UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia "Galileo Galilei"

Master Degree in Astrophysics and Cosmology

Master Thesis

Unveiling the lowest surface brightness regions in dwarf galaxies: a machine learning approach with Gaia early DR3

Supervisor

Dr. Giulia Rodighiero

Co-supervisor

Dr. Giuseppina Battaglia

Co-supervisor

Dr. Giuliano Iorio

Candidate

Marco Boscato

Matricola 2096921

Accademic Year 2024/2025

Abstract

Recent observations suggest that dwarf galaxies appear to host extended tidal structures or debris from disrupted systems that contribute to their stellar halos. These features are extremely challenging to identify, yet they are relics of past interactions and provide crucial constraints on the nature of dark matter and the evolutionary history of galaxies.

In this thesis, I investigated different methods to calculate membership probabilities for individual stars in dwarf galaxies of the Local Group, with a particular focus on identifying members in external structures such as tidal tails and outer halos.

The analysis was performed on four mock catalogs constructed from Gaia eDR3, designed to reproduce different foreground conditions and dwarf galaxies, including cases with tidal tails. Three approaches were explored, all implemented within a machine learning framework but differing in the way the data were modeled: (i) an updated, machine learning based version of the probabilistic method used in Battaglia et al. (2022), (ii) a dimensionality-reduction approach, and (iii) a normalizing flow model capable of mapping complex distributions into simpler ones.

All methods demonstrated a high capacity for detecting true members; however, the first two primarily identified stars in the central regions of the galaxies, whereas the normalizing flow method was the only one able to consistently recover the external structures, highlighting its potential as a powerful tool for probing the outskirts of dwarf galaxies.

Contents

1 Introduction

Dwarf galaxies are small stellar systems containing from a few thousand up to several billion stars, a negligible number compared to massive galaxies such as the Milky Way. There is no universally precise definition of what constitutes a dwarf galaxy. However, since 1993 a commonly adopted working definition has been that dwarf galaxies are systems with absolute magnitudes fainter than $M_V < -17$ and with spatial extensions larger than globular clusters (Tolstoy et al. 2009 and references therein). Within this broad classification, several distinct types of dwarf galaxies have been identified: early-type dwarf spheroidals (dSphs), late-type star-forming dwarf irregulars (dIs), ultra-faint dwarfs (UFDs), centrally concentrated blue compact dwarfs (BCDs) with active star formation, and the more recently recognized ultra-compact dwarfs (UCDs). Dwarf galaxies are believed to have formed through gravitational processes during the early stages of galaxy assembly (Jensen et al. 2023) or, as suggested by recent studies, as by-products of interactions and mergers, arising from streams of baryonic matter and dark matter stripped from larger systems. However, both the fraction of dwarfs formed in this way and their long-term survival remain unclear (Zaragoza-Cardiel et al. 2024).

Although morphologically simple, dwarf galaxies provide fundamental insights into galaxy formation and evolution. In particular, irregular dwarfs, characterized by low metallicity and relatively high gas fractions, are considered the closest local analogues to the first galaxies that populated the early Universe[1]. Moreover, their typically high mass-to-light ratios, often exceeding 100 in the faintest systems, reveal that they are strongly dark matter dominated (Simon 2019 and references therein). Because of their low intrinsic luminosities, the study of nearby dwarfs is especially powerful: their proximity and their high central dark matter densities (Simon et al. 2007), makes them ideal laboratories for testing models of galaxy evolution and the nature of dark matter.

The Local Group (LG), which includes the Milky Way and M31, provides the richest environment for such studies. It hosts the faintest galaxies known to date, with luminosities as low as $L \lesssim 10^5\, L_\odot$ (Simon 2019; McConnachie 2012; Pace 2024), and currently counts nearly one hundred identified dwarfs, of which 68 are satellites of the Milky Way, 9 of M31, and 15 beyond the LG (Goater et al. 2023). Since most of the well-studied dwarfs are satellites of the Milky Way, they also provide unique constraints on the distribution of dark matter subhalos, setting lower limits on their abundance. Their small physical sizes probe dark matter clustering on scales of $\sim 20 - 30$ pc for the most compact ultra-faint dwarfs, a regime inaccessible in any other astrophysical context (Simon 2019).

Within the hierarchical ΛCDM framework[2], galaxies are expected to form and grow through the accretion of smaller systems. Dwarf galaxies therefore represent the smallest-scale environment where these processes can be studied (Deason et al. 2021). Simulations suggest that dwarf galaxies formed within extended dark matter halos and experienced early galaxy mergers and supernova feedback, making them the surviving remnants of the earliest galaxies as confirmed by their ancient and chemically primitive stars ($\sim$ 13 billion years old), even though these events should lie outside their core regions (>2 half-light radii) (Chiti et al. 2021).

[1] *ESA/Hubble*

[2] The Λ Cold Dark Matter model is the current cosmological mathematical model of the Big Bang theory that incorporating a cosmological constant Λ associated with dark energy, cold dark matter and ordinary matter, assuming general relativity as the theory of gravity on large scales.

Recent observations have detected the presence of these external structures in various LG's dwarf galaxies. A recent study of the Fornax dwarf galaxy by Yang et al. (2022) discovered a significant second component that extends out to 2.1 degree (seven time its half-light radius) and about 10% of the total mass of Fornax, which resembles a stellar halo due to its nearly symmetric morphology. Another evidence for this kind of structure has been found by Chiti et al. (2021) for the dwarf galaxy Tucana II where they identify members in a outer region up to 9 half-light radii, demonstrating the system to be even more spatially extended than previous observations.
Simulations support the ubiquity of these structures: in the Auriga cosmological suite, all isolated dwarfs with stellar masses in the range $10^6 \leq M_\star \leq 10^8 M_\odot$ form stellar halos containing 2-20% of their stellar mass, typically via dwarf-dwarf mergers. These halos are systematically more metal-poor than the central regions and exhibit steep density profiles, reflecting their distinct origins (Deason et al. 2021).

Another fundamental property of dwarf galaxies is their relatively shallow potential wells, which make them highly susceptible to tidal interactions and other environmental processes that can alter their morphology.
A subset of these systems, known as tidal dwarf galaxies (TDGs), is thought to represent only a small fraction ($\leq 10\%$) of the dwarf galaxy population in the local universe (Zaragoza-Cardiel et al. 2024 and references therein). TDGs can form either by material pulled from a parent galaxy, evolving into independent systems that are typically dark matter-deficient and may exhibit higher metallicities reflecting their disk origin, or via the tidal disruption of pre-existing dwarfs. Their long-term survival, however, remains uncertain: processes such as mass loss driven by stellar winds and supernovae, tidal stripping, or eventual infall back into the parent galaxy can compromise their persistence. This uncertainty stands in contrast with the fact that the stellar populations of many dwarf galaxies indicate they are among the oldest galaxies (Simon 2019).
One of the galaxies most likely to have such structures is the Carina galaxy, which, as demonstrated by Battaglia et al. (2012), has a clear presence of ancient MSTO (Main Sequence Turn-Off) stars at R $>$ 25', which demonstrates that Carina extends well beyond its literature nominal tidal radius (r_t = 28.8').
On the other hand, the search for tail regions in most dwarf galaxies is rather complicated, where direct detection has not yet been made. In Fornax there was no evidence of structure as tidal tails (Yang et al. 2022) and simulations, based in a frame for which it is dark matter dominated and a long-lived satellite of the MW, shows how the stellar component is very stable against tidal stripping (Battaglia et al. 2015).
EDGE cosmological simulations (Goater et al. 2023) shows that within our local volume of the Universe, multitude of the faintest and most dark matter dominated galaxies, despite their tidal isolation, exhibit anisotropic extended stellar outskirts that masquerade as tidal tails but are instead natal. This lead to dwarf galaxies that form later are fainter and more extended and have more elliptical stellar distributions. The simulations matches the observed ellipticity distributions of Milky Way satellites and isolated dwarfs, and that many dwarfs may be tidally intact, meaning their structures are not primarily shaped by interactions with massive galaxies, but instead reflect intrinsic formation processes.
An important aspect of studying dwarf galaxies is that knowledge of their orbits around the Milky Way allows us to assess the impact of tidal forces on their morphology and evolution. Recent studies (Battaglia et al. (2022); Pace et al. 2022) have leveraged systemic proper motions of Local Group dwarfs derived from Gaia data, have enabled detailed investigations of these tidal effects and their role in shaping dwarf galaxy

properties.
A recent survey of dwarf galaxies in the LG, conducted by Jensen et al. (2023) on the detection of external structures using data from Gaia, identified nine dwarfs (Boötes I, Boötes III, Draco II, Grus II, Segue I, Sculptor, Tucana II, Tucana III, and Ursa Minor) that show evidence for a secondary, low-density outer component in their stellar profiles. For example, kinematic evidence suggests that Sculptor may currently be losing stellar mass due to tidal effects, even though simulations indicate that the Milky Way's tidal influence is not expected to be strong enough to significantly distort its morphology. However, these simulations also show that up to 60% of Sculptor's dark matter halo may have already been stripped away (Iorio et al. 2019). Or in the case of Tucana II, its outer stars are likely the result of a past dwarf-dwarf merger, consistent with the result obtained by Chiti et al. (2021). Among the systems identified as having extended structures, many have been reported as possibly disrupted systems, in particular Boötes I, Boötes III, Tucana III and Ursa Minor, which are consistent with the literature.
The discovery of these outer stellar components suggests that extended halos and tidal debris are common features of Milky Way dwarfs. In some systems, such as Carina, the outskirts appear consistent with tidal stripping by the Milky Way, while in others, such as Tucana II, the extended halo may represent the remnant of a dwarf-dwarf merger. These results highlight that dwarf galaxies cannot be fully understood by examining only their central populations: their outer regions provide the key evidence for hierarchical assembly and tidal processing.

Precisely because of the importance of these external structures and the difficulty in identifying them, the search for new methods is of particular current interest. For this reason, in this work we were interested in the identification of dwarf galaxies outskirts members and so we present three alternative methodologies, all relying on machine learning tools, and tested on mock catalogs constructed from Gaia eDR3 data and synthetic stellar populations. In Section 2, we provide an overview of the Gaia eDR3 astrometric data used, describe the qualitative cuts applied to prepare the final catalog for analysis, and in Sec. 2.2 we explain the construction of the mock catalogs that simulate different scenarios of dwarf galaxies with extended structures, such as tidal tails.
In Section 3, we briefly introduce the methodology of Battaglia et al. (2022) (hereafter referred to as **B22**) and present our alternative machine learning-based approaches for calculating membership probabilities, where the first method (Sec. 3.2) builds on the framework of **B22**, but with important modifications. Here, a machine learning model is used to construct the likelihood and priors of the properties of the contaminants directly from the data, without adopting analytic assumptions. In particular, the spatial, proper motion, and color-magnitude diagram distributions of the contaminants are modeled from the real field stars, thereby producing a faithful representation of the Milky Way foreground in the absence of any dwarf galaxy. The galaxy likelihood is then obtained by subtracting this contaminant model from the observed distribution. The only distribution that remains untouched is the galaxy's density profile, for which we could not identify a robust replacement.
The second and third methods are more machine learning oriented. The second (Sec. 3.3) relies entirely on a more direct clustering algorithm, where sources that share similar feature are labeled as members or contaminants. The third (Sec. 3.4) employs a generative approach using normalizing flows, which transform the complex distributions of the contaminants in proper motion and color-magnitude diagram

space into simpler, nearly Gaussian forms. This enables a more straightforward computation of the likelihoods and hence simplifies the Bayesian inference.
We evaluate these three approaches separately on four mock catalogs: a Sculptor-like dwarf galaxy embedded in the real Sextans foreground contamination, a Sextans-like galaxy with tidal tails in its own contamination field and in Draco's real contamination field and a Draco-like dwarf galaxy with thin tidal tails in its own contamination environment.
Section 4 presents the outcomes obtained from all machine learning methodologies across the different mock catalogs and finally, in Section 5, we compare the results from each methodology and mock scenario, analyzing the strengths and limitations of each approach, and discuss the potential applications of these methods in future studies. This setup allows us to test the performance of the three methodologies across a range of contamination levels and galaxy densities, highlighting their strengths and limitations in the identification of external members.

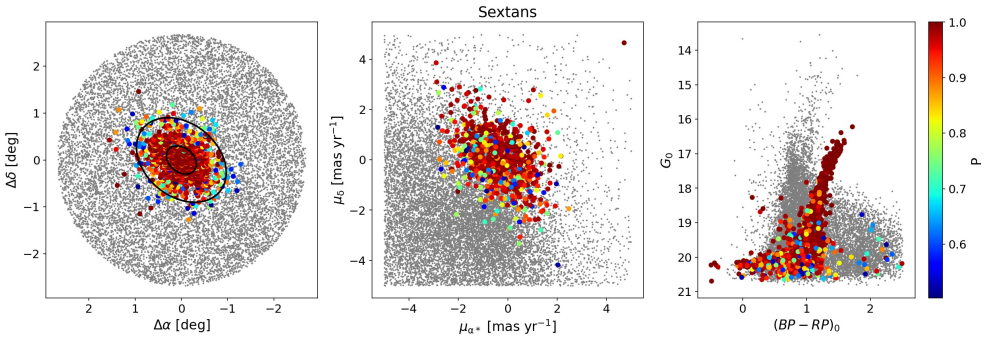


Figure 1: Probable member stars of the Sextans dwarf spheroidal galaxy. Distribution of *Gaia* eDR3 sources within 4° of the optical center. Stars are color-coded by their membership probability from **B22**, with only those having $P > 0.5$ shown in color; gray points correspond to stars with $P < 0.5$. *Left*: spatial distribution on the tangent plane centered on Sextans, with ellipses marking 1× and 3× the half-light radius (see Tab. 1). *Middle*: proper motion plane. *Right*: color-magnitude diagram of the same sample. The figure is reported from **B22**.

2 Data

As the primary objectives of this work is to enhance the **B22** methodology for estimating probabilities of membership of individual stars along the line-of-sight to LG dwarf galaxies (see Sec. 3.1 for a description of the methodology). Since we wish to carry out a direct comparison to the work by **B22**, we tailor our approach to the use of Gaia data, in terms of observables and quality cuts applied.

Therefore this study is based on astrometric and photometric data from Gaia Early Data Release 3 (eDR3)[3], focusing on sources with full parameter astrometric solutions. In addition to the Gaia dataset, we constructed and utilized a series of mock catalogs, produced by Dr. G. Thomas, each tailored to represent specific scenarios such as different dwarf galaxies, foreground contamination, and structural configurations, to systematically evaluate the performance of the proposed algorithms.

2.1 Local Group dwarf galaxies in Gaia eDR3

Within the Gaia magnitude detection limit (up to ~ 21 mag), at the typical distances of Milky Way satellites and, more generally, LG dwarf galaxies, the dominant source of contamination comes from stars belonging to the Milky Way that lie along the line-of-sight between the observer and the target system. For simplicity, throughout this work we will refer to the Milky Way foreground as *contaminants*.
We adopt the Sextans dwarf spheroidal galaxy as a template, since it is a diffuse, low surface brightness system projected against a dense Milky Way foreground (see Tab. 1), making it a demanding benchmark for testing the performance of membership identification algorithms. This difficulty is especially pronounced in the outer regions of the galaxy, where the density contrast between member stars and contaminants is lowest.
The data extracted from the Gaia archive correspond to a circular region with an 8-degree radius centered on the literature coordinates of the target object. We expect this area to be large enough to enclose the galaxy and any extended structures, such

[3]https://gea.esac.esa.int/archive/

Parameter	Sextans	Sculptor	Draco	Reference
RA (deg)	153.27	15.02	260.07	(1)
Dec (deg)	-1.62	-33.72	57.92	(1)
l (deg)	243.5	287.53	86.4	(2);(3);(4)
b (deg)	+42.3	-83.16	+34.7	(2);(3);(4)
dm	19.64 ± 0.01	19.62 ± 0.04	19.53 ± 0.07	(1)
Rh (arcmin)	21.40 ± 0.10	12.43 ± 0.18	9.61 ± 0.10	(1)
ell	0.27 ± 0.03	0.36 ± 0.01	0.30 ± 0.01	(1)
PA (deg)	52.0 ± 3.0	92.0 ± 1.0	87.0 ± 1.0	(1)
$< v_{los} >$ ($\mathrm{km\,s^{-1}}$)	226.0 ± 0.6	110.6 ± 0.5	-292.3 ± 0.4	(1)
σ_{los} ($\mathrm{km\,s^{-1}}$)	8.4 ± 0.4	10.1 ± 0.3	9.0 ± 0.3	(1)
Σ_0 ($\mathrm{mag\,arcmin^{-1}}$)	18.2 ± 0.5	$22 - 26$	25.5 ± 0.5	(2);(3);(5)
L ($L_\odot$)	$4.1 \pm 1.9 \times 10^5$	$2.03 \pm 0.79 \times 10^6$	2.2×10^5	(2);(3);(4)
Stellar Mass ($M_\odot$)	4.4×10^4	$6.3 \pm 0.4 \times 10^6$	6.6×10^5	(2);(3);(4)

Table 1: Summary of the structural and physical properties of the dwarf galaxies analyzed in this work. In order of row we have the coordinates of the optical center first expressed in the equatorial system (RA, Dec) and then in Galactic coordinate system (l, b), the distance module, the half-light radius along the projected major axis, ellipticity defines as $1 - b/a$, with a and b the projected major and minor axes of the stellar component, position angle measured from north to east, heliocentric systemic l.o.s velocity, velocity dispersion along the l.o.s, the central surface brightness, luminosity (in V-band), and the stellar mass. References are listed in the same order as the galaxies: Sextans, Sculptor, and Draco. When a single reference is provided, it applies to all three galaxies.
The corresponding numeric code are: (1) **B22**, (2) Battaglia et al. (2011), (3) Bettinelli et al. (2019), (4) Lokas et al. (2005), (5) Aparicio et al. (2001).

as stellar halos or tidal features, and contains a significant number of foreground contaminants, which are also useful for training purposes (see Sec. 3.2).

To define our analysis sample, we follow the same method used in **B22** by selecting sources that are not flagged as duplicates, possess a complete astrometric solution (`astrometric_params_solved` ≥ 31), and meet the criterion for high-quality astrometry, characterized by a Renormalized Unit Weight Error (`ruwe`) less than 1.4. Additionally, we excluded all objects that are cross-matched with Gaia AGN catalogs.
To further ensure the reliability of the sample and eliminate extended or non-isolated sources, we imposed the following quality cuts: `ipd_frac_multi_peak` ≤ 2, `ipd_gof_harmonic_amplitude` < 0.2, and a corrected photometric excess factor within 5σ of the expected value at a given G-band magnitude, following the procedure described by Riello et al. (2021).
To minimize contamination from clearly non-member stars, we applied a parallax selection, retaining only those sources with `parallax` consistent with that of the dwarf galaxy within $3\sigma_\pi$, where σ_π is the quadrature sum of the parallax error of the star and the uncertainty on the galaxy's parallax. Parallax values were corrected by applying a global zero-point offset of -0.017 mas. The residual Gaia parallax systematics were deemed negligible for the scope of this analysis.
After these cuts, the dataset used for this analysis included the spatial coordinates `ra` and `dec`, transformed into standard coordinates on the projected plane of the galaxy; the proper motion components along right ascension (`pmra`) and declination (`pmdec`); and color-magnitude diagram (CMD) quantities: the extinction and reddening corrected magnitude G_0 and color $(BP - RP)_0$, derived from the Gaia catalog values `phot_g_mean_mag`, `phot_bp_mean_mag`, and `phot_rp_mean_mag`, with corrections applied using the $G_{\rm factor}$ and `phot_bp_rp_excess_factor` for the color.
This filtering process, which serves in particular to eliminate a large number of contaminants and other sources of noise, is performed each time on the raw dataset obtained from Gaia in order to initialize the dataset before performing the analysis.

2.2 Mock catalogs

In order to rigorously evaluate the performance of our algorithms, it was necessary to construct mock catalogs in which the true membership of each star, whether a galaxy member or a contaminant, was explicitly known. To construct these catalogs, Dr. G. Thomas simulated a mock dwarf galaxy by injecting synthetic member stars into a real region of the Gaia eDR3 sky that contains only Milky Way foreground contamination. Each synthetic member was assigned a unique identifier, setting the `source_id` to a fixed value of 42. This allowed for an unambiguous separation between true members and contaminants during analysis and evaluation.
A two-step approach was followed to construct realistic mock dwarf galaxies.

1. **Stellar Component Generation**: the stellar distribution of a dwarf galaxy was modeled by assuming a gravitational potential and either a distribution function or a stellar density profile that ensures dynamical self-consistency. This process yields the positions and velocities of individual stellar particles.
   Each stellar population (e.g., Sculptor has two components, metal-rich and metal-poor components) was defined by a stellar age and a metallicity distribution function. Using the Kroupa Initial Mass Function (IMF) and theoretical isochrones, the model assigns each particle a corresponding absolute magnitude. The output is a physically motivated representation of the stellar content of a dwarf galaxy.

2. **Projection and Realistic Mock Construction**: The synthetic galaxy was projected onto the sky at a specified location, applying a chosen distance modulus, inclination, and position angle.
   To emulate realistic observational conditions, we queried the Gaia archive to retrieve real stellar data within an 8 degrees radius of the same coordinates. These Gaia data were used to train a machine learning model that captures the local dependencies of photometric, proper motion, and parallax uncertainties as functions of stellar magnitude. These empirical relations were then applied to the synthetic dwarf galaxy stars to simulate observational effects, such as parameter uncertainties and detection completeness. Finally, the processed mock galaxy was embedded into the real Gaia background field to produce a comprehensive and realistic mock observation of a dwarf galaxy as it would appear in Gaia eDR3.

In total, four mock catalogs were generated, each representing a distinct observational scenario. In three of the catalogs an additional component was incorporated to simulate a tidal tail. These were designed to test the robustness and consistency of our methods under varying conditions of source compactness and background contamination.

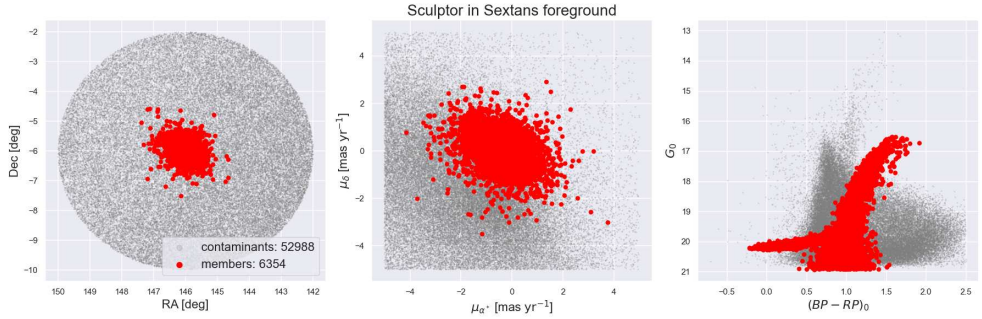
- **Sculptor in Sextans background**: This mock simulates the case of a compact and luminous object like Sculptor embedded in the dense contaminant field characteristic of Sextans. It allows us to assess performance under challenging background conditions with a well-defined galaxy (Fig. 2a). Hereafter we refer to this mock catalog as `SculptorInSextans`.
- **Sextans in Sextans background**: A diffuse system modeled after Sextans, including simulated tidal tails, placed in its own naturally dense background field. This setup closely reproduces the real observational conditions of Sextans (Fig. 2b). Hereafter refer as `SextansInSextans`.
- **Sextans in Draco background**: A Sextans-like system with tidal tails placed in the comparatively cleaner Draco foreground. This case allows us to test how the methods perform when the same diffuse system is observed against a less contaminated background (Fig. 2c). Hereafter we refer to this mock catalog as `SextansInDraco`.
- **Draco in Draco background**: A Draco-like system with a faint tidal feature placed in its native, relatively low-contamination environment. This scenario was used to evaluate performance on low surface brightness galaxies and to test the detectability of less prominent external structures (Fig. 2d). Hereafter we refer to this mock catalog as `DracoInDraco`.

This controlled setup enabled the computation of confusion matrices to quantitatively assess classification performance:

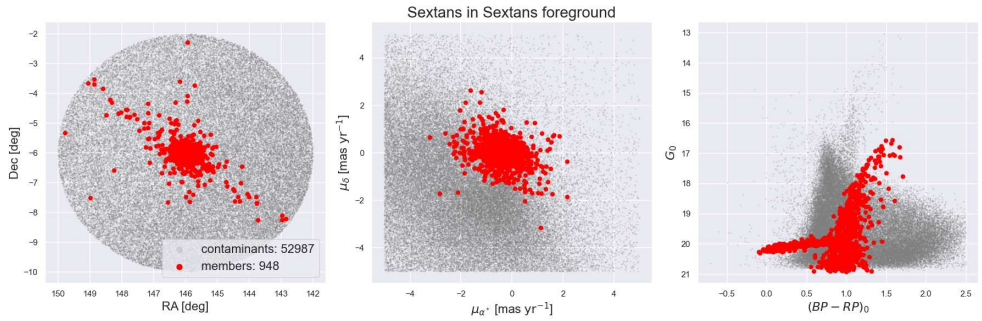
- **True Positives (TP)**: member stars correctly identified as such;
- **False Negatives (FN)**: member stars misclassified as non-members;
- **False Positives (FP)**: non-member (contaminant) stars incorrectly labeled as members;

- **True Negatives (TN)**: contaminant stars correctly identified as non-members.

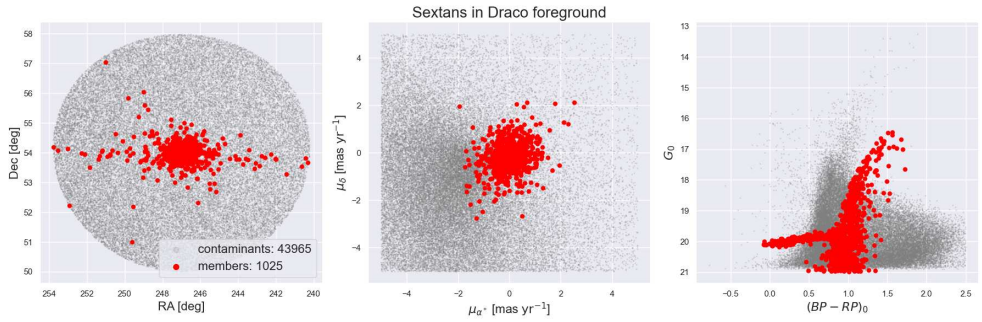
The results obtained for each dataset, both real and mock catalogs, using the various methods presented in this analysis are reported in Sec. 4.



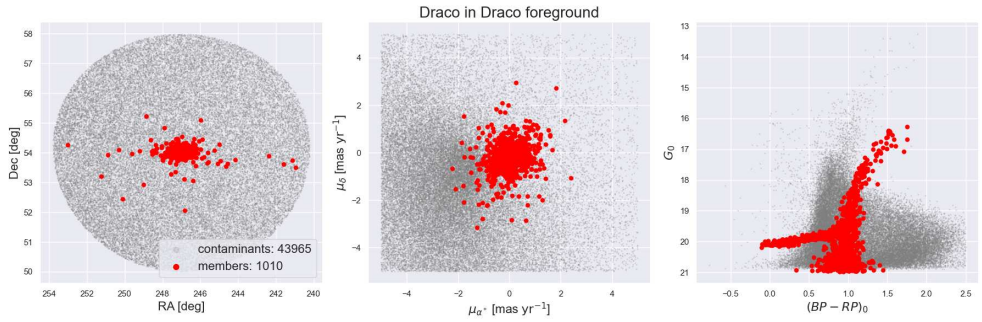
(a)



(b)



(c)



(d)

Figure 2: Mock galaxies preview: For visualization purposes, only sources within 4 degrees of the galaxy center are shown. Each panel reports, in the plot label, the total number of synthetic member stars (red dots) and the number of foreground contaminants (gray dots) within this region for the corresponding mock catalog.
Left: spatial distribution projected on the tangent plane centered on the dwarf galaxy. *Middle*: the proper motion space. *Right*: the color-magnitude diagram.

3 Methods

In this section, we present the various methods employed in this work to compute the probability of membership for stars in dwarf galaxies using Gaia eDR3 data.
The original objective of this thesis was to build upon and extend the methodology proposed in **B22**, with the specific aim of minimizing the use of prior information fed into the model, particularly in order to identify stars in the outer regions of the analyzed galaxies that might otherwise remain undetected due to the influence of strong priors. In particular, we sought to reduce the dependence on foreground density distributions and prior knowledge of the target dwarf galaxy, such as its density profile. For example, in **B22**, an exponential density profile was assumed and calibrated based on previously known properties of each analyzed galaxy.
With this in mind, our goal was to develop a robust and data-driven approach capable of identifying member stars of a dwarf galaxy, especially those located in the outer regions, which are key to probing possible stellar halos or tidal tails. Such features could indicate past interactions with other dwarf galaxies or with the Milky Way itself. To address this challenge, we adopted a machine learning (ML) framework. This allowed us to construct models capable of learning and estimating complex distributions directly from the data. These learned distributions can then be used as likelihood functions within a Bayesian inference framework, replacing theoretical assumptions with data-driven evidence.

In this work, we explore three different methodologies, each offering a distinct approach to the problem:

B22 with a Machine Learning Approach: In this method, we developed a ML model designed to learn how foreground contaminants are distributed across the sky region under analysis. This model helps us predict the distribution of contaminants not only in the outer regions but also in the central part of the image, where the dwarf galaxy lies. It also provides their distribution of the proper motion (PM) space, proper motion in right ascension and declination, and in the color-magnitude diagram (CMD). By learning these distributions, we can effectively model and isolate the contribution of foreground contaminants, enabling a more realistic estimate of the contaminant likelihood term.
We tested three levels of information provided by the ML model, based on how many components of the final likelihood are derived from its predictions:

- **First Level:** At this stage, the model predicts only the contaminant component of the final likelihood (see Eq. 1), while the remaining terms are kept unchanged from the original **B22** approach. The key difference is that in **B22**, the contaminants are assumed to be uniformly distributed across the image, whereas our ML model allows for spatially varying contamination, especially important in regions overlapping with the galaxy.
- **Second Level:** This level builds on the first, with the additional use of the ML model to infer the CMD of the galaxy. Specifically, we compute the galaxy CMD by subtracting the ML-predicted CMD of the contaminants from the observed CMD in the central region. In this setup, only the CMD component of the galaxy is derived from the ML outputs.
- **Third Level:** At this most advanced level, almost all terms in the likelihood are derived from the ML model, except for the galaxy's density profile, which

remains fixed as an exponential profile. Here, we also compute the galaxy's proper motion distribution by subtracting the contaminant proper motion map predicted by the ML from the observed distribution. Moreover, the membership ratio term, representing the relative fraction of member stars to total stars, is estimated from the ML outputs as well.

Through this multi-level ML approach, we aim to evaluate whether incorporating a more realistic and spatially-aware model of the contaminants has a significant impact on the computed membership probabilities for dwarf galaxies in the Local Group.

Dimensionality Reduction Approach: In the previous method, we employed models that are first trained on labeled data and then used to make predictions, a framework known in machine learning as "supervised learning". In this alternative approach, we instead adopt an "unsupervised learning" strategy, where the model does not undergo a traditional training phase with labels.
Here, the goal is to reduce the dimensionality of the input feature space, which includes spatial coordinates, proper motion, CMD information, and parallax, into a lower-dimensional latent space, typically two dimensions. In doing so, the model attempts to group stars with similar features together. This naturally leads to the formation of *clusters* in the reduced space, which ideally correspond to the two main populations in the data: foreground contaminants and dwarf galaxy members.
The clustering itself is performed automatically using a clustering algorithm. The identification of which cluster corresponds to members or contaminants is based on their relative size, under the reasonable assumption that the contaminant population is more numerous than the member population.
To convert this clustering-based binary classification into posterior membership probabilities, we repeat the entire process over multiple synthetic datasets generated via a Monte Carlo approach. Each synthetic dataset is built by perturbing the original data: for each star, its features are randomly drawn from Gaussian distributions centered on the observed values with standard deviations equal to their associated measurement errors. After a large number of such realizations, the final membership probability of a given star is computed as the fraction of times it is classified as a member, yielding an empirical posterior probability estimate.

Normalizing Flow Approach: This method represents the most experimental of the three proposed, as it employs a machine learning architecture not commonly used in this domain. The central idea is to simplify the probability density functions of both the contaminant and member populations, enabling easier construction of likelihoods and more tractable Bayesian inference, essentially, to assess whether a given source is more likely to belong to one distribution or the other.
To achieve this, we adopt a *Normalizing Flow* architecture. As the name implies, these models are designed to transform complex, high-dimensional distributions into simpler ones, typically normal distributions. The objective is to train the Normalizing Flow model to map the complex distribution of the contaminants in the PM and CMD spaces into a multivariate Gaussian with zero mean and identity covariance matrix.
Once trained on a region containing only contaminants, the model learns this transformation exclusively for the contaminant population. When we then apply the model to a region containing both contaminants and dwarf galaxy members, the contaminants will still be mapped to the learned Gaussian in the latent space, whereas the members, having not been seen during training, will be transformed into an unknown,

non-Gaussian distribution.
This setup allows us to isolate the distribution of the member stars in the latent space. We then model this complex distribution using a Gaussian Mixture Model, composed of multiple components, one of which is fixed to match the known Gaussian corresponding to the contaminants. With both distributions now defined, we can apply Bayesian inference to estimate posterior membership probabilities for each source in the catalog.

With these methods briefly introduced, the following sections will present each of them in greater detail. We begin with an overview of the approach proposed in **B22**, to provide the necessary background for understanding the subsequent developments.

3.1 Battaglia 2022

In this section, we provide a comprehensive summary of the methodology used in **B22**, which serves as the foundation for the membership analysis framework used in this work.
The goal is to outline the key steps that lead to the formulation of the membership probability and its associated likelihood components, as applied to stars along the line of sight to LG dwarf galaxies using the Gaia eDR3 catalog.
These membership probabilities are derived using a maximum likelihood approach based on three free parameters: the systemic proper motion of the dwarf galaxy in right ascension ($\mu_{\alpha^*,\mathrm{sys}}$[4]), in declination ($\mu_{\delta,\mathrm{sys}}$), and the fractional contribution of the galaxy's stars relative to the total number of stars in the field, denoted as f_{gal}.
The parameter f_{gal} quantifies the expected fraction of stars belonging to the dwarf galaxy and is determined through the likelihoods associated with each star: L_{gal}, the likelihood of a star being a member of the galaxy, and L_{cont}, the likelihood of it being a contaminant from the background. These likelihoods are computed by modeling the observed stellar population as a mixture of two distinct components: stars belonging to the dwarf galaxy and stars from the foreground. For each star, that falls within the parallax selection previously defined, the likelihood of being a member or a contaminant is evaluated based on its spatial position, proper motion (PM), and location in the color-magnitude diagram (CMD). With these components, the probability that a given star is a member of the dwarf galaxy is computed using the following expression:

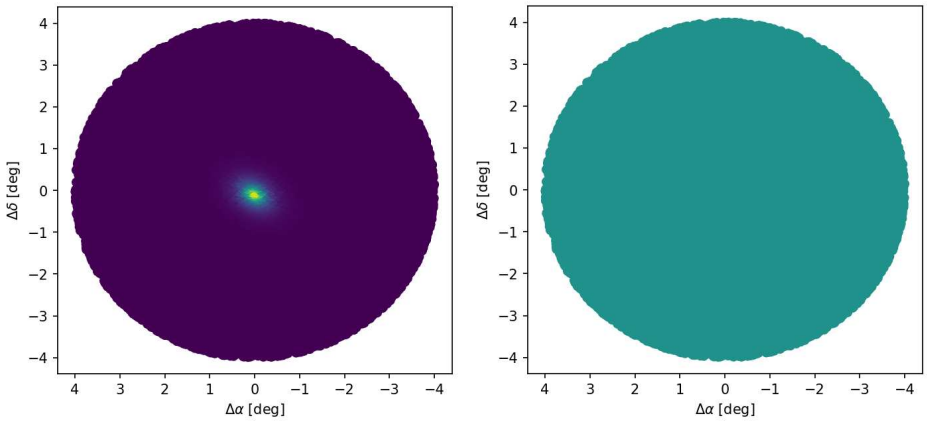
$$P_{\mathrm{gal}} = \frac{f_{\mathrm{gal}} L_{\mathrm{gal}}}{f_{\mathrm{gal}} L_{\mathrm{gal}} + (1 - f_{\mathrm{gal}}) L_{\mathrm{cont}}} \tag{1}$$

The computation of the three main likelihood components is modeled separately for the dwarf galaxy members and the contaminants, employing distinct methodologies tailored to capture the expected differences between the two populations.
In the following paragraph, we describe the construction of each likelihood term in detail.

3.1.1 Spatial distribution

The spatial distribution likelihood term is computed differently for L_{gal} and L_{cont}.
For the contaminant population, a uniform spatial distribution (i.e. a constant surface

[4]In the *Gaia* eDR3 catalog, the proper motion in right ascension is provided in its declination-corrected form, such that $\mu_{\alpha^*} = \mu_\alpha \cos\delta$.



(a) **(b)**

Figure 3: On the left (3a) is the spatial distribution likelihood term of the `SextansInSextans` dwarf galaxy calculate the co-addiction of the 1000 Monte Carlo realizations of the expected 2D surface number density at a given location on the sky; on the right (3b) is the uniform contamination level of the contaminants over the sky. These two plots are output produced by the code used in **B22**

number density of contaminant stars) is assumed across the entire analyzed field. In contrast, the spatial distribution of the dwarf galaxy member stars is assumed by an exponential profile and is modeled through a Monte Carlo-based approach: a two-dimensional lookup map of the surface number density is generated by co-adding 1000 Monte Carlo realizations of the galaxy's exponential profile, and subsequently normalizing the result to represent the expected density distribution. Each realization samples the structural parameters of the galaxy (position angle, half-light radius, and ellipticity) from Gaussian distributions centered on the values reported in Tab. 1, with dispersion equal to the average of the lower and upper respective uncertainties. To prevent unphysical values, the distributions for half-light radius and ellipticity are truncated to exclude negative values.
The final spatial likelihood map for galaxy members is constructed by interpolating the stellar positions from the Gaia dataset onto the exponential profile obtained through this procedure. This map defines the spatial likelihood component $L_{\rm gal}$ used in the membership probability calculation.

3.1.2 Proper motion distribution

The proper motion distribution is also modeled differently for contaminants and for dwarf galaxy members.
For the contaminant population, a 2D histogram is built by dividing the PM space into bins along both μ_{α^*} and μ_δ, using a bin size of $\Delta\mu = 0.2\,\mathrm{mas\,yr^{-1}}$. The histogram is then normalized, and a lookup map is constructed by interpolating the histogram over a grid of predefined points, with each point corresponding to the center of a bin in the 2D histogram.
Details regarding the construction of the 2D histograms are provided in Sec. 3.2.
The contaminant subset used to generate the lookup map is selected from a region where the projected distance from the galaxy center exceeds 5 half-light radii (see Fig. 4). For simplicity, this outer region is hereafter referred to as the "*donut region*," as it will be referenced in the description of other methods.

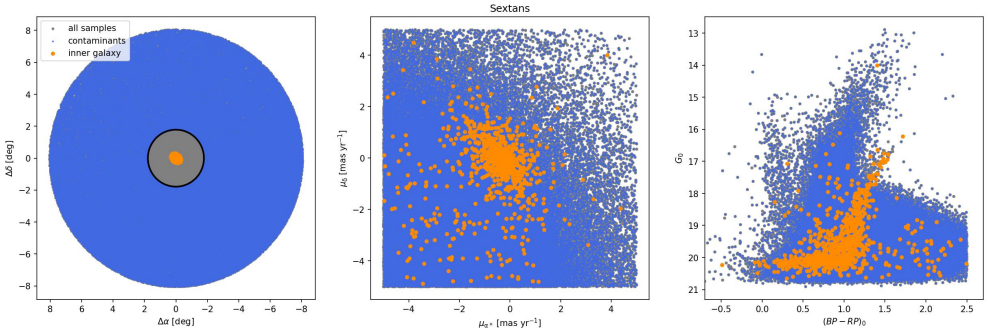


Figure 4: Example of contaminants and inner galaxy subdivision: The figure illustrates the division between sources likely to be contaminants (blue), defined as stars located beyond 5× the half-light radius, and those likely belonging to the inner galaxy region (orange), within 1× the half-light radius. Stars in the intermediate region (gray) are the remaining sources.
Left: spatial distribution of contaminants (blue) and inner members (orange) in the sky frame. *Middle*: corresponding division in proper motion space. *Right*: corresponding division in the CMD.

In contrast, modeling the proper motion distribution for dwarf galaxy members is an important aspect of the method described in **B22**, where one of the main objectives was to measure the systemic proper motion of LG dwarf galaxies.
In this case, the member stars' proper motion distribution is modeled as a bivariate Gaussian, incorporating the correlation between the proper motion components $\mu_{\alpha*}$ and μ_δ of individual stars. The parameter estimation is performed using the `PyMultiNest`[5] package, a Bayesian inference tool that enables both parameter estimation and model selection.
The output consists of the best-fit systemic proper motion parameters and the fractional parameter $f_{\rm gal}$, both of which are used to build the corresponding likelihood term with priors already incorporated.

3.1.3 Color-magnitude distribution

The likelihood terms associated with the CMD for both contaminants and dwarf galaxy members are constructed following the same procedure adopted for the contaminants' proper motion lookup map.
For the dwarf galaxy members, however, the subset of stars used to construct the CMD likelihood is selected within a projected distance of 1 half-light radius from the galaxy center (orange region in Fig. 4). Within this central region, the majority of stars are expected to be genuine members, with only a minor contamination from foreground sources. This approach is appropriate for relatively bright dwarf spheroidal galaxies, as shown in **B22**, where the CMD shape can be reliably recovered with such a selection. However, for ultra-faint dwarf (UFD) galaxies, a different strategy, based on the creation of synthetic CMD in the Gaia photometric filters with stellar evolution models, is adopted since their low stellar density prevents retrieving a well-defined CMD shape through a simple radial cut.
This approach ensures that the CMD distribution of the dwarf galaxy members is dominated by true members, thus minimizing the impact of contaminants on the likelihood modeling.

[5] `https://github.com/JohannesBuchner/PyMultiNest`

At this stage, all three likelihood components required to compute the membership probabilities for each source in the Gaia sky frame have been calculated. We define $L_{\rm gal}$ and $L_{\rm cont}$ as the product of the spatial, proper motion, and CMD likelihood terms for the dwarf galaxy members and contaminants, respectively. By applying Eq. 1, it is then possible to determine, for each star, the probability of being a member of the dwarf galaxy or belonging to the background contamination.

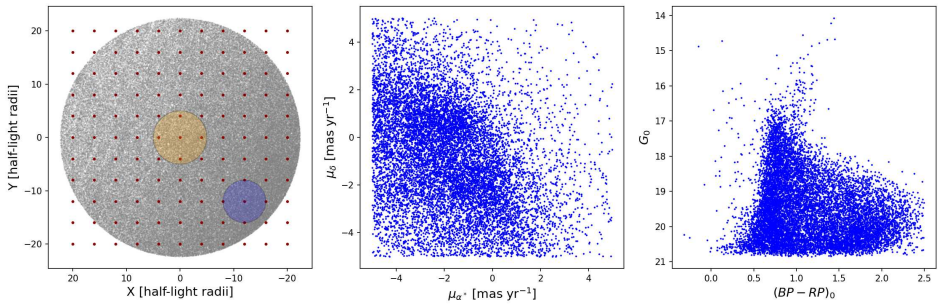
From this perspective, we aim to investigate how strongly the distribution of contaminants influences the final results. Specifically, we ask whether it is sufficient to assume a uniform spatial distribution with constant proper motion and CMD properties across the sky, or whether accounting for their variations in different regions leads to a more reliable identification of contaminants and, consequently, of genuine member stars.

3.2 B22 with ML

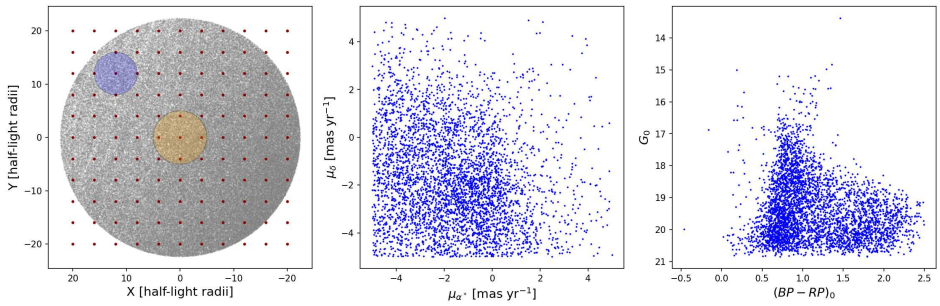
In this section, we present the first method investigated in this work, which constitutes an enhanced version of the framework introduced in Sec. 3.1 for computing membership probabilities of sources in the Gaia eDR3-based mock catalogs. The principal improvement lies in the modeling of the contaminant distribution, which is here derived using a machine learning approach.
A key objective of this modification is to reduce the reliance on strong priors, particularly those imposed on the contaminant distribution, such as assuming a uniform spatial density. Instead, the model is allowed to learn the true contaminant density distribution directly from the data across the entire field of view, including the region surrounding the dwarf galaxy.
This more flexible approach enables a more accurate reconstruction of the spatial distribution of contaminants, as well as their expected PM and CMD features within the same region. With this approach is even possible to reconstruct the features of the dwarf galaxy in the PM and CMD planes and assess whether this yields to better results then the original method.
To achieve this, we employ a neural network trained to predict the morphology and density of the CMD and PM distribution across different regions of the sky. This data-driven strategy enables the construction of a prior-free model for the contaminant population density.

3.2.1 Data preparation

Following the data reduction procedures and quality cuts described in Sec. 2, and after transforming the projected stellar coordinates from degrees to units of the galaxy's half-light radius ($R_{1/2}$), the sky region was partitioned into a uniform `N×N` grid.
Each grid point, hereafter referred to as a "*sky pixel*", represents a circular region with a radius of 2 $R_{1/2}$, and grid points are spaced by 4 $R_{1/2}$ in both directions.
Once the sky pixel coordinates were defined, the dataset was divided into a training set and a test set, which were then used for model training and making prediction, respectively. The test set was selected as the central circular region of the field, with a radius of 5 $R_{1/2}$ (orange area in Fig. 5). The remaining sky pixels were assigned to the training set, with the additional constraint that the area covered by a pixel must



(a)



(b)

Figure 5: Example of sky pixels in the training region: Two regions of the projected sky within the Sextans foreground contamination (see Tab. 1), taken from the `SextansInSextans` mock catalog by selecting only the source with `source_id` $\neq 42$.
Fig. 5a (*upper left*) shows a pixel located in a highly contaminated region, while Fig. 5b (*lower left*) shows one in a less contaminated area. In both figures, the **test region** is highlighted in orange.
Middle: proper motion distribution of the stars within the blue-selected sky pixel on the *left* image. *Right*: CMD of the same stellar selection.

lie entirely within the training region. This constraint was implemented through the following condition:

$$R_{\text{test}} + R_{\text{pix}} \leq d_{\text{pix}} \leq R_{\text{sky}} - R_{\text{pix}} \tag{2}$$

where R_{test} is the radius of the test region, R_{pix} the pixel radius (4 $R_{1/2}$), d_{pix} the distance from the pixel center to the galaxy center, and R_{sky} the total radius of the sky image.
Since this method employs a machine learning approach, having a large training region improves the overall performance of the algorithm. For this reason, we used a wide sky area around each analyzed dwarf galaxy, with a radius of 8 degrees. These images are large enough to contain a substantial number of stars for training, while also including a test region that encompasses the majority of the dwarf galaxy.
It is important to note that the spacing of the sky pixels leads to overlap between individual pixels. This results in a smoothing effect on the estimated density distribution across the sky, and also implies that the pixels are not statistically independent from one another.

For each sky pixel in the training set, we selected all the sources located within its circular region and for those stars, we considered their $\mu_{\alpha,*}$, μ_δ and the magnitude in the G-band and BP-RP color.
As shown in Fig. 5, when comparing two different sky pixels located on opposite sides of the field, we observe distinct differences in both their PM and CMD distributions. This clearly highlights the non-uniformity of the contaminant population across the sky.
For each sky pixel, two separate two-dimensional histograms were generated: one for the PM distribution and one for the CMD. In both cases, the x-axis (PM in right ascension for the former, and color for the latter) and the y-axis (PM in declination and magnitude, respectively) were discretized into 50 equally spaced bins. For the proper motion distributions, both axes span the range [−5.4, 5.4] mas/yr, resulting in a bin size of approximately 0.2 mas/yr. For the CMD, the color spans the range [−0.7, 2.6] with a bin width of about 0.06 mag, while the magnitude ranges from [13.1, 21.2] with bins of roughly 0.16 mag.
Each bin thus represents a rectangular region and the value assigned to the bin is the count of stars that fall in. To mitigate the effects of shot noise and improve the learning performance, we applied a 2D convolutional smoothing kernel of width equal to 3×3 pixels[6] to each histogram. A boxcar kernel was used for the CMD histograms, while a Gaussian kernel was applied to the PM histograms, the latter yielding better empirical performance during training. Any occasional `NaN` values were set to zero.
After generating the smoothed 2D histograms for each training pixel, the data were serialized into a dataset where each row corresponds to the coordinates of the corresponding sky pixel, the 2D coordinates of the current histogram bin (proper motion or CMD), and its value. We can see an example of how the dataset is structured in Tab. 2. In the final dataset, the first four columns represent the input features ($\mathbf{X}$), while the fifth column serves as the target label ($\mathbf{y}$)[7].

[6]Here the 2D histogram is seen as a image so each square bin corresponds to a pixel.

[7]In Machine Learning, particularly in a supervised learning task, $\mathbf{X}$ denotes the input features used by the learning model, typically representing the observed data, while $\mathbf{y}$ corresponds to the target values on which the model is trained to make predictions.

Table 2: Example of a training dataset: $\star$ corresponds to number of stars per square bin:

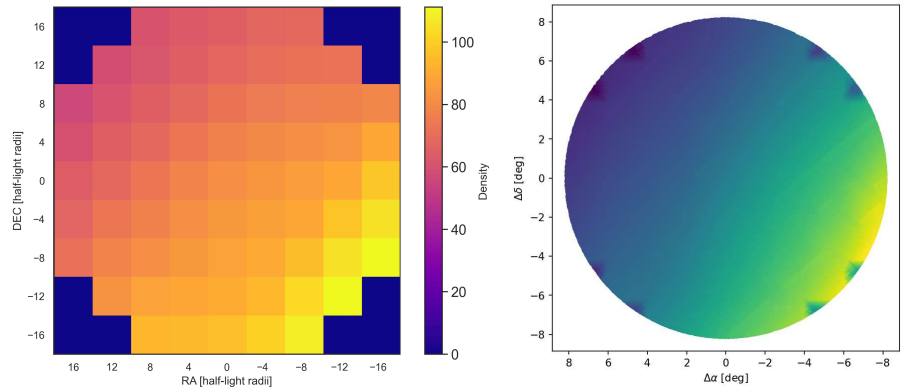
Sky x-coord [$R_{1/2}$]	Sky y-coord [$R_{1/2}$]	hist x-bin [mag]	hist y-bin [mag]	bin value [$\star$]
-16.0	-8.0	-0.67	13.18	0.0
-18.0	-8.0	-0.60	13.18	0.0
...	...	...	...	...
-4.0	-12.0	0.81	20.62	24.67
-4.0	-12.0	0.88	20.62	20.22
-4.0	-12.0	0.95	20.62	19.44
...	...	...	...	...
8.0	16.0	0.95	18.14	12.33
8.0	16.0	1.02	18.14	9.44
8.0	16.0	1.08	18.14	5.89
...	...	...	...	...
16.0	8.0	2.50	21.12	0.0
16.0	8.0	2.57	21.12	0.0

3.2.2 Training the model

Following the preparation of the training dataset, an analogous dataset was generated for the test region, using the same procedure applied to the training sky pixels. The only difference in this case is that the test region consists exclusively of the central sky pixel, located at coordinates (0,0), which corresponds to the expected position of the dwarf galaxy.
The model that was selected for implementation is a Multilayer Perceptron (MLP). This model features a slight structural variation for the cases of CMD or PM analysis and it utilizes Mean Squared Error loss.
Once the model has been trained on the training set, it can be applied to the test pixel to generate predictions. Importantly, providing the model solely with the input features $\mathbf{X}$ of the test dataset, the model will generate an output for the target variable $\mathbf{y}$ under the implicit assumption that the test region should contain only contaminants. This effectively yields a model-predicted 2D histogram representing the expected distribution of contaminants, in the absence of any galaxy signal.

3.2.3 Making prediction

The ability to predict the morphology and number density of the CMD and PM distribution within the central sky pixel enables the construction of a spatial map of the expected contaminant density distribution (see Fig. 6). In this figure, each square represents an individual sky pixel, while the corresponding value denotes the total number of stars predicted by the model within its circular region divide by the pixel area. The total number of predicted stars in each pixel is obtained by summing all the values of the 2D histogram bins, in this case, from the CMD histogram.
This density map is generated using the full dataset, without differentiating between the training and test sets. As such, the resulting contaminant distribution is entirely model-driven, reflecting the spatial variation in stellar density across the field as inferred from the trained network. Predictions over the training region are highly accurate, given that the model was explicitly optimized on this data, resulting in a close match between the predicted and true star counts. It is important to emphasize that the central region, coincident with the expected location of the dwarf galaxy, is also affected by foreground and background contamination, so establishing a reliable model-based estimate of the contaminant contribution in this region is therefore essential.



(a) **(b)**

Figure 6: Predicted Density map of the contaminants: in Fig. 6a each square correspond to an analyzed pixel during the train procedure of `SextansInSextans` and their value is the density of the contaminants predicted by the model. The density correspond to the total number of star inside that pixel divide by its circular area calculated in half-light radii. In this case correspond to an area of $\pi(4R_{1/2})^2$ where for the case of `SextansInSextans` is $R_{1/2} = 0.36$. Fig. 6b correspond to the normalized and interpolate contaminants density map used in the calculation of the likelihood.

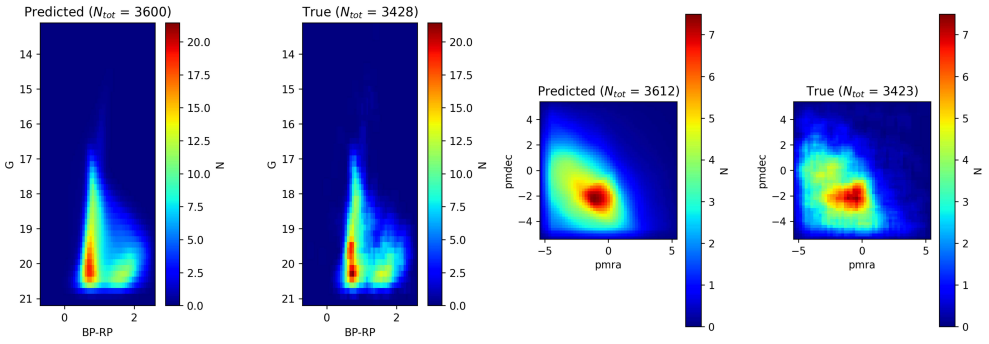


Figure 7: Predicted vs. true 2D histograms: The example figure compares the predicted and true only contaminants 2D histograms for the test region for the Sculptor mock catalog. The first two panels show the CMD, while the last two correspond to the PM distribution. The total number of stars, computed as the sum of all histogram bin values and rounded to the nearest integer, is displayed at the top of each panel. Discrepancies on the total number of stars between the predicted and true histograms are due to rounding and the kernel-smoothing applied during histogram construction.

So, after generating the predictions for the test region, we proceed to construct the likelihood terms used in Eq. 1, as previously described in Sections 3.1.2 and 3.1.3. The specific formulation of these likelihood terms varies depending on the level of model integration described in Sec. 3. Each level reflects the extent to which information provided by the machine learning model is incorporated in the construction of the final likelihoods, specifically which 2D histograms are used to derive the contaminant and galaxy look-up maps.

First Level: In this initial approach, we incorporate the outputs of the ML model exclusively for the contaminant terms in Eq. 1. As discussed earlier, the model predicts 2D histograms for both the PM and CMD across all *sky pixels* (see Fig.7). During the interpolation step used to generate the look-up maps, we select the predicted histogram associated with the sky pixel closest to the position of the star being evaluated. This means that, unlike the original **B22** methods where a single contaminant distribution is assumed, our approach leverages multiple histograms that reflect the spatial variability of the contaminant population. Consequently, the contaminant look-up maps for both PM and CMD inherently account for the local spatial structure of the stellar field.
For the spatial term of the contaminant likelihood, we use the interpolated density map shown in Fig.,6b.

Second Level: This level is similar to the first one, with a slight but important difference: in this case, we construct the CMD look-up map of the galaxy differently. In **B22**, the CMD shape was derived by selecting the elliptical region within 1 R_e (see the orange region in Fig. 4). However, this approach includes a significant number of contaminants, which can distort the true CMD morphology of the galaxy.
To mitigate this, we instead compute the galaxy CMD by subtracting the predicted CMD 2D histogram of the central sky pixel (representing only the contaminants) from the true 2D histogram of the same region. The resulting residual histogram, that should reflects more accurately the galaxy's CMD, is then used to construct the CMD look-up map as usual.
In Fig. 8, we show two examples of CMDs obtained using this subtraction method. The results clearly depend on the level of contamination and therefore on the performance of our model. In Fig. 8a, which corresponds to the Sextans example within a dense background, the high contamination significantly affects the final result. In contrast, in Fig. 8b, where the galaxy is more densely populated, the subtraction is more reliable and the CMD is better preserved. This highlights that both the intrinsic properties of the galaxy and the contamination level in its region are crucial factors influencing the final outcome.

Third Level: This final level builds upon the procedures of the first and second levels, with the key difference that here the galaxy PM look-up map is constructed in the same way as the galaxy CMD map in the second level, by subtracting the predicted 2D histogram of the contaminants from the true 2D histogram of the central sky pixel. As a result, we no longer rely on the statistical modeling described in Sec. 3.1.2 for the galaxy component, since both the CMD and PM maps are now directly obtained through this subtraction approach.
Furthermore, as illustrated in Fig. 7, the total number of stars represented in a histogram can be obtained by summing the values of all bins, since each bin contains

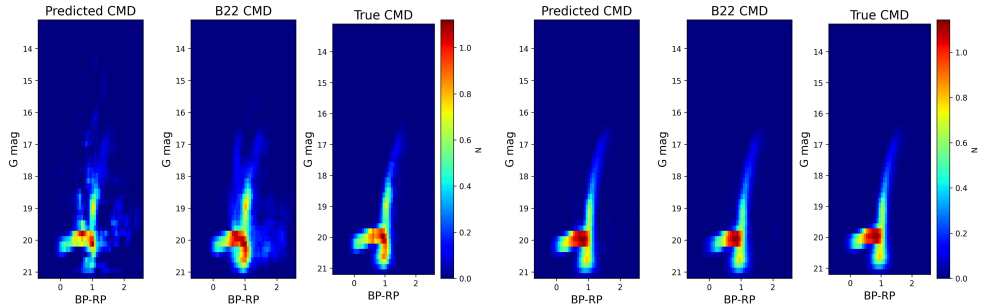


Figure 8: Predicted CMD - B22 CMD - True CMD comparison: In the example figures are shown three different CMD 2D histograms obtained with the ML model by the subtraction of the true histogram of the test region and the predicted one, resulting in only the CMD of the galaxy. The CMD galaxy histogram obtained with the method presented in **B22** and for completeness the actual true CMD of the mock galaxy. The Fig. 8a shows the comparison for the Sextans mock catalog meanwhile Fig. 8b is the example for the Sculptor mock catalog.

the count of stars falling within that interval. This allows us, through the predicted histograms as in Fig. 8, to also estimate the number of stars belonging to the dwarf galaxy, enabling the calculation of the ratio $f_{\rm gal}$ in Eq. 1.
Overall, this level relies almost entirely on outputs from the ML model, with the only exception being the spatial distribution of the galaxy, which continues to follow the exponential profile described in Sec. 3.1.1.

3.3 Dimensional Reduction

We also investigated alternative, prior-free approaches to calculating membership probabilities. In these methods, no assumptions are made about the spatial distribution, proper motions, or CMD of either the contaminants or the galaxy members. The only distinction made is between regions of the sky expected to contain only contaminants and those that include both the dwarf galaxy and contaminants.
As a first test, we employed dimensionality reduction algorithms, specifically using the `UMAP`[8] Python package.
The underlying idea of this method is intuitive: given a dataset with n features, each feature can be interpreted as a separate dimension in the n-dimensional space $\mathbb{R}^n$. Dimensionality reduction techniques aim to project this high-dimensional data into a lower-dimensional space, typically two or three dimensions, while preserving as much of the original data structure as possible. In this reduced space, sources with similar physical properties tend to cluster together, making it easier to identify groups with common characteristics. This projection not only facilitates visualization but also enables the application of clustering algorithms to perform classification tasks.

For this study, we focused on a set of features where physical separations between dwarf galaxy members and background sources are most evident: spatial coordinates, proper motions, color-magnitude information, and parallax. In Sec. 4.2, we will provide a detailed discussion of how these features were used and how the relationships among them were exploited for member identification.

[8] `https://umap-learn.readthedocs.io/en/latest/`

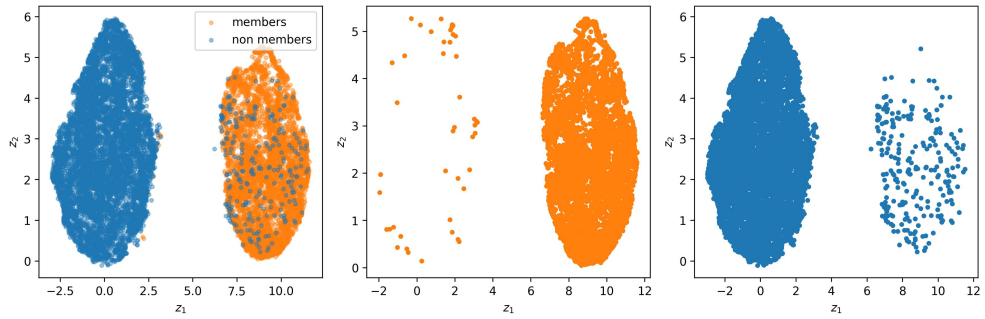


Figure 9: Example result of dimensionality reduction using UMAP: The plots show the 2D latent space produced by UMAP for the Sculptor-like mock galaxy. The blue points corresponds to contaminants stars, while the orange points represents the actual members of the mock galaxy.

3.3.1 Data Preparation

First, it is important to clarify that the dimensionality reduction method, due to UMAP's memory and computational limitations, cannot be applied directly to very large catalogs, as it fails to transform such datasets into the latent space. To overcome this, we restricted our catalog to a 4-degree region and excluded a fixed number of known contaminants (45,000 for the mocks with the Sextans foreground and 37,000 for those with Draco foreground), thereby maintaining a significant number of foreground stars while keeping all member stars. For `SculptorInSextans` we remain with 14,342 sources, 8,937 for `SextansInSextans`, 7,990 for `SextansInDraco` and 7,975 for `DracoInDraco`.
This setup was adopted purely for technical reasons, with the goal of evaluating whether the dimensionality reduction method can distinguish members based on shared features that differ from those of the contaminants. The 4-degree region was chosen both to allow a direct comparison with the other methods described in Section 3.2 and Section 3.4, since for all the three approaches this region was chosen to display the final results on membership probabilities (see Appendix). Moreover, the mock catalogs were constructed such that the synthetic galaxies extend up to 4 degrees, thereby ensuring that all members are included.
Once the dataset containing the selected features was obtained, we applied a *Min-Max* scaler to rescale all features to the [0,1] range. This step optimizes the learning process by placing all features on comparable scales while preserving their relative distances. As a result, the performance of both dimensionality reduction and clustering is improved without distorting the underlying structure of the data.

3.3.2 Single UMAP application

After preparing the dataset, we applied the UMAP dimensionality reduction algorithm, setting the number of output components to 2. This choice was made because the primary goal is binary classification, and working in a 2D Cartesian plane allows for easier interpretation and visualization. In particular, it helps to visually identify the presence of clusters, where sources with similar properties are grouped closely together.
We adopted a relatively high number of `neighbors`, fixing the value at 50. This parameter in UMAP controls the balance between local and global structure preservation: lower values emphasize local clustering at the expense of global organization,

while higher values prioritize maintaining the broader manifold structure. The minimum distance parameter, which governs how closely UMAP is allowed to pack points together in the lower-dimensional space, was set to 0.0 to favor tight groupings, as our primary interest lies in identifying distinct clusters corresponding to galaxy members and contaminants.
The resulting projection, shown in Fig. 9 for the case of the Sculptor mock galaxy, reveals a clear separation between the two populations, facilitating classification through clustering algorithms. We employed a hierarchical clustering approach, using the `linkage` and `fcluster` functions from the `scipy.cluster.hierarchy` package, to partition the UMAP latent space $\mathbf{Z}$ into exactly two clusters. The larger cluster was associated with the field contaminants, while the smaller cluster was assigned to likely galaxy members. This choice is motivated by the expectation that background contaminants dominate the stellar population in the observed field. We opted for a hierarchical clustering method, rather than centroid-based algorithms like `k-means` or Gaussian Mixture Models, because the shape of the two clusters is not necessarily spherical or regular (as will be further discussed in Sec. 4.2).

3.3.3 Multiple UMAP application

As shown in Fig. 9, some contaminants are inevitably grouped within the cluster of members, and vice versa. This misclassification naturally arises from the overlap in observable properties between certain contaminants and true members, which complicates a clean separation. Moreover, a single application of UMAP does not inherently provide a probabilistic assessment of membership.
To address both the issue of misclassification and the lack of probabilistic outputs, we repeated the UMAP-clustering procedure on 500 synthetic realizations of the original catalog. Each realization was generated by perturbing the measured quantities according to their associated uncertainties, assuming Gaussian errors.
The final membership probability for each star was defined as the number of realizations (out of 500) in which it was classified as a member, divided by the total number of realizations.
The choice of 500 realizations was guided by computational efficiency and performance constraints of the CPU used for this analysis. Although this number could be increased to achieve higher resolution in the posterior, we found that 500 realizations were sufficient to recover a stable and informative posterior probability distribution for all sources in the catalog.

3.4 Normalizing Flow

In parallel with the dimensionality reduction approach discussed in Sec. 3.3, we explored an alternative fully prior-free method based on the framework of *Normalizing Flows* (NFs).
NFs are a class of unsupervised learning algorithms commonly used in machine learning tasks such as generative modeling, density estimation, and variational inference. They have found widespread application in fields like image and audio synthesis, dataset compression, and noise modeling.
The core idea behind NFs is to construct a series of bijective (invertible and differentiable) transformations that map a simple, known probability distribution, typically

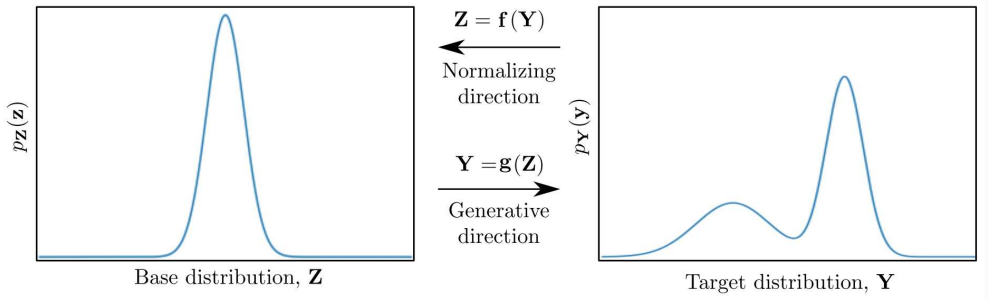


Figure 10: Example plot of the change of variables in Eq. 3 from Kobyzev et al. (2020)

a multivariate Gaussian, into a complex data distribution, and vice versa.

In the context of our analysis, we leverage this property not to generate data, but to simplify it. Specifically, our goal is to learn an invertible transformation that maps the complex, unknown distribution of the contaminants, encoded in PM and CMD space, into a standard Gaussian distribution centered at zero with unit variance.
After training the normalizing flow model on a large sample of contaminants, selection that includes sources more than 4 degrees away from the dwarf optical center (similar to Fig. 4), we apply the learned transformation to the circular region within 4 degrees around the dwarf galaxy. In this region, contaminants are mapped to a standard Gaussian distribution by construction, while the member stars, unseen during training, are transformed into a complex, non-Gaussian distribution in the latent space. As a result, the latent representation in this region consists of a mixture of the known Gaussian background and an unknown structured signal, corresponding to the dwarf galaxy.
This transformation enables a simplification of the likelihood modeling step since we can apply Bayesian inference more effectively and robustly by operating in the transformed latent space. This avoids the complications of modeling likelihoods directly in the original, highly non-Gaussian space.

3.4.1 Basic logic of NFs

To better understand how a NFs architecture operates, let us consider a random variable $\mathbf{Z}$ in $\mathbb{R}^D$ with a known and tractable probability density function (PDF) $p_{\mathbf{Z}} : \mathbb{R}^D \to \mathbb{R}$.
We define an invertible and differentiable function $\mathbf{g}$ that maps $\mathbf{Z}$ to another variable $\mathbf{Y}$ through the transformation $\mathbf{Y} = \mathbf{g}(\mathbf{Z})$ and, denoting $\mathbf{f} = \mathbf{g}^{-1}$ the inverse function such that $\mathbf{Z} = \mathbf{f}(\mathbf{Y})$, we can express the PDF of $\mathbf{Y}$ using the change-of-variables of probability density functions formula as:

$$\begin{aligned} p_{\mathbf{Y}}(y) &= p_{\mathbf{Z}}(\mathbf{f}(y)) \left|\det \mathbf{Df}(\mathrm{y})\right| \\ &= p_{\mathbf{Z}}(\mathbf{f}(y)) \left|\det \mathbf{Dg}(\mathbf{f}(\mathrm{y}))\right|^{-1} \\ &= p_{\mathbf{Z}}(\mathbf{f}(y)) \left|\det \mathbf{Dg}(\mathbf{z})\right|^{-1} \end{aligned} \tag{3}$$

where $\mathbf{Df} = \frac{\partial \mathbf{f}}{\partial y}$ is the Jacobian of $\mathbf{f}$ and $\mathbf{Dg} = \frac{\partial \mathbf{g}}{\partial z}$ is the Jacobian of $\mathbf{g}$

$$
\begin{aligned}
p_{\mathbf{Y}}(y) &= p_{\mathbf{Z}}(z)\,|\det(\frac{\partial \mathrm{z}}{\partial \mathrm{y}})| \\
&= p_{\mathbf{Z}}(\mathbf{g}^{-1}(y))\,|\det(\frac{\partial \mathbf{g}^{-1}(\mathrm{y})}{\partial \mathrm{y}})| \\
&= p_{\mathbf{Z}}(\mathbf{f}(y))\,|\det(\frac{\partial \mathbf{f}(\mathrm{y})}{\partial \mathrm{y}})|
\end{aligned} \tag{4}
$$

where $\det(\frac{\partial \mathbf{f}(\mathrm{y})}{\partial y})$ is the determinant of the Jacobian matrix.
The resulting PDF $p_{\mathbf{Y}}(y)$ is referred to as the *push-forward* of the base density $p_{\mathbf{Z}}$. In the context of generative models, the function $\mathbf{g}$ "pushes forward" the simple base distribution $p_{\mathbf{Z}}$ (e.g., a multivariate Gaussian) into a more complex data distribution. This transformation direction is known as the *generative direction*.

Conversely, the inverse mapping $\mathbf{f}$ transforms complex observed data $\mathbf{Y}$ back into the simple latent space described by $\mathbf{Z}$. This inverse flow is referred to as the *normalizing direction*, if the base measure $p_{\mathbf{Z}}$ is chosen as a normal distribution, and it is this property that gives rise to the name "Normalizing Flows".

In the Normalizing Flows framework, the bijective transformations used must fulfill three key requirements: they should be computationally efficient, analytically invertible, and permit tractable evaluation of the Jacobian determinant.
A common approach is to construct a complex transformation as a composition of simpler invertible functions $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_N$, remains bijective and their Jacobians follow a structured form. Specifically, if we define the composite transformation as

$$
\mathbf{g} = \mathbf{g_N} \circ \mathbf{g_{N-1}} \circ \cdots \circ \mathbf{g_1}
$$

then its inverse becomes

$$
\mathbf{f} = \mathbf{f_1} \circ \mathbf{f_2} \circ \cdots \circ \mathbf{f_N}
$$

and the determinant of the Jacobian is

$$
\det(\frac{\partial \mathrm{f}(\mathrm{y})}{\partial \mathrm{y}}) = \prod_{\mathrm{i}=1}^{\mathrm{N}} \det(\frac{\partial \mathrm{f_i}(\mathrm{y_i})}{\partial \mathrm{y_i}}) \tag{5}
$$

A practical way to construct the bijective transformation $\mathbf{g}$ is by leveraging the change-of-variables formula to compute the log-likelihood of samples $\mathbf{Y} = \mathbf{g}(\mathbf{Z})$, given a known base distribution $p_{\mathbf{Z}}$ and the inverse transformation $\mathbf{f} = \mathbf{g}^{-1}$ (see Eq. 4). This formulation enables direct training of the model using the log-likelihood as the objective function.
Given a sample $\mathbf{y}$ drawn from a complex, unknown distribution $p_{\mathbf{Y}}$, and a neural network parameterized by θ that maps $\mathbf{y}$ to the latent space $\mathbf{Z}$, we express the log-likelihood as:

$$
\begin{aligned}
\log p_{\mathbf{Y}}(y) &= \log\left(p_{\mathbf{Z}}(\mathbf{f}(y)) \cdot \left|\det\left(\frac{\partial \mathbf{f}(y)}{\partial y}\right)\right|\right) \\
&= \log p_{\mathbf{Z}}(\mathbf{f}(y)) + \log\left|\det\left(\frac{\partial \mathbf{f}(y)}{\partial y}\right)\right|
\end{aligned} \tag{6}
$$

Assuming a standard multivariate Gaussian prior for $p_{\mathbf{Z}}$, with independent components of zero mean and unit variance, the expression simplifies to:

$$
\log p_{\mathbf{Y}}(y) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}|\mathbf{f}(y)|^2 + \log\left|\det\left(\frac{\partial \mathbf{f}(y)}{\partial y}\right)\right| \tag{7}
$$

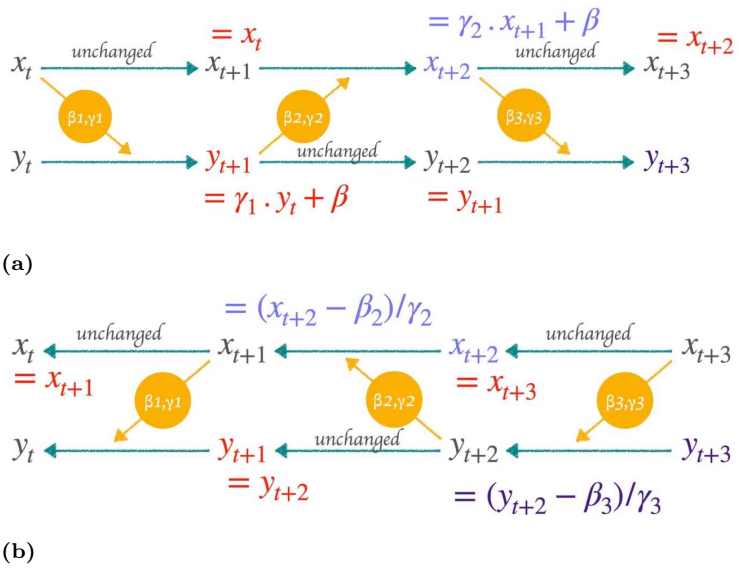


Figure 11: Scheme of the workflow of Real NVP throught the coupling layers. Fig. 11a illustrates the forward transformation, corresponding to the *normalizing direction* of the NFs, through coupling layers in the Real NVP architecture, while Fig. 11b shows the corresponding inverse (*push-forward*) transformation. The subscript t denotes the transformation step, which more precisely corresponds to the output after a single coupling layer. Images from the YouTube channel Tiny Volt

3.4.2 Real NVP

We have seen that NFs are a class of generative models that transform a simple base distribution into a more complex target distribution through a sequence of invertible and differentiable transformations. Several architectures have been developed to construct flow-based models, including the Masked Autoregressive Flow (MAF) and the Coupling and Autoregressive Rational-Quadratic Spline Flows (C-RQS and A-RQS), (Coccaro et al. 2023). In this work, we adopt the Real NVP (Real-valued Non-Volume Preserving) architecture, as it is among the most widely studied in the literature and offers a more straightforward and accessible implementation.

The key feature of the Real NVP architecture is its use of a specific class of bijective transformations known as *affine coupling layers*. These transformations are designed to be computationally efficient, easily invertible, and allow for straightforward computation of the Jacobian determinant, essential for maximum-likelihood training in normalizing flows.
In each affine coupling layer, the input vector is split into two parts: one part remains unchanged, while the other undergoes an element-wise affine transformation, commonly referred to as a *shift-and-scale* operation. The parameters of this transformation are predicted by neural networks conditioned on the unchanged part of the input.
In the subsequent coupling layer, the roles of the two parts are reversed: the previously transformed part is left unchanged, and the other is now transformed.
This alternating process continues across all coupling layers.

To illustrate the mechanism, consider a simple 2D example with input data (x_0, y_0). In Fig. 11a show the steps between each coupling layer to calculate the final output values (x_{t+3}, y_{t+3}). In the first coupling layer, x_0 remains unchanged and is used as

input to a neural network to compute the affine transformation parameters β_1 (shift) and γ_1 (scale) for y_0:

$$\begin{aligned} x_t &= x_0 \\ y_t &= \beta_1 + \gamma_1 \cdot y_0 \end{aligned} \tag{8}$$

In the next layer, the roles are swapped: y_t remains unchanged, and is used to compute new parameters β_2 and γ_2, which are applied to x_t:

$$\begin{aligned} x_{t+1} &= \beta_2 + \gamma_2 \cdot x_t \\ y_{t+1} &= y_t \end{aligned}$$

This alternating transformation continues across the full sequence of coupling layers. The functions β and γ can be any sufficiently expressive nonlinear functions, but are typically modeled using neural networks (NNs). The input to each NN is controlled via a *masking* mechanism. For instance, in the first coupling layer, we wish to ensure that only x_0 is passed to the NN. This is achieved by multiplying the input vector (x_0, y_0) by the mask $(1, 0)$:

$$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} x_0 \\ 0 \end{pmatrix} \longrightarrow \text{NN} \longrightarrow \begin{pmatrix} \beta_1 \\ \gamma_1 \end{pmatrix}$$

In the next coupling layer, the mask is inverted to make the NN depend only on y instead of x. This alternating masking strategy is applied to all coupling layers.
In higher-dimensional cases, such as our 4D setting used in this thesis, it is still possible to design general masks to split the input vector. For example, masks like only the first half $(1, 1, 0, 0)$ or only the odd values $(1, 0, 1, 0)$ can be used, along with their respective inverses.

For the backward propagation, the methodology remains the same. As shown in Fig.,11b, we start from the output values x_{t+3} and y_{t+3}. Using x_{t+3}, we compute the parameters β_3 and γ_3, and then apply the inverse of the shift-and-scale transformation to recover y_{t+2}. By inverting Eq.,8, we obtain:

$$\begin{aligned} x_{t+2} &= x_{t+3} \\ y_{t+2} &= \frac{y_{t+3} - \beta_3}{\gamma_3} \end{aligned}$$

This procedure is repeated layer by layer until reaching the first coupling layer, always using the alternating masking scheme.
By stacking multiple affine coupling layers and interleaving the roles of the variables via masking, Real NVP is capable of modeling highly complex distributions while preserving all the desired properties of a normalizing flow.

To build our Real NVP model, we used as a reference the `Pytorch` implementation presented in Dinh et al. (2016), available on their GitHub page[9], as it was developed with a goal similar to ours.

3.4.3 Application of the NF algorithm

As briefly introduced in Sec.,3.4, our application of the Normalizing Flows (NFs) algorithm differs from the typical usage. While NFs are commonly employed to generate

[9] `https://github.com/xqding/RealNVP`

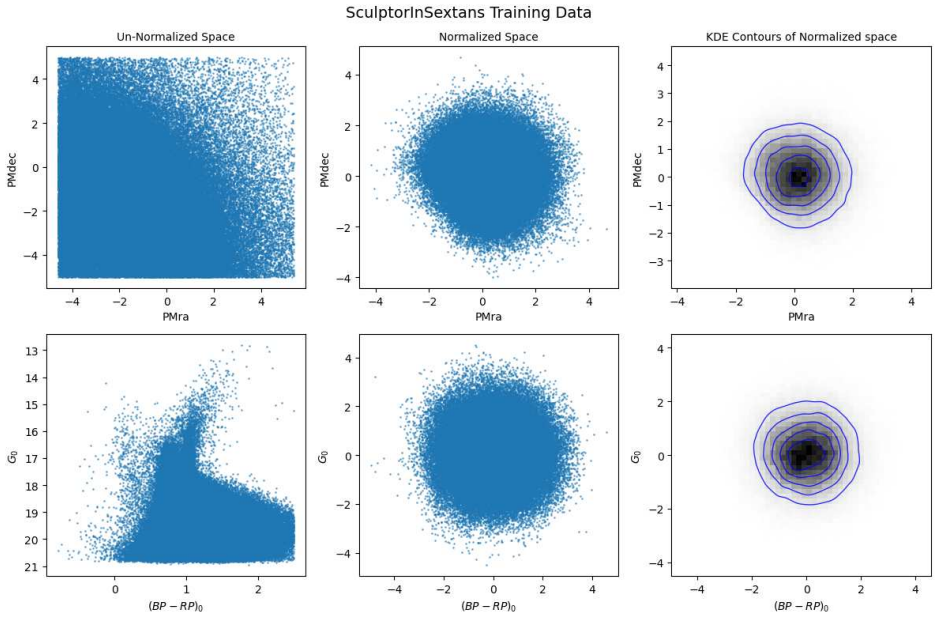


Figure 12: Results of the NF model on the training region: The first column displays the PM and CMD distributions of the contaminant stars before applying the normalizing flow, representing the raw input data used for training. The second column shows the same data after the normalization process, representing the output of the trained model applied to the training data. The last column displays the corresponding contour plots of the normalized data.

complex data from a simple base distribution, such as in image generation, our goal is instead to simplify the observed data distribution.
A distinctive feature of this method is its strong reliance on machine learning. It combines the NF architecture with custom Gaussian Mixture Models (GMMs) to model the probability density functions of both the contaminant population and the dwarf galaxy members in the latent space $\mathbf{Z}$. These density functions are then used to compute membership probabilities through Bayesian inference.
The methodology can be divided into three main stages: the normalization phase, the density fitting phase, and finally, the Bayesian inference phase.

Normalization phase

The training procedure adopted here closely follows the approach described in Sec. 3.2, where the entire dataset is divided into a training region and a test region. However, in this case, the test region is defined as a circular area with a radius of 4 degrees centered on the coordinates of the target galaxy. This choice reflects the aim of treating the test region as an independent catalog to which the trained algorithm will be applied. The radius of 4 degrees was selected to ensure the inclusion of all synthetic stars belonging to the dwarf galaxy, particularly the outermost stars and those forming tidal tails, in accordance with the properties of the mock catalogs created by Dr. G. Thomas (see Sec. 2.2).
The training region will correspond to the already seen *donut-shape region* with only foreground stars and we select only four physical quantities: the PM in RA and Dec and the color and G-magnitude of the CMD, making our analysic only with four features.

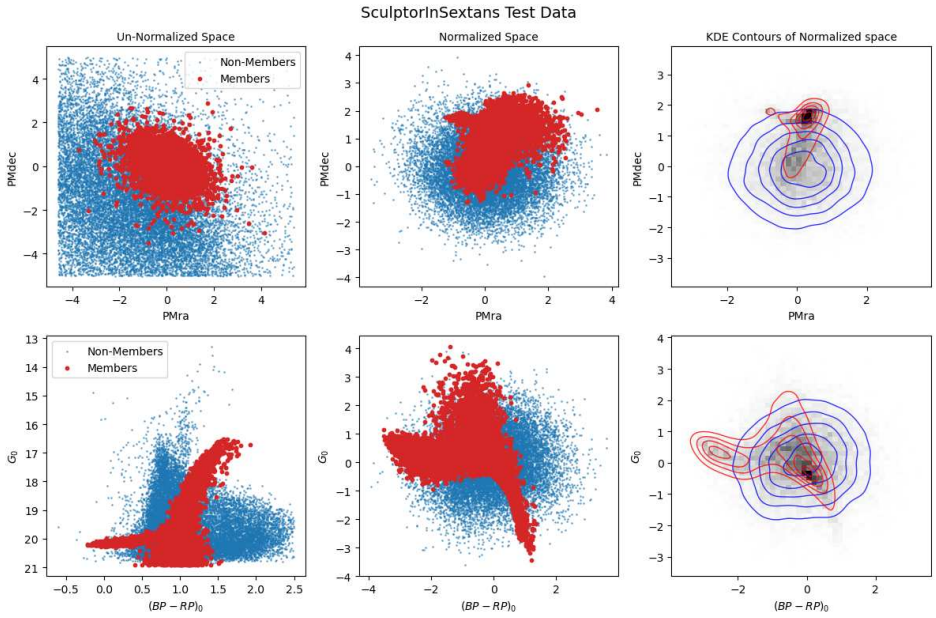


Figure 13: Results of the NF model on the test region of the SculptorInSextans mock galaxy: The first column shows the PM and CMD distributions of both contaminant (blue) and member stars (red) before applying the normalizing flow. The second column presents the same data after the normalization process, representing their mapping into the normalized latent space. The last column displays the corresponding contour plots of the normalized data. For improved visualization, approximately 70% of the contaminant stars in the test region have been removed to better highlight the differences between the contours of contaminants and members.

The choice to use only PM and CMD information is deliberate. In particular, we excluded spatial information for two main reasons. First, the Real NVP model is designed to transform complex distributions into simpler ones for datasets whose features share similar distributions to those used during training. In the case of PM and CMD, the values, and thus the overall shapes, vary within well-defined ranges, as discussed in Sec. 3.2.1 during the construction of the 2D histograms. However, including spatial information in training causes the model to learn how to transform a specific geometry (the annular "*donut-shaped*" training region) and when the model is later applied to the test region, which has a circular shape, it encounters a distribution shift and struggles to perform the transformation correctly.
The second reason is that, as we will explore in more detail in Sec. 4, including spatial information tends to significantly hinder the detection of member stars in the outer regions of the galaxy, particularly those in tidal structures or tidal tails. To address this, we chose to exclude spatial data from the model and focus solely on PM and CMD, ensuring that the resulting membership probabilities are not biased by a star's distance from the dwarf galaxy center.
As shown in Fig.,12, our model is trained on the dataset to simultaneously transform the complex distributions of the PM and CMD into a four-dimensional Gaussian distribution centered at $(0, 0, 0, 0)$ with a identity covariance matrix $\mathrm{diag}(1, 1, 1, 1)$. This transformation is enforced through the model's design, particularly by minimizing the negative log-likelihood loss. Additionally, during the model setup, we imposed bounds on the latent space, constraining each dimension to span the interval $[-4.0, 4.0]$.

Once the model is trained, we proceed by applying the transformation to the latent space to stars in the test region, which we remind the reader contains both contaminants and dwarf galaxy stars. The underlying idea is that the model, now trained to recognize and transform only the contaminant population, will map these sources in the test region to the Gaussian distribution. In contrast, the true member stars, unseen by the model during the training phase, retain a complex non-Gaussian distribution in the latent space.
This leads to a final distribution that is a mixture of the well-understood Gaussian corresponding to the contaminants and a more intricate distribution representing the member stars, which the model could not normalize.
The example in Fig.,13 shows the Sculptor-like dwarf galaxy mock catalog test region before and after the normalization procedure. It is evident that the actual member stars do not follow the distribution imposed by the NF model.

Density fitting phase - GMM

The outcome obtained in the previous section (normalization phase) allows us to proceed with Bayesian inference by computing, for each star individually, the probability of belonging to either the known contaminant distribution or the remaining complex distribution, enabling a direct comparison to determine the more likely membership.
The next step is to characterize this complex distribution associated with the member stars, since the distribution for the contaminants is already known. To do that, we focus on extracting the complex distribution of the member stars from the latent space $\mathbf{Z}$, which is obtained after applying the Normalizing Flow model to the test region.
From the NF model, we know that the latent space distribution of the test data is a mixture of two components:

- A standard multivariate Gaussian distribution $\mathcal{N}(0, \mathbb{I})$, representing the transformed contaminant stars, which the model has learned during training.
- A more complex, unknown distribution corresponding to the member stars, which the model could not fully map to the Gaussian form because it was never exposed to them during training.

The key insight here is that this complex distribution of the members can itself be approximated as a mixture of multiple Gaussian components. To make the explanation more intuitive, we illustrate the idea using a 1-dimensional case, though in practice we deal with a 4-dimensional latent space (proper motions in RA and Dec, color, and magnitude).
We can model the member star distribution as a Gaussian Mixture Model (GMM) of n components:

$$p_{\text{memb}}(x) = \sum_{i=1}^{n} w_i \cdot \mathcal{N}(x|\mu_i, \sigma_i^2) \tag{9}$$

where w_i are the mixture weights (with $w_i > 0$ and $\sum_i w_i = 1$), μ_i and σ_i are the mean and standard deviation of the i-th Gaussian component, respectively. Each component is defined by the parameter set $\Theta_i = \{w_i, \mu_i, \sigma_i\}$, and the full set of trainable parameters across all components is denoted by $\Theta = \{\Theta_1, ..., \Theta_n\}$.
The total probability distribution in the latent space, which includes both contami-

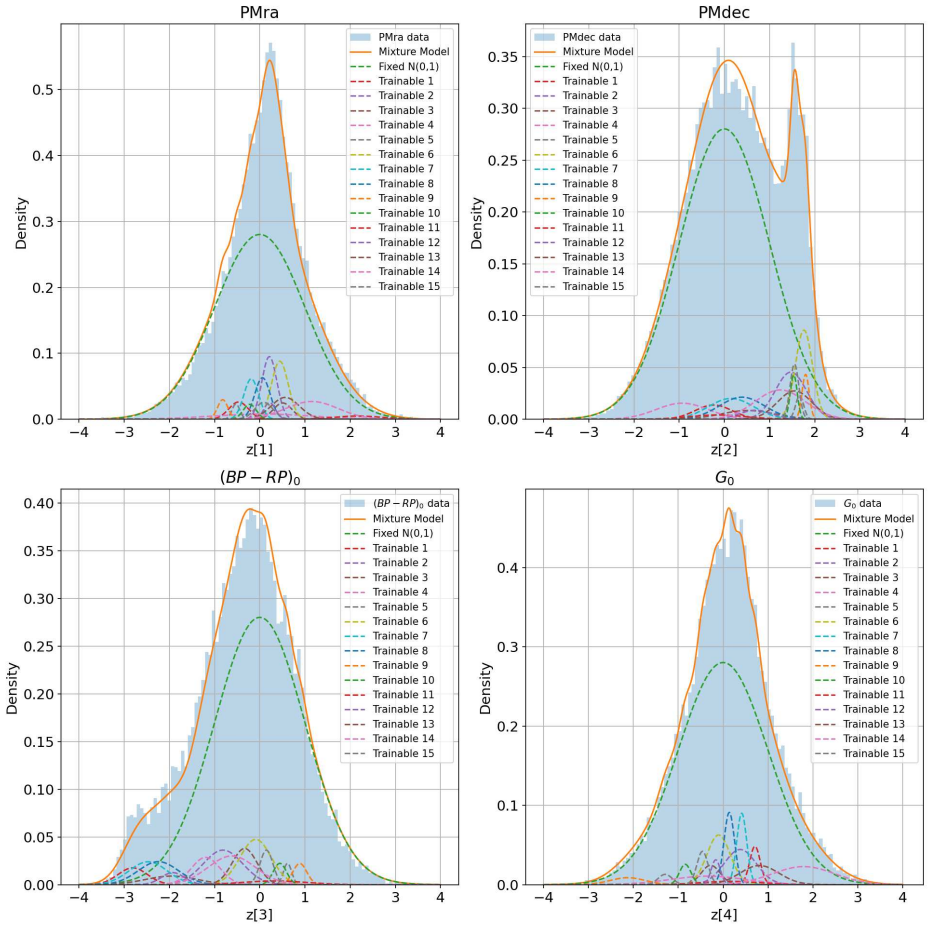


Figure 14: Example of the GMM fitting for the Sculptor mock galaxy: The model consists of a custom Gaussian Mixture Model with one fixed component and 15 additional trainable components. Each panel shows the fit to a different feature of the input data: proper motion in right ascension (1), proper motion in declination (2), color (3), and magnitude (4).

nants and members, can thus be expressed as:

$$p_{\text{tot}}(x) = w_0 \cdot \mathcal{N}(x|0,1) + \sum_{i=1}^{n} w_i \cdot \mathcal{N}(x|\mu_i, \sigma_i^2) \tag{10}$$

Here, w_0 represents the weight of the fixed Gaussian component corresponding to the known contaminant distribution. It is the only prior explicitly introduced in this method. This weight can be estimated approximately by multiplying the surface density of contaminants, previously measured in the training region, by the area of the test region.
It is important to emphasize that w_0 is not "a prior" used later in the final Bayesian membership probability calculation, since the contaminants likelihood is defined by $\mathcal{N}(0,1)$, but instead, it is used only during the training of the Gaussian Mixture Model to anchor the fixed gaussian component for the contaminants.
By doing so, we ensure that the model retains the correct representation of the known Gaussian; otherwise, the trainable components (modeling the member stars distribution) could overfit and absorb the known Gaussian structure, making it impossible to reliably extract the member distribution parameters Θ. This constraint helps maintain the interpretability and accuracy of the GMM fit.
To fit this model to the test region's latent space, the GMM with one fixed component $\mathcal{N}(0,1)$, for the contaminants, and several trainable components for the members, is trained to optimize the parameters Θ of the member components to best fit the overall data distribution in the latent space.
After training, the parameters of the GMM corresponding to the member components define the shape of the member distribution. To isolate this member component and ensure it is properly normalized, we use:

$$p_{\text{memb}}(x) = \frac{\sum_{i=1}^{n} w_i \cdot \mathcal{N}(x|\mu_i, \sigma_i^2)}{\sum_{i=1}^{n} w_i} \tag{11}$$

Extending this logic to our 4-dimensional case with $z \in \mathbf{Z}$ is possible to verify that this method gives a good result of extraction the complex distribution of the members. In Fig. 14 is reported the result for the Sculptor mock galaxy where is easy to see that the trainable components of the GMM togheter with the fixed one produce a nice fit of the total data. The choice of using 15 trainable components is only keep light and simple the code and avoid the risk of overfitting but keeping a good fitting result. In Fig. 15 is show how the complex probability function found with the Θ by the GMM and with Eq. 11 follows the real distribution of the members.

Bayesian inference phase

The normalized form of the member distribution allows us to proceed with Bayesian inference: for any star in the latent space, we can now compute the probability that it belongs to either the known Gaussian contaminant distribution or the complex member distribution. This enables a probabilistic classification, with the Bayesian inference, of stars as likely members or contaminants based on their location in latent space.
For this step, we take inspiration from the method proposed by Rinaldi et al. (2024), hereafter referred to as **R24**, as their objective closely aligns with ours. The goal is to evaluate, for each star, which of the two distributions (members or contaminants)

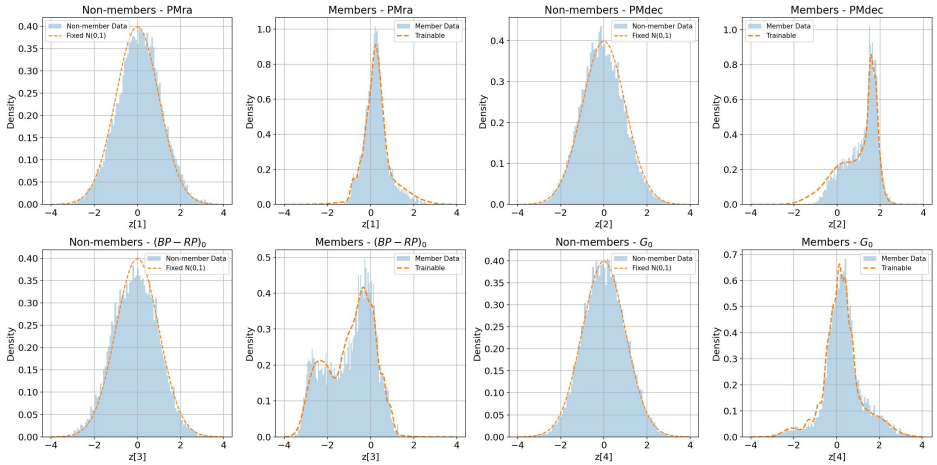


Figure 15: Example of the inferred member distribution for the Sculptor mock galaxy: For each dimension of the latent space, the histograms of both actual contaminants and actual member stars are shown. The theoretical $\mathcal{N}(0,1)$ distribution is overplotted on the contaminant histogram, representing the expected normalized distribution. For the member stars, the fitted probability density function—computed using Eq.,11 and the GMM-derived parameters Θ overlaid on the corresponding histogram.

it is more likely to belong to.

In **R24**, Bayesian inference is implemented through a Gibbs sampling procedure to estimate the full posterior distribution of each star. This approach is particularly powerful as it allows us to track how the membership probabilities evolve across iterations, providing not only individual star posteriors but also a global probabilistic understanding of the dataset.

At the core, this method is based on classical Bayesian inference, which provides a principled way to estimate the probability that a given star belongs to the dwarf galaxy population (members) or to the Milky Way foreground (contaminants), based on its position in the normalized latent space $\mathbf{Z}$.

Given a star with normalized coordinates $\mathbf{z} \in \mathbb{R}^4$, the posterior probability of membership is computed using Bayes' theorem:

$$P(\text{memb/cont} \mid \mathbf{z}) = \frac{p_{\text{memb/cont}}(\mathbf{z}) \cdot \pi_{\text{memb/cont}}(\mathbf{z})}{p_{\text{tot}}(\mathbf{z})} \tag{12}$$

where

- in the case of $p_{\text{memb}}(\mathbf{z})$ is the likelihood of the star under the member distribution (obtained from the trained GMM).
- in the case of $p_{\text{cont}}(\mathbf{z})$ is the likelihood under the known contaminant distribution, modeled as a multivariate Gaussian $\mathcal{N}(0, \mathbb{I})$.
- $\pi_{\text{memb/cont}}(\mathbf{z})$ is the prior probability of membership or contamination
- $p_{\text{tot}}(\mathbf{z}) = p_{\text{memb}}(\mathbf{z}) \cdot \pi_{\text{memb}}(\mathbf{z}) + p_{\text{cont}}(\mathbf{z}) \cdot \pi_{\text{cont}}(\mathbf{z})$ is the total marginal probability.

The prior $\pi_{\text{memb/cont}}$ follows the formulation given in Eq. 15 in the **R24**, which we report here:

$$\pi(l = \mathcal{M}/\mathcal{C} \mid \mathbf{l_{-i}}) = \begin{cases} \dfrac{N_{\mathcal{M}} + \beta/2}{N + \beta} & \text{if } l = \mathcal{M} \\ \dfrac{N_{\mathcal{C}} + \beta/2}{N + \beta} & \text{if } l = \mathcal{C} \end{cases} \tag{13}$$

To apply this, as done in **R24**, a new auxiliary variables $\mathbf{l} = \{l_1, ..., l_N\}$ is introduced, assigning each star a label: $\mathcal{M}$ for members and $\mathcal{C}$ for contaminants. During inference, the prior for the i-th star is calculated by excluding its own label, denote as $\mathbf{l}_{-i}$, and computing the frequency-based prior from the remaining labels. For example, if star i is labeled as a "member", its prior becomes:

$$\pi_{\text{memb}}(\mathbf{z_i} \mid \mathbf{l_{-i}}) = \frac{N_{\mathcal{M}} + \beta/2}{N + \beta}$$

where $N_{\mathcal{M}}$ is the number of stars currently labeled as members in $\mathbf{l}_{-i}$, $N_{\mathcal{C}}$ is corresponding number of contaminants and $N = N_{\mathcal{M}} + N_{\mathcal{C}}$. The parameter β controls the strength of the prior. In our case, we set $\beta = 2$, corresponding to a uniform (non-informative) prior. This choice ensures that, before considering the data, both populations are treated equally, leading to balanced and robust posterior probabilities.

The prior probabilities π are the only components in Eq.,12 that change during the Gibbs sampling iterations. At each step, we compute the posterior probabilities $P(\text{memb} \mid \mathbf{z}_i)$ and $P(\text{cont} \mid \mathbf{z}_i)$ for each star. If, for example, the probability of being a contaminant is higher than that of being a member, we update the corresponding label to $l_i = \mathcal{C}_i$; otherwise, we set $l_i = \mathcal{M}_i$.
By doing this, the number of stars labeled as members or contaminants changes at each iteration, and consequently, the priors π_{memb} and π_{cont} are also updated according to Eq.,13. This iterative update process allows the model to refine its classification over time.
One of the advantages of the Gibbs sampling approach is that the initial labels $\mathbf{l}$ can be assigned randomly (or split evenly 50/50), and the algorithm will naturally converge to a stable distribution. The final state of the labels not only provides membership probabilities for individual stars but also yields an estimate of the total number of dwarf galaxy members present in the sample.

For this work as the final posteriors of the stars in the catalog analyzed with the NF algorithm we took the last result of the final Gibbs iteration since is the one where we reach the convergence in the labels variable $\mathbf{l}$, and the results are reported in Sec. 4.3.

thresholds	TP	FP	FN	TN
0.02 - 0.95	6005	36	3	51410
0.05 - 0.95	6005	36	4	51975
0.1 - 0.95	6005	36	10	52251
0.2 - 0.9	6113	57	18	52519
0.5 - 0.5	6308	226	46	52758

(a) SculptorInSextans: m=6354 c=52988

thresholds	TP	FP	FN	TN
0.02 - 0.95	634	7	12	51736
0.05 - 0.95	634	7	25	52211
0.1 - 0.95	634	7	36	52469
0.2 - 0.9	714	15	54	52679
0.5 - 0.5	852	115	96	52868

(b) SextansInSextans: m=948 c=52987

thresholds	TP	FP	FN	TN
0.02 - 0.95	763	9	7	42977
0.05 - 0.95	763	9	16	43344
0.1 - 0.95	763	9	28	43552
0.2 - 0.9	836	26	43	43709
0.5 - 0.5	947	115	78	43849

(c) SextansInDraco: m=1025 c=43965

thresholds	TP	FP	FN	TN
0.02 - 0.95	894	6	6	43543
0.05 - 0.95	894	6	13	43719
0.1 - 0.95	894	6	20	43795
0.2 - 0.9	931	13	25	43846
0.5 - 0.5	979	55	31	43909

(d) DracoInDraco: m=1010 c=43965

Table 3: Confusion values from B22: Each table presents the confusion values (TP, FP, FN, TN) for each mock galaxy obtained using the **B22** method, evaluated at different threshold choices in a 4 degrees circular area around the galaxy center. The symbol m indicates the total number of actual members in the mock catalog, while c denotes the total number of contaminants.

4 Results

In this section, we present the outcomes of the different workflows developed in this study. Special attention is given to the analysis of the confusion matrices, emphasizing false positives and false negatives, since the primary goal of this work is to enhance the identification of member stars in the outer regions of dwarf galaxies.
We begin with the results of the method described in Sec. 3.2, which allows a direct comparison with the approach of **B22** (see Sec. 3.1). We then proceed to the dimensionality reduction method (Sec. 3.3), again using **B22** as a reference due to the comparable results, even though the methodologies are fundamentally different. Finally, we present the Normalizing Flow results (Sec. 3.4), which are analyzed independently, as this method is conceptually and structurally distinct from the **B22** framework.

4.1 B22 vs B22 ML

We have shown how the two approaches differ in the modeling of the contaminants' distribution, and how this improvement could also enable a more accurate recovery of the dwarf galaxy signal.
We present the results for each mock catalog used in the analysis, and for each of the three levels introduced earlier. This comparison helps assess whether the method performs consistently across the four mock galaxy cases and whether common properties emerge across them.
This comparison is not only to evaluate the improvement introduced by the ML-based modeling of the contaminants, but also to verify whether the new method maintains robustness and generality across different dwarf galaxy configurations. Since mock catalogs were designed with varying properties such as density, structure, and foreground contamination, they provide a controlled setting to benchmark the reliability and adaptability of the method.

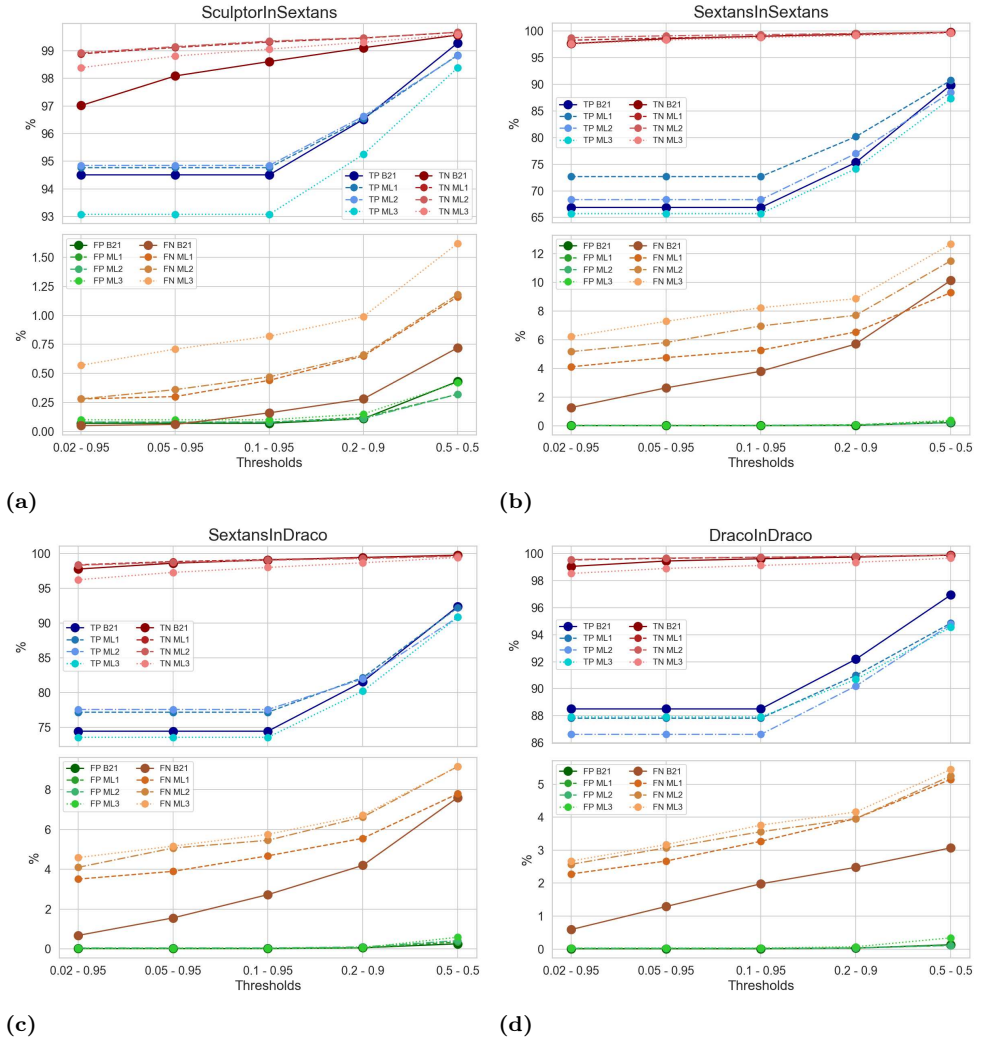


Figure 16: Confusion results comparison between B22 and the ML implementation: Each panel presents a schematic summary of the TP, FP, FN and TN for each mock catalog, comparing the original **B22** method with the results from the ML-enhanced implementation in all the three level analyzed. The y-axis shows the percentage of each classification outcome relative to the total number of actual members (for TP and FN) and actual contaminants (for FP and TN). The x-axis displays five different threshold combinations: each pair defines the probability range below which stars are classified as contaminants and above which they are classified as members.
Thick and continuous lines represent results from the **B22** method, while dashed lines correspond to ML-based method results for the First Level (ML1), dashdot lines correspond to the Second Level (ML2) and dotted lines to the Third Level (ML3). In Fig. 16a are reported the results relative to the `SculptorInSextans` mock galaxy, in Fig. 16b are reported the results relative to the `SextansInSextans` mock galaxy, in Fig. 16c are reported the results relative to the `SextansInDraco` mock galaxy and in Fig. 16d are reported the results relative to the `DracoInDraco` mock galaxy

The evaluation is based on confusion matrices, which report the number of true positives, false positives, false negatives, and true negatives, together with their relative fractions with respect to the total sample. Rather than adopting a simple binary criterion, classifying stars with $P > 0.5$ as members and those with $P < 0.5$ as contaminants, we applied two thresholds at the extremes of the probability distribution. This approach allows us to separate, with (almost) certainty, genuine members from clear contaminants, while leaving ambiguous cases unclassified. For example, in Fig. 35, we can see both histograms exhibit two prominent peaks near probability values of 0 and 1, indicating that the majority of non-members and members are correctly classified. However, a non-negligible fraction of stars still falls within the intermediate membership probability range, for example between 0.3 and 0.8.
For an initial assessment, we define candidate member stars as those with a membership probability $P > 0.95$, and candidate contaminants as those with $P < 0.02$. These thresholds are intentionally strict and will be relaxed in the following sections to enable a more inclusive classification.
In this way, we can better quantify the improvement in recovering true member stars, particularly in the low-density outskirts of the galaxies, where contamination is most severe and the benefits of a refined model are expected to be most evident.
The complete set of confusion plots, showing the detailed classification outcomes for each mock catalog and each membership level, is presented in the Appendix (Sec. A), where for visualization purposes the results presented correspond to a region of 10 times the $R_{1/2}$ of the corresponding mock galaxy, which is sufficiently large to encompass both the main body of the galaxy and any potential tidal structures.
It is interesting to note that in the ML-based result, in this case of the mock Sextans but as the others too, there is a more pronounced peak at low probability values for actual members. This suggests that the ML method labels a higher number of true members as contaminants compared to **B22**. This behavior, and its implications, will be analyzed in more detail in the next section.

The Summary figures in Fig. 16, provide a more compact visualization of the results obtained. These diagrams schematically show how the numbers of True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN) vary as a function of the probability thresholds chosen for assigning membership and non-membership. This approach allows us to assess the model's robustness and sensitivity to different classification criteria. In practice, choosing a narrow membership probability threshold reduces FP but also reduce the TP, similar happens to FN and TN in the case of small contaminants threshold. These threshold curves are therefore essential tools for calibrating the model to achieve the desired trade-off between completeness and purity depending on the scientific objective.
The ideal scenario, although highly challenging, is to achieve both FP and FN values close to zero. Realistically, the desired outcome is to maintain low levels of FP and FN, and to have FP and FN preferentially located within the central regions of the galaxies, since we are particularly interested in identifying member stars in the outermost regions.

4.1.1 First Level

We now analyze the results for each mock catalog, with particular emphasis on the False Positives (FP) and False Negatives (FN).

thresholds	TP	FP	FN	TN
0.02 - 0.95	6022	44	18	52402
0.05 - 0.95	6022	44	19	52521
0.1 - 0.95	6022	44	28	52626
0.2 - 0.9	6135	62	41	52704
0.5 - 0.5	6280	172	74	52813

(a) SculptorInSextans: m=6354 c=52988

thresholds	TP	FP	FN	TN
0.02 - 0.95	689	10	39	52068
0.05 - 0.95	689	10	45	52290
0.1 - 0.95	689	10	50	52464
0.2 - 0.9	760	35	62	52619
0.5 - 0.5	860	159	88	52825

(b) SextansInSextans: m=948 c=52987

thresholds	TP	FP	FN	TN
0.02 - 0.95	791	14	36	43246
0.05 - 0.95	791	14	40	43446
0.1 - 0.95	791	14	48	43574
0.2 - 0.9	842	28	57	43685
0.5 - 0.5	945	153	80	43811

(c) SextansInDraco: m=1025 c=43965

thresholds	TP	FP	FN	TN
0.02 - 0.95	887	7	23	43754
0.05 - 0.95	878	7	27	43806
0.1 - 0.95	878	7	33	43843
0.2 - 0.9	919	11	40	43865
0.5 - 0.5	958	50	52	43914

(d) DracoInDraco: m=1010 c=43965

Table 4: Confusion values from the First Level by the ML implementation The tables follow the same structure of Tab. 3

SculptorInSextans mock catalog: As shown in Fig. 2a at the end of Sec. 2, the mock galaxy used in this case is a Sculptor-like dwarf galaxy, highly populated and compact, superimposed on the foreground of the real Sextans dwarf galaxy. We note the absence of tidal features or gravitational tails, meaning that the outer stars remain relatively close to the galaxy's center. Thanks to the high stellar density, both the CMD and PM distributions exhibit well-defined and compact clumps, which in principle should make the task of the model easier, allowing for more accurate predictions. A first look at the confusion results in Fig. 16a reveals a slight overall improvement: the number of TP and TN for the ML implementation are higher compared to those from **B22**, while the FP are only slightly increased but is an acceptable trade-off. However, the situation changes when using the $0.5 - 0.5$ probability threshold (i.e., equal cutoff for member and contaminant classification): in this case, the ML approach yields higher TNs but also lower TPs and FPs. The most notable difference is in the FN, which are consistently higher for the ML method across all threshold choices compared to **B22**. This isn't necessarily detrimental, especially if the FN stars are mostly located in the inner regions of the galaxy, where their classification is less critical for our primary objective, identifying outer region members.
A more detailed analysis, through visualization of the TP, FP, FN, and TN distributions in sky coordinates, PM, and CMD space (see Fig. 36 for the ML case), shows that TPs effectively cover the spatial extent of the dwarf galaxy. FPs are mostly concentrated toward the center, while FNs dominate the outer regions. In the **B22** case (Fig. 31), FNs are fewer but also predominantly found in the outskirts. Notably, at higher threshold settings, the **B22** method also starts to miss more stars in the outer regions, increasing its FN count there as well.

SextansInSextans mock catalog: This mock scenario is particularly relevant, as the Sextans dwarf galaxy is characterized by a highly dense foreground contamination and a spatially diffuse stellar population. These conditions make the task of assigning reliable membership probabilities especially challenging, particularly in the outer regions of the galaxy.

Looking at Fig. 16b, we observe a similar trend to the Sculptor case, where the FN in the ML method are consistently higher than in **B22**. In this specific mock, Sextans was modeled to include tidal tails, clearly visible in Fig. 4. These tails correspond to the majority of the FN, indicating that outer stars are often missed by the ML classification.
In the case of **B22**, the FN also mostly coincide with the stars in the tails (see Fig. 37). However, their membership probabilities are slightly higher than in the ML case, though still too low to be classified as members under the strict threshold. As observed in the Sculptor mock, these stars begin to appear as members when the contaminant threshold is increased.

Sextans in Draco mock catalog: An almost identical result, consistent with the two previous cases, is obtained for the `SextansInDraco` mock catalog, as shown in Fig.,16c and the corresponding confusion plots in the Appendix.
The Draco background is less dense compared to Sextans, so when placing the same Sextans-like mock galaxy over it, the results confirm the trends observed in the `SextansInSextans` case. This could bring the idea that the model's performance is strongly influenced by the properties of the mock galaxy rather than by the background density alone.

DracoInDraco mock catalog: For the final mock catalog, shown in Fig. 16d, we observe a different outcome compared to the previous Sculptor and Sextans cases.
In the case of a Draco-like galaxy embedded in its real foreground, we find that the TP count from the ML method is lower than that of **B22**, the first instance where this occurs. On the other hand, the FP, FN, and TN follow the same general trends observed in the previous mocks.
Despite the low number of TPs, both methods succeed only in recovering member stars located in the central region of the dwarf galaxy, failing to identify stars in the outer parts and the gravitational tails. This limitation is clearly visible in the confusion plots reported in the Appendix.

4.1.2 Second Level

As previously described in Sec. 3.2.3, the difference between the First Level and the Second Level is how we calculate the galaxy CMD look-up map, extracted no more from the elliptical $1\,R_e$ region around the galaxy center but by the subtraction of the predicted contaminants CMD histogram from the true CMD histogram of the test region, as reported as example in Fig. 8.

SculptorInSextans mock catalog: This Sculptor-like mock galaxy is characterized by a high stellar density, which results in a sharply defined CMD compared to the other mock galaxies. This clarity makes the subtraction method used to obtain the general CMD 2D histogram of the galaxy particularly effective, as illustrated in Fig. 8b.
In fact, the output shown in Fig. 16a is nearly identical to the First Level result as we can see that both lines almost overlap each others. However, upon closer inspection a Tab. 4a and Tab. 5a, we note a slight general increase in both TP and TN.
This difference is too small to be considered a significant improvement, and thus we can conclude that the performance of both levels is essentially equivalent in this case. As before, the FNs correspond to stars located in the outer regions of the dwarf galaxy.

thresholds	TP	FP	FN	TN
0.02 - 0.95	6027	44	18	52421
0.05 - 0.95	6027	44	23	52538
0.1 - 0.95	6027	44	30	52646
0.2 - 0.9	6140	57	42	52705
0.5 - 0.5	6279	171	75	52814

(a) SculptorInSextans: m=6354 c=52988

thresholds	TP	FP	FN	TN
0.02 - 0.95	648	13	49	52325
0.05 - 0.95	648	13	55	52512
0.1 - 0.95	648	13	66	52623
0.2 - 0.9	730	23	73	52744
0.5 - 0.5	839	111	109	52873

(b) SextansInSextans: m=948 c=52987

thresholds	TP	FP	FN	TN
0.02 - 0.95	795	24	42	43217
0.05 - 0.95	795	24	52	43401
0.1 - 0.95	795	24	56	43536
0.2 - 0.9	840	45	68	43642
0.5 - 0.5	931	183	94	43781

(c) SextansInDraco: m=1025 c=43965

thresholds	TP	FP	FN	TN
0.02 - 0.95	875	8	26	43769
0.05 - 0.95	875	8	31	43816
0.1 - 0.95	875	8	36	43842
0.2 - 0.9	911	11	40	43872
0.5 - 0.5	957	50	53	43914

(d) DracoInDraco: m=1010 c=43965

Table 5: Confusion values from the Second Level by the ML implementation The tables follow the same structure of Tab. 3

The small increase in TPs in the Second Level is confined to the central regions, indicating no enhancement in identifying outer members.

SextansInSextans mock catalog: A discussion similar to the one presented for the First Level can be made for the `SextansInSextans` mock catalog. The overall trends in TP, FP, FN, and TN remain consistent with the First Level results. However, unlike the Sculptor case, the outcome here is slightly worse in the Second Level. As shown in Tab. 4b and Tab. 5b, and also visible in Fig. 16b, the number of TPs is slightly lower compared to the First Level, while both FP and FN are marginally higher. On the other hand, the number of TNs increases by approximately 300, indicating a better identification of contaminants in this second stage.

SextansInDraco mock catalog: As shown in Fig. 16c and in Tables 4c and 5c, the overall trend remains consistent with the First Level. The main differences are observed in the FP and FN values, which are higher in the Second Level. However, the differences are small and do not significantly alter the interpretation of the final result. The TP values show a modest improvement but remain nearly identical overall.

DracoInDraco mock catalog: In this case, the ML implementation performs slightly worse than the **B22** methodology. As shown in Fig. 16d, the blue dashed-dot line from the ML method remains consistently below the solid dark blue line of **B22** and to the First Level case. The number of FPs remain almost unchanged, although there is a general small increase in FNs and TNs between the First and Second Levels, but this improvement is marginal.
Overall, the results are comparable to the First Level. As seen in Fig. 44 in the Appendix, the TPs are still concentrated in the central region of the galaxy, along with the FPs, while most of the FNs are located in the outer regions. This again leads to the failure to identify the gravitational tails present in the mock catalog.

thresholds	TP	FP	FN	TN
0.02 - 0.95	5914	51	36	52133
0.05 - 0.95	5914	51	45	52360
0.1 - 0.95	5914	51	52	52492
0.2 - 0.9	6052	77	63	52621
0.5 - 0.5	6251	220	103	52765

(a) SculptorInSextans: m=6354 c=52988

thresholds	TP	FP	FN	TN
0.02 - 0.95	623	10	59	51711
0.05 - 0.95	623	10	69	52084
0.1 - 0.95	623	10	78	52339
0.2 - 0.9	703	32	84	52542
0.5 - 0.5	828	202	120	52782

(b) SextansInSextans: m=948 c=52987

thresholds	TP	FP	FN	TN
0.02 - 0.95	754	10	47	42300
0.05 - 0.95	754	10	53	42766
0.1 - 0.95	754	10	59	43076
0.2 - 0.9	822	31	69	43373
0.5 - 0.5	931	259	94	43705

(c) SextansInDraco: m=1025 c=43965

thresholds	TP	FP	FN	TN
0.02 - 0.95	888	15	27	43313
0.05 - 0.95	888	15	32	43476
0.1 - 0.95	888	15	38	43574
0.2 - 0.9	916	30	42	43677
0.5 - 0.5	955	151	55	43813

(d) DracoInDraco: m=1010 c=43965

Table 6: Confusion values from the Second Level by the ML implementation The tables follow the same structure of Tab. 3

4.1.3 Third Level

The Third Level includes the PM look-up histogram, derived using the subtraction method described in Sec. 3.2.3, analogous to what was done for the CMD in the Second Level, and the calculation of the ratio $f_{\rm gal}$ as the sum of the values in the galaxy CMD 2D histogram. As shown in the corresponding figures in the Appendix, the results of this level are overall similar to those of the previous levels and the **B22** method: stars labeled as members are confined to the central regions, while the outer stars are again classified as contaminants. However, we observe fluctuations in the TP, FP, FN, and TN values across different mock catalogs and levels.
Notably, for all mock catalogs, the Third Level yields worse results compared to **B22**. As seen in Fig. 16, the number of TPs is consistently lower than in **B22**, and for the first time, even the TNs are lower, except in the `SculptorInSextans` case, which remains slightly above **B22**. The number of FNs increases significantly compared to both **B22** and previous levels, and the FPs are also slightly higher. These results suggest that the Third Level is the least reliable among the three methods.

ScluptorInSextans mock catalog: As mentioned, this is the only case in which the Third Level maintains a higher TN count than **B22**, despite a marked drop in TPs, the lowest among all levels. Both the FP and FN values increase substantially, with the FN rising dramatically, as seen in Fig. 16a.

SextansInSextans mock catalog: This mock represents one of the worst-performing scenarios. Alongside the expected drop in TP, the TNs do not improve and are in fact lower than in previous levels. The FP values remain in line with those seen in **B22** and earlier levels, offering no notable improvement.

SextansInDraco mock catalog: In this case, the Third Level shows no improvement whatsoever compared to **B22**. Both TN and TP are lower, while FN and FP reach their highest values across all methods. This confirms that the subtraction

method applied to PMs is not effective in this foreground configuration.

DracoInDraco mock catalog: While Draco foregrounds typically yield the worst performance, this particular case offers an interesting exception. At low threshold values, the TP count is actually the highest across all levels, though this is no longer true at the $0.2 - 0.9$ threshold range. However, both FN and FP are again at their highest levels. This might indicate that the increase in TP comes at the cost of also increasing the FP rate, possibly due to a broader classification of stars as members.

4.2 Dimensional Reduction: UMAP

In the UMAP workflow, we explored several combinations of feature sets, as introduced in Sec. 3.3.1, to determine which configuration provides the most effective separation between member and foreground stars. The baseline feature set included:

- Spatial coordinates of the tangent plane relative to the dwarf galaxy center,
- Systemic-corrected proper motions in RA and Dec
  (defined as $\mathrm{PM}_{\mathrm{source}} = \mathrm{pm}_{\mathrm{source}} - \mathrm{pm}_{\mathrm{system}}$),
- Inverse of the relative error in systemic proper motion,to take into account the level of accuracy in determining the proper motion of a given star (PM errors given by Gaia `pmra_error` and `pmdec_error`),
- CMD features using $\mathrm{BP0} - \mathrm{RP0}$ and Gaia G_0-band magnitude,
- The parallax-over-error ratio, which, like proper motion, we expect for the true members to be consistent with the parallax corresponding to the distance of the galaxy.

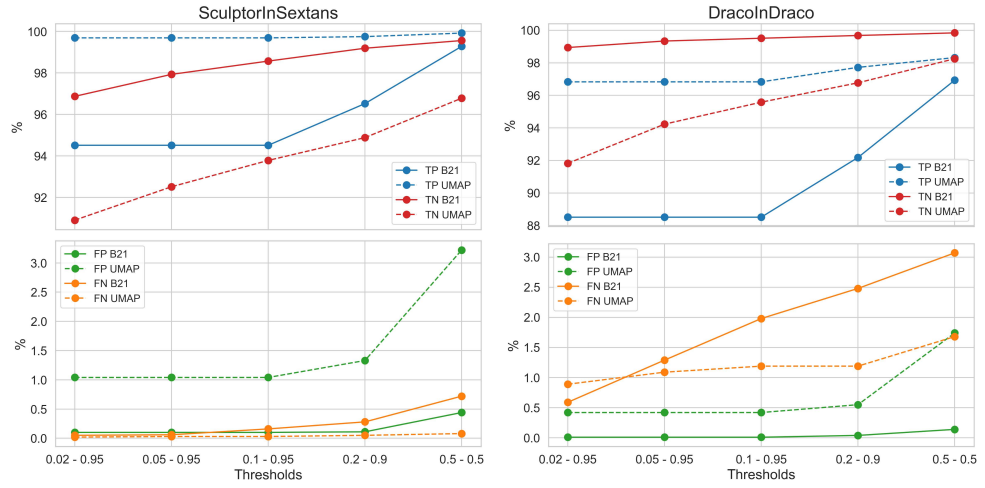
As mentioned in Sec.,3.3.1, the results discussed in this section are those relating to the sample resized due to the UMAP limitation.
As we will show in the following sections, alternative feature combinations were also tested to evaluate how the choice of input data affects the outcome of the method.
We proceed by presenting the results for each mock catalog separately. For each case, we assess the performance of UMAP when applied:

1. Once to a single mock realization,
2. Across N synthetic realizations, where we track how frequently each star is classified as a member.

We compare these results directly with those of **B22**, which serves as our benchmark for evaluating performance gains and identifying the strengths and limitations of this dimensionality reduction approach.

4.2.1 SculptorInSextans mock catalog

If we apply the UMAP algorithm only once to the catalog (Sec. 3.3.2), we already obtain an interesting result. As shown in Tab. 7a and Fig. 51 (Appendix), the algorithm provides a first good separation between member stars and background contaminants,



(a) **(b)**

Figure 17: Confusion results comparison between B22 and the UMAP implementation: Each panel presents a schematic summary of the TP, FP, FN and TN for each mock catalog, comparing the original **B22** method with the results from the dimensional reduction, after the 500 iteration of UMAP. The y-axis shows the percentage of each classification outcome relative to the total number of actual members (for TP and FN) and actual contaminants (for FP and TN). The x-axis displays five different threshold combinations: each pair defines the probability range below which stars are classified as contaminants and above which they are classified as members.
Thick and continuous lines represent results from the **B22** method, while dashed lines correspond to UMAP method results.

despite having no prior information. This suggests that the physical parameters of members and contaminants differ sufficiently for UMAP to distinguish them. Naturally, the separation is not perfect: some contaminants, particularly those located in the inner region of the dwarf galaxy, exhibit similar values to the members. These contaminants tend to fall into the same proper motion clump and follow the same CMD sequence as the members, leading to a high number of FP in the central region. Conversely, FN are mostly located in the outer regions, indicating that member stars far from the galaxy center are more likely to be classified as contaminants.
It is also worth noting that even in the single application case (last row of Tab. 7a), the number of TP is significantly higher than in **B22** (see Tab. 3a). However, this comes at the cost of a higher number of FPs, which, as shown in Fig. 51, spread across both the central and outer parts of the galaxy. This highlights the typical trade-off: a higher FP count can artificially boost the TP count.
To address this, we ran the method across the 500 catalogs with the same structure as the original one. This aims to reduce the variability in TP and FN, which are typically the hardest classes to correctly label. As shown in Tab. 7a and Fig. 17a, the TP count remains nearly unchanged, while the number of FPs drops significantly, still higher than in **B22**, but much closer. Additionally, the FN count is lower compared to **B22**. As visible in Fig. 55, the remaining FN is located in the outer region of the galaxy, similarly to what was observed in **B22** (Sec. 4.1.1), where relaxing the contaminant threshold also introduced FNs in that region.
Finally, we observe that TN are less well recovered in this method. As seen in Fig. 55, there is a less covered TN region in the galaxy's central area. This too could be improved by adjusting the threshold used to define contaminants.

thresholds	TP	FP	FN	TN
0.02 - 0.95	6334	83	1	7261
0.05 - 0.95	6334	83	1	7390
0.1 - 0.95	6334	83	2	7491
0.2 - 0.9	6338	106	3	7579
0.5 - 0.5	6349	257	5	7731
single UMAP	6345	252	9	7736

(a) SculptorInSextans: m=6354 c=7988

thresholds	TP	FP	FN	TN
0.02 - 0.95	0	0	0	0
0.05 - 0.95	0	0	0	0
0.1 - 0.95	0	0	0	0
0.2 - 0.9	0	0	0	0
0.5 - 0.5	0	0	0	0
single UMAP	902	172	48	7815

(b) SextansInSextans: m=948 c=7987

thresholds	TP	FP	FN	TN
0.02 - 0.95	0	0	0	0
0.05 - 0.95	0	0	0	0
0.1 - 0.95	0	0	0	0
0.2 - 0.9	0	0	0	0
0.5 - 0.5	0	0	0	0
single UMAP	986	138	39	6827

(c) SextansInDraco: m=1025 c=6965

thresholds	TP	FP	FN	TN
0.02 - 0.95	978	29	9	6395
0.05 - 0.95	978	29	11	6563
0.1 - 0.95	978	29	12	6657
0.2 - 0.9	987	38	17	6740
0.5 - 0.5	993	121	17	6843
single UMAP	998	122	12	6843

(d) DracoInDraco: m=1010 c=6965

Table 7: Confusion values from the UMAP implementation The tables follow the same structure of Tab. 3 with the exception of the last row where is reported the values for a single application of UMAP to the dataset. The number of members m and contaminants c is proportionate to the all sources analyzed in the UMAP method.

4.2.2 Sextans mock catalogs

Here we examine both the Sextans mock catalogs, one with Sextans as foreground and the other with Draco as foreground, due to a particular failure mode observed during the UMAP iterations. In both cases, UMAP consistently failed to clearly separate the data into two distinct clusters across all 500 synthetic catalogs. As a result, the subsequent clustering step also fails, leading to ineffective probability distributions. Figures 50b and 50c clearly show that the output probabilities are concentrated in a narrow range, making it impossible to apply any meaningful threshold as done in previous cases.
This failure is illustrated in Fig. 18, where the clustering effectively divides the data in half without regard to the underlying structure, sometimes assigning all members as contaminants, and other times misclassifying large numbers of contaminants as members. Consequently, in Tables 7b and 7c, all performance metrics are zero.

An exception arises for the single application of UMAP, where the dimensionality reduction does succeed, yielding results somewhat comparable to the Sculptor case. However, this scenario still suffers from a very high number of FP, which makes it nearly impossible to confidently identify members in the outer regions. Furthermore, the FN continue to correspond to stars located in the gravitational tails, failing the main objective of detecting them.

4.2.3 DracoInDraco mock catalog

Also in the case of the mock Draco catalog, the full-dimensional UMAP method successfully performs dimensionality reduction, allowing for the computation of posterior probabilities over 500 iterations, similarly to the Sculptor mock galaxy.
Starting with the single application of UMAP, we observe in Fig. 54 (Appendix) that

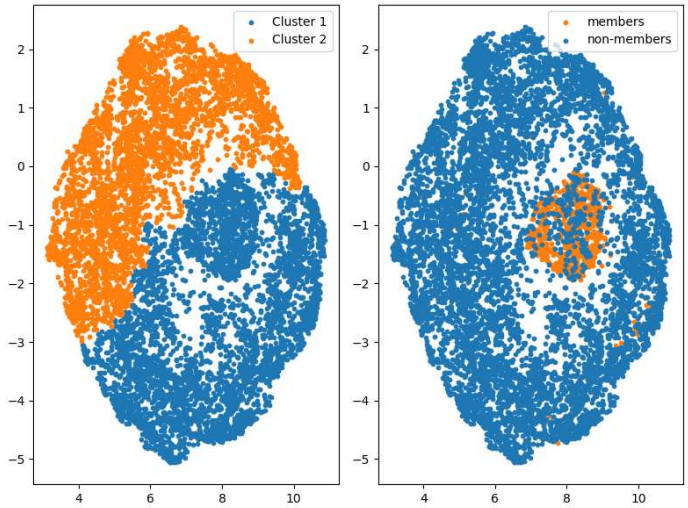


Figure 18: Failed dimensionality reduction: The figure shows an unsuccessful UMAP projection into latent space, making it difficult to separate the data into two distinct clusters. The left panel displays the clustering results produced by the algorithm, while the right panel shows the true underlying clusters.

the result exhibits the same issue found in previous cases: FPs are scattered around the galaxy's center, making it challenging to identify outer region members. Additionally, the FNs correspond to stars located in the gravitational tails.
When analyzing the 500 synthetic catalogs, the results significantly improve in terms of reducing both FPs and FNs. Compared to the **B22** results in Tab. 3d, we achieve a higher number of TPs, while the FP count, although still slightly higher, is comparable and more concentrated in the central region (see Fig. 56). The FNs, on the other hand, remain nearly unchanged and, as shown in Fig. 17b, they remain constant across different threshold values, consistently corresponding to stars in the gravitational tails.
As for the TNs, the outcome mirrors the Sculptor case, showing a less dense TN region near the galaxy center. However, as previously discussed, this limitation can be mitigated by adopting a more relaxed threshold for contaminants.

4.2.4 UMAP without spatial information

As discussed in earlier sections, a recurring issue with dimensionality reduction techniques is their inability to identify stars located in the outer regions and gravitational tails of the galaxy, one of the primary objectives of our work. A first hypothesis is that the inclusion of spatial information introduces a bias in the final classification: stars that share similar PM and CMD features with true members may still be labeled as contaminants solely because of their distance from the galaxy's center.
To investigate this, we performed a UMAP analysis excluding the spatial coordinates, retaining only PM, CMD, and parallax as input features. After a single application of UMAP on the mock catalogs, the resulting latent space (Fig. 19) does show some degree of separation, but it is much less distinct compared to the case where spatial information is included (Fig. 9). The outcome is a high number of misclassified contaminants identified as members, spatially scattered across the field. This renders the results unusable, even after 500 iterations.

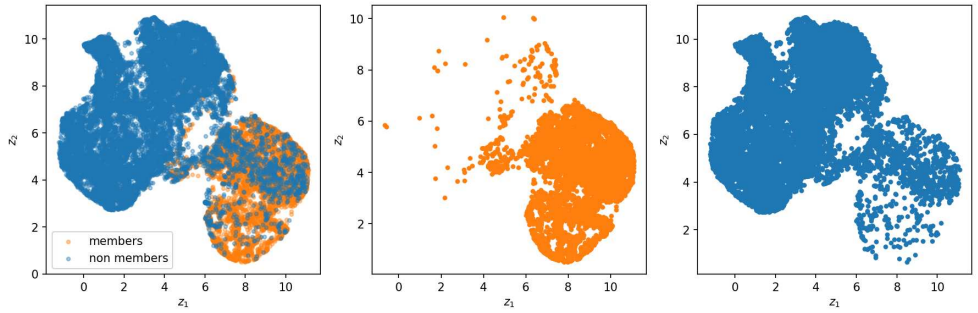


Figure 19: Example of UMAP latent space without spatial information: The plots show the 2D latent space produced by UMAP for the Sculptor-like mock galaxy. The figure follows the same logic of Fig: 9.

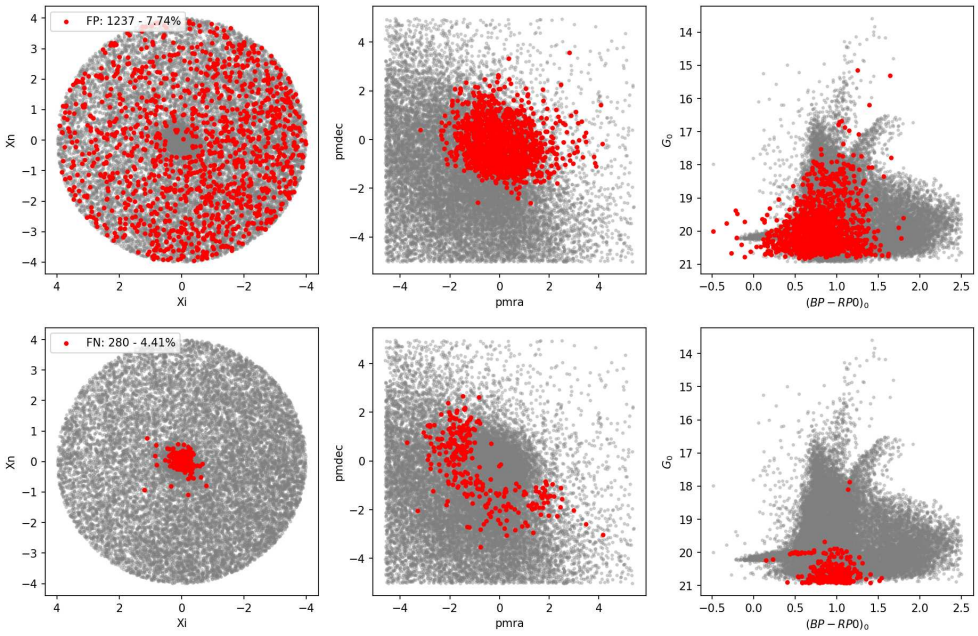


Figure 20: False Positive and False Negative distributions: The plots show the spatial, proper motion, and color–magnitude diagram distributions of the True Positives (TP) and False Negatives (FN) obtained when spatial information is excluded, after a single application of UMAP. These results correspond to the latent space shown in Fig. 19. The first column shows the tangent plane centered on the galaxy center, the second column shows the PM space, and the third column shows the CMD space.

As shown in Fig. 20, the number of FNs increases drastically. In particular, the method systematically fails to recognize member stars fainter than 20 mag.
These results suggest that spatial information is essential for achieving reliable performance with dimensionality reduction techniques in this context.

4.3 Normalizing Flow

The normalizing flow results are discussed differently from those of the previous methods. This is not only because the normalizing flow does not incorporate spatial information and their associated uncertainties, but also because its primary goal is distinct. Here, even if the optimization of the confusion matrix would be the ideal outcome, as say maximizing TP and TN while minimizing FP and FN, rather is assessing whether the method can reliably identify stars in the outer regions of the galaxy and reveal the presence of possible tidal features. This can be translated as the key requirement is that stars in the tidal tails must have high membership probabilities, while it is less critical to recover all member stars.
A key limitation of this method is that it becomes ineffective when the number of contaminant sources significantly exceeds that of the members. In such cases, the distribution of contaminants may dominate and obscure the more complex distribution of the true members, making it difficult for the model to isolate and learn the latter. To mitigate this, as we did for the dimensionality reduction approach, we manually removed a reasonable number of contaminants for each mock catalog. The contaminant population was adjusted so that it remained larger than the number of members, but with a maximum ratio of approximately 3:1. This balance ensures that the model still experiences a realistic foreground population while maintaining its ability to learn the members' distribution. For `SculptorInSextans` we remain with 21,342 sources, 3,935 for `SextansInSextans`, 3,990 for `SextansInDraco` and 3,975 for `DracoInDraco`.

In the following sections, we analyze each mock catalog individually, with particular focus on the Sextans and Draco mocks, which exhibit clear signs of gravitational tails. Specifically, we investigate the spatial distribution of TP and FP in the projected sky frame. Even if the FP count is relatively high, a uniform distribution of these contaminants across the field could still allow the detection of the tails, provided that the TP population includes the stars that trace them.

4.3.1 SculptorInSextans mock catalog

The mock catalog of Sculptor lacks gravitational tails and contains only a few stars in the outer regions (see Fig. 2a). However, we include the results here for completeness. Following the contaminant reduction step, the final dataset analyzed consists of 6,354 members stars and 14,988 contaminants.
In Fig. 21, we show the stars with final membership probability $P > 0.95$ as classified by the NF algorithm. It is clear that the FPs in this case are distributed uniformly across the sky frame. This is quickly explained by Fig. 58 in the Appendix, where it is shown that most FPs fall within the galaxy's PM clump and CMD sequence. This is a direct consequence of not including spatial information in the input features.
The final output allows us to distinguish the dwarf galaxy, although the outer stars are difficult to identify due to the noise introduced by the FPs. The number of TPs

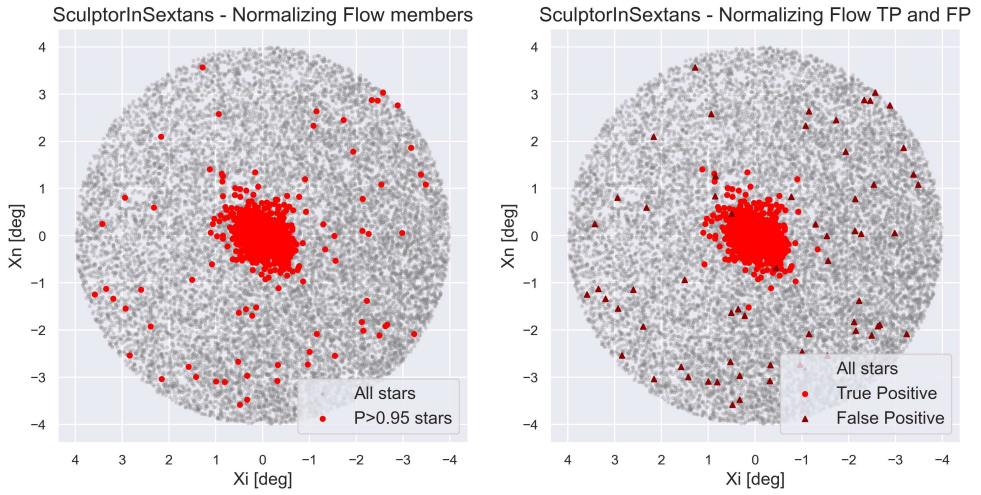


Figure 21: Member Classification by the Normalizing Flow Algorithm: The plots show the final output of the Normalizing Flow method, where stars with membership probability greater than 0.95 are classified as members. The left panel displays the raw NF output, while the right panel separates the TP and FP classifications to facilitate a clearer distinction between true and false positives.

is relatively low compared to previous methods (only 48% of the total members), but the number of FPs is also quite low (0.42% of the total contaminants), which helps preserve higher, density regions and may still enable the detection of gravitational tails.
A promising result is observed in the FN distribution: the FNs are concentrated only in the central region of the galaxy (see Fig. 58). This suggests that, unlike the previous methods, stars in the outer regions are now assigned a higher probability of membership. This outcome indicates that the GMM fitting (see Fig. 14 and Fig. 15) performed well, likely due to the informative structure in the PM and CMD features, which are the only inputs used in this setup.

4.3.2 SextansInSextans mock catalog

The `SextansInSextans` mock catalog is among the most relevant, as the mock galaxy was explicitly constructed to exhibit gravitational tails extending up to 4 degrees. Additionally, the high-density foreground of the real Sextans galaxy makes identifying outer-region members particularly challenging.
Following the contaminant reduction step, the final dataset analyzed consists of 948 members stars and 2,987 contaminants.
Fig. 22 presents very promising results for identifying gravitational tails. The spatial distribution of FPs is similar to the previous Sculptor case, comprising only 0.64% of the contaminants, but the key result is the clear detection of the tails, now distinguishable within the TP population.
This is an exciting outcome: despite the approximations involved in the extracted PDF (see Fig. 23), the method successfully identifies most of the stars in the tails. As shown in the PM space (Fig. 59, Appendix), some FPs are clearly offset from the galaxy clumps. This opens the possibility of post-processing steps, such as removing FPs spatially, to further increase the contrast between TP and FP.
Most importantly, the detection of the tails is achieved. The FNs remain concentrated

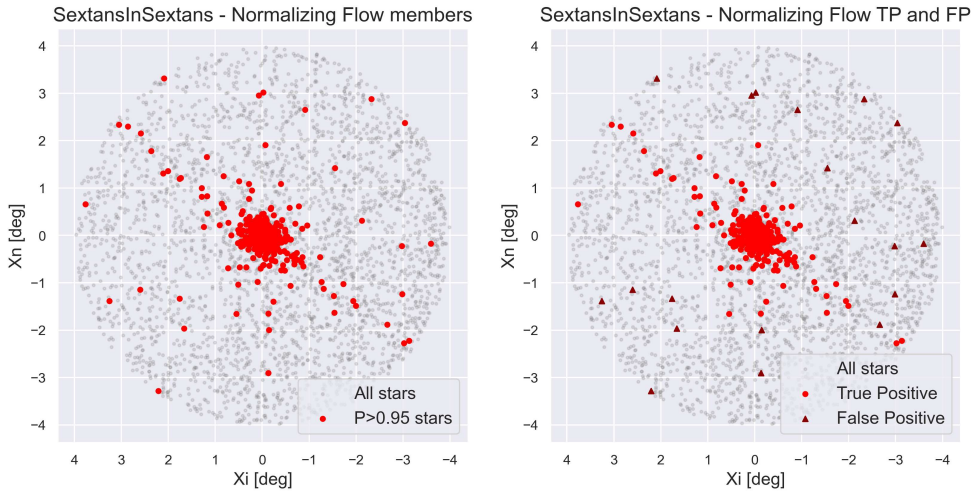


Figure 22: Member Classification by the Normalizing Flow Algorithm: Same as Fig. 21 but for the `SextansInSextans` mock data-set.

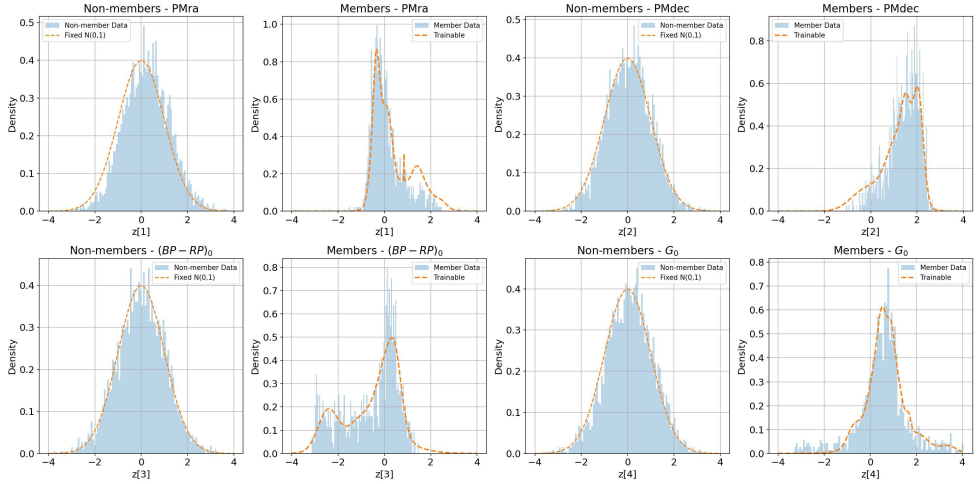


Figure 23: Inferred member distribution for the Sextans in Sextans foreground mock galaxy: The figure shows the probability density functions for the contaminants and the inferred distribution for the members across each of the four dimensions in the `SextansInSextans` mock catalog. The layout of the figure follows the same logic as Fig. 15 in Sec. 3.4.3.

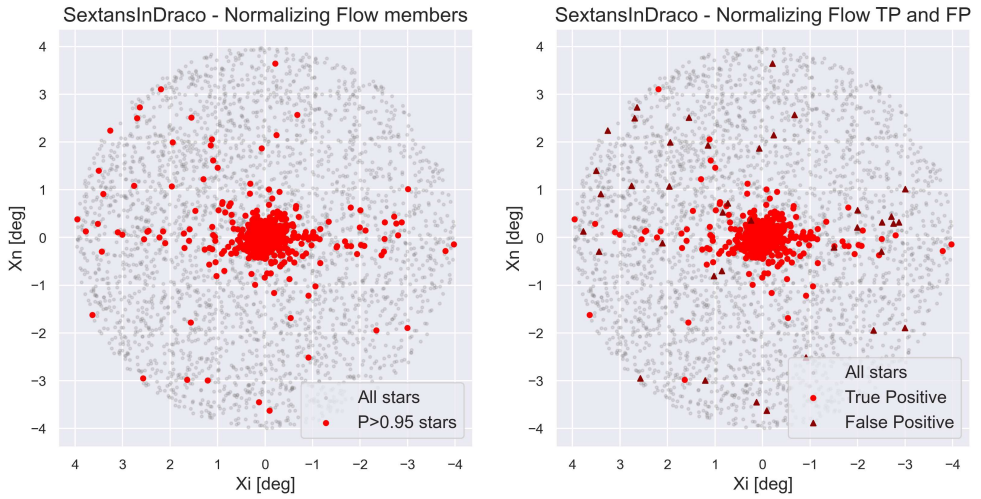


Figure 24: Member Classification by the Normalizing Flow Algorithm: Same as Fig. 21 but for the `SextansInDraco` mock data-set.

in the central region, while TNs are spread uniformly across the sky frame. Even relaxing the membership threshold (e.g. to $0.1 - 0.9$) still allows tail detection, as both TP and FP increase, but not to a degree that obscures the overdensities.

4.3.3 SextansInDraco mock catalog

This case is similar to the previous `SextansInSextans` mock galaxy, but now the foreground is modeled after Draco, which exhibits different properties compared to Sextans. This makes it an interesting test case to assess the robustness of the NF model against varying foreground conditions and to evaluate whether its performance remains stable or is affected.

Following the contaminant reduction step, the final dataset consists of 1,025 members stars and 2,965 contaminants.

As in the `SextansInSextans` case, this mock galaxy includes gravitational tails, although oriented in a different direction on the sky. In Fig. 24, the gravitational tails are still visible, although the detection is less distinct than in the previous case. Interestingly, the FPs, still spread across the sky frame, show a higher concentration in the right portion of the field, which corresponds to a region with higher foreground contamination, as visible in Fig. 2c. But this is more likely to be explained as the algorithm identifies all stars with fainter magnitudes as members, which can be seen in Fig. 60.

Relaxing the membership threshold increases the number of TP in the tails, but also significantly raises the total number of FPs, making the visual identification of the tails less reliable due to the increased background noise.

From Fig. 60 (Appendix), it is evident that the FP rate is higher than in the previous case, reaching 1.28%, approximately double the value observed for `SextansInSextans` despite a comparable number of contaminants.

It is interesting to note that the poorer TP results in this case may not stem from an inadequate fit of the member stars' complex PDF. In fact, as shown in Fig. 25, the member PDF appears better defined and more closely aligned with the true member distribution than in Fig. 23. This suggests that the performance degradation is likely

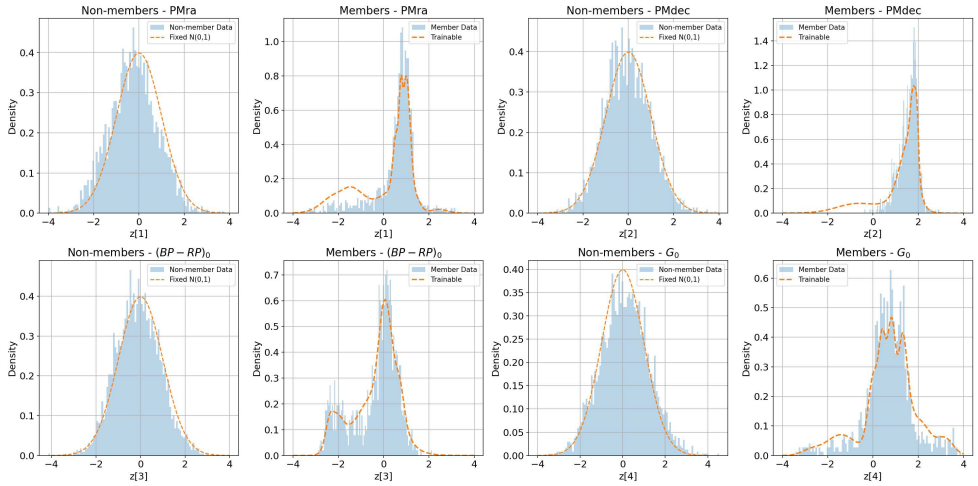


Figure 25: Inferred member distribution for the Sextans in Draco foreground mock galaxy: The figure shows the probability density functions for the contaminants and the inferred distribution for the members across each of the four dimensions in the `SextansInDraco` mock catalog. The layout of the figure follows the same logic as Fig. 15 in Sec. 3.4.3.

due to the properties of the Draco foreground, where the separation between the member and contaminant PDFs is less distinct.
The FNs remain concentrated primarily in the central region of the galaxy, while TNs are broadly distributed across the sky. However, in the PM space, TNs appear more concentrated to the right of the galaxy clump. In the CMD, there is a notable uncovered region at fainter magnitudes, particularly between ~ 20.5 and 21 mag, where TNs are sparse.

4.3.4 DracoInDraco mock catalog

This represents the most challenging scenario for the NF method, as the mock Draco galaxy features very faint and narrow gravitational tails, and, as seen in the previous case, the real Draco foreground tends to increase the number of FPs.

After the contaminant reduction step, the final dataset consists of 1,010 member stars and 2,965 contaminants.
In Fig. 26, a slightly higher density region is still visible, corresponding to the galaxy's left-side tail. However, compared to the `SextansInDraco` case, the number of TP is higher. This may be due to a more precise extraction of the member PDF in this case (Fig. 27) compared to Fig. 25, which leads to a increase in final membership probabilities during Gibbs sampling. This interpretation is supported by the TN distribution shown in Fig. 61 (Appendix), where, due to a strict threshold on the contaminant probability (0.02), no FNs are identified, marking the first time this occurs.
In any case, it is very difficult to detect the presence of tails in this mock with certainty without any some prior knowledge due to their low density.

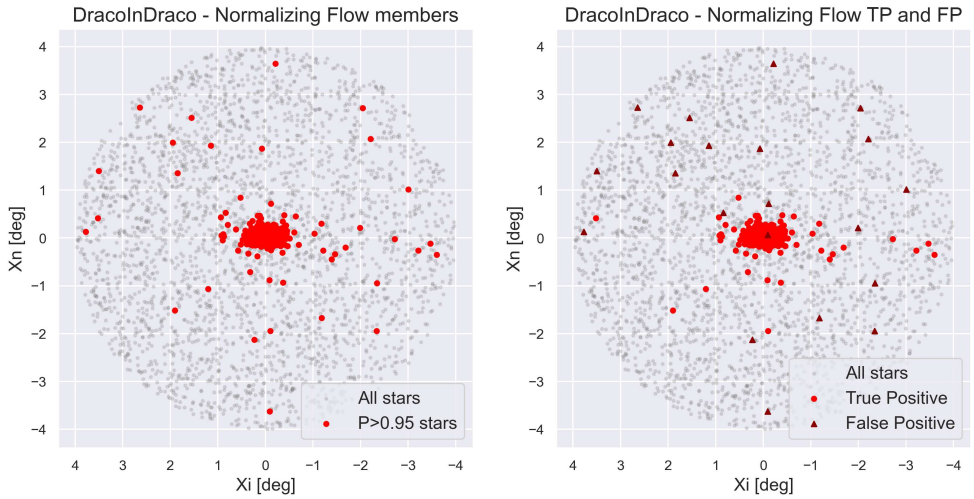


Figure 26: Member Classification by the Normalizing Flow Algorithm: Same as Fig. 21 but for the `DracoInDraco` mock data-set.

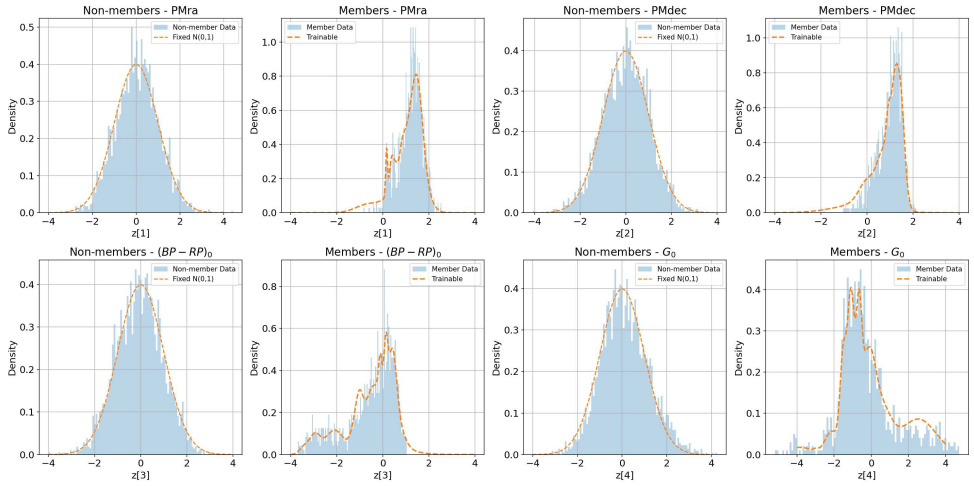
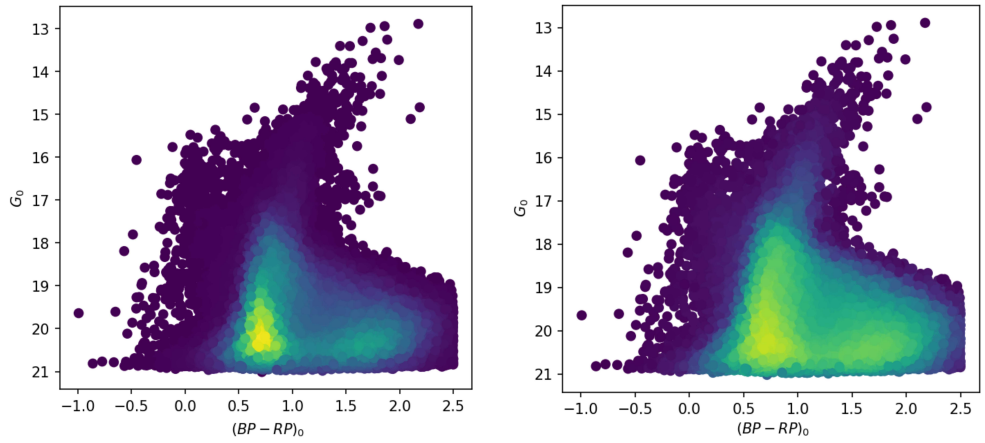


Figure 27: Inferred member distribution for the Draco in Draco foreground mock galaxy: The figure shows the probability density functions for the contaminants and the inferred distribution for the members across each of the four dimensions in the `DracoInDraco` mock catalog. The layout of the figure follows the same logic as Fig. 15 in Sec. 3.4.3.



(a) **(b)**

Figure 28: Difference in the contaminants CMD for DracoinDraco: The figure show the difference between the contaminants CMD from **B22** (Fig. 28a) and the one from the ML implementation (Fig. 28b). For each source is reported their probability values from the look-up map.

5 Discussion and conclusions

In this section, we delve into the results of the various methods presented in Sec. 4, with a closer look at the confusion values and a discussion of the strengths and limitations of each approach in the context of the primary objective of this work, that is detecting tidal tails or signs of past interactions with other galaxies. Therefore, the detection of outer members is critical to our analysis.

We begin with an in-depth discussion of the results obtained from the machine learning implementation of **B22** (Sec. 4.1), then turn to the results obtained using the two alternative methodologies: dimensionality reduction (Sec. 4.2) and Normalizing Flows (Sec. 4.3).

B22 ML: At first glance, the machine learning (ML) implementation of the **B22** method appears to slightly improve the overall results, though not significantly. However, this is not a trivial achievement, as the original method in **B22** is already robust and exhibits very high performance. As shown in the confusion metrics presented in Sec. A, with a probability threshold range of 0.02–0.95, a narrow window centered on the two peaks in the posterior probability histograms (see Fig. 35), the method already correctly classifies the majority of member and contaminant stars, even without including the sources that fall between the threshold interval.
For the Sculptor mock galaxy, which is relatively dense and compact, **B22** achieves excellent results, correctly identifying 94.77% of members and 98.89% of contaminants, with very low misclassification rates: only 0.08% False Positives and 0.28% False Negatives. Results differ for the mock Sextans and Draco galaxies, where the total number of correctly identified members is lower for the same threshold, typically ranging between 70% and 90%. This outcome is expected, as these mock galaxies are more diffuse and embedded in a denser field of contaminants, making classification more challenging. Nevertheless, the True Negatives remain well identified, consis-

tently above 98%, and misclassification rates (FP and FN) remain low.

When comparing these results with those of the ML implementation, we observe that TP and TN values generally improve (see Fig. 16), while the FN values are systematically higher than those of **B22**, and the FP values remain nearly unchanged. This is not inherently negative: if the FNs were concentrated in the central region of the galaxy, the impact, given our goal, would be minimal. However, as discussed in Sec. 4.1, the primary issue in both approaches is that most FNs correspond to stars located in the outer regions of the galaxy, particularly those in the tidal tails.
As seen in Fig. 35, and consistently across other levels, the ML implementation exhibits a higher peak at low probabilities for actual member stars compared to **B22**, indicating a greater number of FNs. From the corresponding confusion plots, it is evident that these FNs are predominantly stars in the outskirts and tidal features, leading to a failure in achieving one of the key goals of this analysis: detecting outer members.
However, this does not necessarily imply that the ML model performs worse than **B22**; rather, it suggests that the ML assigns lower membership probabilities to these stars. Indeed, if we relax the threshold for contaminants, for example by including all stars with probabilities lower than 0.2, these same FN stars begin to appear as FNs in the **B22** method as well. This indicates that **B22** assigns higher probabilities to these sources than the ML model, though still not high enough for them to be classified as members.
A potential explanation lies in the treatment of the contaminant distribution. In **B22**, this distribution is assumed to be uniform, while in the ML implementation, it reflects the actual density distribution of contaminants, particularly in the PM and CMD spaces. As shown in Fig. 6b, the contaminant density varies across the field, creating regions where the probability of being a contaminant is higher. However, spatial plots of FN stars (see Sec. A.2) show that these stars are not necessarily located in areas of higher contaminant density. Instead, they appear to be uniformly distributed across the galaxy's outer regions, with no clear preferential direction.
This observation suggests that the main difference in classification performance may stem from the way the ML model incorporates the contaminant look-up maps in PM and CMD space. As shown in Fig. 28, the CMD generated with the ML implementation displays a more extended region with higher probabilities of being contaminants. Focusing on the mock Draco case, the majority of FNs in the CMD, illustrated in Fig. 39, tend to lie on the central region at faint magnitude of the total CMD, corresponding to the yellow-shaded region in Fig. 28b, where the contaminant density is greater.
These could lead stars to received a higher probability to be a contaminant.
Despite this, the construction of PM and CMD look-up maps using the ML model leads to generally better performance in terms of TP and TN. This enhances the model's ability to correctly identify both member and contaminant stars, particularly in the central regions of the galaxy. Among the three tested configurations, Levels 1 and 2 perform better than **B22**, reinforcing the idea that incorporating the true contaminant distribution significantly improves the performance of the probabilistic membership since A variable density is predicted depending on the position.
However, Level 3 shows overall worse results across all mock cases. The primary methodological change at this level was the use of the subtraction method to construct the galaxy PM look-up map, similar to the CMD construction in Level 1. This outcome suggests that both the PM and CMD look-up maps determined with the

ML, may be too noisy and, in some regions, may not capture the general behavior of contaminant distribution. Despite this, the construction of PM and CMD look-up maps using the ML model generally leads to improved performance in terms of TP and TN, enhancing the model's ability to correctly identify both member and contaminant stars, particularly in the central regions of the galaxy. Among the three tested configurations, Levels 1 and 2 slightly outperform **B22**, reinforcing the idea that incorporating the true contaminant distribution significantly improves the probabilistic membership assignment, since a variable density is predicted as a function of position. In contrast, Level 3 yields systematically worse results across all mock cases. The main methodological difference at this level was the adoption of the subtraction method to construct the galaxy PM look-up map, analogous to the CMD construction in Level 2. These results suggest that look-up maps for PM and CMD determined purely with ML may be too noisy and, in some regions, may fail to capture the underlying contaminant distribution accurately, and that the PM map calculation method used in **B22** is more effective and has a stronger influence on the final membership probability estimation.
Considering the overall performance across all levels, Level 1 emerges as the most consistent and reliable. This indicates that combining a machine learning approach to model the contaminant distribution with the statistical framework presented in **B22** can yield improved general results. However, this method still falls short when the primary goal is the detection of external structures such as tidal tails, where outer members are often misclassified.

UMAP: First of all, considering the primary objective of identifying stars in the outer regions, the results in Sec. 4.2 clearly show that UMAP, in the configuration applied here, is not the most suitable methodology for this purpose.
Aside from not always providing consistent or robust results, particularly in the case of the 500 iterations applied to the mock Sextans catalogs, where UMAP fails entirely, the majority of FNs correspond to stars located in the outer regions.
When applying UMAP only once to the catalogs, the resulting classification is not reliable enough to be used for final analysis. The large number of FPs, scattered across the field, prevents an accurate identification of members and distorts the apparent shape of the galaxy, making any visual morphological study impractical.
Nevertheless, UMAP could still provide useful insights in future work. For example, extracting the correlation matrix of the features used here could help identify patterns among the physical quantities that might strengthen the classification model. It should be noted that our use of UMAP in this work was intentionally simple, focusing solely on dimensionality reduction. The UMAP documentation indicates the possibility of developing more complex approaches, such as combining multiple UMAP models or embedding the latent space in non-Cartesian geometries (e.g. spherical spaces) or with custom metric definitions.

Applying UMAP to the 500 synthetic catalogs was intended to emulate its application to multiple distinct observations of the same object, where measurement uncertainties cause values to fluctuate around a given mean. In such cases, each dimensionality reduction run effectively "sees" each star differently, potentially leading to different classifications. However, as shown in Sec. 4.2, this approach fails for the Sextans mock catalog. This may be due to the complexity of the true Sextans contaminant field, where, in the synthetic catalogs, UMAP does not effectively separate the data. The precise cause was not investigated in detail, as the method already proved far from meeting the primary goal of detecting outer-region members.

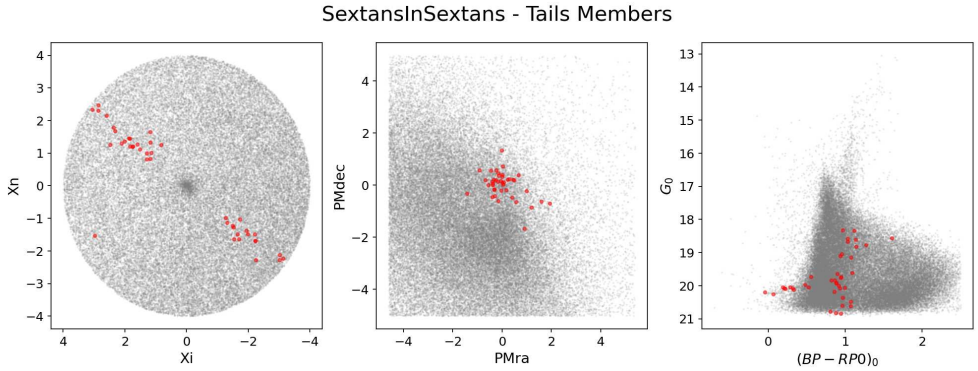


Figure 29: Distribution of true tidal-tail member stars for the `SextansInSextans` mock galaxy: The panels show the positions of stars belonging to the tidal tails (distance from the galaxy center > 0.8 degrees) in three planes: (left) sky coordinates, (middle) proper motion (PM) space, and (right) color–magnitude diagram (CMD) space. These stars are identified using the ground truth of the mock catalog to illustrate the spatial and kinematic extent of the tails.

On the other hand, for the Sculptor and Draco mock catalogs (Fig. 55 and Fig. 56), the results are more promising, and in fact not far from those of **B22** and **B22 ML**. It is important to note that, for computational reasons, the number of actual contaminants included in the UMAP analysis is smaller; however, in relative terms, UMAP performs better at detecting members, with only a small increase in FPs, mainly located in the galaxy's central regions. Additionally, inspection of the CMD shows that a non-negligible fraction of these FPs lies outside the galaxy's CMD shape, making their manual removal possible.
Contaminant detection is less effective. There is a noticeable drop in density in the region corresponding to the galaxy's position, likely because UMAP, when given spatial coordinates as input features, tends to classify inner stars as members even if their CMD positions are inconsistent with true galaxy members.

This leads us to the same conclusion drawn from the **B22 ML** results: spatial information has a very strong influence on the classification outcome, which is detrimental when the goal is to detect outer-region members. This is why we attempted a UMAP configuration without spatial features; however, in this case, the remaining features alone did not provide sufficient discriminatory power for the model to perform effectively.

Normalizing Flow: For the detection of tidal–tail stars, the Normalizing Flow (NF) algorithm was the only method that achieved this objective. Therefore, we discuss its results from a different perspective, without focusing on the TP, FP, FN, and TN values as in the previous methodologies. In this case, we are not interested in detecting the majority of the stars, but rather in determining whether the tidal tails are present. Our discussion concentrates mainly on the `SextansInSextans`, `SextansInDraco` and `DracoInDraco` mock galaxies, since they contain tidal tails, while we also include `SculptorInSculptor` for completeness and to verify result consistency.

The overall results for the mocks with tidal tails are fairly consistent with each other, as in all three cases the tidal tails members are detected for the first time, albeit not entirely. As shown in Sec. 4.3, the tidal tails are clearly visible, even though some FPs introduce noise in the sky frame. As seen in Fig. 59, the FPs are uniformly distributed across the field but are generally far from the galaxy. This behaviour is

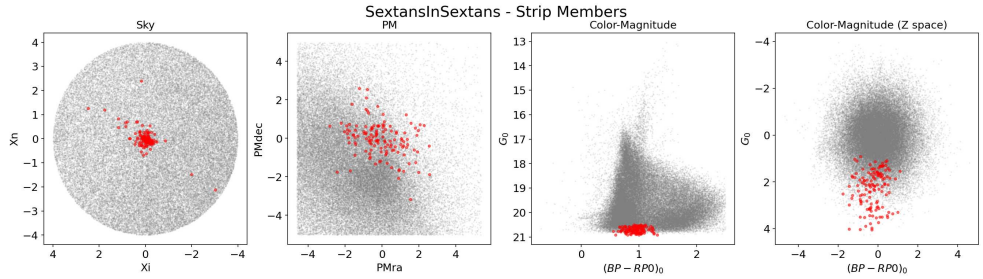


Figure 30: Faint magnitude "strip" members detected by the Normalizing Flow model in the `SextansInSextans` mock galaxy: The panels display the distribution of stars in (first) sky coordinates, (second) proper motion (PM) space, (third) color–magnitude diagram (CMD), and (fourth) the two coordinate of the Normalizing Flow latent space corresponding to the CMD. The strip corresponds to stars around $G \approx 21$.

likely due to the fact that, during the NF analysis, we did not include any spatial information, but only PM and CMD features were used. Consequently, contaminants in the outer region with PM and CMD values similar to members can be misclassified as members. This feature of the method, which does not rely on spatial information, is precisely what seems to enable it to recover the members of the tidal tails, but without introducing the heavy noise seen in the UMAP run without spatial features. Moreover, many FPs can be removed manually by discarding objects located too far from galaxy's PM clumps or CMD sequences to be true members. For example, this applies to stars far from the horizontal branch region of the CMD, where stars have almost identical magnitudes, or from the red giant branch, as the two branches are easily recognizable as evolutionary stages of the galaxy's CMD. Similarly, some FPs in the PM space are far from the galaxy's PM clump and can be confidently removed as contaminants.

A positive result is that the FN distribution, across all mock catalogs, is largely concentrated in the galaxy's central region. This means that very few (or any) tidal–tail stars receive low posterior probabilities, in contrast to previous methods. The contaminant distribution is also relatively uniform, without density gradients or local underestimations near the galaxy's position.

The smaller number of TP, FP, FN, TN that we get reflects our choice of probability threshold: as with the other methods, the $0.02 - 0.95$ cut removes a significant fraction of stars from the final classification (see Fig. 57). However, this choice has little impact on the primary goal, detecting tidal tails. The adopted threshold also explains the gap observed in the CMD at magnitudes $20.0 \lesssim G_0 \lesssim 20.5$, just above the faint strip. This region coincides with the highest density of contaminants (see Fig. 58 and following), making it particularly difficult for genuine members to reach high membership probabilities. By relaxing the membership threshold from 0.95 to lower values, stars in this gap begin to be classified as members, but at the cost of introducing a larger number of FPs. We therefore conclude that a threshold of 0.95 provides a good compromise, being sufficient to identify tidal tails while still yielding a reliable reconstruction of the galaxy's CMD morphology.

We took `SextansInSextans` as example in this discussion because its tidal tails are the most visible, by construction of the mock. However, in `SextansInDraco` the tails remain easily visible despite a higher FP count, and in `DracoInDraco` they are thin and harder to detect but still recoverable after a manual contaminants cleaning on PM and CMD plots.

One concern that emerged from the analysis of the four mock galaxies is the presence of a consistent feature in the CMD of the TP: a strip of faint stars around $G \sim 21$ mag that are systematically identified as members. This may be a consequence of the mock construction, in which mock galaxy stars extend to fainter magnitudes than the contaminants, making this CMD region stand out and therefore easier for the NF model to classify.
A possible concern could be that tidal-tail members are primarily drawn from this faint strip. To test this, we selected only stars located in the tidal tails (i.e., beyond 1° from the galaxy center) and examined their distribution in PM and CMD space. The resulting plots in Fig. 29 show that tidal-tail stars populate nearly all evolutionary phases of the CMD and are not confined to faint magnitudes.
Conversely in Fig. 30, when restricting the sample to stars with $G > 20.5$, corresponding to the faint TP strip, we find that these stars are mainly concentrated in the central regions of the galaxy and do not correspond to the tails. This behavior indicates that their correct identification arises from the NF latent space, where this faint strip is clearly separated from the Gaussian distribution of contaminants and more likely associated with the complex distribution of true members.

Summarizing the results obtained from the analysis of the performance of the three methods and their ability to identify true members and, in particular, their spatial distribution, we can draw some final considerations.
Starting with the dimensionality reduction approach, as implemented here with UMAP, we found that it is not a robust method: it does not always succeed in clearly separating members from contaminants in the two-dimensional latent space. This limits its applicability, as the algorithm would need to be adapted each time to the specific dataset under study, making it an impractical tool for general use. However, in cases where UMAP does converge to a meaningful separation, it provides a reasonable estimate of the total number of stars belonging to the galaxy, recovering more than 95% of true members. The drawback is that this success comes at the expense of missing member stars in the outer regions, which are systematically not identified and, equally importantly, to run this algorithm it is essential to remove a significant number of known contaminants.

A similar conclusion can be drawn for the machine learning implementation of **B22**. The results are broadly comparable to UMAP in terms of overall performance, but the method is significantly more robust and consistent. In particular, by explicitly incorporating the distribution of contaminants, it systematically improves the identification of true members. Nevertheless, this refinement also leads to a deterioration in the recovery of members in the outermost regions. These stars are often assigned low membership probabilities: lower than those given by **B22**, which are still not high enough to exceed the membership threshold, thereby **B22+ML** does not altering the final balance when the method is applied to real catalog data.

The NF method appears to be the most promising among those tested, as it was the only approach capable of consistently identifying tidal tails in all mock samples where such structures were present. Despite relying on a reduced set of features, namely PM and CMD, the model successfully produced a bimodal posterior membership probability distribution, with contaminants concentrated at low probabilities and true members at high probabilities. This outcome is visibly comparable to the results

obtained through the more complex analysis of **B22+ML** (see Fig. 35 and Fig. 57 for comparison). Crucially, unlike the previous methods, the stars belonging to the tidal tails are now associated with the high-probability peak of the members, whereas before they were associated with the low-probability peak of the contaminants.
The main limitation of the NF approach is its reduced ability to recover the largest possible number of member stars. However, the majority of these missed members lie in the inner regions of the dwarf galaxy, which is less critical when the primary objective is the detection of tidal features. A further drawback is that some contaminants dispersed across the sky are misclassified as members, since spatial information relative to the galactic center is not explicitly considered in the model. This issue can nonetheless be mitigated by applying post-selection criteria in PM and CMD space, effectively removing such spurious classifications. The final limitation, similar to that of UMAP, is that in order to run this algorithm, it is essential to remove a significant number of known contaminants, so that the complex distribution of members can be visible compared to that of contaminants.

When opting for the NF approach, a natural question arises: how can its performance and final results be further improved? A straightforward step would be to incorporate additional observational features beyond Gaia's PM and CMD, such as magnitudes and colors in other photometric bands, or even radial velocity information. Spatial information, we remember, should not be included, both because of the intrinsic structure of the algorithm, where training and testing data must remain comparable, despite the strong positional variations between the two, and because including position would introduce a bias against detecting members at large distances. The inclusion of extra features could also benefit UMAP, since increasing the amount of discriminating information would likely improve its ability to separate members from contaminants. This remains, however, only a hypothesis, as such extensions were not investigated in this work.
Another avenue for improvement concerns the NF model itself. The implementation adopted here is based on a basic Real NVP architecture, which could be refined to achieve more accurate normalizing transformations and reduce the loss function more effectively. Enhancing the architecture in this way would likely improve the reliability of the probability estimates.
Finally, the Gaussian Mixture Model component could also be optimized: as shown in figures such as Fig. 15, the fitted GMM does not perfectly capture the true distribution of the extracted members, resulting in a probability density function that only approximates the underlying structure. A better fit would translate into higher probabilities during Gibbs sampling, thereby increasing the number of identified members in a more realistic manner.

A major strength of the NF approach is its ability to operate directly on raw data, or on corrected values such as the color $(BP - RP)_0$ and magnitude G_0, without requiring additional statistical preprocessing steps like the construction of look-up maps, as in **B22+ML**. This flexibility makes it possible to extend the method to data from multiple catalogs, rather than being restricted solely to Gaia. Looking further ahead, the release of Gaia DR4, expected in December 2026[10], will provide even more precise astrometric and photometric information, offering an ideal opportunity to reassess and refine the methodology presented in this thesis.

As mentioned above, an important limitation of the NF method is that its application requires a reasonable balance between the number of sources belonging to the

[10]https://www.cosmos.esa.int/web/gaia/data-release-4

dwarf galaxy and the number of contaminants. While this condition can be enforced in mock catalogs by directly removing contaminants, in real datasets alternative strategies must be adopted to reliably identify and exclude contaminating stars from the sample. A promising approach would be to combine the **B22+ML** method with NF, where **B22+ML** is first applied to identify contaminants, for example by selecting sources with membership probability < 0.001. Indeed, tests have shown that choosing such a threshold ensures that no member star is mistakenly classified as a contaminant in the mock samples. This result can be understood by noting that **B22+ML**, even though includes the spatial component, assigns a non-zero membership probability even to stars at large distances from the galactic center if they belong to the PM cluster or the CMD sequence of the dwarf galaxy. As a concrete example, in the case of `SextanInSextans`, applying this threshold allows for the identification of 91.34% of true contaminants, after which one can decide whether to remove the entire set or only a fraction of the selected stars.

In conclusion, this thesis has provided valuable insights into the applicability and limitations of the three methods investigated, each of which offers different advantages depending on the final objective of the analysis. Among them, Normalizing Flow has emerged as a particularly promising tool for detecting structures in the outer regions of dwarf galaxies. Its ability to identify tidal features and extended components, despite relying only on PM and CMD information, highlights its potential for uncovering signatures of past dynamical interactions. With further refinement and optimization of the model, it will be especially compelling to reapply this analysis to real systems such as Sculptor, Sextans, and Draco with the Gaia eDR3 data, and to extend the methodology to a larger sample of Local Group dwarfs. This would enable a more systematic investigation of their morphology and provide new insights into their formation pathways and evolutionary histories.

Acknowledgements

I would like to express my deepest gratitude to Dr. Giuseppina Battaglia and her entire team, in particular Dr. Thomas Guillaume and José María, for welcoming, supporting, and guiding me throughout this long yet wonderful thesis journey. My experience at the Instituto de Astrofísica de Canarias has been extremely formative, allowing me to closely experience the daily life of a research group and inspiring in me dedication, passion, curiosity, and the constant drive to give my best every day. This experience, which has made me fully understand the reasons why I chose this academic path, will always stay with me, and I hope it will be just the beginning of a long and exciting adventure in the world of astrophysics.

A Appendix

A.1 B21 Confusion plots

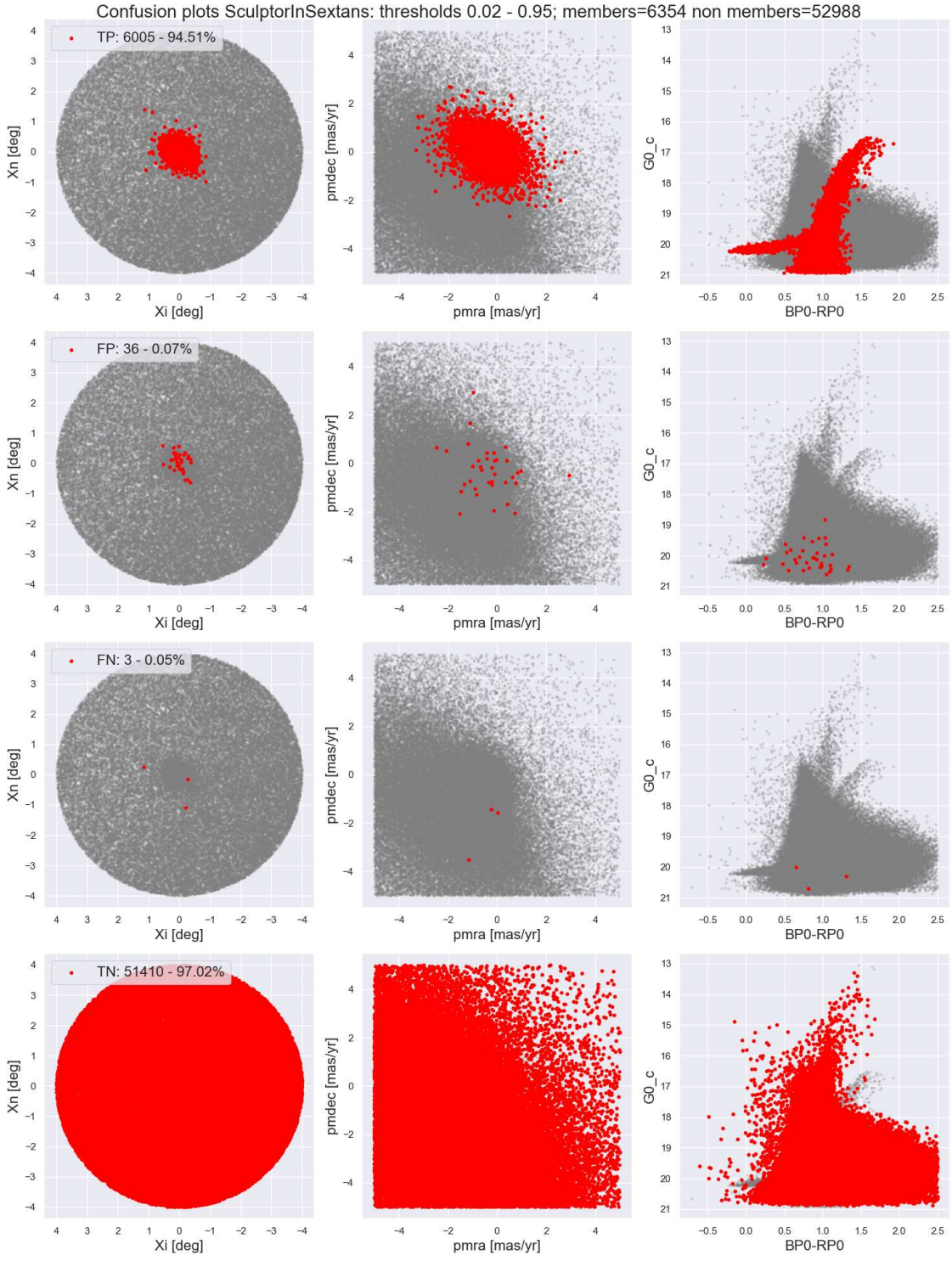


Figure 31: Sculptor In Sextans confusion plots: the figure shows the confusion results for choice of threshold 0.02-0.95

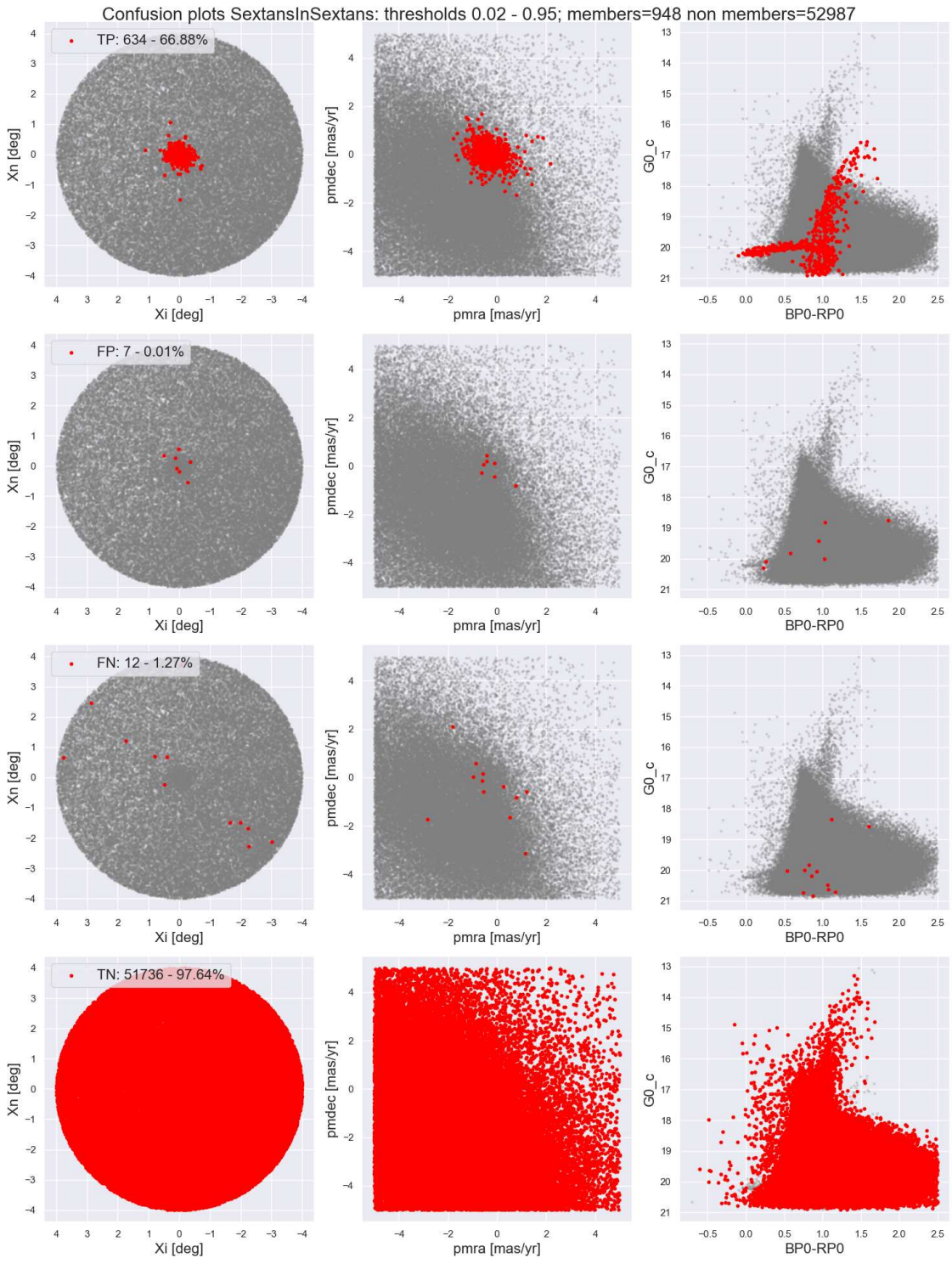


Figure 32: Sextans In Sextans confusion plots: the figure shows the confusion results for choice of threshold 0.02-0.95

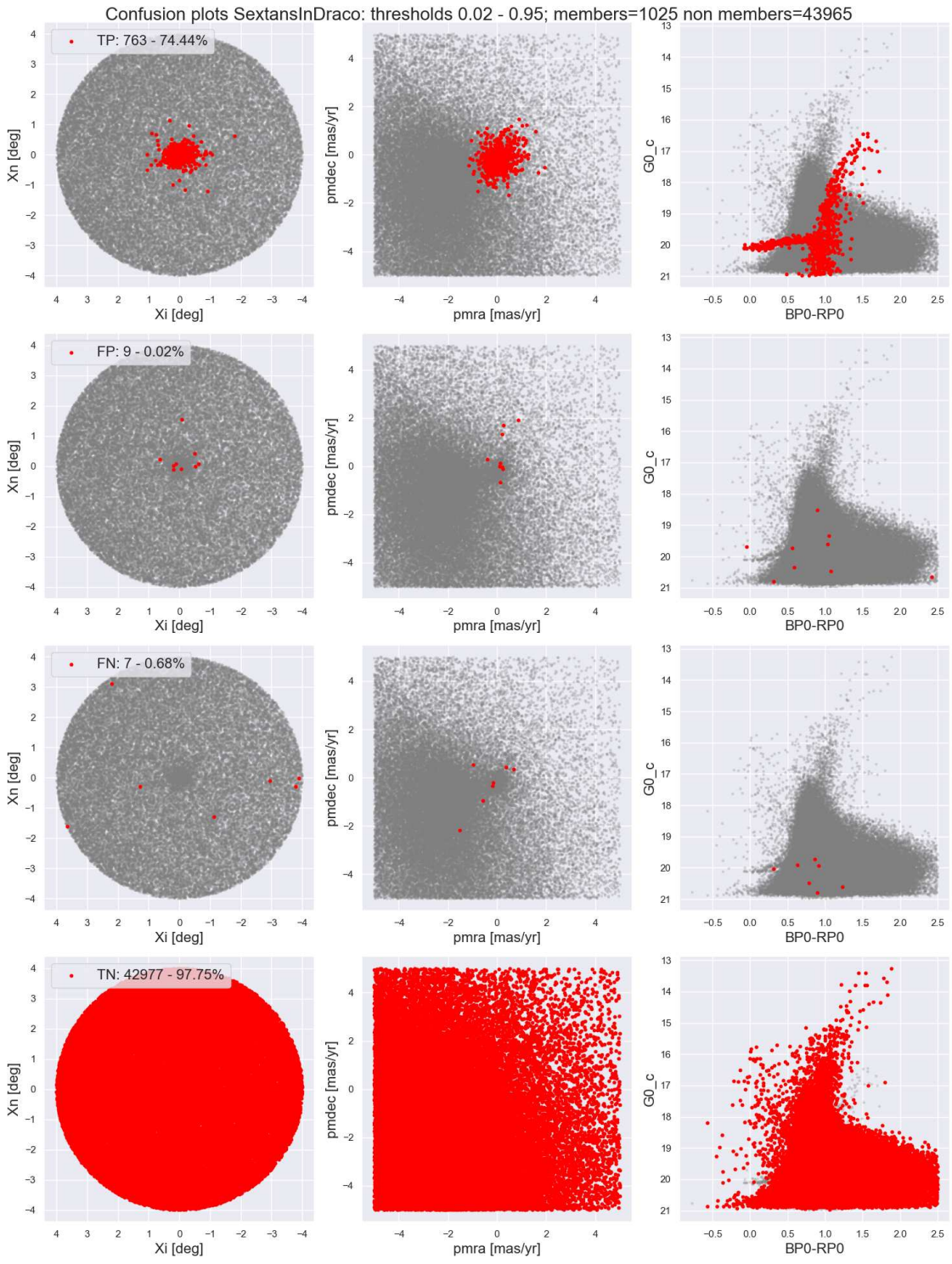


Figure 33: Sextans In Draco confusion plots: the figure shows the confusion results for choice of threshold 0.02-0.95

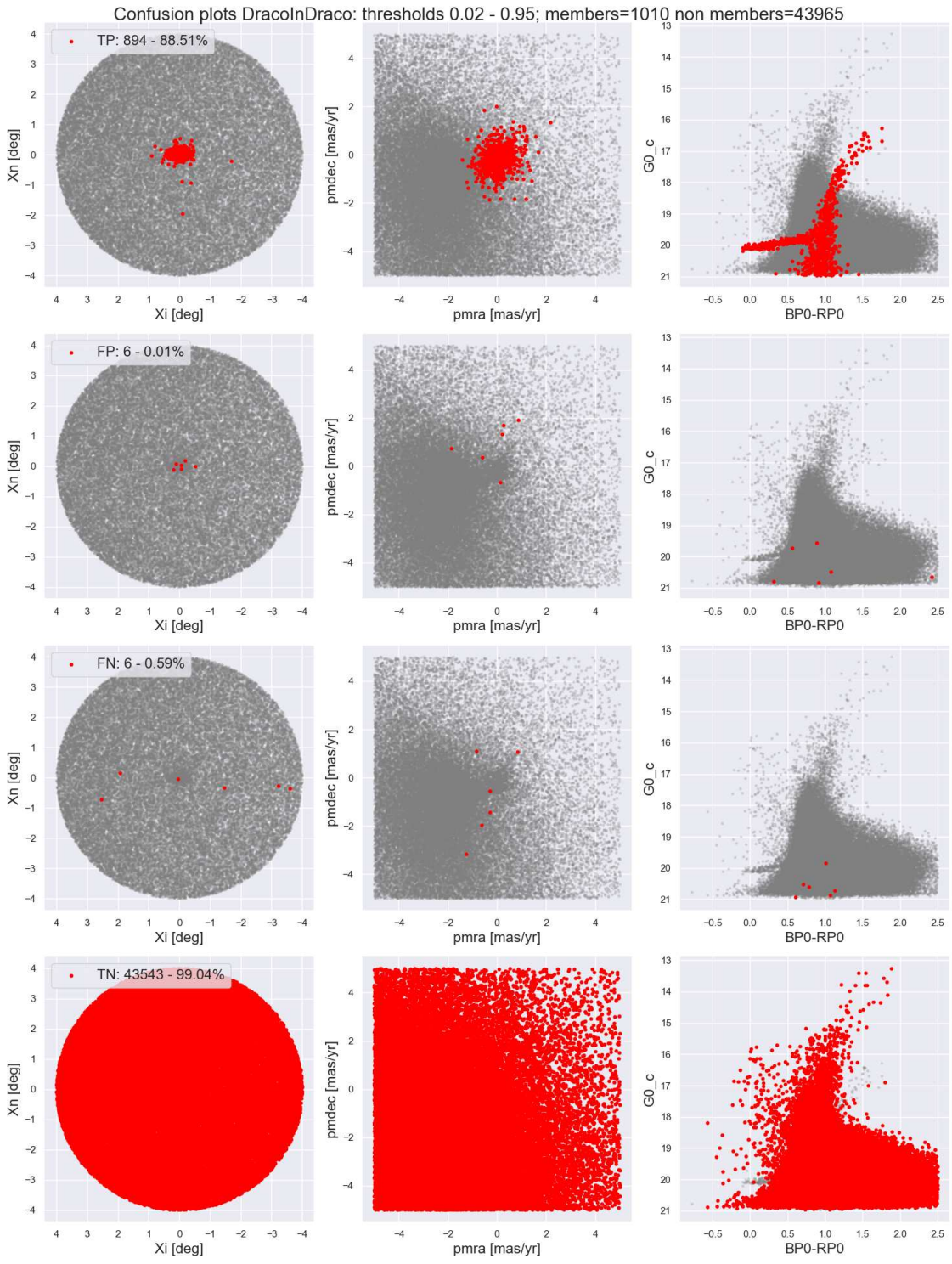
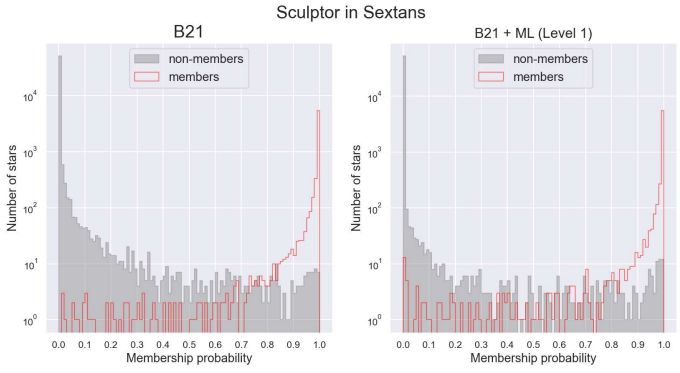


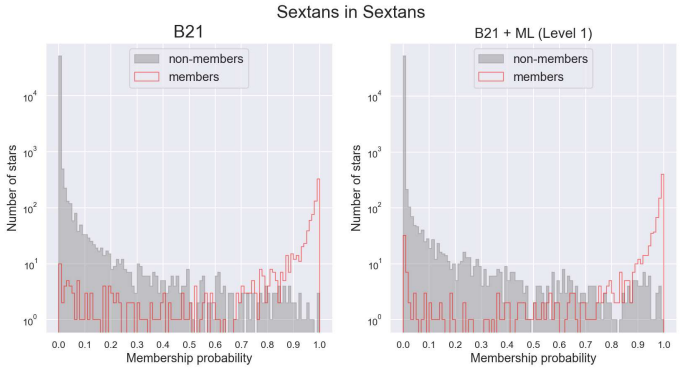
Figure 34: Draco In Draco confusion plots: the figure shows the confusion results for choice of threshold 0.02-0.95

A.2 B21 + ML results

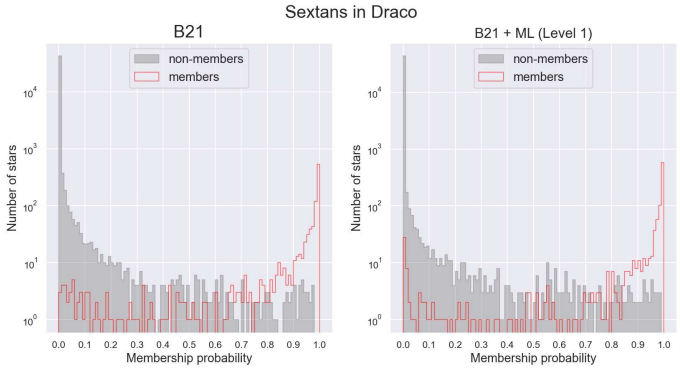
A.2.1 First Level



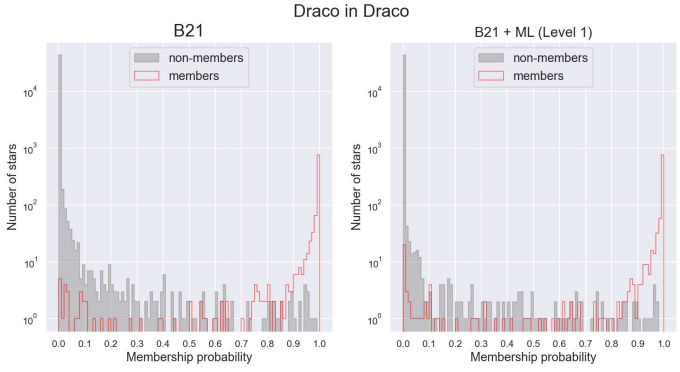
(a)



(b)



(c)



(d)

Figure 35: B21 + ML posterior probability histograms: Each plot shows the posterior probabilities obtained at the First Level in Sec. 4.1.1. Member stars are indicated by the thin red line, while contaminants are shown as a grey histogram.

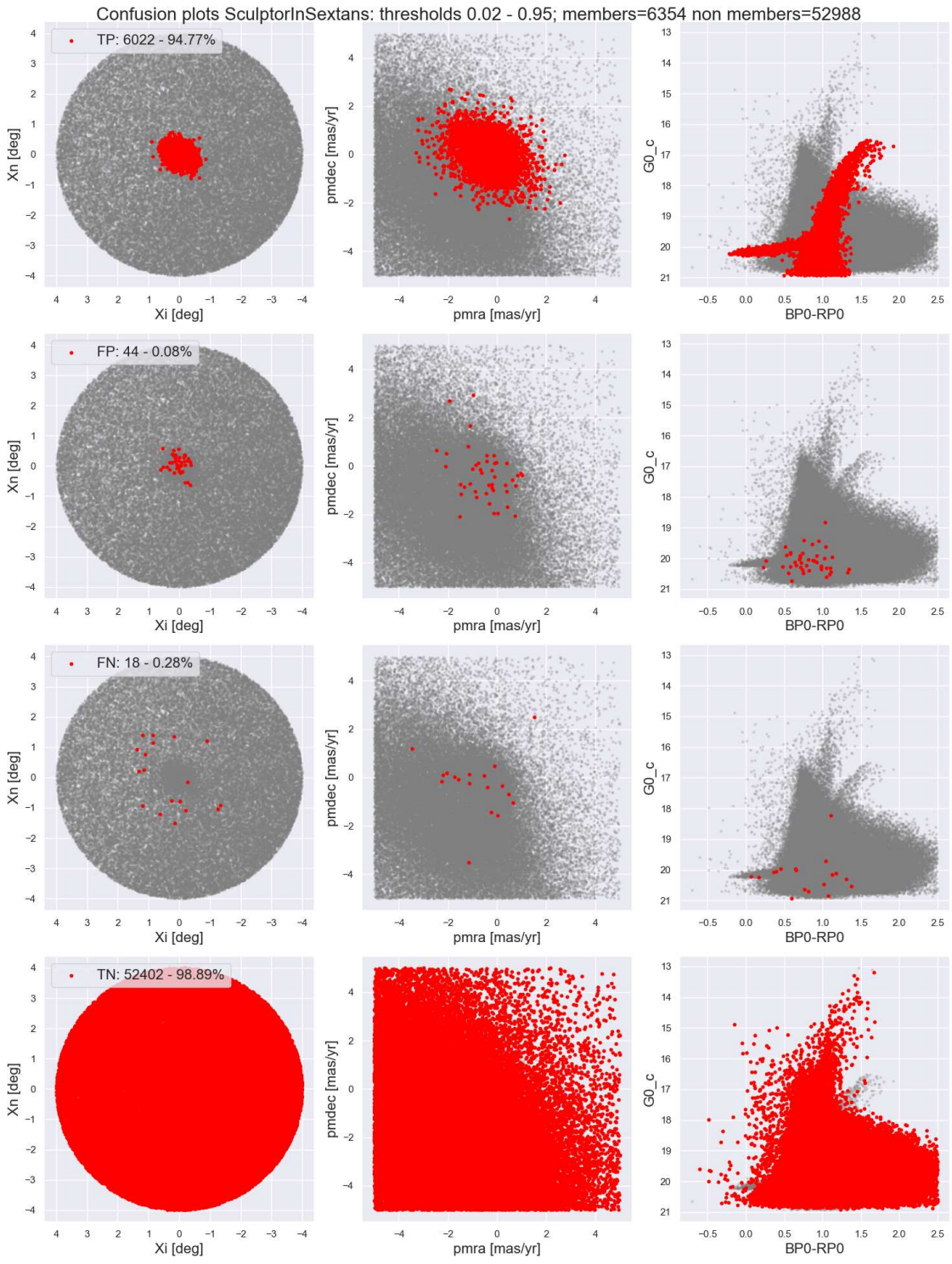


Figure 36: Sculptor In Sextans confusion plots: the figure shows the confusion results for choice of threshold 0.02-0.95 corresponding to Fig. 35a

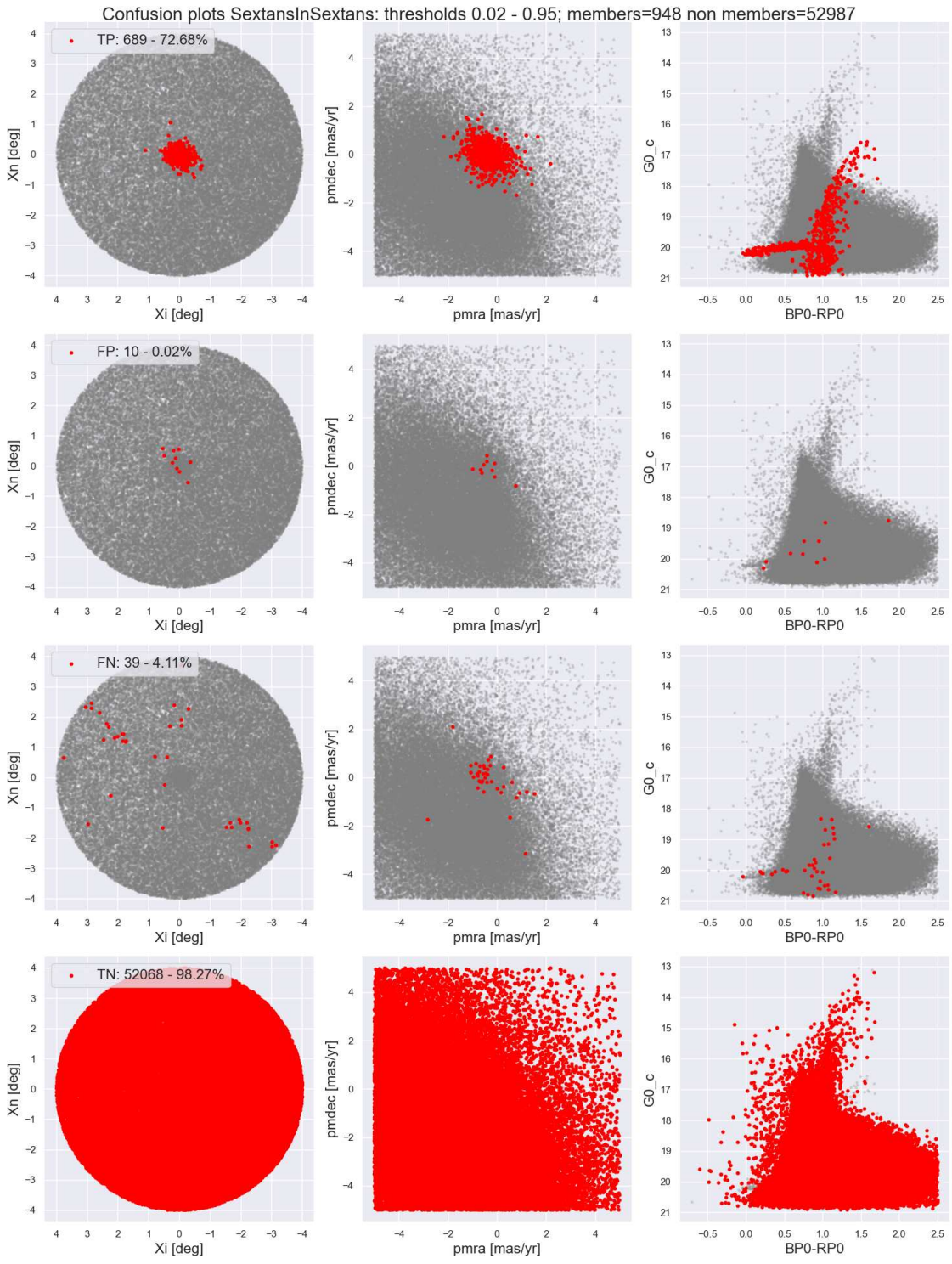


Figure 37: Sextans In Sextans confusion plots ML: the figure shows the confusion results for choice of threshold 0.02-0.95 corresponding to Fig. 35b

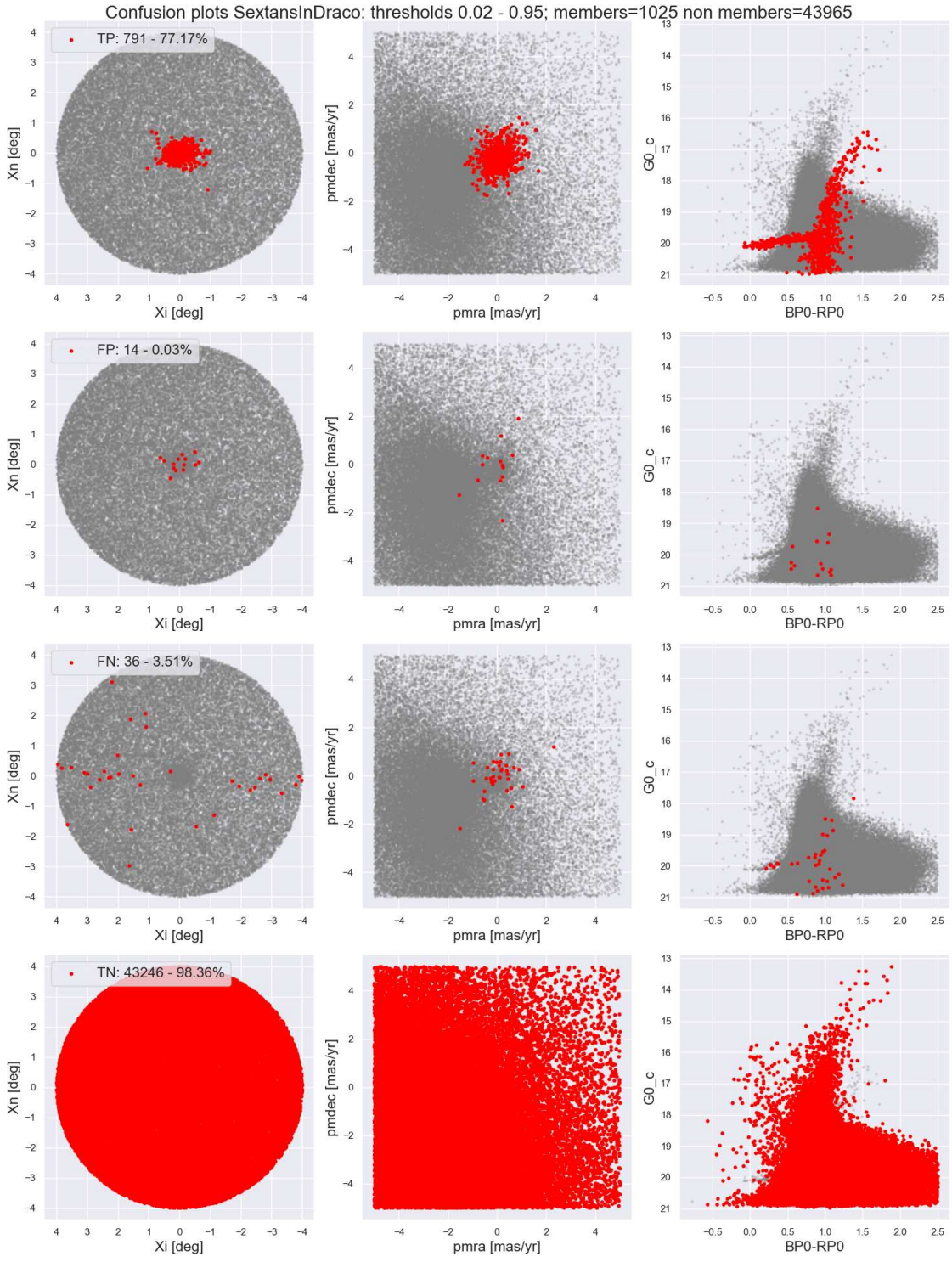


Figure 38: Sextans In Draco confusion plots ML: the figure shows the confusion results for choice of threshold 0.02-0.95 corresponding to Fig. 35c

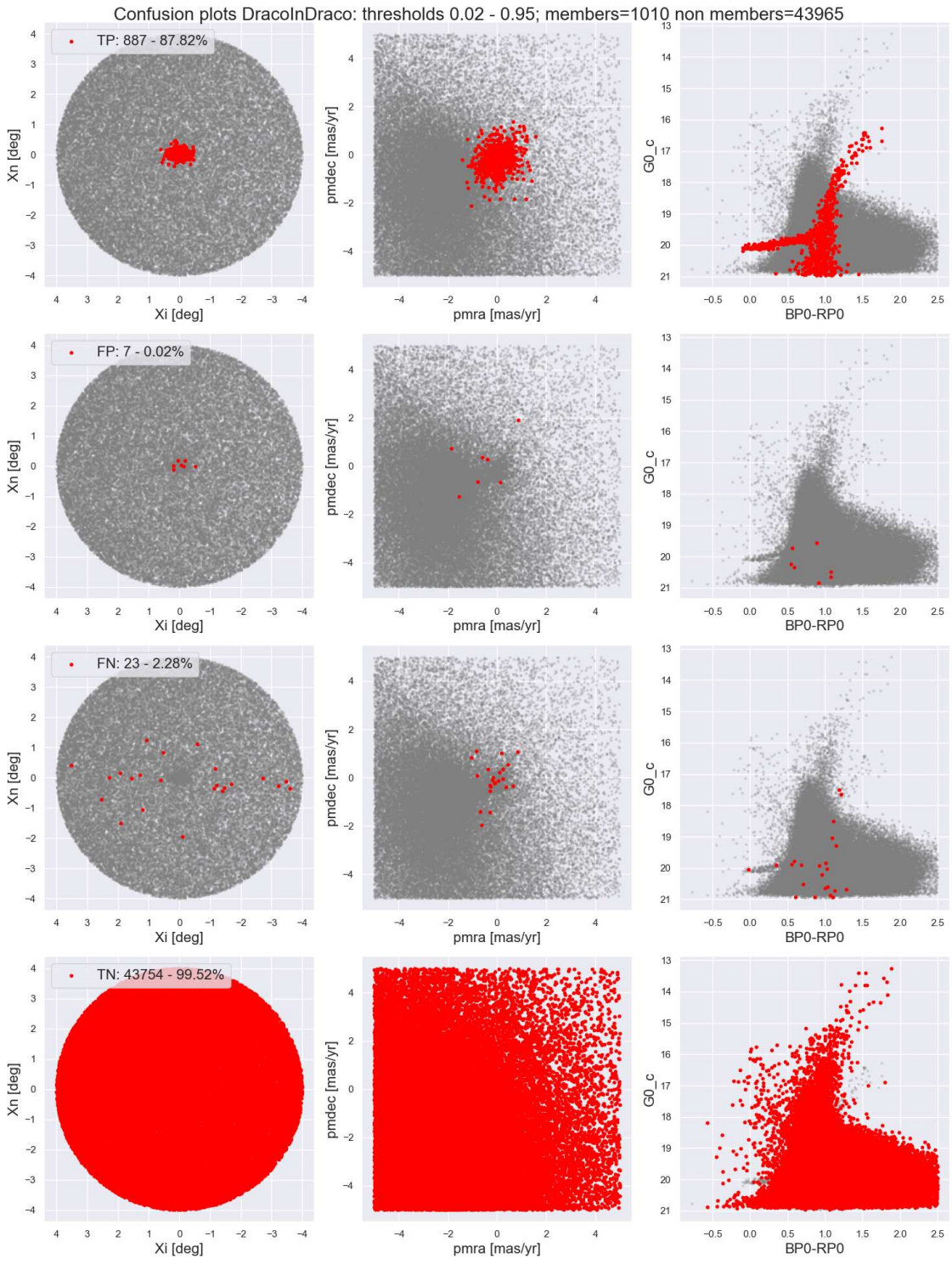


Figure 39: Draco In Draco confusion plots ML: the figure shows the confusion results for choice of threshold 0.02-0.95 corresponding to Fig. 35d

A.2.2 Second Level

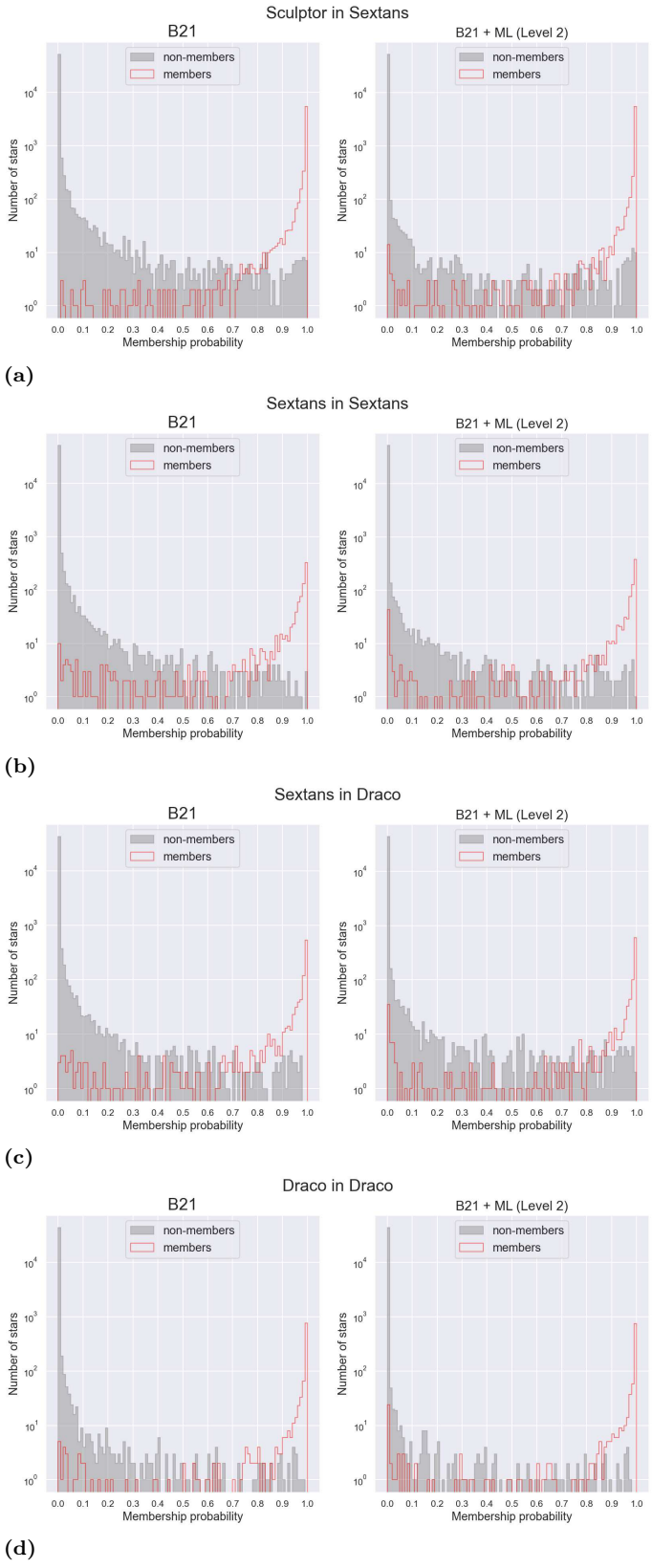


Figure 40: B21 + ML posterior probability histograms: Each plot shows the posterior probabilities obtained at the Second Level in Sec. 4.1.2. Member stars are indicated by the thin red line, while contaminants are shown as a grey histogram.

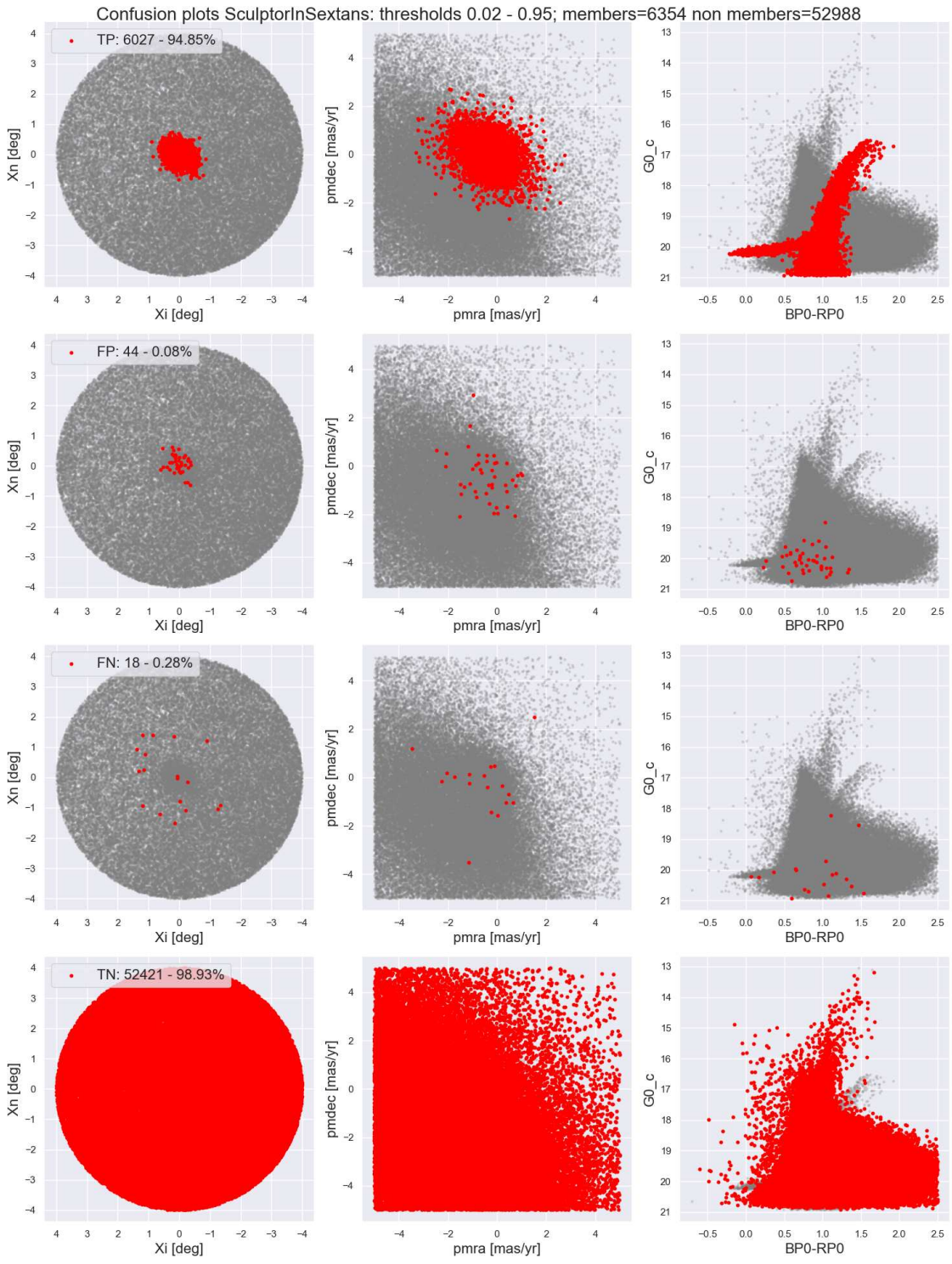


Figure 41: Sculptor In Sextans confusion plots: the figure shows the confusion results for choice of threshold 0.02-0.95 corresponding to Fig. 40a

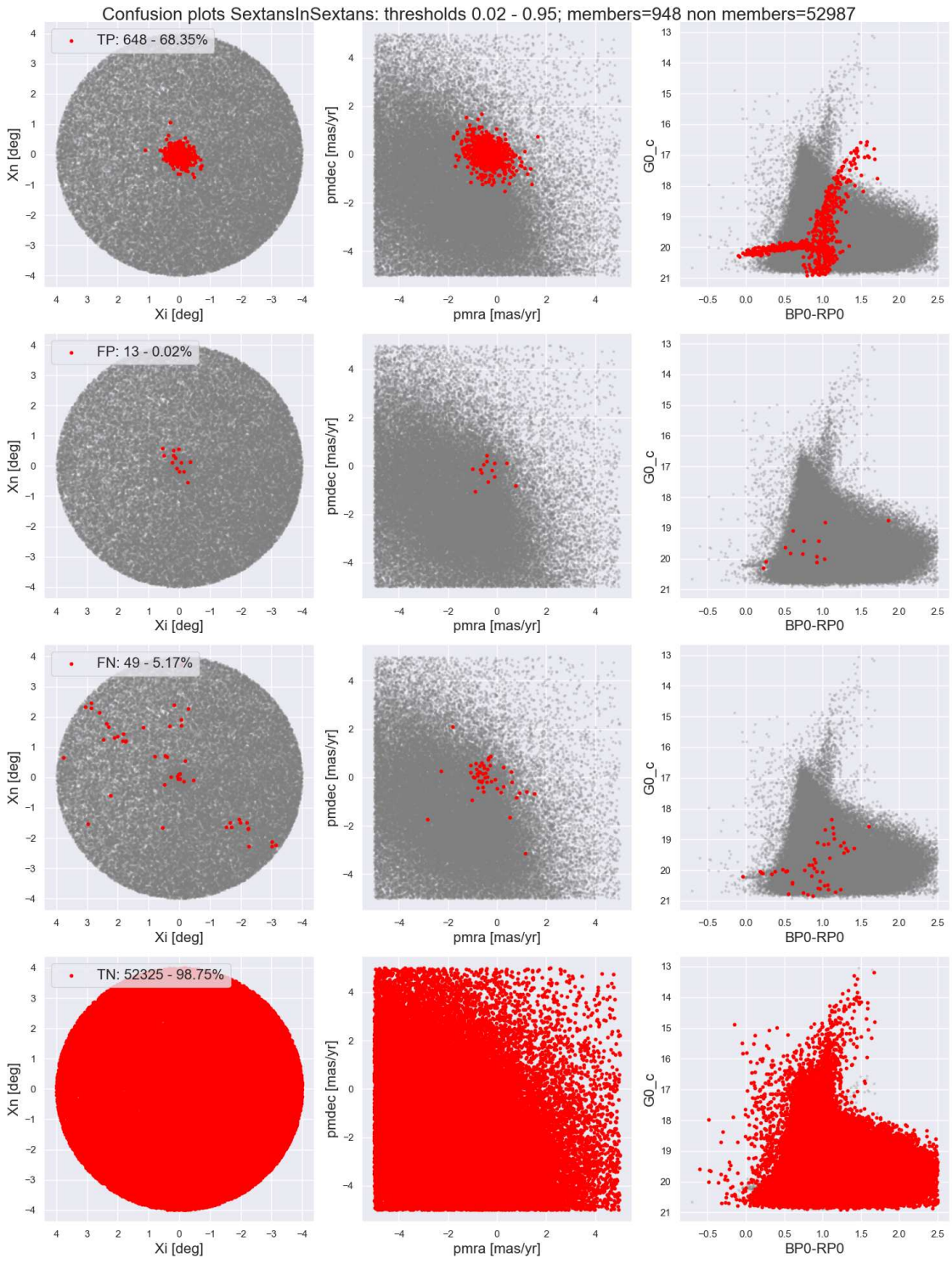


Figure 42: Sextans In Sextans confusion plots ML: the figure shows the confusion results for choice of threshold 0.02-0.95 corresponding to Fig. 40b

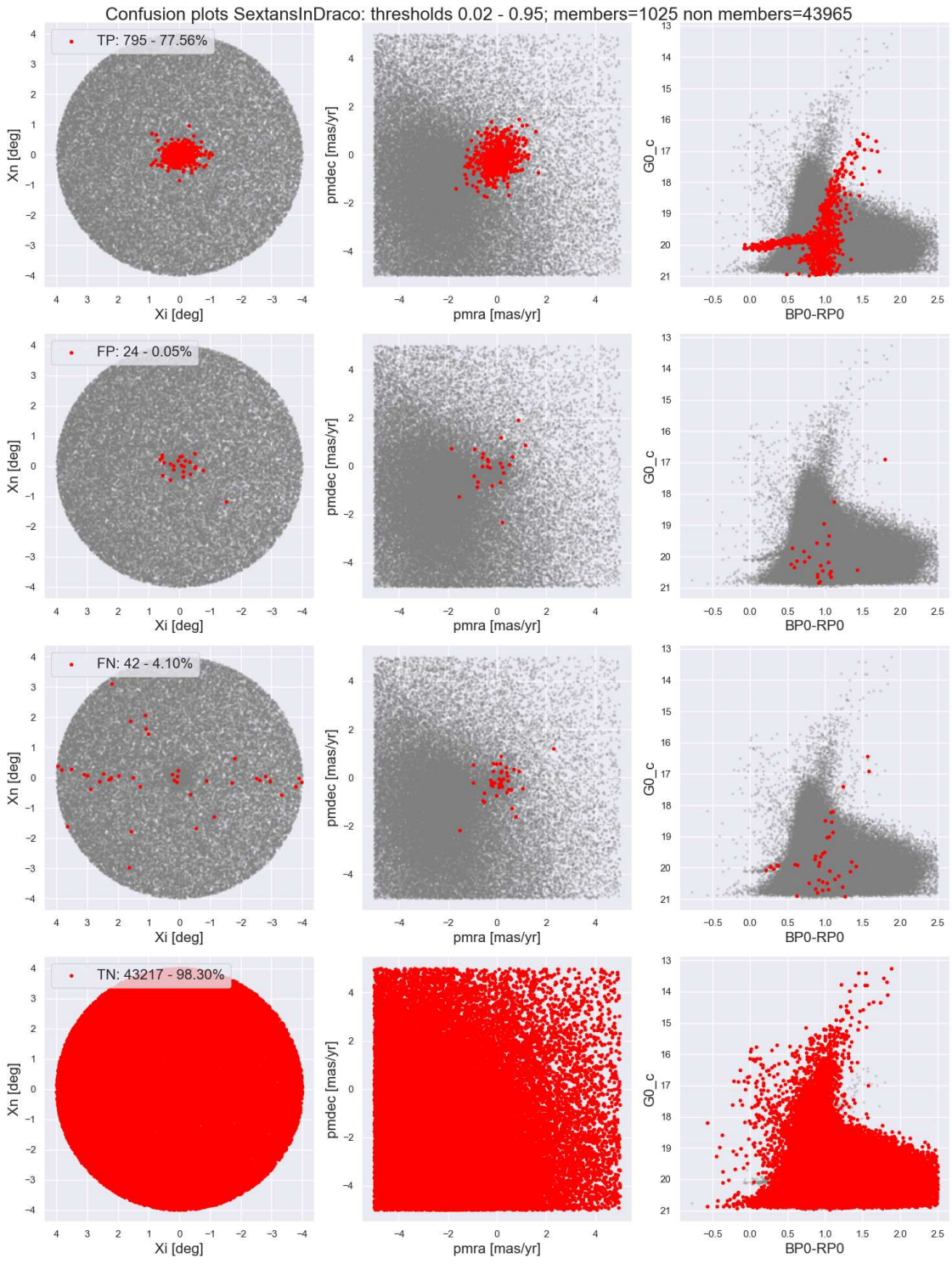


Figure 43: Sextans In Draco confusion plots ML: the figure shows the confusion results for choice of threshold 0.02-0.95 corresponding to Fig. 40c

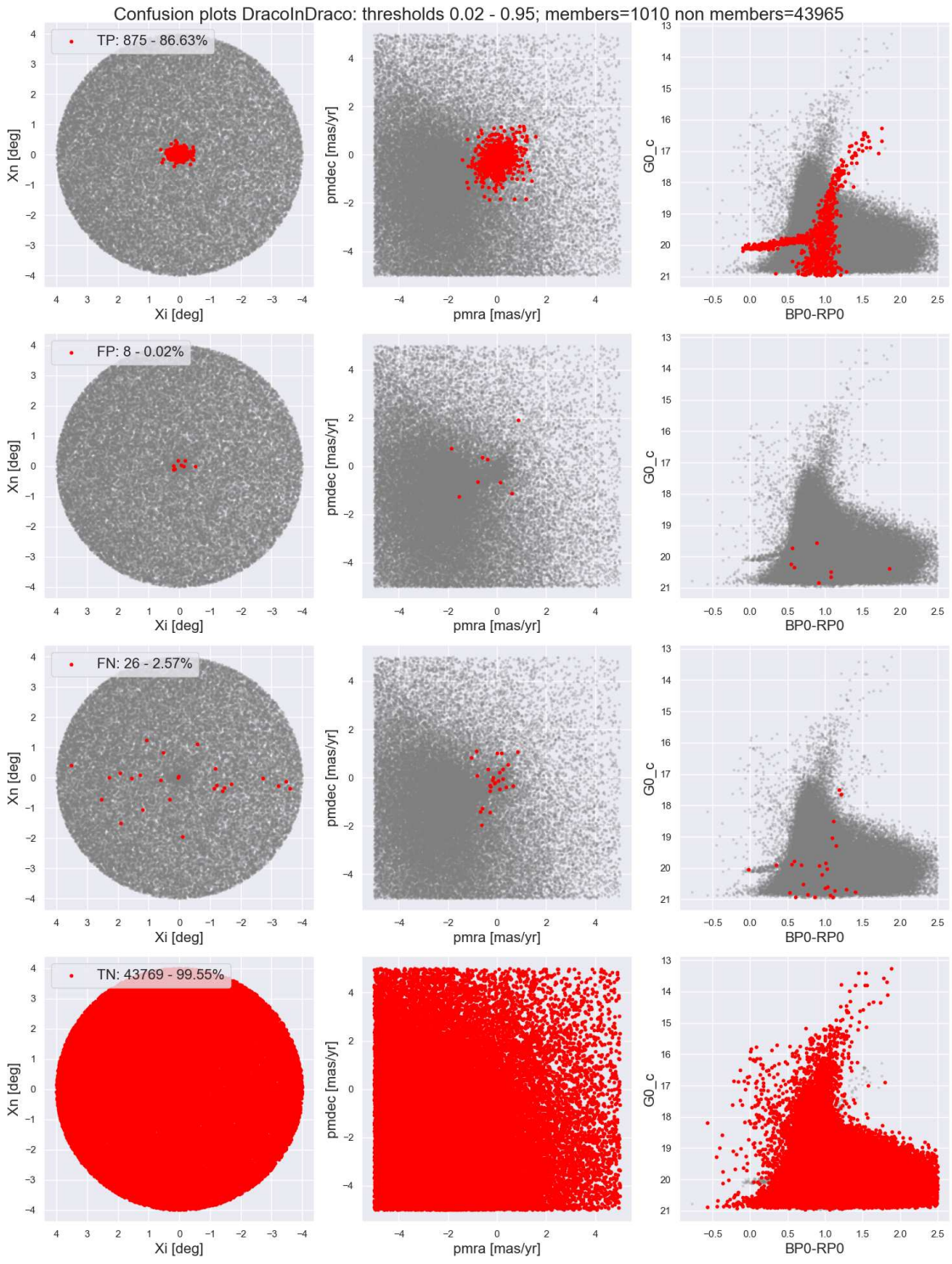


Figure 44: Draco In Draco confusion plots ML: the figure shows the confusion results for choice of threshold 0.02-0.95 corresponding to Fig. 40d

A.2.3 Third Level

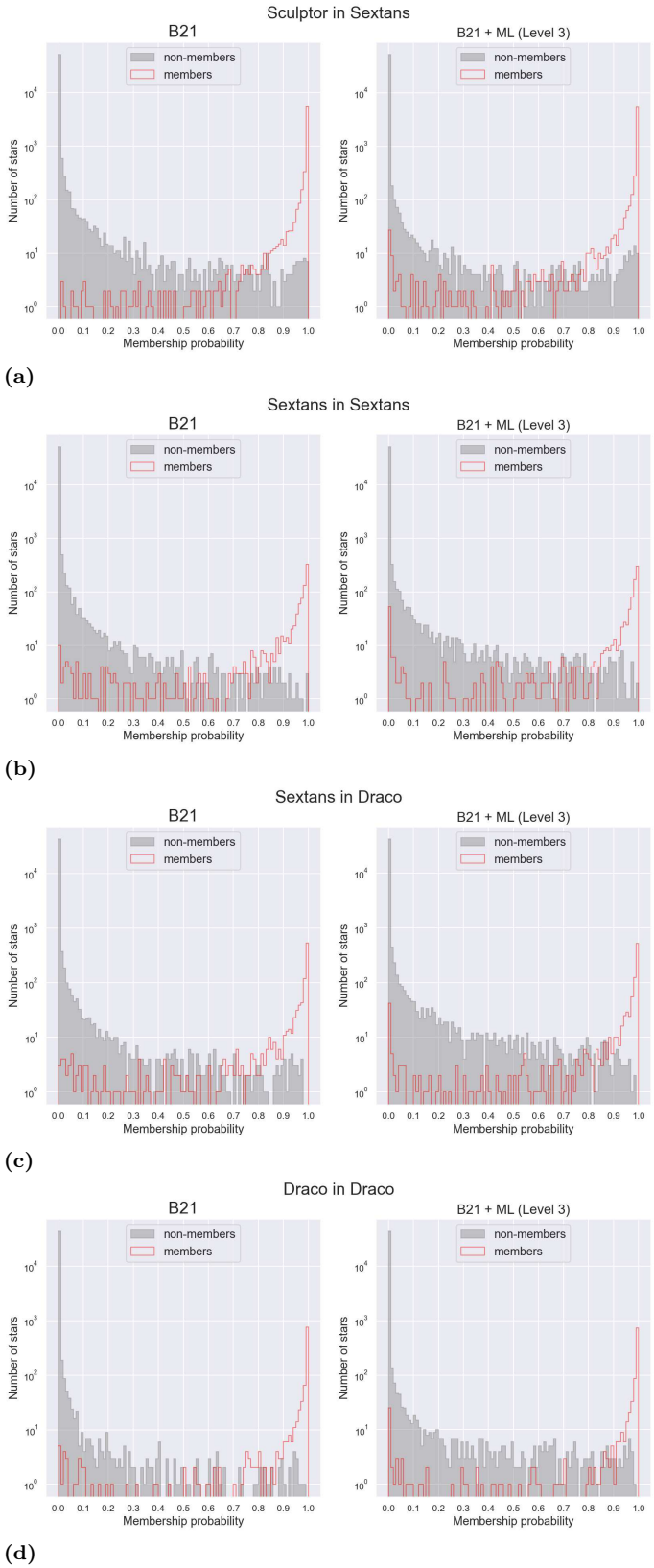


Figure 45: B21 + ML posterior probability histograms: Each plot shows the posterior probabilities obtained at the Second Level in Sec. 4.1.3. Member stars are indicated by the thin red line, while contaminants are shown as a grey histogram.

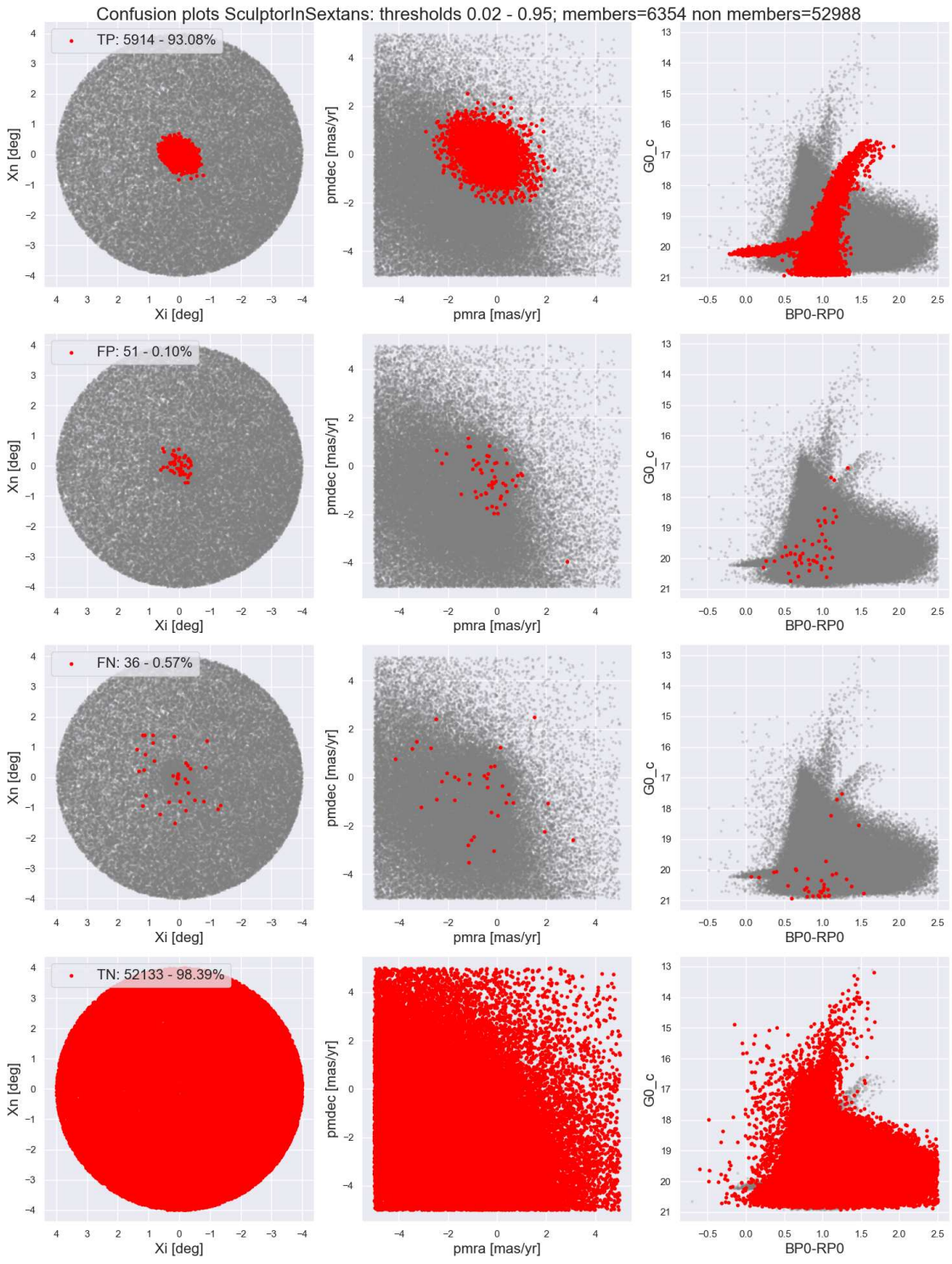


Figure 46: Sculptor In Sextans confusion plots: the figure shows the confusion results for choice of threshold 0.02-0.95 corresponding to Fig. 45a

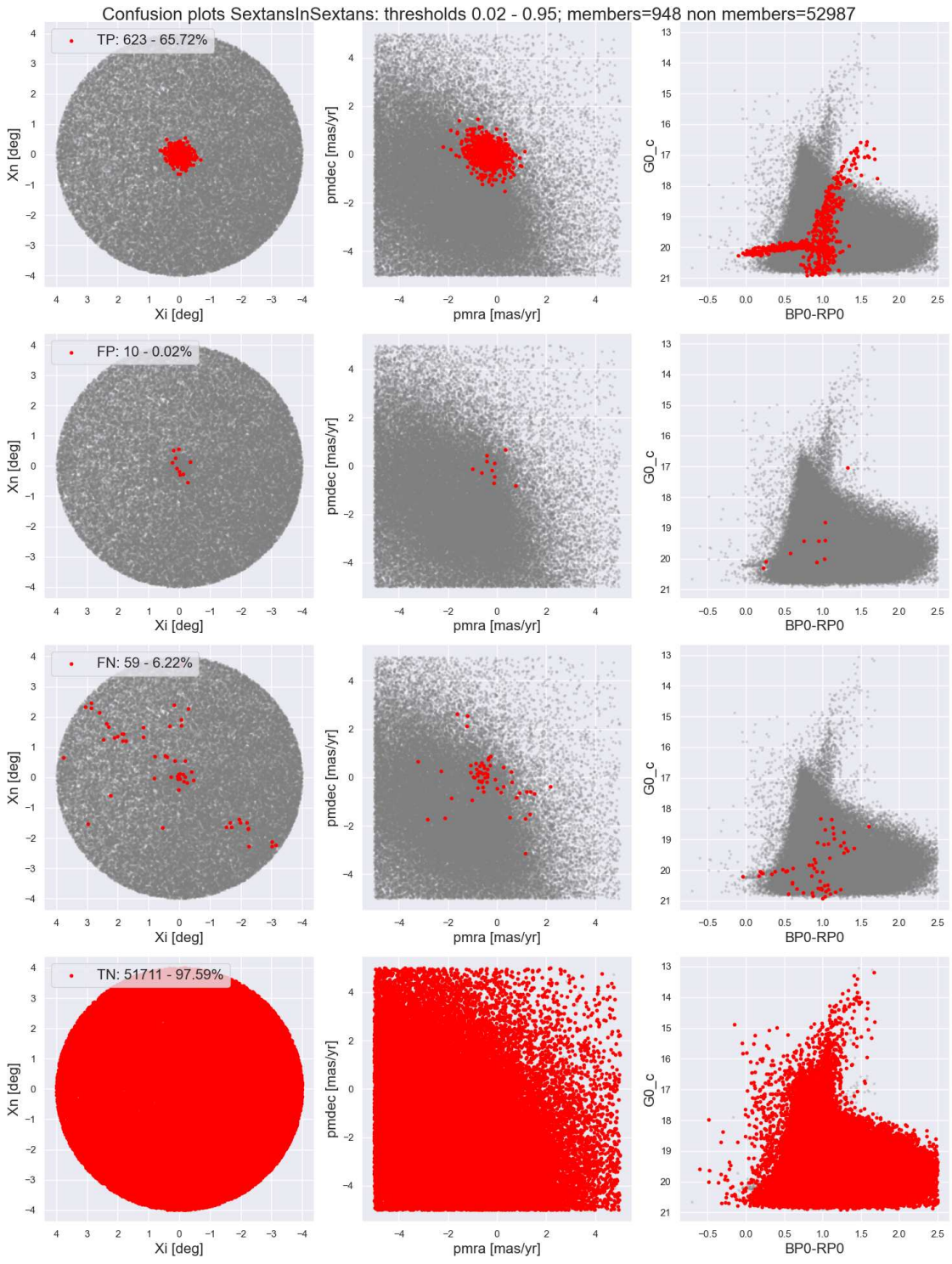


Figure 47: Sextans In Sextans confusion plots ML: the figure shows the confusion results for choice of threshold 0.02-0.95 corresponding to Fig. 45b

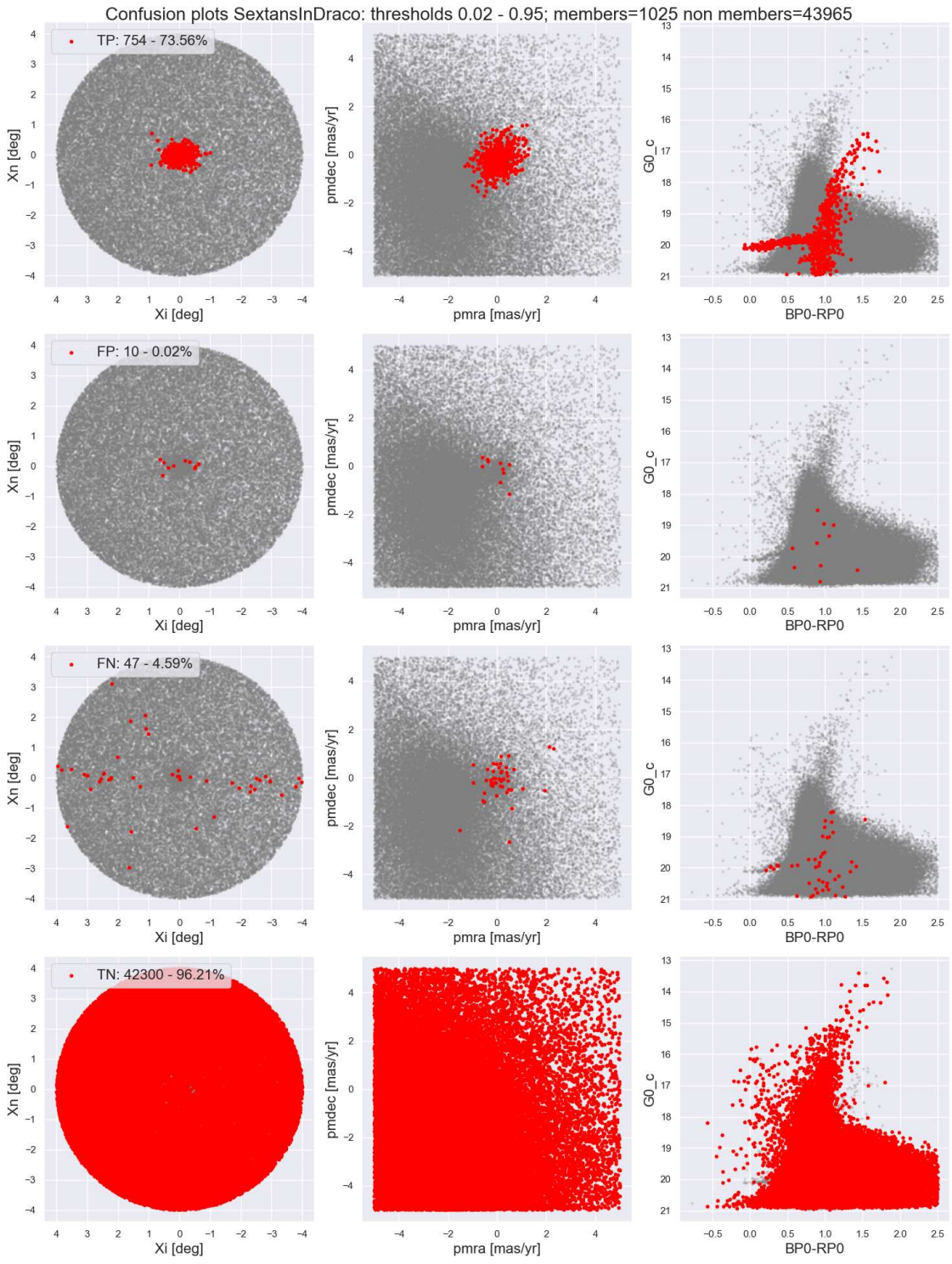


Figure 48: Sextans In Draco confusion plots ML: the figure shows the confusion results for choice of threshold 0.02-0.95 corresponding to Fig. 45c

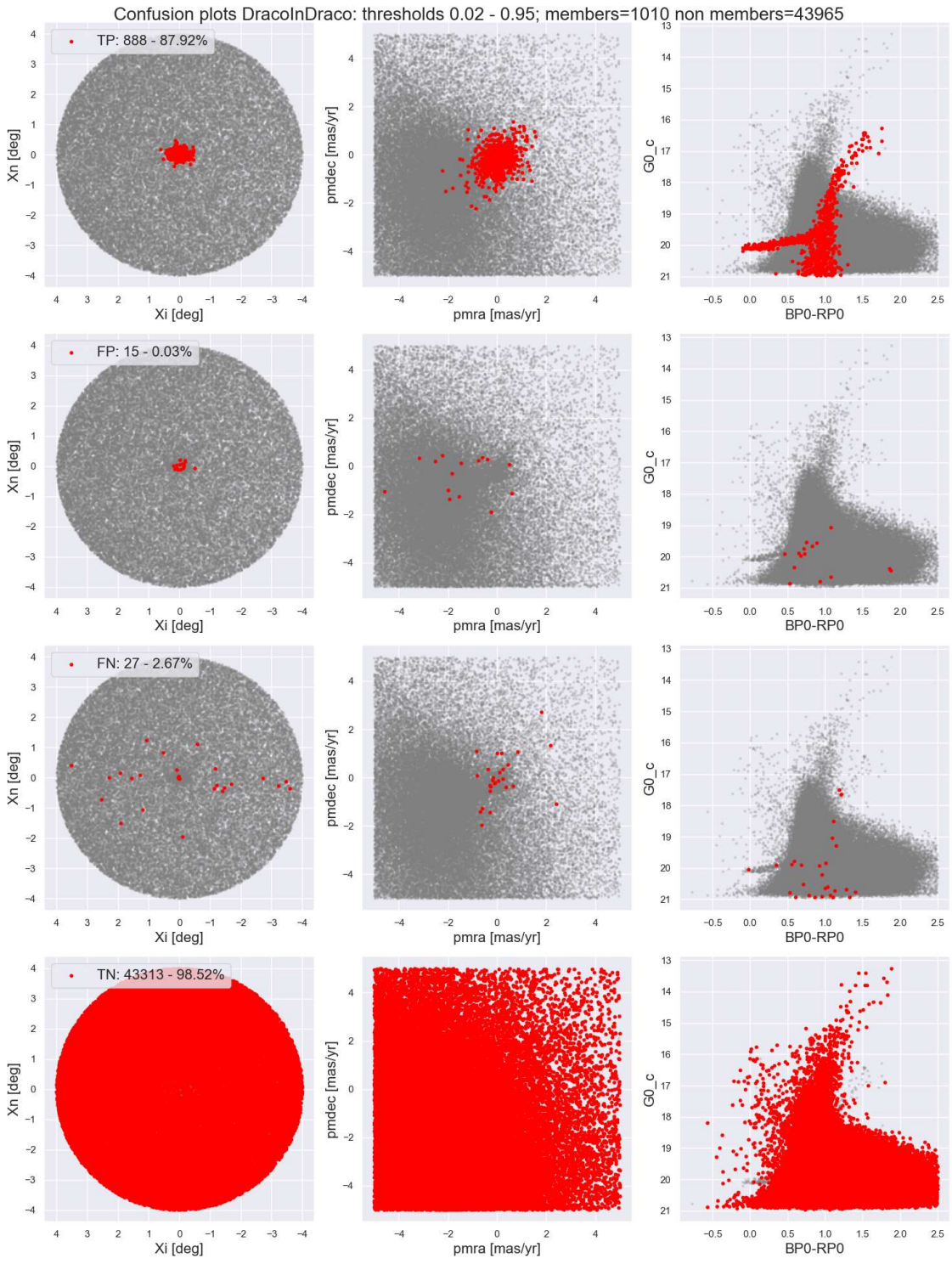
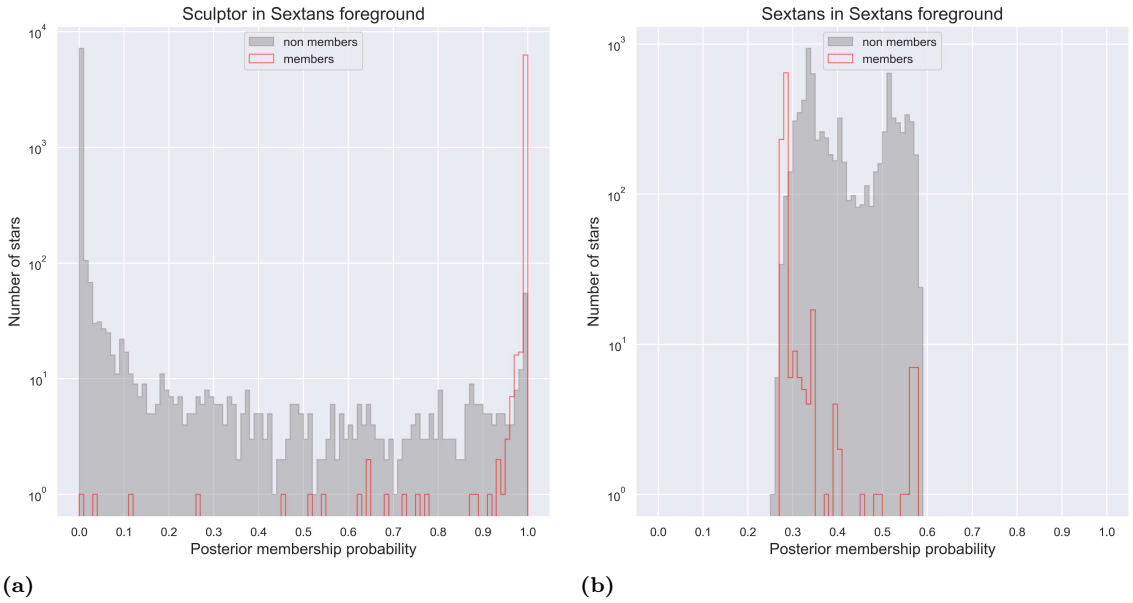


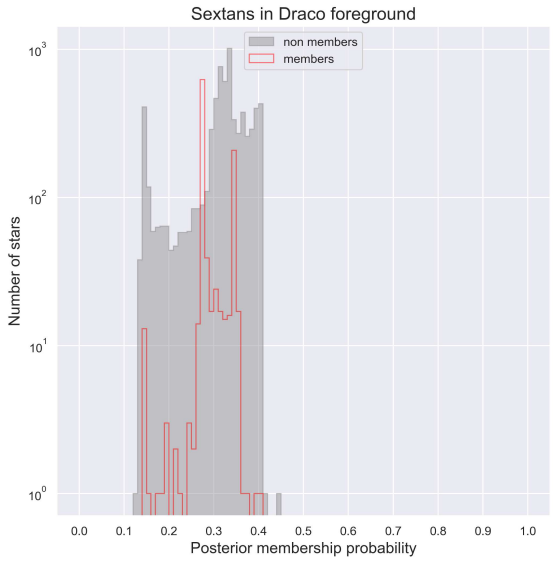
Figure 49: Draco In Draco confusion plots ML: the figure shows the confusion results for choice of threshold 0.02-0.95 corresponding to Fig. 45d

A.3 Dimensional Reduction results

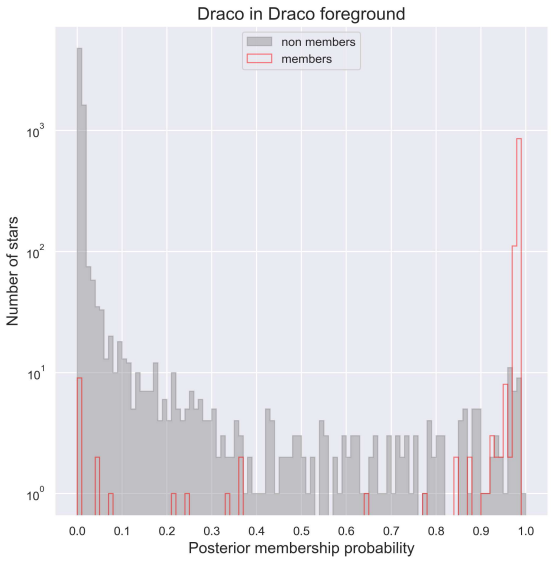


(a)

(b)



(c)



(d)

Figure 50: NF posterior probability histograms: Each plot shows the posterior probabilities obtained after the 500 iteration, where the probability value correspond to count of how many times a stars has been labeled as member divided by the total number of iteration. Member stars are indicated by the thin red line, while contaminants are shown as a grey histogram. This visualization highlights how the probability distributions differ between members and contaminants. Results are shown for `SculptorInSextans` (Fig. 50a), `SextansInSextans` (Fig. 50b), `SextansInDraco` (Fig. 50c), and `DracoInDraco` (Fig. 50d)

A.3.1 Single UMAP application

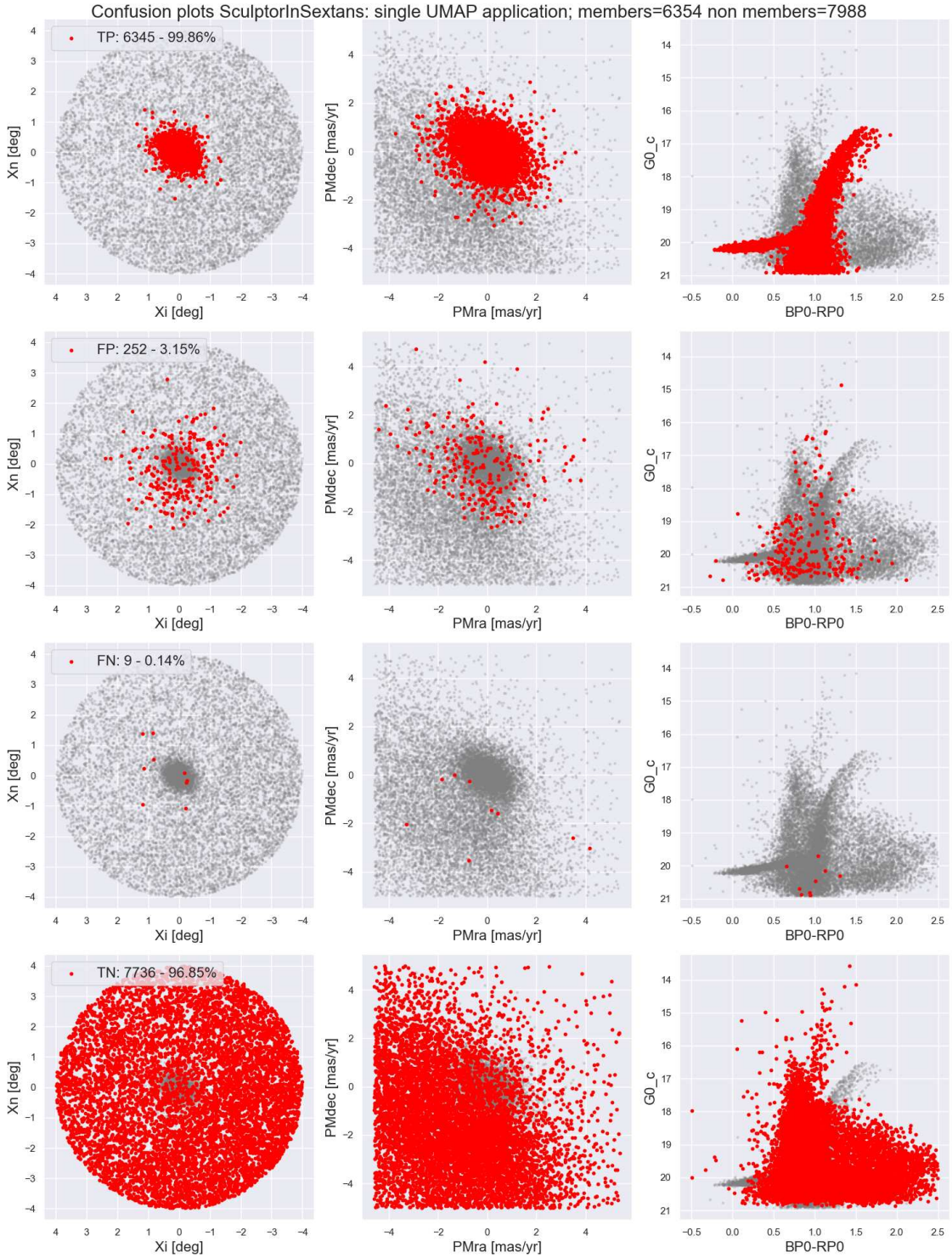


Figure 51: Sculptor In Sextans confusion plots: the figure shows the confusion results for single application of UMAP

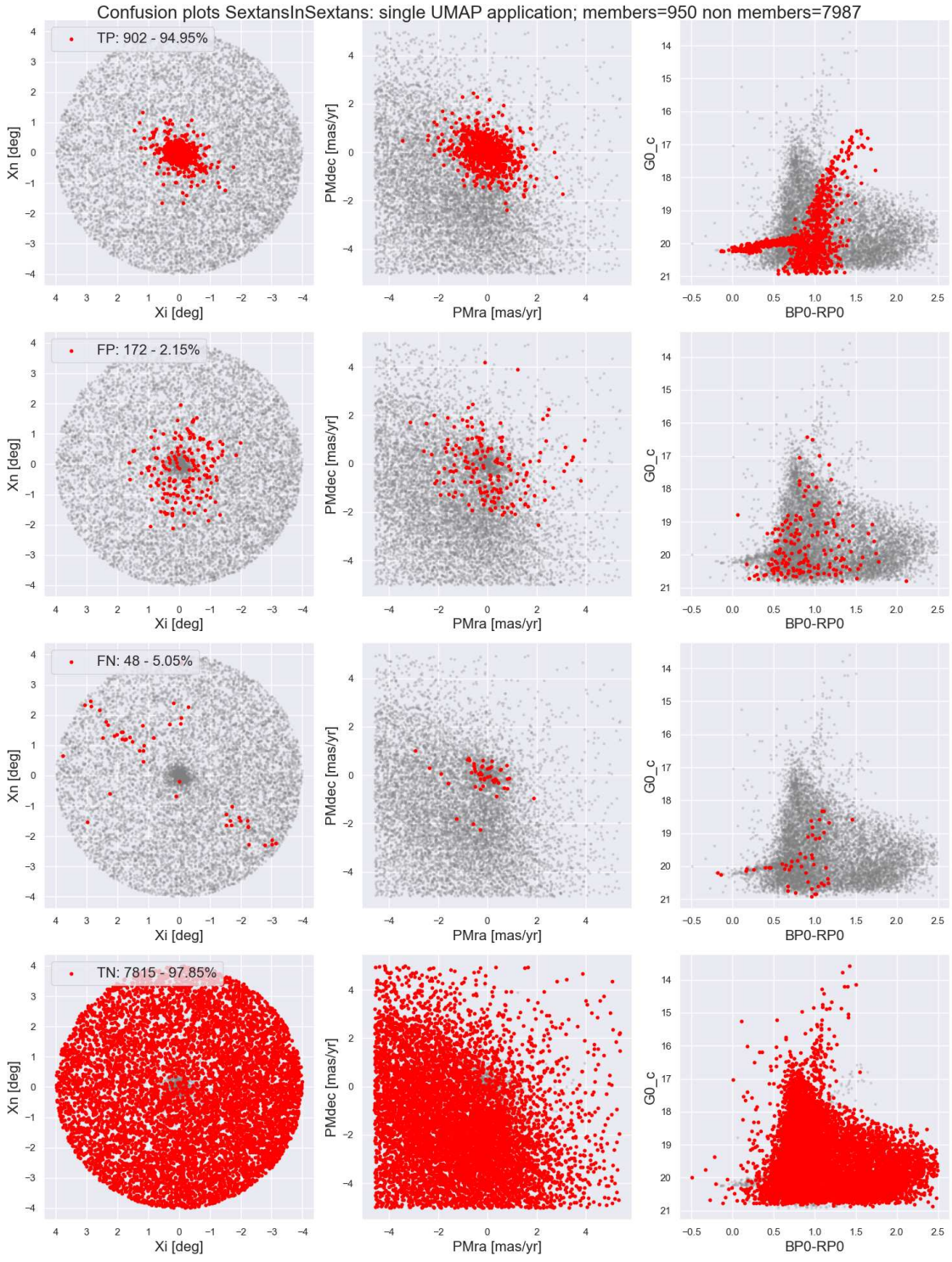


Figure 52: Sextans In Sextans confusion plots: the figure shows the confusion results for single application of UMAP

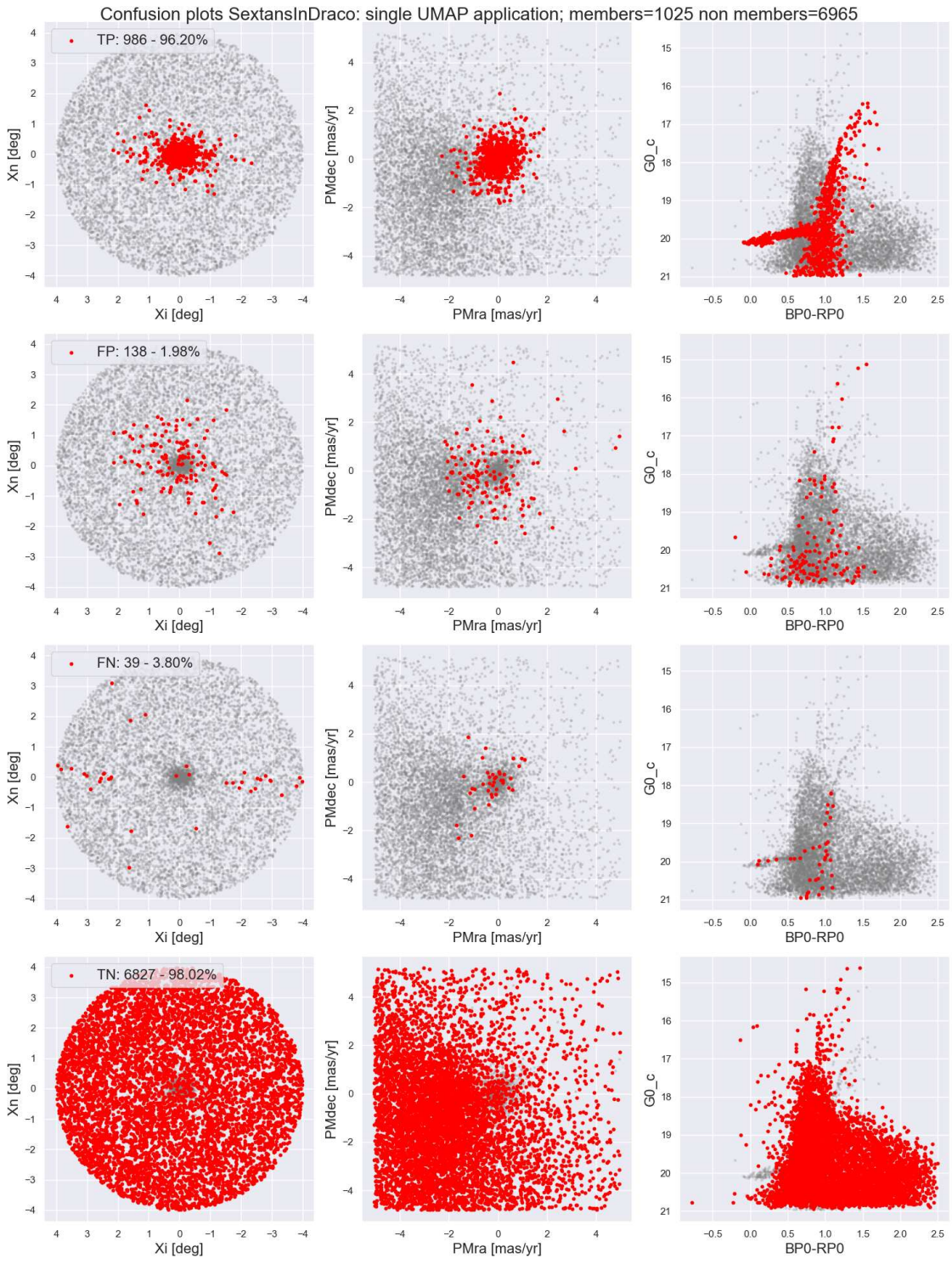


Figure 53: Sextans In Draco confusion plots: the figure shows the confusion results for single application of UMAP

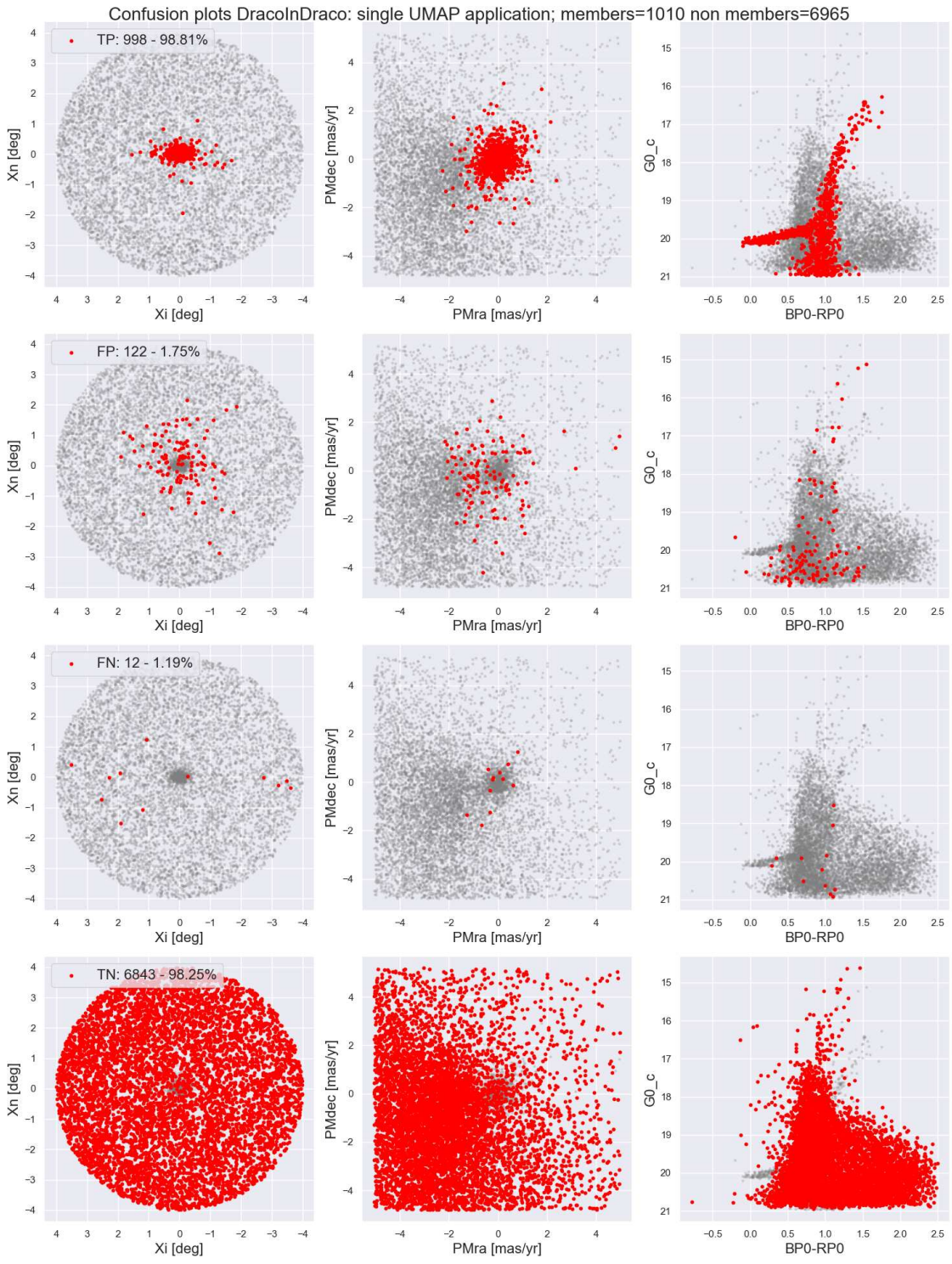


Figure 54: Draco In Draco confusion plots: the figure shows the confusion results for single application of UMAP

A.3.2 Iterations

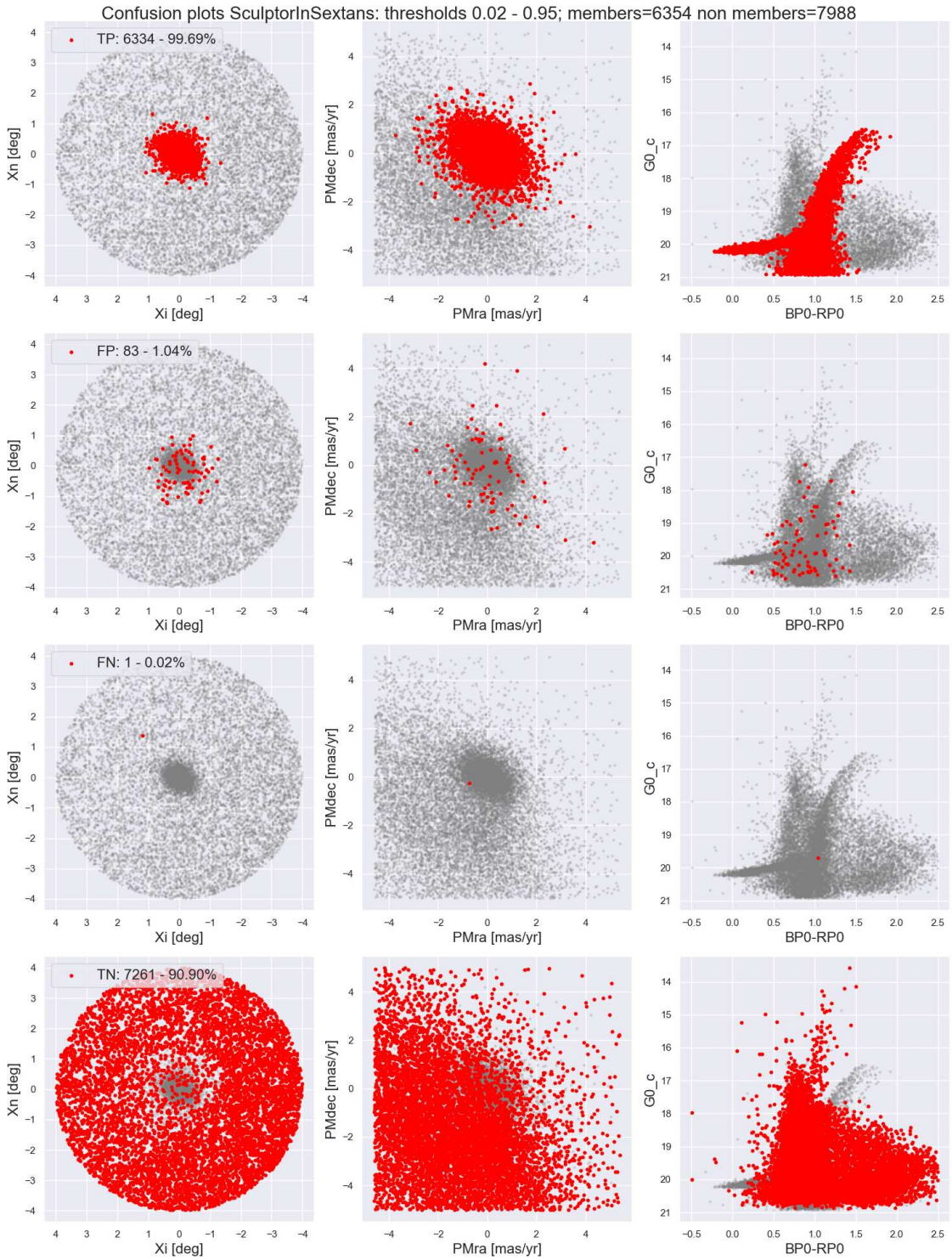


Figure 55: Sculptor In Sextans confusion plots: the figure shows the confusion results after the 500 applications of UMAP

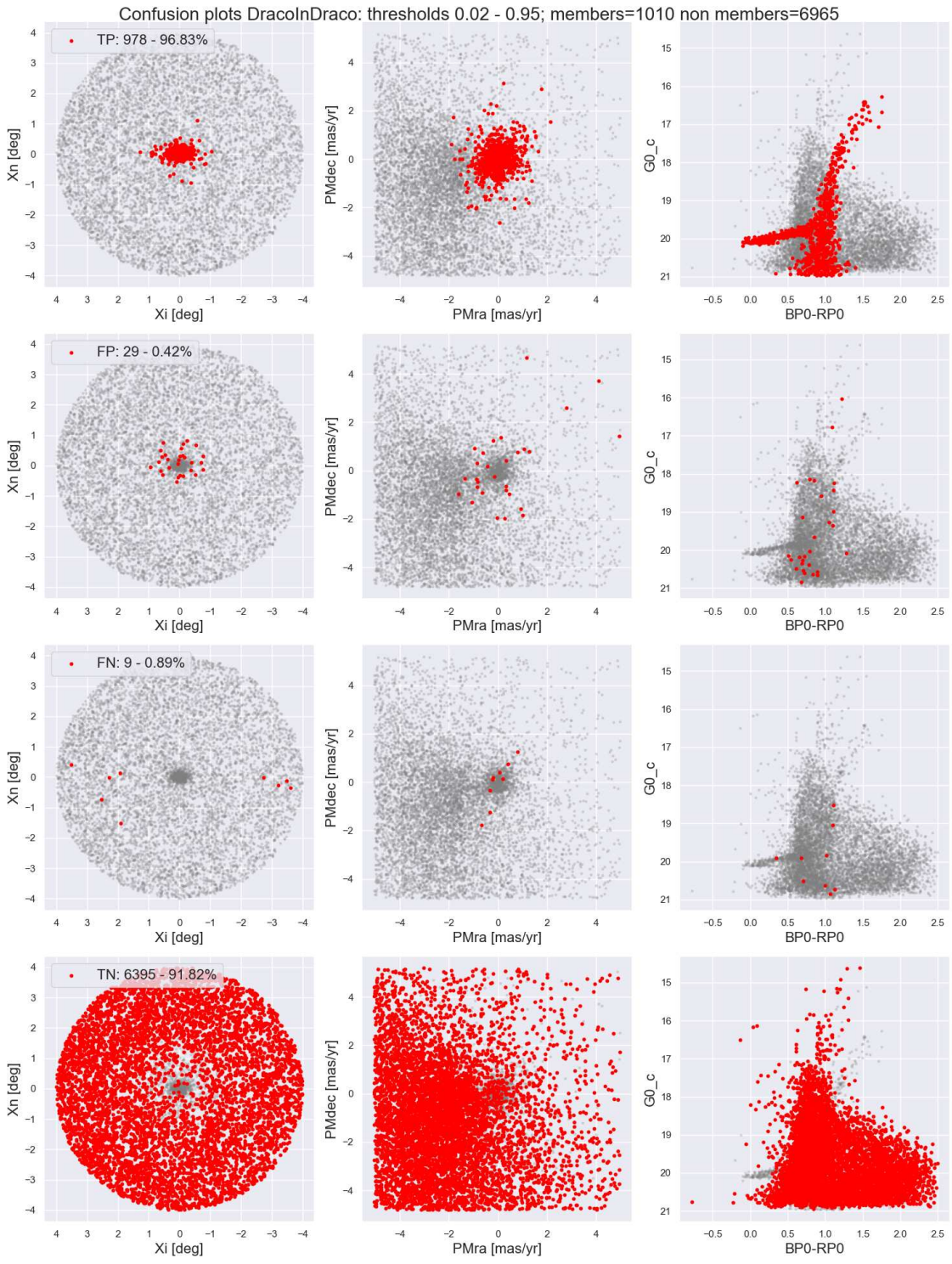


Figure 56: Draco In Draco confusion plots: the figure shows the confusion results after the 500 applications of UMAP

A.4 Normalizing Flows results

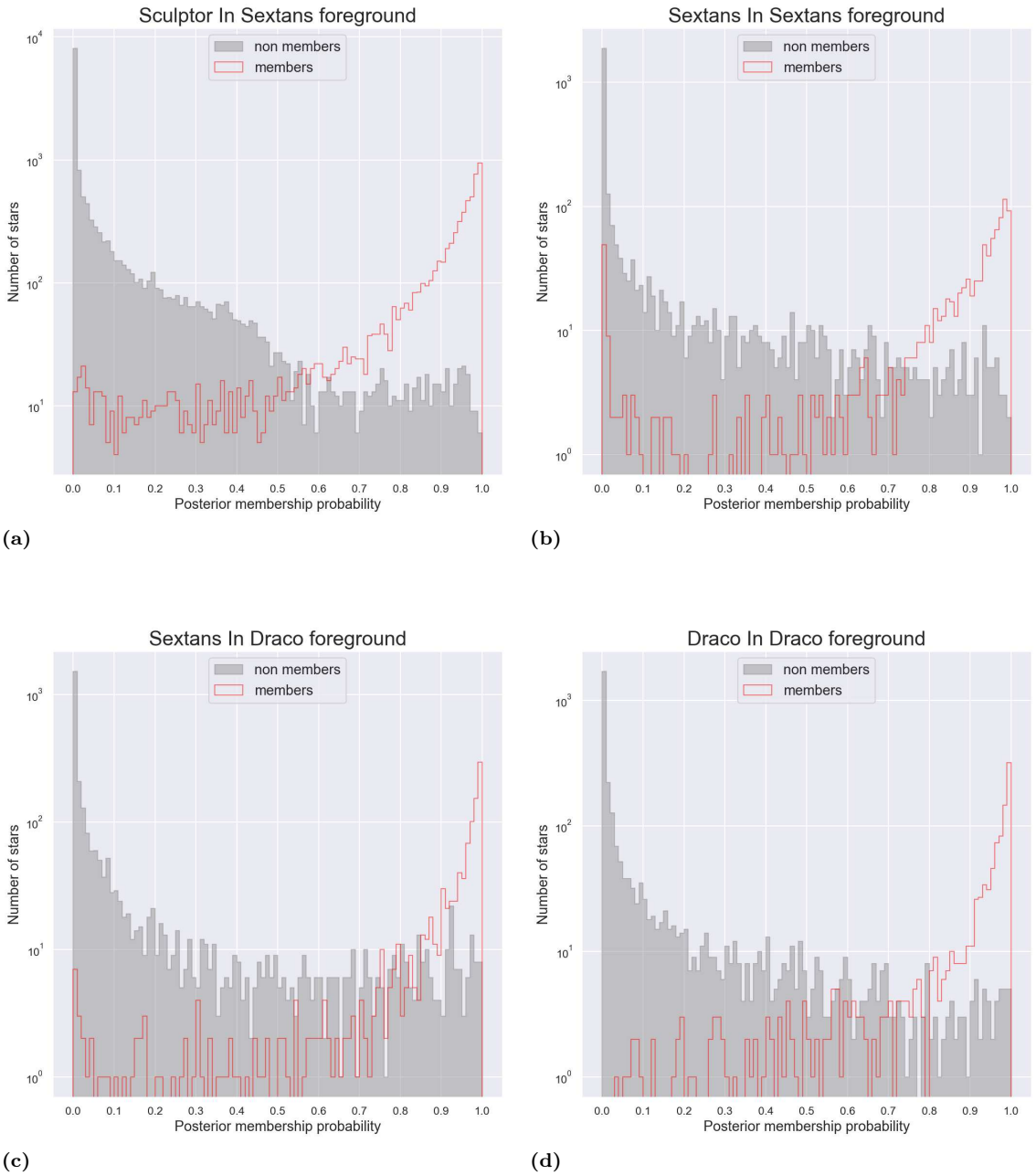


Figure 57: NF posterior probability histograms: Each plot shows the posterior probabilities obtained at the final Gibbs iteration, where convergence was reached. Member stars are indicated by the thin red line, while contaminants are shown as a grey histogram. This visualization highlights how the probability distributions differ between members and contaminants. Results are shown for `SculptorInSextans` (Fig. 57a), `SextansInSextans` (Fig. 57b), `SextansInDraco` (Fig. 57c), and `DracoInDraco` (Fig. 57d)

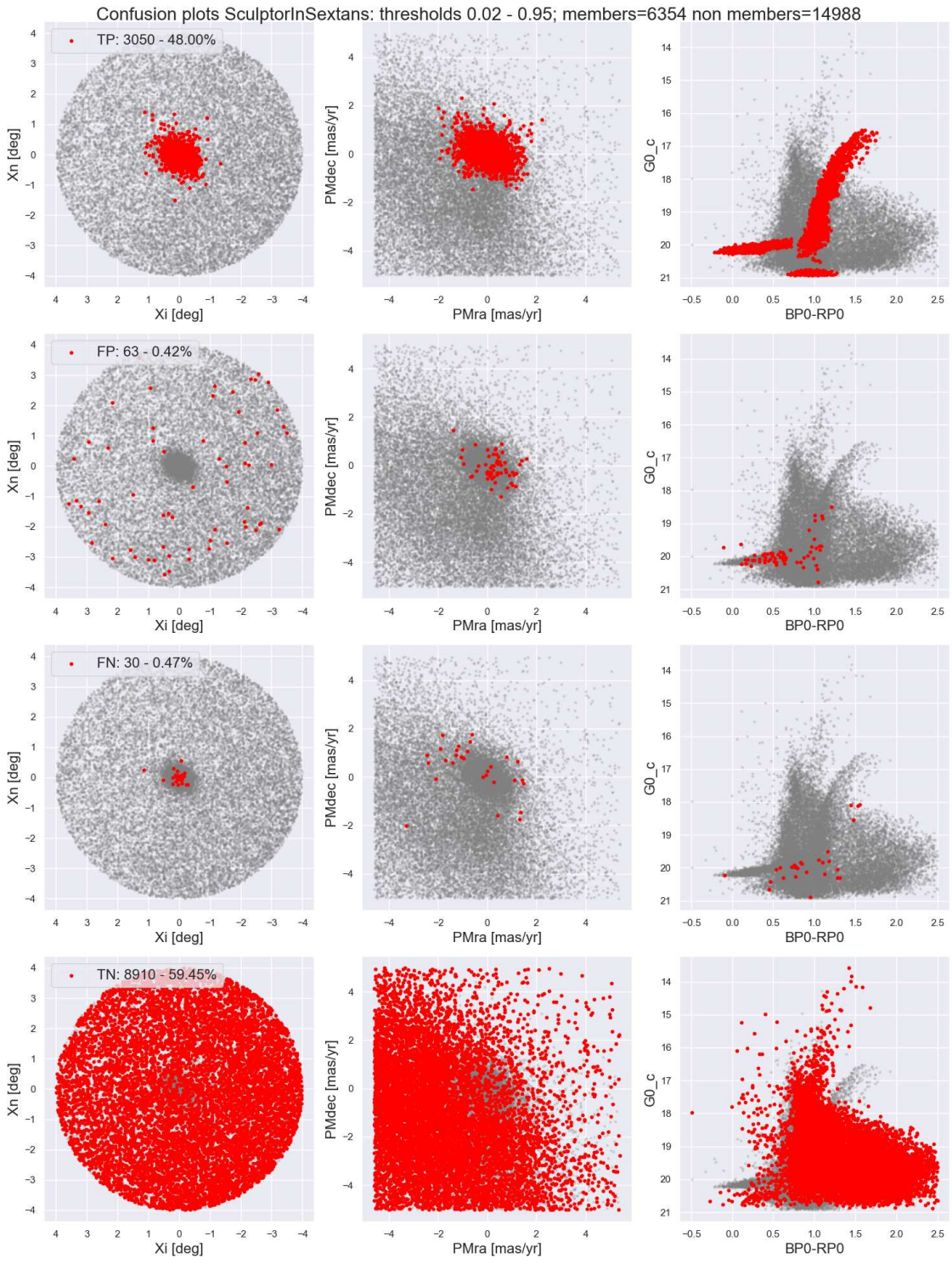


Figure 58: Sculptor In Sextans confusion plots: the figure shows the confusion results for choice of threshold 0.02-0.95

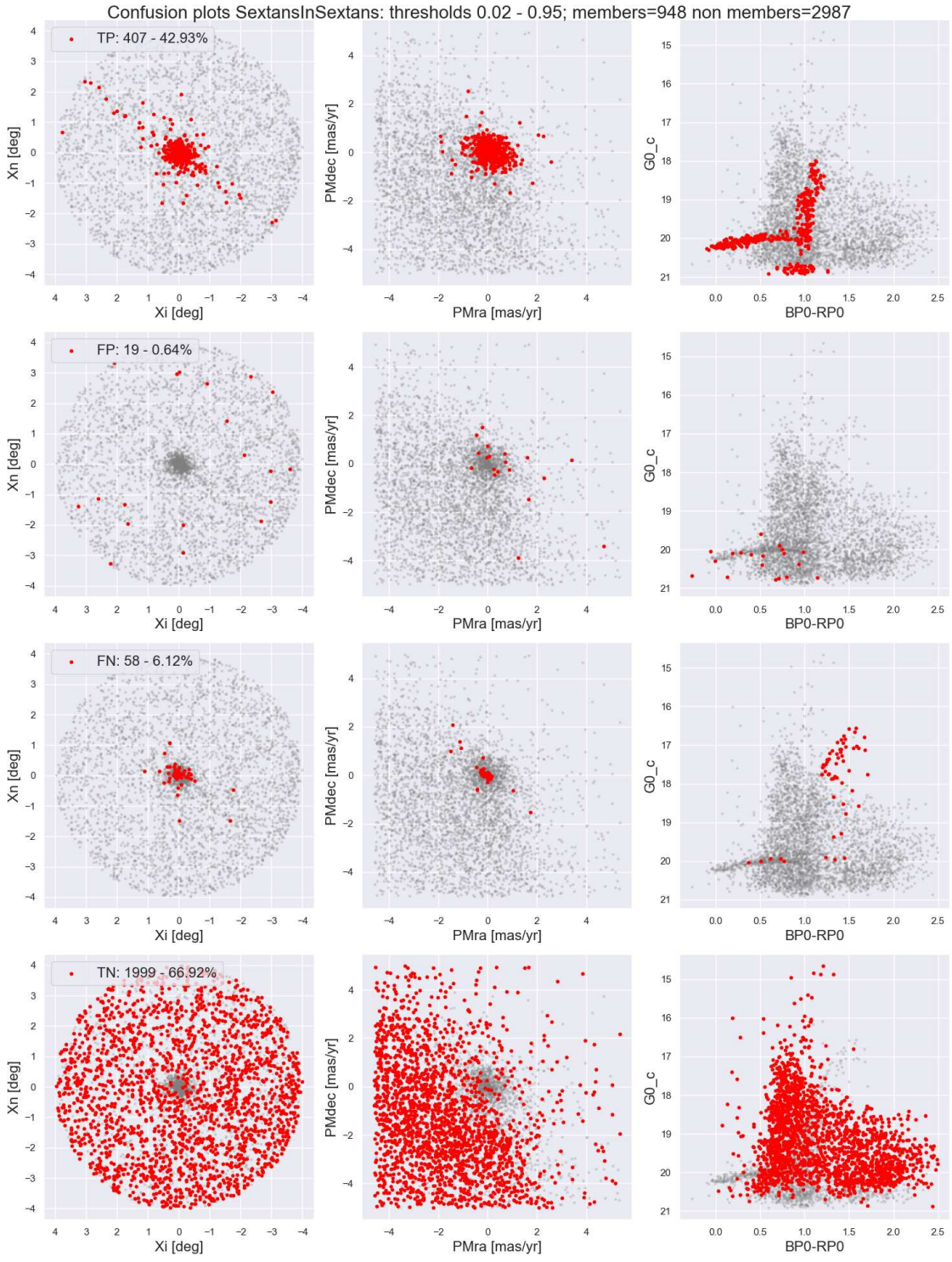


Figure 59: Sextans In Sextans confusion plots: the figure shows the confusion results for choice of threshold 0.02-0.95

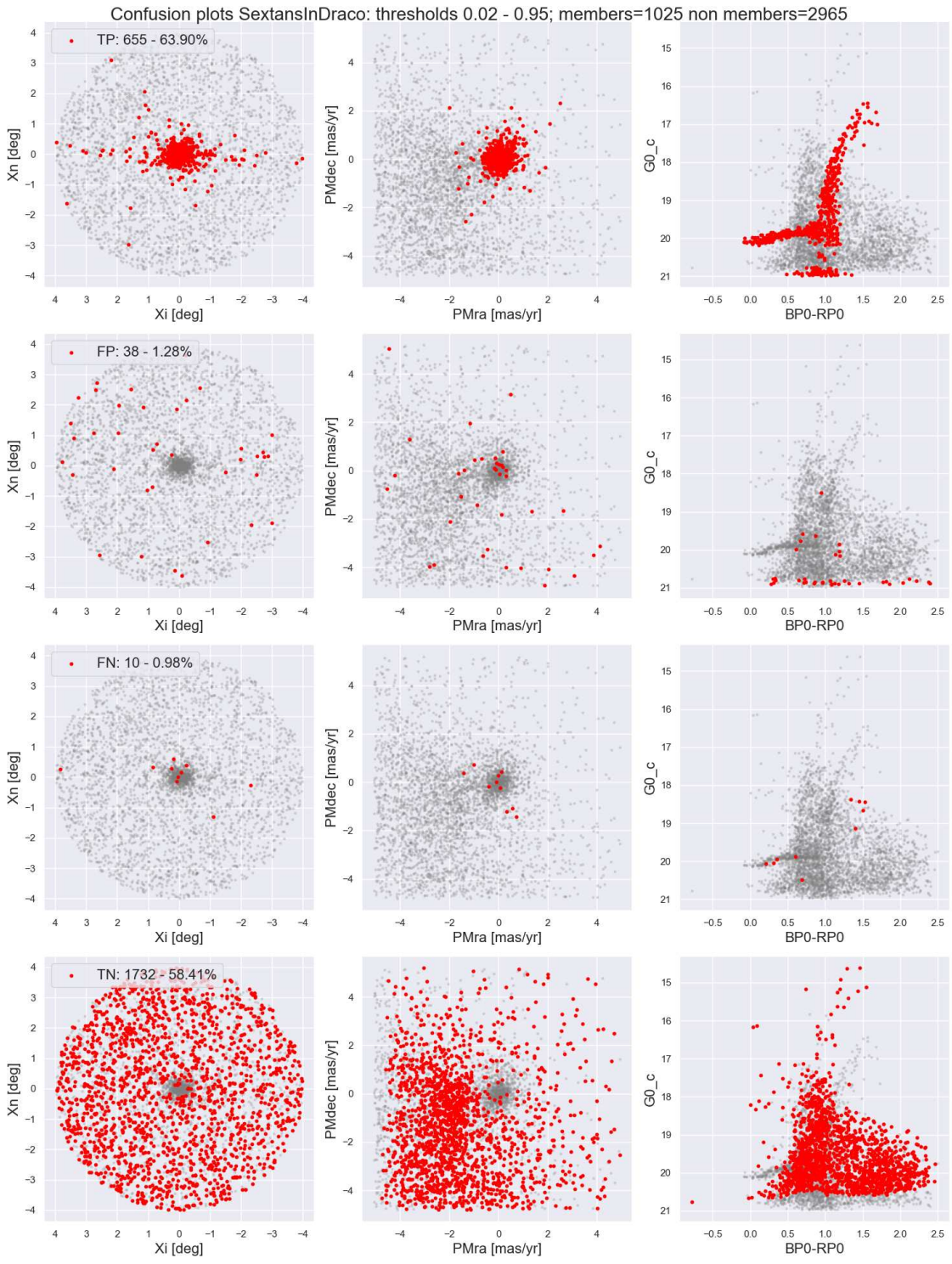


Figure 60: Sextans In Draco confusion plots: the figure shows the confusion results for choice of threshold 0.02-0.95

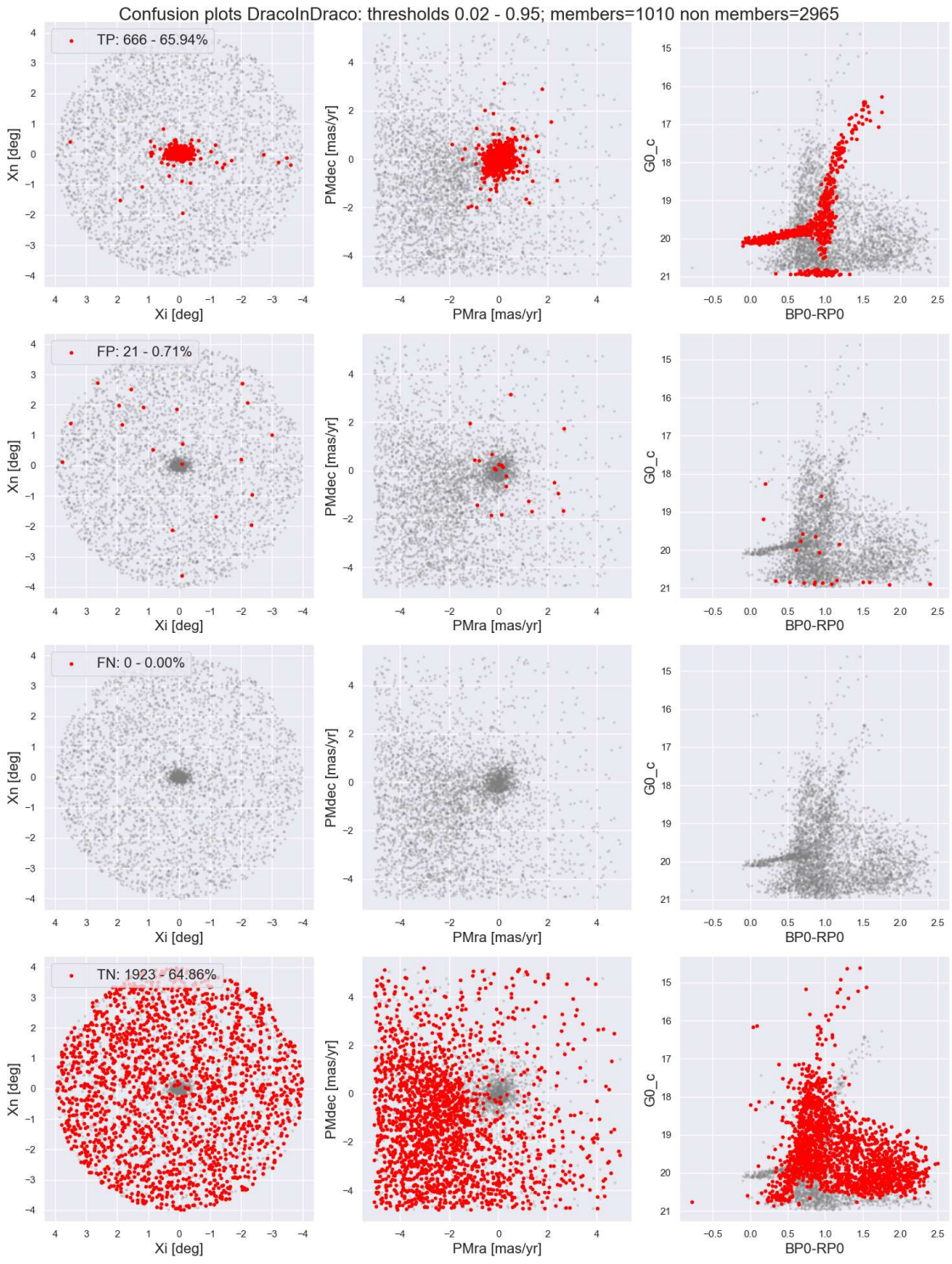


Figure 61: Draco In Draco confusion plots: the figure shows the confusion results for choice of threshold 0.02-0.95

References

Aparicio, Antonio, Ricardo Carrera, and David Martínez-Delgado (2001). "The star formation history and morphological evolution of the Draco dwarf spheroidal galaxy". In: *The Astronomical Journal* 122.5, p. 2524.

Battaglia, G. et al. (2011). "Study of the Sextans dwarf spheroidal galaxy from the DART Ca II triplet survey". In: *Monthly Notices of the Royal Astronomical Society* 411.2, pp. 1013–1034.

Battaglia, G. et al. (Dec. 2012). "THE EXTENSIVE AGE GRADIENT OF THE CARINA DWARF GALAXY". In: *The Astrophysical Journal* 761.2, p. L31. ISSN: 2041-8213. DOI: 10.1088/2041-8205/761/2/l31. URL: http://dx.doi.org/10.1088/2041-8205/761/2/L31.

Battaglia, G. et al. (Jan. 2022). "Gaia early DR3 systemic motions of Local Group dwarf galaxies and orbital properties with a massive Large Magellanic Cloud". In: *Astronomy & Astrophysics* 657, A54. ISSN: 1432-0746. DOI: 10.1051/0004-6361/202141528. URL: http://dx.doi.org/10.1051/0004-6361/202141528.

Battaglia, Giuseppina, Antonio Sollima, and Carlo Nipoti (Oct. 2015). "The effect of tides on the Fornax dwarf spheroidal galaxy". In: *Monthly Notices of the Royal Astronomical Society* 454.3, pp. 2401–2415. ISSN: 1365-2966. DOI: 10.1093/mnras/stv2096. URL: http://dx.doi.org/10.1093/mnras/stv2096.

Bettinelli, Margherita et al. (2019). "The star formation history of the Sculptor dwarf spheroidal galaxy". In: *Monthly Notices of the Royal Astronomical Society* 487.4, pp. 5862–5873.

Chiti, Anirudh et al. (Feb. 2021). "An extended halo around an ancient dwarf galaxy". In: *Nature Astronomy* 5.4, pp. 392–400. ISSN: 2397-3366. DOI: 10.1038/s41550-020-01285-w. URL: http://dx.doi.org/10.1038/s41550-020-01285-w.

Coccaro, A et al. (2023). "Comparative Study of Coupling and Autoregressive Flows through Robust Statistical Tests". In: *arXiv preprint arXiv:2302.12024*.

Deason, Alis J et al. (Dec. 2021). "Dwarf stellar haloes: a powerful probe of small-scale galaxy formation and the nature of dark matter". In: *Monthly Notices of the Royal Astronomical Society* 511.3, pp. 4044–4059. ISSN: 1365-2966. DOI: 10.1093/mnras/stab3524. URL: http://dx.doi.org/10.1093/mnras/stab3524.

Dinh, Laurent, Jascha Sohl-Dickstein, and Samy Bengio (2016). "Density estimation using real NVP". In: *arXiv preprint arXiv:1605.08803*.

Goater, Alex et al. (2023). *EDGE: The direct link between mass growth history and the extended stellar haloes of the faintest dwarf galaxies*. arXiv: 2307.05130. URL: https://arxiv.org/abs/2307.05130.

Iorio, G et al. (May 2019). "The effect of tides on the Sculptor dwarf spheroidal galaxy". In: *Monthly Notices of the Royal Astronomical Society* 487.4, pp. 5692–5710. ISSN: 0035-8711. DOI: 10.1093/mnras/stz1342. eprint: https://academic.oup.com/mnras/article-pdf/487/4/5692/28901729/stz1342.pdf. URL: https://doi.org/10.1093/mnras/stz1342.

Jensen, Jaclyn et al. (2023). *Small-scale stellar haloes: detecting low surface brightness features in the outskirts of Milky Way dwarf satellites*. arXiv: 2308.07394 [astro-ph.GA]. URL: https://arxiv.org/abs/2308.07394.

Kobyzev, Ivan, Simon JD Prince, and Marcus A Brubaker (2020). "Normalizing flows: An introduction and review of current methods". In: *IEEE transactions on pattern analysis and machine intelligence* 43.11, pp. 3964–3979.

Lokas, Ewa L, Gary A Mamon, and Francisco Prada (2005). "Dark matter distribution in the Draco dwarf from velocity moments". In: *Monthly Notices of the Royal Astronomical Society* 363.3, pp. 918–928.

Martin, Nicolas F. et al. (2013). "The PAndAS View of the Andromeda Satellite System. I. A Bayesian Search for Dwarf Galaxies Using Spatial and Color-Magnitude Information". In: DOI: 10.1088/0004-637X/776/2/80. arXiv: 1307.7626 [astro-ph.GA].

McConnachie, Alan W. (June 2012). "THE OBSERVED PROPERTIES OF DWARF GALAXIES IN AND AROUND THE LOCAL GROUP". In: *The Astronomical Journal* 144.1, p. 4. ISSN: 1538-3881. DOI: 10.1088/0004-6256/144/1/4. URL: http://dx.doi.org/10.1088/0004-6256/144/1/4.

Pace, Andrew B. (2024). *The Local Volume Database: a library of the observed properties of nearby dwarf galaxies and star clusters*. arXiv: 2411.07424 [astro-ph.GA]. URL: https://arxiv.org/abs/2411.07424.

Pace, Andrew B., Denis Erkal, and Ting S. Li (Nov. 2022). "Proper Motions, Orbits, and Tidal Influences of Milky Way Dwarf Spheroidal Galaxies". In: *The Astrophysical Journal* 940.2, p. 136. ISSN: 1538-4357. DOI: 10.3847/1538-4357/ac997b. URL: http://dx.doi.org/10.3847/1538-4357/ac997b.

Papamakarios, George, Theo Pavlakou, and Iain Murray (2017). "Masked autoregressive flow for density estimation". In: *Advances in neural information processing systems* 30.

Papamakarios, George et al. (2021). "Normalizing flows for probabilistic modeling and inference". In: *Journal of Machine Learning Research* 22.57, pp. 1–64.

Riello, Marco et al. (2021). "Gaia early data release 3-photometric content and validation". In: *Astronomy & astrophysics* 649, A3.

Rinaldi, Stefano and María Claudia Ramírez-Tannus (2024). "Non-parametric identification of single-lined binary candidates in young clusters using single-epoch spectroscopy". In: *Astronomy & Astrophysics* 692, A173.

Simon, Joshua D. (Aug. 2019). "The Faintest Dwarf Galaxies". In: *Annual Review of Astronomy and Astrophysics* 57.1, pp. 375–415. ISSN: 1545-4282. DOI: 10.1146/annurev-astro-091918-104453. URL: http://dx.doi.org/10.1146/annurev-astro-091918-104453.

Simon, Joshua D. and Marla Geha (Nov. 2007). "The Kinematics of the Ultra-faint Milky Way Satellites: Solving the Missing Satellite Problem". In: *The Astrophysical Journal* 670.1, p. 313. DOI: 10.1086/521816. URL: https://dx.doi.org/10.1086/521816.

Tolstoy, Eline, Vanessa Hill, and Monica Tosi (Sept. 2009). "Star-Formation Histories, Abundances, and Kinematics of Dwarf Galaxies in the Local Group". In: *Annual Review of Astronomy and Astrophysics* 47.1, pp. 371–425. ISSN: 1545-4282. DOI: 10.1146/annurev-astro-082708-101650. URL: http://dx.doi.org/10.1146/annurev-astro-082708-101650.

Yang, Yanbin et al. (Mar. 2022). "An extended stellar halo discovered in the Fornax dwarf spheroidal using Gaia EDR3". In: *Monthly Notices of the Royal Astronomical Society* 512.3, pp. 4171–4184. ISSN: 1365-2966. DOI: 10.1093/mnras/stac644. URL: http://dx.doi.org/10.1093/mnras/stac644.

Zaragoza-Cardiel, Javier et al. (Sept. 2024). "Detection and characterization of detached tidal dwarf galaxies". In: *Astronomy &; Astrophysics* 689, A206. ISSN: 1432-0746. DOI: 10.1051/0004-6361/202450349. URL: http://dx.doi.org/10.1051/0004-6361/202450349.