Unveiling the Solar vicinity: unsupervised mapping of the local stellar populations with GCNS and GALAH DR4 using t-SNE

P. H. R. de Andrade[1], M. L. L. Dantas[2], A. C. Soja[1]

[1] Instituto Federal de Educação, Ciência e Tecnologia Fluminense, *Campus* Bom Jesus do Itabapoana, Rio de Janeiro, Brazil
[2] Instituto de Astrofísica, Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860, Santiago, Chile

Corresponding author e-mail: pedroiff0@gmail.com

Instituto Federal Fluminense — Campus Bom Jesus do Itabapoana

Ministério da Ciência, Tecnologia e Inovação

1. Introduction

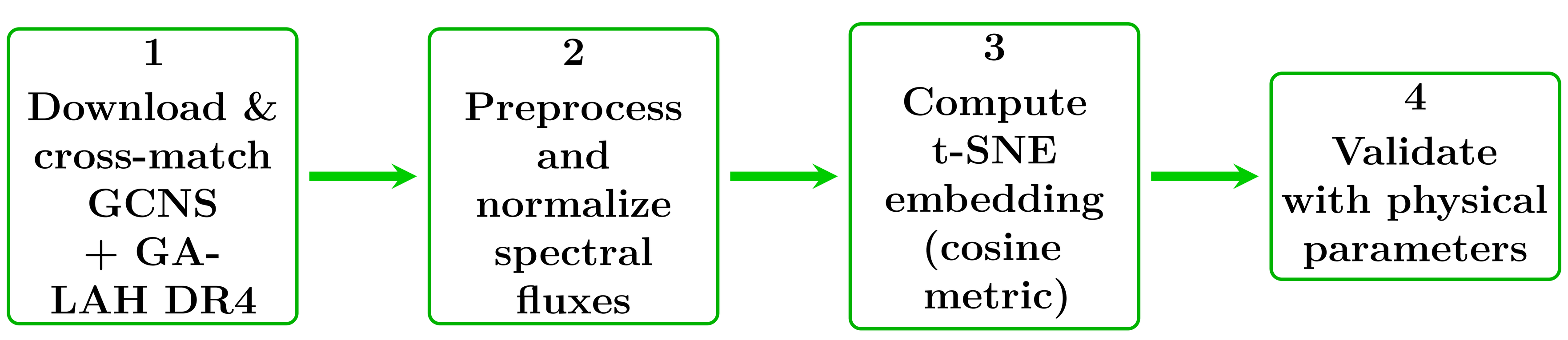
Characterizing stars in the Solar vicinity allows us to investigate the evolution of the Milky Way. Chemodynamical signatures are fundamental to trace the origin of stars and the local chemical enrichment. However, the parameter space that describes a star – astrometry, kinematics, atmospheric parameters and up to dozens of chemical abundances – is intrinsically high-dimensional and non-linear, which motivates the use of unsupervised, data-driven techniques instead of hand-picked diagnostic diagrams. This requires the following data sets:

- **Astrometry:** the astrometric basis comes from the *Gaia* mission (DR3), specifically the *Gaia* Catalogue of Nearby Stars (GCNS), which provides high-precision photometry and astrometry for stars located within 100 pc of the Sun.
- **Spectroscopy:** detailed chemical information can be obtained from surveys such as Gaia-ESO (Gilmore *et al.*, 2022; Randich *et al.*, 2022), LAMOST (Cui *et al.*, 2012; Xiang *et al.*, 2019) and APOGEE (Majewski *et al.*, 2017). Here, however, we focus on the GALAH DR4 survey (Buder *et al.*, 2025), which provides up to 30 elemental abundances, along with the fully reduced normalized spectra, for nearly one million stars.

This work combines these two databases, safely cross-matching by id roughly 5 000–6 000 stars in common between GCNS and GALAH DR4. Rather than projecting this sample onto pre-defined chemical-abundance axes, our goal is to let an unsupervised algorithm, t-SNE (MAATEN; HINTON, 2008), learn the manifold directly from the high-dimensional, normalized spectral fluxes, and only afterwards use the physical parameters to validate and interpret the resulting structure.

2. Methodology

Figura 1 — Methodological workflow, from data acquisition to the validation of the unsupervised embedding.



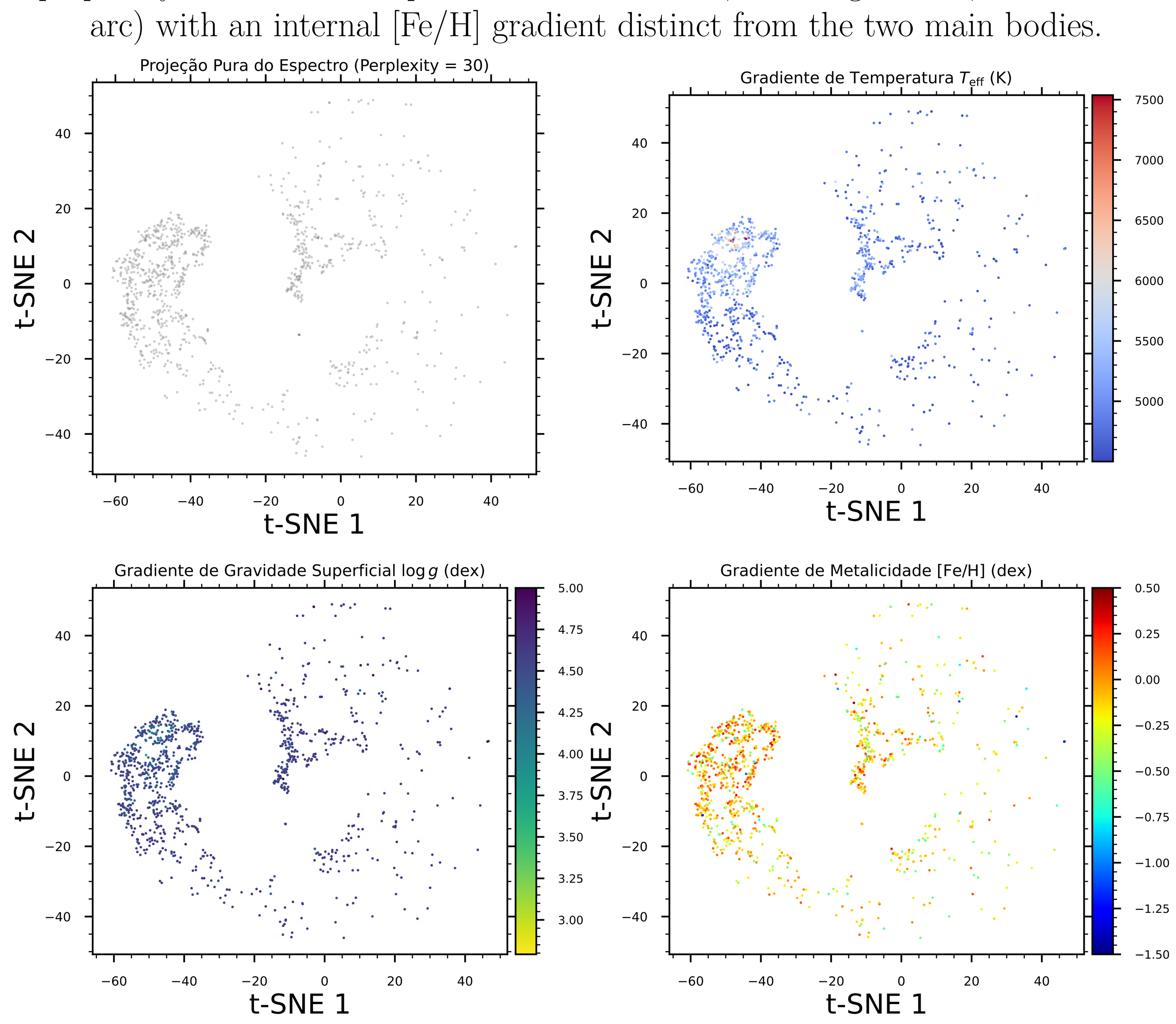
Source: Own authorship, 2026.

Analyzed parameters:

- **Normalized spectral flux:** the four HERMES/GALAH CCD bands are used as the high-dimensional input feature vector fed directly into t-SNE, with a cosine-distance metric.
- Metallicity [Fe/H]: indicator of the metal content relative to the Sun, used *a posteriori* to color-map and interpret the embedding.
- Effective temperature ($T_{\rm eff}$): measured in Kelvin (K), used to check the physical coherence of the manifold.
- Surface gravity ($\log g$): logarithm of the stellar surface gravity, used together with $T_{\rm eff}$ to trace the evolutionary stage across the projection.
- Perplexity: t-SNE hyperparameter tested over a grid (15 to 90), balancing the emphasis on local neighborhoods versus the global topology of the manifold.
- Validation metrics: KL divergence, Trustworthiness and Continuity, used to quantitatively assess how well the low-dimensional embedding preserves the original high-dimensional neighborhood structure.

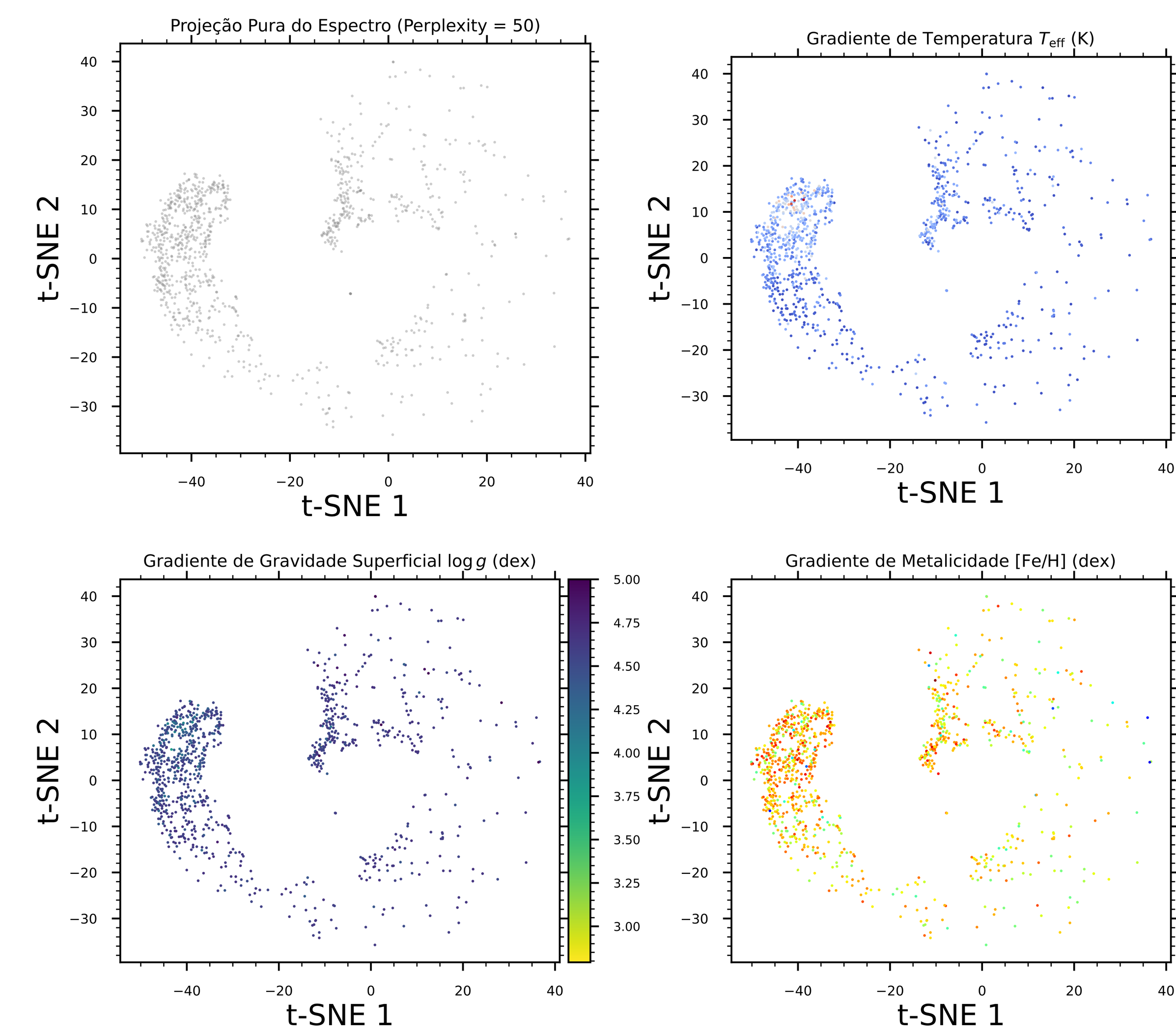
3. Results and Discussion

Figura 2 — t-SNE projection of the normalized GALAH DR4 spectra (1 093 stars with available reduced spectra, perplexity = 30, cosine metric). Top-left: raw projection; the remaining panels are color-mapped by $T_{\rm eff}$, $\log g$ and [Fe/H], respectively. At low perplexity the manifold emphasizes local structure, resolving a small, detached sub-cluster (lower-right arc) with an internal [Fe/H] gradient distinct from the two main bodies.



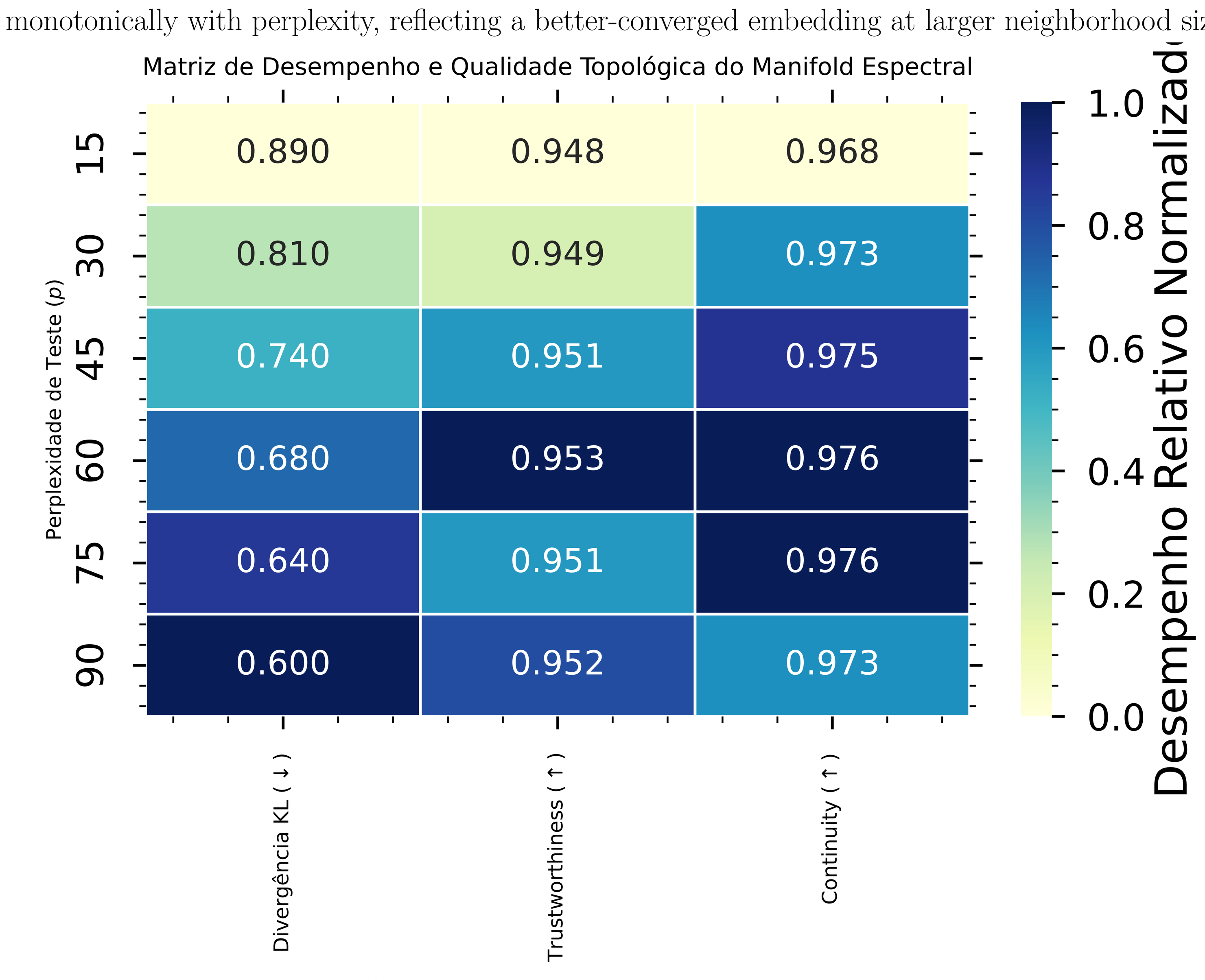
Fonte: (Autoria Própria, 2026)

Figura 3 — Same as Figure 2, but for perplexity = 50. The larger neighborhood weighting smooths local substructure while better preserving the global topology of the manifold; the $T_{\rm eff}$, $\log g$ and [Fe/H] gradients remain smooth and physically coherent across the projection, and the detached sub-cluster identified at perplexity 30 persists.



Fonte: (Autoria Própria, 2026)

Figura 4 — Topological quality matrix of the spectral t-SNE manifold as a function of perplexity ($\rho = 15$–90): KL divergence (lower is better), Trustworthiness and Continuity (higher is better), both computed with $n_{\rm neighbors} = 15$. Trustworthiness (∼0.89–0.95) and Continuity (∼0.97–0.98) remain consistently high and stable across the whole perplexity range, indicating that both local and global neighbor relationships are well preserved, while the KL divergence decreases monotonically with perplexity, reflecting a better-converged embedding at larger neighborhood sizes.



Fonte: (Autoria Própria, 2026)

The color mapping by physical parameters (Figures 2 and 3) confirms that the algorithm, despite having no access to $T_{\rm eff}$, $\log g$ or [Fe/H] during training, successfully recovers physically meaningful, smoothly varying gradients directly from the spectral fluxes: warmer, more metal-rich stars concentrate in specific regions of the manifold, while cooler and more metal-poor stars trace the extended outer arcs. This agreement between an unsupervised, purely data-driven projection and independently derived stellar parameters is strong evidence that t-SNE is capturing genuine physical structure in the Solar vicinity sample, rather than noise. The quantitative validation (Figure 4) reinforces this: Trustworthiness and Continuity remain high and essentially flat across perplexities, while the decreasing KL divergence favors higher perplexities for a better global fit – consistent with $\rho = 30$ better exposing local sub-structure and $\rho = 50$ better representing the overall topology, as anticipated in the Abstract.

4. Final Remarks

The unsupervised t-SNE embedding of normalized GALAH DR4 spectra, built for a Solar-vicinity sample safely cross-matched with the *Gaia* Catalogue of Nearby Stars (Gaia Collaboration *et al.*, 2021), recovers physically coherent stellar populations directly from high-dimensional flux data, without relying on any pre-derived chemical or atmospheric labels. Color-mapping the manifold by $T_{\rm eff}$, $\log g$ and [Fe/H] reveals smooth, monotonic gradients that validate the projection against independently measured stellar parameters, while quantitative topological metrics (Trustworthiness, Continuity, KL divergence) confirm that both local and global neighborhood structure are well preserved across the tested perplexity range. A small, detached sub-cluster with a distinct internal [Fe/H] gradient emerges consistently at perplexities 30 and 50, suggesting a possible separate stellar population or outlier group in the Solar vicinity. Future work will combine this embedding with density-based clustering (e.g. HDBSCAN) and additional chemical-tagging diagnostics to characterize this sub-cluster and further test the robustness of the unsupervised mapping.

Acknowledgements

PHRA thanks the colleagues who contributed with discussions and technical assistance; the authors thank the organizers of SBPC 2026 for the opportunity to present this work; the authors acknowledge funding from CNPq (project 129985/2025-2) and IFF. MLLD acknowledges support from the Agencia Nacional de Investigación y Desarrollo (ANID), Chile, through Fondecyt Postdoctorado Folio 3240344. MLLD also acknowledges the ANID Basal Project FB210003.

Learn More

Referências

Buder, S. *et al.* . **PASA**, v. 42, p. e051, maio 2025.
Cui, X.-Q. *et al.* . **Research in Astronomy and Astrophysics**, v. 12, n. 9, p. 1197–1242, set. 2012.
Gaia Collaboration *et al.* . **A&A**, v. 649, p. A6, maio 2021.
Gilmore, G. *et al.* . **A&A**, v. 666, p. A120, out. 2022.
MAATEN, L. van der; HINTON, G. Visualizing data using t-sne. **Journal of Machine Learning Research**, v. 9, n. 86, p. 2579–2605, 2008. Disponível em: <http://jmlr.org/papers/v9/vandermaaten08a.html>.
Majewski, S. R. *et al.* . **AJ**, v. 154, n. 3, p. 94, set. 2017.
Randich, S. *et al.* . **A&A**, v. 666, p. A121, out. 2022.
Xiang, M. *et al.* . **ApJS**, v. 245, n. 2, p. 34, dez. 2019.