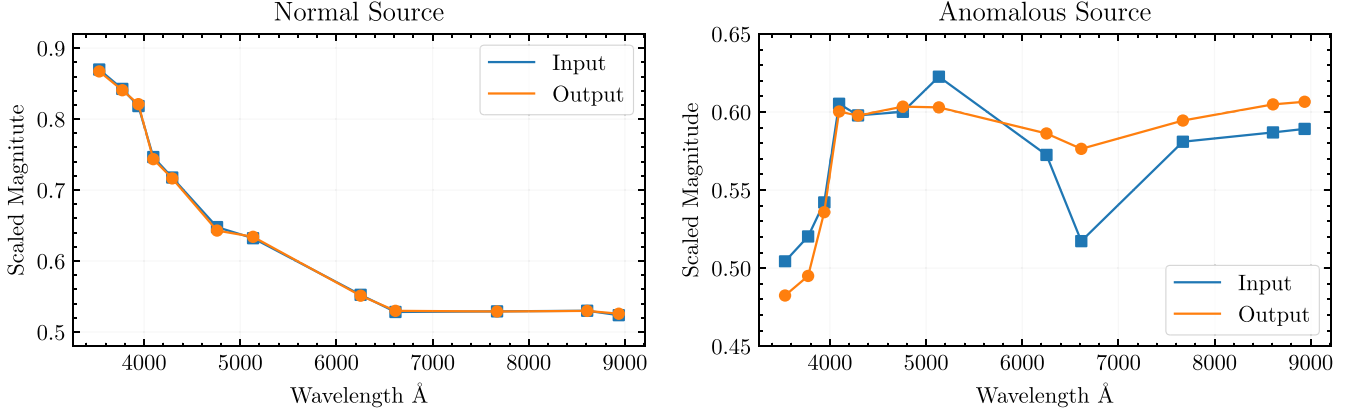


Figure 3. Comparisons between input spectral energy distribution (SED), in blue, and the reconstructed SED, in orange, for a normal source (left panel) and for an anomalous source (right panel).

for both the encoder and decoder, neuron per layer, and the dimension of the latent space, were determined with automatic hyperparameters tuning using the `RandomSearch` algorithm implemented in `KerasTuner` (T. O’Malley et al. 2019). The final architecture of the model, after the hyperparameters tuning, consists of two hidden layers with eight nodes each for both the encoder and the decoder layers, and a latent space representation of dimension 7. During the training process, we use the Adam optimizer (D. P. Kingma & J. Ba 2014), with the learning rate defined by default in Keras ($\eta=0.001$), apply the ReLU activation function, and employ the mean squared error as the objective function.

3.2. Autoencoder Applied to the S-PLUS Data

Giving as input for the autoencoder the 12 magnitudes from the S-PLUS filters, previously scaled between 0 and 1 using the `MinMaxScaler()` function provided by `Scikit-learn` (F. Pedregosa et al. 2011), we detected approximately 19,000 stars (anomalous sources) that surpassed three standard deviations from the mean global reconstruction error found during the training phase. The left panel in Figure 3, illustrates the input and output in blue and orange, respectively, for a source classified as normal (Gaia DR3 5659571422048804224). In contrast, the right panel displays the input and output for a source (Gaia DR3 232931789599827840) identified as an anomalous star. It is evident that the autoencoder fails to reconstruct the input SED for the anomalous source. It is worth noting that the photometric errors of the input data are not taken into account in the comparison between the input and the output data. This is unnecessary because the photometric errors are, in general, much smaller than the threshold applied to the

reconstruction error to classified the anomalous sources (see Section 6).

4. t-SNE

To segregate stellar populations within the sample of objects identified as anomalous by the autoencoder, we applied the t-distributed Stochastic Neighbor Embedding algorithm (t-SNE; L. van der Maaten & G. Hinton 2008), which is an extension of the Stochastic Neighbor Embedding algorithm (G. E. Hinton & S. Roweis 2002). t-SNE is a nonparametric unsupervised machine learning technique for dimensional reduction used for visualization, that has been applied in different areas of science, including astronomy (G. Traven et al. 2017; F. Anders et al. 2018; K. Čotar et al. 2019, among others). The main idea of this non-linear algorithm is to place objects described by a high-dimensional vector (features) into a low-dimensional space (map), preserving the local neighbor property defined in the high-dimensional space through a probabilistic approach. In this case, the similarity between points in the high-dimensional space is described by a joint probability p_{ij} given by

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}, \quad (1)$$

where n is the number of data points and

$$p_{j|i} = \frac{\exp(-||x_i - x_j||^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-||x_i - x_k||^2 / 2\sigma_i^2)} \quad (2)$$

is interpreted as the probability that the points x_j are neighbors of the points x_i based on the probability density of a Gaussian distribution with variance σ_i , centered at x_i . The value of σ_i for

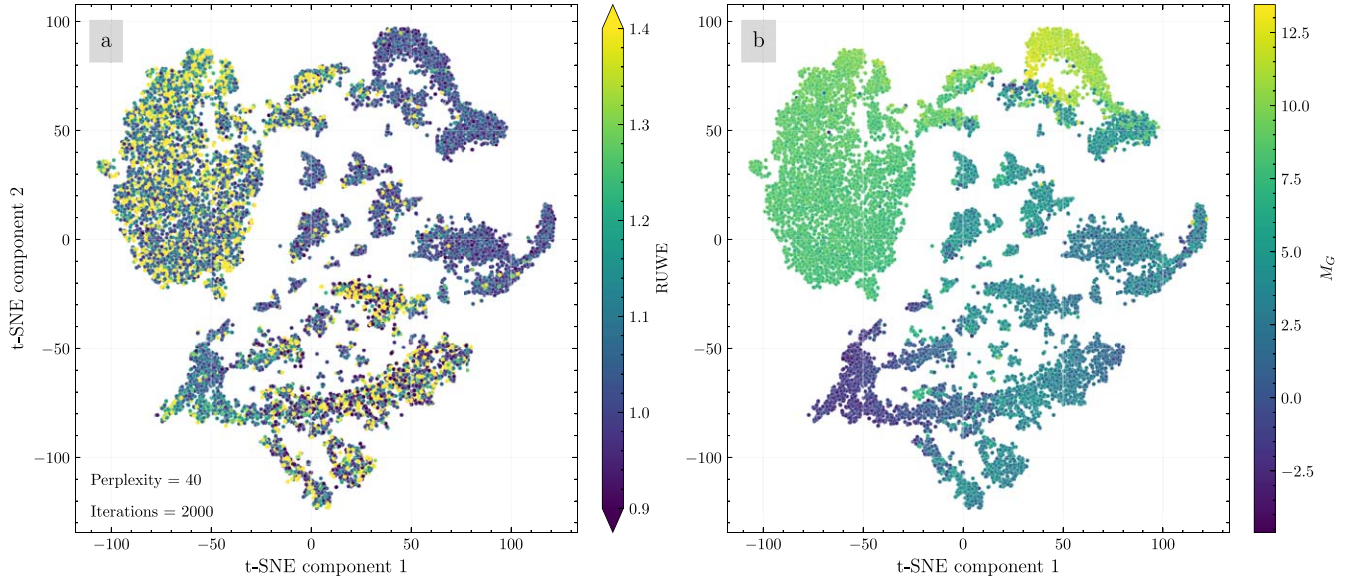


Figure 4. t-SNE 2D projections. (a) Colored by Gaia *ruwe* parameter. (b) Colored by Gaia absolute magnitude. The maps contain all the sources identified using the autoencoder. Except for the perplexity and the number of iterations, which were set to 40 and 2000, respectively, all the other hyperparameters were kept at their default values in the *Scikit-learn* implementation of t-SNE. To compute the absolute magnitudes, we consider the distances provided by C. A. L. Bailer-Jones et al. (2021).

each x_i is determined internally by t-SNE through a binary search algorithm based on the user-defined perplexity parameter (μ) given by:

$$\mu = 2^{-\sum_j p_{ji} \log_2 p_{ji}}, \quad (3)$$

which can be interpreted as the effective neighbors around x_i . This procedure allows the algorithm to set smaller σ_i values for dense regions and larger σ_i values for more sparse regions.

Once p_{ij} is defined, the next step is to define the joint probability q_{ij} in the low-dimensional space using Student's t-distribution with one degree of freedom:

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq i} (1 + \|y_k - y_j\|^2)^{-1}}. \quad (4)$$

Considering that the initial positions of the points in the low-dimensional space are placed randomly or using Principal Component Analysis, t-SNE minimizes the Kullback-Leibler divergences (cost function) between p_{ij} and q_{ij} , through the gradient-descent method, in order to preserve p_{ij} .

4.1. t-SNE Applied to S-PLUS Colors

We apply the *Scikit-learn* implementation of the t-SNE in PYTHON¹³ to a set of color indices, each one obtained as the difference between two magnitudes following the convention that the first magnitude always corresponds to the shorter-wavelength bandpass, and the second magnitude always corresponds to the longer-wavelength bandpass. By combining the 12 magnitudes provided by S-PLUS, we obtained a set of 66 color indices. These 66 indices are further normalized using the *StandardScaler()* function, also from *Scikit-learn*, to have zero mean and unit variance, and these constitute the input for t-SNE. This analysis is performed on the sample of stars identified by the autoencoder as anomalous objects. Except

for the perplexity and the number of iteration hyperparameters, which were set to 40 and 2000, respectively, all the other hyperparameters were left at their default values in the implementation. Initially, we adopted a value of 30 for the perplexity, as recommended in L. van der Maaten & G. Hinton (2008), and kept the other hyperparameters at their default settings. We then tested values of 25, 35, 40, and 45. Although there are some variations in the clumps' distribution and densities on the resulting 2D map, the changes are subtle for the tested values. As for the number of iterations, we tested 1000 (the default value), 2000, and 5000. Again, these values did not have a significant effect on the resulting map. This means that the results do not change significantly over a wide range of these two parameters. Choosing a perplexity value of 40, in particular, was mostly motivated because it yielded the highest number of carbon stars candidates. Figure 4 shows the t-SNE map, colored by the *ruwe* parameter and the absolute magnitude M_G . Overdensities appear and some regions are dominated by *ruwe* > 1.4, suggesting the potential presence of binary systems in the sample. By running the t-SNE several times, we have verified that, despite some subtle variations in the map, the overall structure and densities remain mostly unchanged for the values adopted in this work.

To determine whether the clumps in the map correspond to previously classified objects, we crossmatched our data with the SIMBAD database, identifying approximately 8000 cataloged sources. To identify the clumps, we use the primary classification in SIMBAD based on the *main_type* label. To ensure a reliable classification and focus on specific populations, we excluded sources labeled as candidate, as well as those generically classified as “Star” (*main_type* ≠ ‘Star’). This refinement resulted in a final sample of 2644 objects with specific classifications.

Figure 5 shows the t-SNE maps based on the specific populations identified in SIMBAD, with the labels reflecting their classifications and an asterisk (*) denoting the source as a star.¹⁴

¹³ <https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html>

¹⁴ Information about the classification in SIMBAD can be found at <http://simbad.u-strasbg.fr/guide/otypes.htx>.

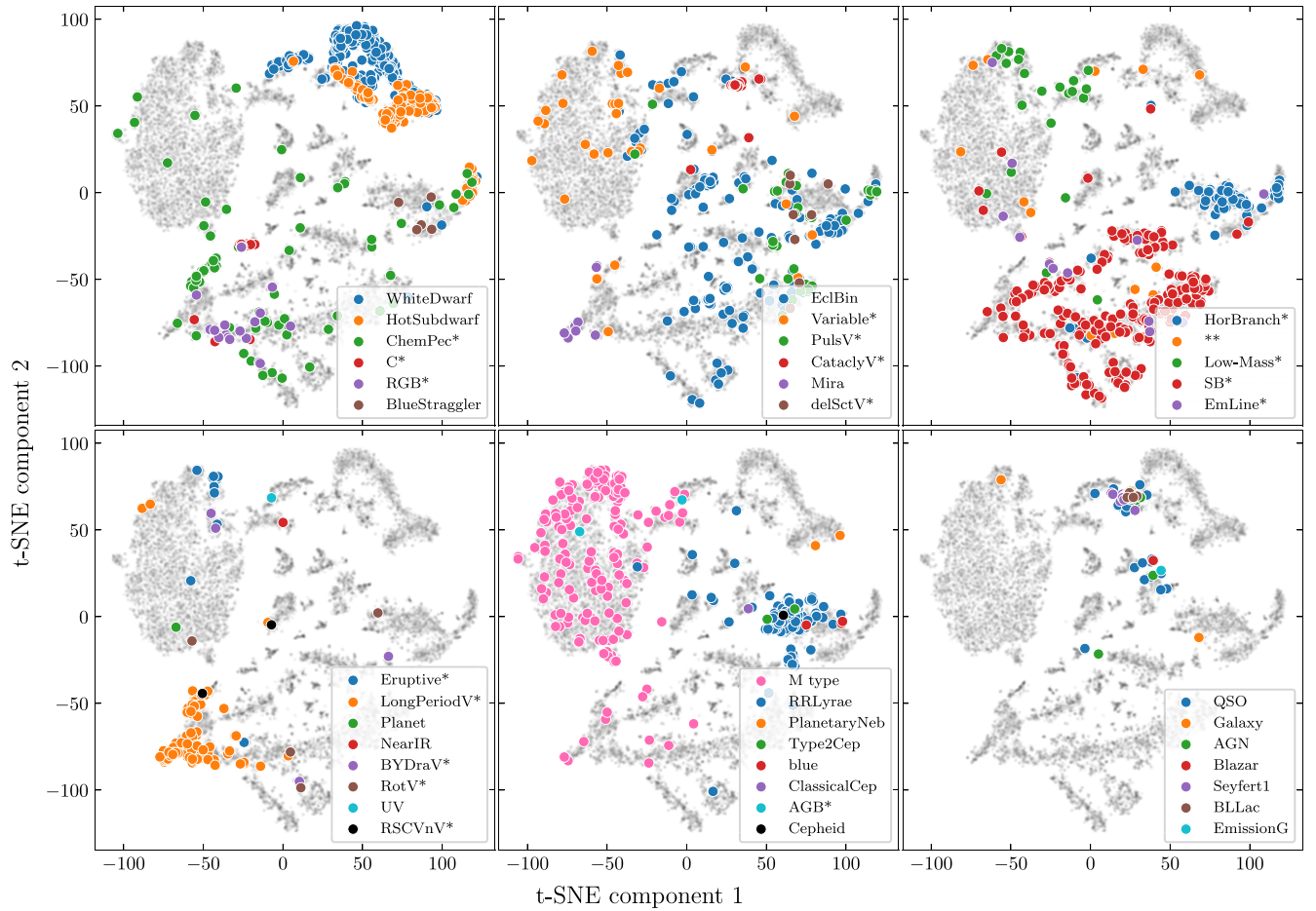


Figure 5. t-SNE 2D maps: in all panels, the background in gray represents sources identified by the autoencoder. The colored points indicate sources that overlap with sources classified in the SIMBAD database. The ChemPec* label encompasses a variety of chemically peculiar objects, including CEMP, alpha2 CVn variables, R CrB variables, CH stars, barium stars, and dwarf carbon stars.

The sources detected by the autoencoder are shown in gray, while the colored points represent the objects with information in SIMBAD. Some populations that are clearly grouped in these maps will be described in the next section. As a reference, Figure 6 presents the color–magnitude diagrams based on Gaia photometry for the sources shown in Figure 5 that have positive parallax and photogeometric distances provided by C. A. L. Bailer-Jones et al. (2021). The x -axis in this figure is the $G_{BP} - G_{RP}$ color, and the y -axis is the absolute G magnitude (M_G). The latter has been determined using the apparent G magnitude of Gaia, corrected by extinction using the SFD2D dust map (E. F. Schlafly & D. P. Finkbeiner 2011), via the `dustmapsPYTHON` package (G. Green 2018), along with the photogeometric distances. The extinction coefficients were calculated using the parameters listed in Table 1 of Gaia Collaboration et al. (2018), and applying a code by S. E. Kposov (2025, private communication), which just implements the polynomials described in that paper.

5. Stellar Populations Identified Using Autoencoder and t-SNE

In this section, we focus on four distinct stellar classes that notably cluster in specific regions of the map after applying the t-SNE algorithm. The selection of these clusters has been done manually on the t-SNE map. Other populations, such as cataclysmic variables, or those that do not fit clearly into

categories like the RR Lyrae stars, extragalactic sources, and others identified in this study, will be further analyzed in upcoming publications. Figure 7 shows, from top to bottom, the SEDs for sources selected as: a normal star, a carbon-rich star, a white-dwarf star, an M-type star, and a single-line spectroscopic binary (SB1) system, respectively.

5.1. Carbon Stars

Carbon stars are defined by an overabundance of carbon relative to oxygen in their stellar atmospheres ($C/O > 1$). Normal C stars, also known as C-N stars, are cool objects ($T_{\text{eff}} < 3500$ K) in the thermally pulsing asymptotic giant-branch phase. These stars become intrinsically carbon-rich as they enrich their atmospheres with carbon through third dredge-up events (F. Herwig 2005). In addition to these intrinsic C stars, various types of C stars can form through extrinsic mechanisms, such as mass transfer in binary systems. Other scenarios, including mergers and anomalous He flashes that mix carbon to the surface (e.g., O. Zamora et al. 2009; X. Zhang et al. 2020), have also been proposed to explain the formation of different types of C stars.

Among these types, C-R stars (early and late) are warmer than C-N stars (see O. Zamora et al. 2009). Early-R stars have luminosities similar to red clump stars and show no s-process elements enhancement (J. F. Dominy 1984; G. Knapp et al. 2001). Late-R stars are likely AGB stars, though not as evolved

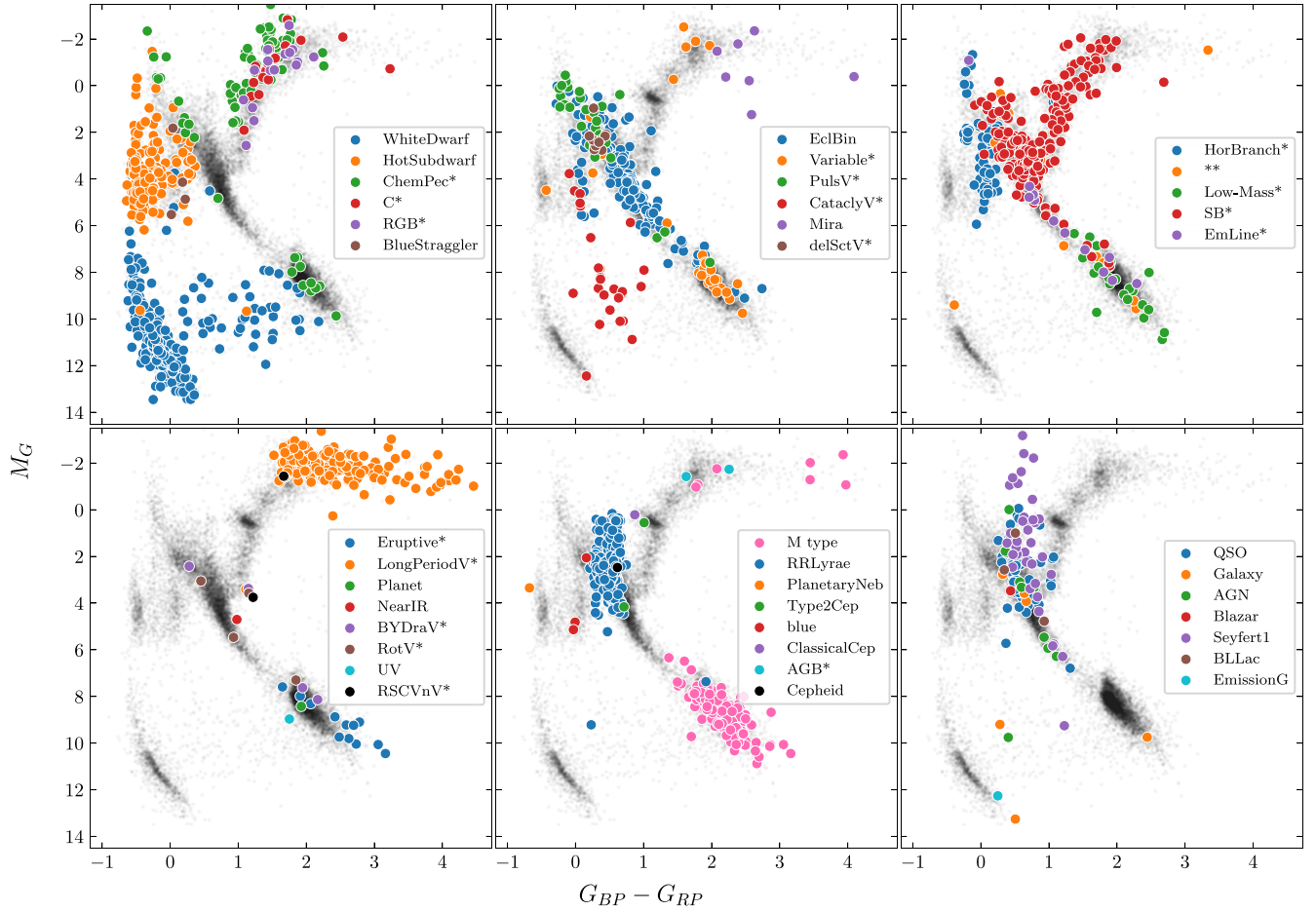


Figure 6. Color–magnitude diagram. In all panels, the background in gray represents sources identified by the autoencoder. The colored points indicate sources that overlap with sources classified in the SIMBAD database. The ChemPec* label encompass a variety of chemically peculiar objects, including CEMP, α 2 CVn variables, R CrB variables, CH stars, barium stars, and dwarf carbon stars.

as C-N stars. C-J stars are characterized by very low isotopic carbon ratios, in some cases as low as $^{12}\text{C}/^{13}\text{C} \approx 3.5$, and high lithium abundances, which are even higher than those of the early-R stars. Their luminosities are also consistent with the AGB phase (C. Abia & J. Isern 2000). CH stars, on the other hand, are found in a different metallicity regime ($-2 \leq [\text{Fe}/\text{H}] \leq -1$) and exhibit a very strong G band due to the CH molecule. They also have relatively high galactic velocities, typical of thick disk and halo stars. These population are expected to result from mass transfer from an AGB companion. At even lower metallicities, carbon-enhanced metal-poor stars (CEMPs; see the complete classification by T. C. Beers & N. Christlieb 2005) are thought to be the more metal-poor counterparts of CH stars. While most CEMP stars originate from mass transfer, some of them may have different formation pathways. Additionally, dwarf C (dC) stars, which are found on the main-sequence phase, can be identified using JHK color–color diagrams due to the weakness of their $2.3 \mu\text{m}$ CO bands compared to those of C giant stars (G. Wallerstein & G. R. Knapp 1998).

With our method, we have identified 96 candidates C stars, with 27 of them previously reported in the literature (C. B. Stephenson 1973; G. Bothun et al. 1991; N. Christlieb et al. 2001; C. R. Hayes et al. 2018). Assuming $M_G > 5.0$ mag ($\log g > 4$) as a criterion, four of the newly identified objects are likely dC stars.

To refine the classification of these candidates, we compare our sample in detail with the C stars studied by L. Li et al. (2024), who identified and classified 3546 C stars using the LAMOST DR7 data (G. Zhao et al. 2012). Their classification relied on line indices, color–color diagrams, spatial distribution, and visual inspection of spectra. Figure 8 presents a series of diagrams to differentiate various types of C stars and compare them with our candidates. Panel (a) shows a color–magnitude diagram that clearly identifies dwarf stars and distinguishes C-N stars, while the separation between CH and C-R stars is less distinct (a subtle difference also noted by C. Abia et al. 2022 in another set of C stars). Panel (b) presents a kinematic diagram based on Gaia astrometry, including radial velocity measurements, which distinguishes between CH (thick disk) and C-R (thin disk) populations. In principle, Gaia provides radial velocities primarily for the brightest stars, meaning that we may be overlooking the fainter stars in our analysis. However, by applying a quality cut in parallax to the stars with radial velocity, we verified that the overall velocity pattern remained unaltered, implying that ignoring the fainter stars does not introduce an obvious bias toward any specific kinematic population. Panel (c) further supports that the C star candidates identified in this study are likely CH and CEMP stars with characteristics of thick disk and halo populations. By performing a crossmatch with a recent catalog of candidate CEMP stars (M. Lucey et al. 2023), we verified that 73 objects in our sample had already been classified as CEMP star

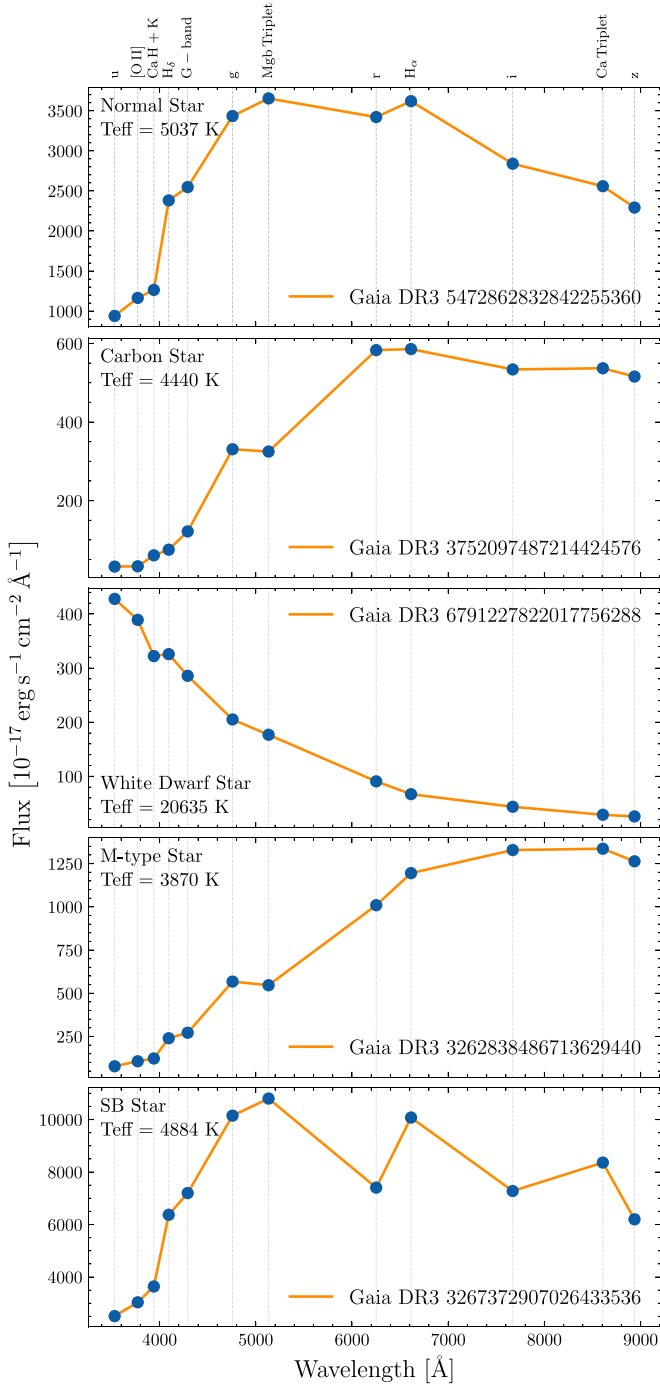


Figure 7. Spectral energy distribution of the populations described in this work

candidates using BP/RP spectra from Gaia DR3, reinforcing their classification.

Figure 9 shows a zoom in of the top-left panel in Figure 5, where these populations of carbon stars cluster in a specific region on the t-SNE map. This region overlaps with four stars classified as peculiar members of the CEMP_s population and one RGB star. Table 2 lists some of the candidate C stars; the complete sample is available in machine readable form.

5.2. White-dwarf Stars

White dwarfs (WDs) mark the final evolutionary stage of stars with initial masses below $8M_{\odot}$. Classified as compact

objects, they typically feature an extremely dense core ($\sim 10^6 \text{ g cm}^{-3}$), primarily composed of carbon and oxygen, surrounded by a thin atmosphere (D. Saumon et al. 2022; S. Blouin 2024). Based on the chemical composition observed in their atmospheres, the main populations of white dwarfs can be classified as follows: DA (spectrum dominated by hydrogen lines), DB (spectrum dominated by neutral helium lines), DC (absence of spectral lines in their spectrum), DO (spectrum dominated by ionized helium), and DZ (spectrum with spectral lines from heavier elements). Since heavy elements are not expected in the atmospheres of WD, their presence indicates pollution with accreted planetary debris, and for this reason, such stars are known as polluted WD (M. Jura & E. D. Young 2014).

Unlike other stellar populations, which occupy specific loci in the t-SNE map, the region of WDs partially overlaps with that of the hot subdwarfs (Figure 5; top-left panel). However, these populations are segregated in the color–magnitude diagram (Figure 6; top-left panel). In view of this, we add the M_G magnitude from Gaia (for those stars with available photogeometric distances from C. A. L. Bailer-Jones et al. 2021) to the corresponding S-PLUS colors and normalize the data using the `StandardScaler()` function; then used it as the input for the t-SNE. The resulting map is shown in Figure 10 (left panel), and we can see that the two populations are well segregated: the WD population is represented in blue and the hot subdwarfs population is represented in orange.

Additionally, we can note that there are three regions of overdensity in the WD population, labeled as 1, 2, and 3 in Figure 10. To determine whether these overdensities correspond to specific subpopulations of WDs, we rely on sources that have spectral type information. The result is shown in Figure 10 (middle panel), where region 1 consists mostly of binary systems (WD + main-sequence stars), region 2 mainly includes DB and DC type stars, and region 3 is primarily composed of DA-type stars. In the right panel, we represent these populations in the color–magnitude diagram. Consistently with their classification, most of the binaries are located between the white-dwarf sequence and the main sequence. In contrast, the DB and DC populations are not clearly separated from the DA stars in the color–magnitude diagram. We conclude that we can reliably apply the t-SNE to the S-PLUS colors to segregate the main subpopulations of WDs. Nevertheless, we must mention that using color–color diagrams C. López-Sanjuan et al. (2022) show segregation of the DA and DB populations in the J-PLUS data. Besides, although a clear segregation is not appreciable in the color–magnitude diagram, a large number of WDs in Gaia have revealed substructures associated with the WD subpopulations in the WD loci (Gaia Collaboration et al. 2018).

5.3. Low-mass Stars

One of the most numerous population (~ 5000 stars) identified by the autoencoder is that of the low-mass stars ($M < 0.8M_{\odot}$), located in the upper left part of the t-SNE map (Figure 4). According to data from SIMBAD (Figure 5), this region contains low-mass stars, variable stars, and eruptive stars. Based on the spectral types of these populations, M-type stars are the predominant group in this area. Consistently with their spectral classification, these sources occupy the region of low-mass stars in the main sequence in the color–magnitude diagram (Figure 6). From this sample, 416 stars have data from

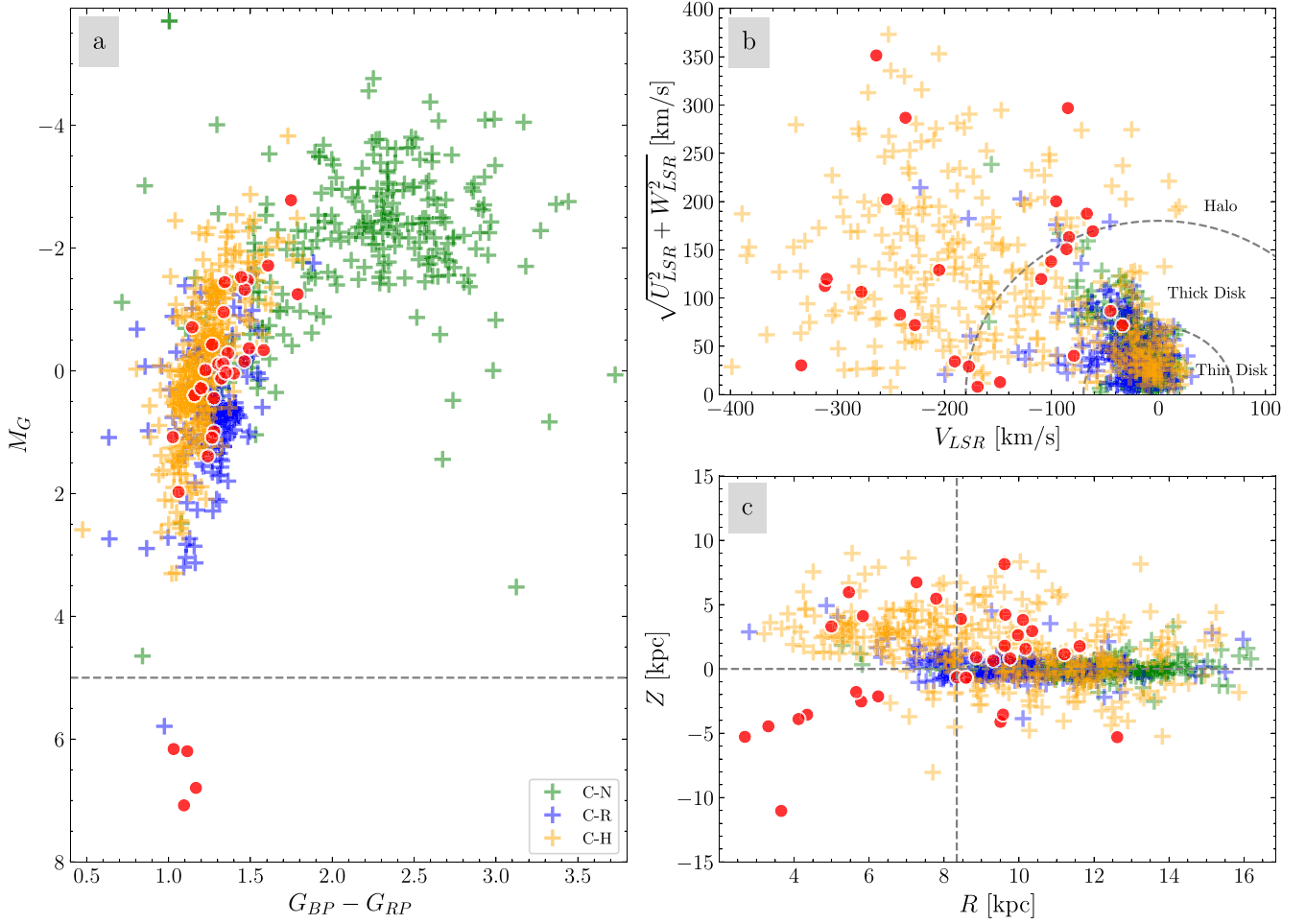


Figure 8. Comparison of C star candidates (red circles) with the C stars analyzed by L. Li et al. (2024). Panel (a) presents a color–magnitude diagram with different types of C stars color coded: green for C-N, orange for C-H, and blue for C-R. A dashed line at $M_G = 5.0$ mag marks the boundary for stars in the main-sequence phase (dC). Panel (b) shows a Toomre diagram with dashed curves at $v_t = 180 \text{ km s}^{-1}$ and 70 km s^{-1} , indicating the thick and thin disk regions, respectively. Panel (c) illustrates the distribution of objects by their Z coordinate as a function of the Galactocentric radius, R , assuming $R_\odot = 8.34 \text{ kpc}$ (M. J. Reid et al. 2014). Stars with a relative parallax error greater than 30% were excluded from this analysis.

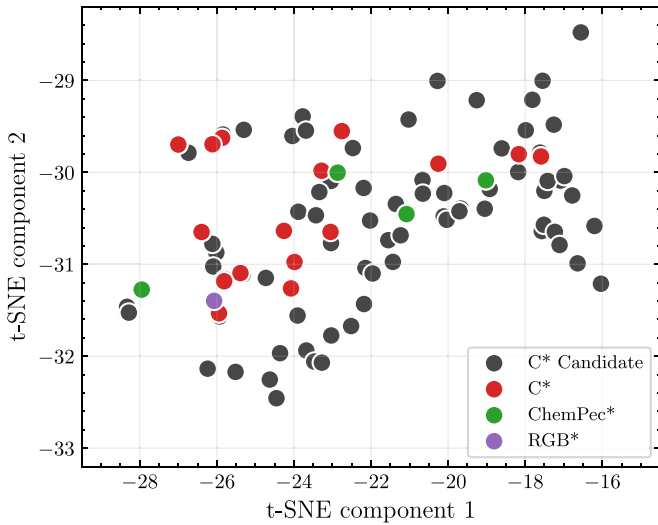


Figure 9. Zooming in on the region of carbon-rich stars in the t-SNE map. The carbon-rich stars identified in this study are shown in gray. The red, green, and purple points indicate stars classified as carbon, peculiar, and RGB in SIMBAD, respectively.

APOGEE. Considering that the effective temperatures are reliable in APOGEE for low-mass stars (D. Souto et al. 2021, 2022), the reported temperatures are consistent with those expected for this population. Due to their low surface temperature, their spectra exhibit a complex structure, characterized by the presence of water molecules, methane, titanium oxide, and other compounds in their atmospheres. As illustrated in Figure 7, their SED differs significantly from that of the stars classified as normal in this study.

It is well known in the literature that some low-mass stars exhibit X-ray emission due to intense stellar activity (S. Mohanty & G. Basri 2003; M. Güdel 2004). The origin of this activity is linked to the age, presence of strong magnetic fields and their rapid rotation, which contributes to the magnetic dynamo mechanism (E. R. Newton et al. 2017). Since X-ray data provide insights into coronal and chromospheric activity, as well as accretion processes in young low-mass stars, we present in Figure 11 (upper panel) the sources that have X-ray information from the eRosita-DE Data Release 1 (A. Merloni et al. 2024), highlighted in orange. The lower panel in the same figure shows the “X-ray main-sequence” diagram proposed by A. C. Rodriguez (2024). This diagram distinguishes between accreting compact X-ray-emitting

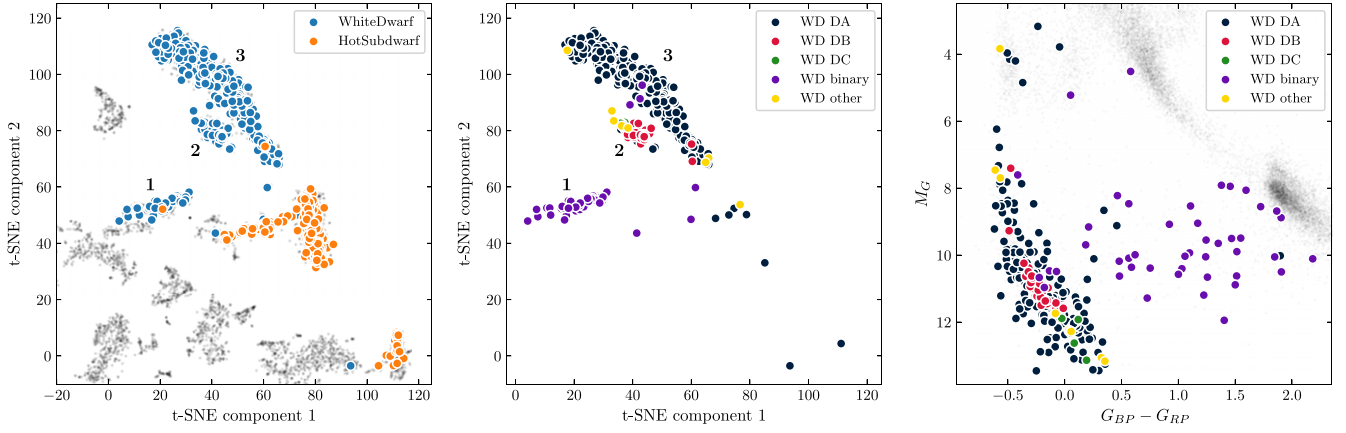


Figure 10. White-dwarf populations. The left panel shows the t-SNE map, illustrating the separation between WDs (blue dots) and hot subdwarf stars (orange dots) after using the M_G as an additional feature for the t-SNE input. The regions 1, 2, and 3 in this panel are overdensities on the t-SNE map. The middle panel shows the t-SNE map with spectral type information in SIMBAD for the same population of WD as in the left panel. In relation to the left panel, region 1 corresponds to binary systems (purple points), region 2 mainly corresponds to DB and DC stars (red and green points), and region 3 predominantly corresponds to DA stars (dark blue points). The right panel presents the color-magnitude diagram for WD. The locations of the WD populations in this diagram are consistent with their spectral classification.

Table 2
Astrometric Parameters and G Apparent Magnitude from Gaia DR3 for Carbon-rich Star Candidates

Gaia DR3 ID	α (deg \pm mas)	δ (deg \pm mas)	ϖ (mas)	μ_α (mas yr $^{-1}$)	μ_δ (mas yr $^{-1}$)	$V_{\text{rad}}(\text{Gaia})$ (km s $^{-1}$)	G (mag)
5658943497829715712	149.836	-26.121	0.034	0.247	-0.867	155.423	14.658
5451431500088634240	162.409	-31.364	0.128	-2.096	-0.931	NaN	16.307
5458744764242802816	150.850	-34.025	-0.001	-0.422	-0.398	NaN	15.559
5391795020192080896	161.673	-41.746	0.040	-4.519	1.264	NaN	14.456
5391637446431500928	162.265	-41.460	0.104	2.579	-9.015	236.292	13.419

(This table is available in its entirety in machine readable form in the [online article](#).)

objects, and stars that emit X-rays due to coronal activity, based on their optical color and X-ray-to-optical flux ratio (F_X/F_{opt}). Most of the sources are located in the active star region, below the empirical cutoff represented by the blue dashed line, as defined by A. C. Rodriguez (2024). On the other hand, it is worth noting that in the t-SNE map, a large portion of the stars with X-ray information are clustered in the upper right part, slightly separated from the rest of the group. This result suggests that photometric data from S-PLUS can be used to identify candidates for highly active stars. This is not surprising, given that some of the S-PLUS filters are located over specific spectral lines such as Ca H + K, H δ , H α , and the Ca triplet, which are indicators of stellar activity (C. Cincunegui et al. 2007).

5.4. Binary Systems

Because quality cuts, such as `astrometric_gof_all`, `ruwe` and `non_single_star`, have been applied to select the normal sources sample, binary systems are expected to be identified as anomalous sources. According to the `ruwe` parameter in Figure 4, the region occupied by the potential binary objects has a complex pattern in the t-SNE map. They are represented by the more yellowish colors, and are located mostly in the upper left and lower right regions. From the crossmatch with SIMBAD (Figure 5), most of the systems classified as spectroscopic binaries (SBs) are located in the

bottom part of the map, while the eclipsing binaries appear more dispersed throughout the map.

Since binary systems are classified through their detection methods rather than by their physical properties, it is not reasonable to anticipate an overdensity based on these classifications. Instead, to identify binary subpopulations within the anomalous sample, we search for stars that overlap with the `nss_two_body_orbit` table from the nonsingle stars catalog provided by Gaia, which includes the orbital parameters for astrometric, spectroscopic, and eclipsing binary systems (Gaia Collaboration et al. 2023b). After the cross-match, we found ~ 400 sources that are classified by solution type as SB1, EclipsingBinary, Orbital, or AstroSpectroSB1. Taking into account the information from the orbital parameters, it is not possible to observe a clear correlation with the overdensities in the map. This is likely due to the fact that multiple configurations of the orbital parameters can yield similar SEDs, although this issue would deserve a deeper analysis in the future.

Another issue that deserves investigation refers to the capacity of the autoencoder to detect binary systems where the primary is a normal star (i.e., compatible with the training set), and the secondary is too faint and contributes very little to the joint SED. That could be the case for many SB1 systems. To address this issue, we have performed a t-SNE analysis including in the input data a subset of the autoencoder training sample, i.e., a subset of normal stars.

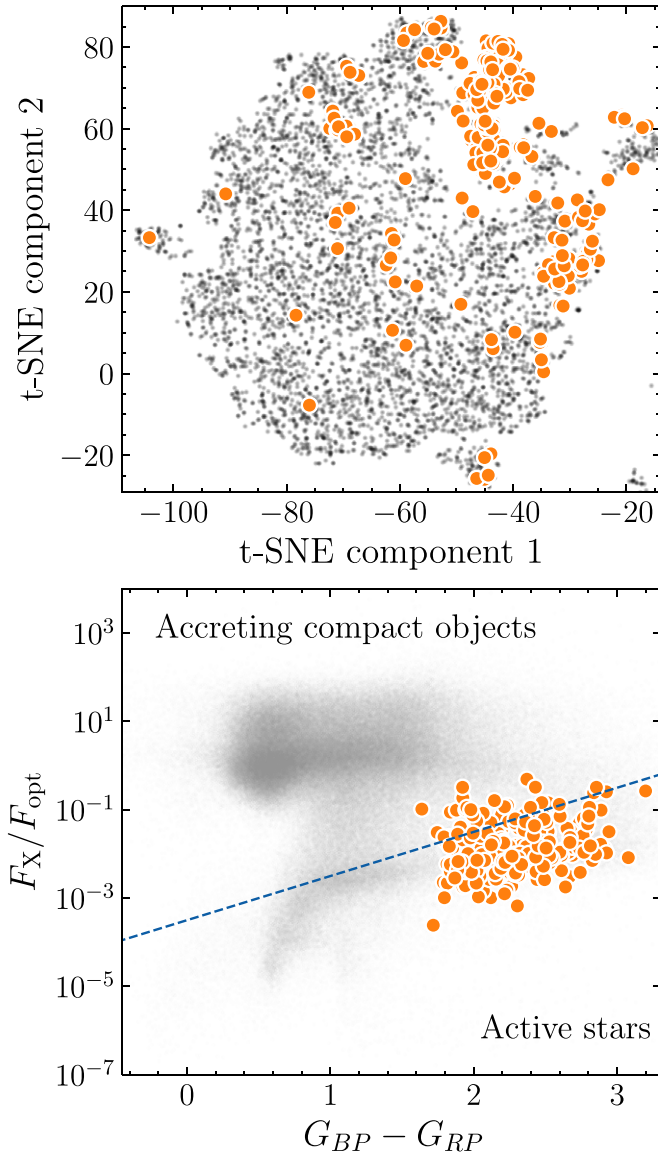


Figure 11. The upper panel shows a zoomed-in view of the low-mass star region in the t-SNE map, with sources containing X-ray information highlighted in orange. The bottom panel displays the X-ray main-sequence diagram, where objects with X-ray data are positioned in the expected region for active stars.

A total of 2000 stars with known stellar parameters (temperature, gravity, and metallicity) have been considered, which represents 10% of the training sample with known data contained in SIMBAD. The resulting t-SNE map is shown in Figure 12, which shall be compared, for example, to Figure 4. Despite the flipping along the horizontal axis, we can see that the distribution of clumps in the periphery of the map is very well preserved, and the identification of specific populations should be straightforward. Only in the central part of the map, where the normal stars are concentrated, the differences become more appreciable. In particular, there is a partial overlap between the normal stars population and part of the SB population, confirming, as expected, that many SBs would be indistinguishable from normal stars. However, there is still a large fraction of SBs that are clearly distinguished from the normal sample. These objects would deserve a detailed analysis in the future.

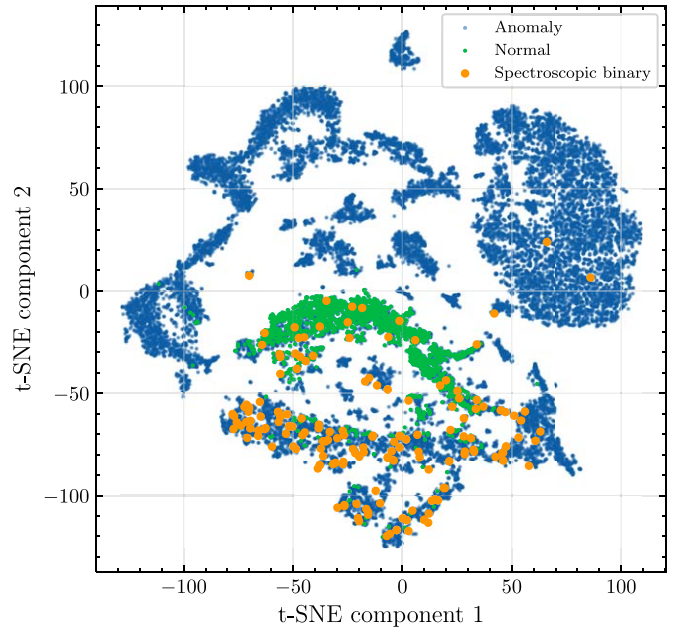


Figure 12. t-SNE 2D map constructed for a set that includes all the anomalous stars detected by the autoencoder (blue dots) and a subset of the normal stars used to train the autoencoder (green dots). The location of the anomalous stars identified as spectroscopic binaries in SIMBAD (orange dots) is plotted for reference.

6. Discussion and Conclusions

In this work, we demonstrate the potential of combining unsupervised machine learning methods, the photometric data from the broadband and narrowband filters system of S-PLUS, and stellar population information from different databases to detect potentially peculiar stars and classify stellar subpopulations.

Although the methodology presented in this study has proven effective for identifying and classifying some stellar populations, it is worth noting that the results reported here are expected to be sensitive to the cutoff values adopted for selecting the nonanomalous objects to train the autoencoder, to the threshold used to select the anomalous objects detected by the autoencoder, and to the parameters used as input for the t-SNE method. In particular, since the autoencoder has been trained using sources with magnitude <19 , it is probable that sources with higher magnitudes will be identified as anomalous objects even if they are not.

As for the threshold to select the anomalous objects, there is no value established in the literature, since it will depend on the objects of interest. However, a high threshold should be used if we aim to identify potentially atypical objects. In our case, we use a conservative value of 3σ above the mean reconstruction error. This limit represents a trade-off, enabling us to obtain a sample with minimal contamination from objects classified as normal in this work, while also ensuring that we do not overlook any interesting objects.

As for the t-SNE parameters, although the hyperparameters may influence the configuration in lower dimensions, the final result primarily depends on the perplexity value. As indicated in L. van der Maaten & G. Hinton (2008), the choice depends on the study's objective. If the focus is on local clusters without regard to the overall configuration, small values of perplexity should be used. Conversely, if maintaining the global structure is important, large values are recommended. In our case, the

perplexity was adjusted by capitalizing information from the SIMBAD database for some sources found in the overdensity regions of the t-SNE map, along with visual inspection of the spectral energy distributions. While this approach may not be effective for clumpy areas without known sources in the literature, analyzing the spectral energy distributions can assist in identifying these groups.

Although a discussion on the completeness of the populations identified in this study is not the primary focus, we should stress that the size of the populations could slightly vary depending on the configuration of the input parameters in t-SNE. On the other hand, the size of the whole anomalous sample is highly dependent on the chosen threshold for the autoencoder reconstruction error. Therefore, we should expect that some sources, which could be considered anomalous, were not actually detected as such because their reconstruction error is slightly lower than the threshold.

In the end, we report 69 carbon star candidates, identified as likely CH or CEMP carbon stars based on their spatial and kinematic characteristics. We clearly identify the main subpopulations of white-dwarf stars, which can be used to select objects of interest for spectroscopic follow-up. Leveraging X-ray data, we demonstrate that very active low-mass stars tend to differ from the majority of their counterparts. We also detect a large number of binary systems, although it is not possible to identify specific subpopulations.

We believe that the kind of analysis presented here can serve as an alternative method for identifying candidate active stars, using multiband photometry with certain filters placed over spectral lines associated with activity. Nevertheless, we must realize that indicators derived from spectra, such as the H-alpha and Ca II H and K lines, or the analysis of epoch photometry, are undoubtedly more effective for identifying and characterizing active stars. In particular, Gaia XP data could be used to identify active stars, since the relevant lines are within its spectral range, and Gaia DR3 epoch photometry could also be applied for this purpose.

Finally, recognizing the importance of medium- and high-resolution spectroscopic follow-up to validate our findings and to better understand the nature of these anomalous objects, we are currently requesting telescope time to observe a small sample of carbon-rich star candidates reported in this work.

Acknowledgments

F.Q.H. acknowledges the support of a fellowship (313979/2025-1) from the PCI Program-MCTI and CNPq. F.R. acknowledges support from CNPq grant 312429/2023-1. N.H. acknowledges a fellowship 301126/2024-0 of the PCI Program—MCTI and CNPq. V.L.T. acknowledges a fellowship 302195/2024-6 of the PCI Program—MCTI and CNPq. R.E. acknowledges the support from CAPES (Finance Code 001: 88887.914708/2023-00). R.L.O. was partially supported by the Brazilian institution *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq; PQ-315632/2023-2 and 445047/2024-0). P.K.H. acknowledges support from the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) grant 2023/14272-4.
















The authors acknowledge comments from A. Alvarez-Candal and from an anonymous referee.

The S-PLUS project, including the T80-South robotic telescope and the S-PLUS scientific survey, was founded as a partnership between the Fundação de Amparo à Pesquisa do

Estado de São Paulo (FAPESP), the Observatório Nacional (ON), the Federal University of Sergipe (UFS), and the Federal University of Santa Catarina (UFSC), with important financial and practical contributions from other collaborating institutes in Brazil, Chile (Universidad de La Serena), and Spain (Centro de Estudios de Física del Cosmos de Aragón, CEFCA). We further acknowledge financial support from the São Paulo Research Foundation (FAPESP), Fundação de Amparo à Pesquisa do Estado do RS (FAPERGS), the Brazilian National Research Council (CNPq), the Coordination for the Improvement of Higher Education Personnel (CAPES), the Carlos Chagas Filho Rio de Janeiro State Research Foundation (FAPERJ), and the Brazilian Innovation Agency (FINEP). The authors who are members of the S-PLUS collaboration are grateful for the contributions from CTIO staff in helping in the construction, commissioning and maintenance of the T80-South telescope and camera. We are also indebted to Rene Laporte and INPE, as well as Keith Taylor, for their important contributions to the project. From CEFCA, we particularly would like to thank Antonio Marín-Franch for his invaluable contributions in the early phases of the project, David Cristóbal-Hornillos and his team for their help with the installation of the data reduction package jype version 0.9.9, César Iñiguez for providing 2D measurements of the filter transmissions, and all other staff members for their support with various aspects of the project.

This work used of data from the European Space Agency (ESA) mission Gaia (<https://www.cosmos.esa.int/gaia>), processed by the Gaia Data Processing and Analysis Consortium (DPAC; <https://www.cosmos.esa.int/web/gaia/dpac/consortium>). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the Gaia Multilateral Agreement.

ORCID iDs

F. Quispe-Huaynasi  <https://orcid.org/0000-0001-8741-8642>
 F. Roig  <https://orcid.org/0000-0001-7059-5116>
 N. Holanda  <https://orcid.org/0000-0002-8504-6248>
 V. Loaiza-Tacuri  <https://orcid.org/0000-0003-0506-8269>
 Romualdo Eleutério  <https://orcid.org/0009-0003-6830-8044>
 S. Daflon  <https://orcid.org/0000-0001-9205-2307>
 V. M. Placco  <https://orcid.org/0000-0003-4479-1265>
 R. Lopes de Oliveira  <https://orcid.org/0000-0002-6211-7226>
 F. Sestito  <https://orcid.org/0000-0002-3182-3574>
 P. K. Humire  <https://orcid.org/0000-0003-3537-4849>
 M. Borges Fernandes  <https://orcid.org/0000-0001-5740-2914>
 A. Kanaan  <https://orcid.org/0009-0007-8005-4541>
 C. Mendes de Oliveira  <https://orcid.org/0000-0002-7736-4297>
 T. Ribeiro  <https://orcid.org/0000-0002-0138-1365>
 W. Schoenell  <https://orcid.org/0000-0002-4064-7234>

References

- Abia, C., de Laverny, P., Romero-Gómez, M., & Figueras, F. 2022, *A&A*, **664**, A45
- Abia, C., & Isern, J. 2000, *ApJ*, **536**, 438
- Allende Prieto, C. 2016, *A&A*, **595**, A129
- Anders, F., Chiappini, C., Santiago, B. X., et al. 2018, *A&A*, **619**, A125
- Bailer-Jones, C. A. L., Rybizki, J., Fouesneau, M., Demleitner, M., & Andrae, R. 2021, *AJ*, **161**, 147
- Bank, D., Koenigstein, N., & Giryas, R. 2023, *Autoencoders* (Cham: Springer), 353
- Beers, T. C., & Christlieb, N. 2005, *ARA&A*, **43**, 531
- Benitez, N., Dupke, R., Moles, M., et al. 2014, arXiv:1403.5237

- Berahmand, K., Daneshfar, F., Salehi, E. S., Li, Y., & Xu, Y. 2024, *Artif. Intell. Rev.*, 57, 28
- Blouin, S. 2024, arXiv:2409.03941
- Bothun, G., Elias, J. H., MacAlpine, G., et al. 1991, *AJ*, 101, 2220
- Cenarro, A. J., Moles, M., Cristóbal-Hornillos, D., et al. 2019, *A&A*, 622, A176
- Chambers, K. C., Magnier, E. A., Metcalfe, N., et al. 2016, arXiv:1612.05560
- Chollet, F., et al. 2015, Keras, <https://Keras.io>
- Christlieb, N., Green, P. J., Wisotzki, L., & Reimers, D. 2001, *A&A*, 375, 366
- Cincunegui, C., Díaz, R. F., & Mauas, P. J. D. 2007, *A&A*, 469, 309
- Čotar, K., Zwitter, T., Kos, J., et al. 2019, *MNRAS*, 483, 3196
- Cui, X.-Q., Zhao, Y.-H., Chu, Y.-Q., et al. 2012, *RAA*, 12, 1197
- Da Costa, G. S., Bessell, M. S., Mackey, A. D., et al. 2019, *MNRAS*, 489, 5900
- De Silva, G. M., Freeman, K. C., Bland-Hawthorn, J., et al. 2015, *MNRAS*, 449, 2604
- Dominy, J. F. 1984, *ApJS*, 55, 27
- Evans, D. W., Riello, M., De Angeli, F., et al. 2018, *A&A*, 616, A4
- Ferreira Lopes, C. E., Gutiérrez-Soto, L. A., Ferreira Alberice, V. S., et al. 2025, *A&A*, 693, A306
- Finlator, K., Ivezić, Ž., Fan, X., et al. 2000, *AJ*, 120, 2615
- Gaia Collaboration, Arenou, F., Babusiaux, C., et al. 2023b, *A&A*, 674, A34
- Gaia Collaboration, Babusiaux, C., van Leeuwen, F., et al. 2018, *A&A*, 616, A10
- Gaia Collaboration, Vallenari, A., Brown, A. G. A., et al. 2023a, *A&A*, 674, A1
- Green, G. 2018, *JOSS*, 3, 695
- Güdel, M. 2004, *A&ARv*, 12, 71
- Hayes, C. R., Majewski, S. R., Shetrone, M., et al. 2018, *ApJ*, 852, 49
- Helmi, A., Ivezić, Ž., Prada, F., et al. 2003, *ApJ*, 586, 195
- Herpich, F. R., Almeida-Fernandes, F., Oliveira Schwarz, G. B., et al. 2024, *A&A*, 689, A249
- Herwig, F. 2005, *ARA&A*, 43, 435
- Hinton, G. E., & Roweis, S. 2002, Stochastic Neighbor Embedding, in NIPS 2002, Vol. 15 (Cambridge, MA) ed. S. Becker, S. Thrun, & K. Obermayer (Cambridge, MA: MIT Press)
- Hinton, G. E., & Salakhutdinov, R. R. 2006, *Sci*, 313, 504
- Huang, Y., Beers, T. C., Xiao, K., et al. 2024, *ApJ*, 974, 192
- Jura, M., & Young, E. D. 2014, *AREPS*, 42, 45
- Kao, M. L., Hawkins, K., Rogers, L. K., et al. 2024, *ApJ*, 970, 181
- Kingma, D. P., & Ba, J. 2014, arXiv:1412.6980
- Knapp, G., Pourbaix, D., & Jorissen, A. 2001, *A&A*, 371, 222
- Laroche, A., & Speagle, J. S. 2025, *ApJ*, 979, 5
- Lenz, D. D., Newberg, J., Rosner, R., Richards, G. T., & Stoughton, C. 1998, *ApJS*, 119, 121
- Li, L., Zhang, K., Cui, W., et al. 2024, *ApJS*, 271, 12
- López-Sanjuan, C., Tremblay, P. E., Ederoclite, A., et al. 2022, *A&A*, 658, A79
- Lucey, M., Al Kharusi, N., Hawkins, K., et al. 2023, *MNRAS*, 523, 4049
- Majewski, S. R., Schiavon, R. P., Frinchaboy, P. M., et al. 2017, *AJ*, 154, 94
- Martin, D. C., Fanson, J., Schiminovich, D., et al. 2005, *ApJL*, 619, L1
- Masana, E., Jordi, C., & Ribas, I. 2006, *A&A*, 450, 735
- Mauron, N., Azzopardi, M., Gigoyan, K., & Kendall, T. R. 2004, *A&A*, 418, 77
- Mendes de Oliveira, C., Ribeiro, T., Schoenell, W., et al. 2019, *MNRAS*, 489, 241
- Merloni, A., Lamer, G., Liu, T., et al. 2024, *A&A*, 682, A34
- Mienye, I. D., & Swart, T. G. 2025, Archives of Computational Methods in Engineering, doi:10.1007/s11831-025-10260-5
- Mohanty, S., & Basri, G. 2003, *ApJ*, 583, 451
- Nakazono, L., Mendes de Oliveira, C., Hirata, N. S. T., et al. 2021, *MNRAS*, 507, 5847
- Nazábal, A., Olmos, P. M., Ghahramani, Z., & Valera, I. 2020, *PatRe*, 107, 107501
- Newton, E. R., Irwin, J., Charbonneau, D., et al. 2017, *ApJ*, 834, 85
- O'Malley, T., Bursztein, E., Long, J., et al. 2019, KerasTuner, <https://github.com/keras-team/keras-tuner>
- Oussidi, A., & Elhassouny, A. 2018, in Int. Conf. on Intelligent Systems and Computer Vision (ISCV) (New York: IEEE), 1
- Pang, G., Shen, C., Cao, L., & Hengel, A. V. D. 2021, *ACM Comput. Surv.*, 54, 1
- Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, *JMLR*, 12, 2825
- Portillo, S. K. N., Parejko, J. K., Vergara, J. R., & Connolly, A. J. 2020, *AJ*, 160, 45
- Reid, M. J., Menten, K. M., Brunthaler, A., et al. 2014, *ApJ*, 783, 130
- Rodriguez, A. C. 2024, *PASP*, 136, 054201
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. 1986, *Natur*, 323, 533
- Rybizki, J., Green, G. M., Rix, H.-W., et al. 2022, *MNRAS*, 510, 2597
- Saumon, D., Blouin, S., & Tremblay, P.-E. 2022, *PhR*, 988, 1
- Schlaflly, E. F., & Finkbeiner, D. P. 2011, *ApJ*, 737, 103
- Sharma, K., Kembhavi, A., Kembhavi, A., Sivarani, T., & Abraham, S. 2019, *BSRSL*, 88, 174
- Sichevskij, S. G., Mironov, A. V., & Malkov, O. Y. 2014, *AstBu*, 69, 160
- Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, *AJ*, 131, 1163
- Souto, D., Cunha, K., & Smith, V. V. 2021, *ApJ*, 917, 11
- Souto, D., Cunha, K., Smith, V. V., et al. 2022, *ApJ*, 927, 123
- Starkenburg, E., Martin, N., Youakim, K., et al. 2017, *MNRAS*, 471, 2587
- Steinmetz, M., Zwitter, T., Siebert, A., et al. 2006, *AJ*, 132, 1645
- Stephenson, C. B. 1973, Cleveland Publications of the Warner and Swasey Observatory (Cleveland, OH: Case Western Reserve Univ.)
- Szkody, P., Anderson, S., Agüeros, M., & Covarrubias, R. 2002, in ASP Conf. Ser. 261, The Physics of Cataclysmic Variables and Related Objects, ed. B. T. Gänsicke, K. Beuermann, & K. Reinsch (San Francisco, CA: ASP), 297
- Traven, G., Matijević, G., Zwitter, T., et al. 2017, *ApJS*, 228, 24
- van der Maaten, L., & Hinton, G. 2008, *JMLR*, 9, 2579
- Wallerstein, G., & Knapp, G. R. 1998, *ARA&A*, 36, 369
- Whitten, D. D., Placco, V. M., Beers, T. C., et al. 2021, *ApJ*, 912, 147
- Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, *AJ*, 140, 1868
- Yang, T., & Li, X. 2015, *MNRAS*, 452, 158
- Yao, Y., Ji, A. P., Koposov, S. E., & Limberg, G. 2024, *MNRAS*, 527, 10937
- York, D. G., Adelman, J., Anderson, J. E. J., et al. 2000, *AJ*, 120, 1579
- Zamora, O., Abia, C., Plez, B., Domínguez, I., & Cristallo, S. 2009, *A&A*, 508, 909
- Zhang, X., Jeffery, C. S., Li, Y., & Bi, S. 2020, *ApJ*, 889, 33
- Zhao, G., Zhao, Y.-H., Chu, Y.-Q., Jing, Y.-P., & Deng, L.-C. 2012, *RAA*, 12, 723