

# Representation Learning for Gaia XP DR3

Bernd Doser, Kai L. Polsterer, and Sebastian Trujillo-Gomez

*Heidelberg Institute for Theoretical Studies, Heidelberg, Baden-Württemberg, Germany; bernd.doser@h-its.org*

**Abstract.** We present a novel representation learning framework for the Gaia XP DR3 dataset that leverages two advanced data exploration tools: Spherinator and HiPSter. Spherinator provides a method for learning compact representations of high-dimensional data, including images, point clouds, data cubes, time series, and spectra. Our training process uses variational autoencoders with hyper-spherical latent spaces to efficiently and robustly extract physically meaningful parameterizations of data properties. Our approach explicitly incorporates uncertainties from experiments into representation learning, which produces more robust and physically consistent latent representations. HiPSter generates and serves HiPS-based representations of learned features, enabling the scalable visualization and exploration of the latent space using Aladin-Lite. We demonstrate the scientific potential of our method using the largest spectral dataset available from Gaia XP DR3 and showcase the effectiveness of cross-disciplinary tools developed under the EU SPACE initiative to enhance data-driven astronomy.

## 1. Introduction

Machine learning provides an effective approach to reduce data to a lower number of dimensions while preserving the similarity of objects in a compressed representation. As part of the EU-funded project "Scalable Parallel Astrophysical Codes for Exascale" (SPACE), we developed the tools Spherinator and HiPSter (Polsterer et al. 2024). Spherinator utilizes a variational autoencoder to reduce high dimensional data including images or spectra to a 2-dimensional spherical space. HiPSter generates Hierarchical Progressive Survey (HiPS) images, which allows for visualization of the latent space using Aladin. Together, Spherinator and HiPSter are essential components of a machine-learning workflow that covers all stages, from data collection and preprocessing to training, prediction, and final deployment (Doser et al. 2025).

This contribution presents the results obtained from the Gaia XP spectra. The Gaia Data Release 3 (DR3) (De Angeli, F. et al. 2023) represents the largest spectroscopic survey ever conducted, encompassing approximately 220 million low-resolution spectra corresponding mostly to stars in the Milky Way.

The Gaia XP spectra have many scientific uses, including determining the atmospheric properties of stars, mapping the distribution of interstellar dust in the Milky Way, and identifying rare and exotic objects. Our goal is to test the capabilities of our tools for exploration and knowledge discovery using the compact representations learned from extremely large datasets.

This survey captures data across two photometric channels: the blue photometer (BP, 330-680 nm) and the red photometer (RP, 630-1050 nm). Together, these are

referred to as XP spectra. Each spectrum is a time-averaged mean spectrum, parameterized using Hermite polynomial basis functions with 55 coefficient amplitudes per channel. This approach allows for efficient storage and transmission of spectral information.

## 2. The Training

The Preprocessing Engine for Spherinator Training (PEST) is a comprehensive data preparation tool that handles diverse input formats. The continuous data was calibrated using GaiaXPy from 336 to 1021 nm in 2 nm increments (343 data points) and stored in an Apache Parquet<sup>1</sup> format for optimal performance and portability.

The Spherinator employs a variational autoencoder (VAE) operating in a (hyper-)spherical latent space and is trained with PyTorch Lightning. This architecture provides exceptional flexibility and can accommodate diverse data modalities, including images, spectra, data cubes, graphs, and point clouds, through its modular encoder-decoder architecture. For the Gaia spectroscopic data, feature extraction was performed using a one-dimensional convolutional neural network (1D-CNN) with eight layers. This network compresses the initial 343 spectral data points into a 128-dimensional feature map. The intermediate representation is then passed through a single fully connected layer to reduce it to a final 3-dimensional spherical bottleneck.

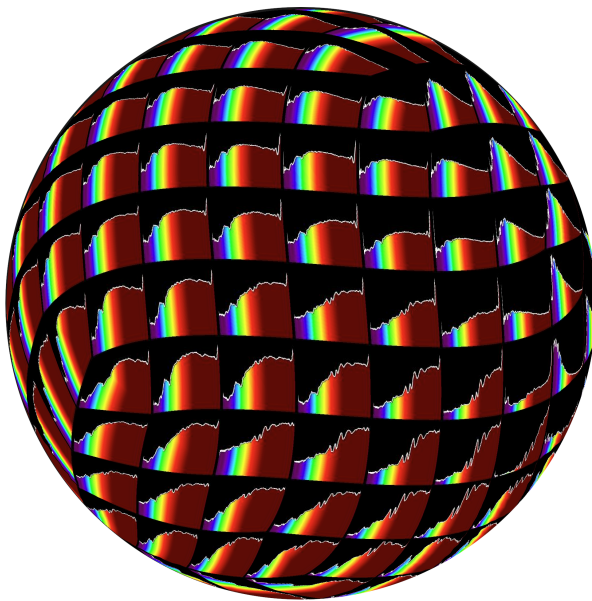


Figure 1. The spherical latent space learned from the Gaia XP DR3 dataset using Spherinator and visualized using HiPSter and Aladin-Lite. Each tile shows a spectrum representing all the similar objects in that region of the latent space. The colors show the simulated visual appearance of each spectrum.

---

<sup>1</sup><https://parquet.apache.org/>

The VAE loss function combines the Kullback-Leibler (KL) divergence with a reconstruction term, balancing regularization of the latent space and data fidelity. The balancing factor is set to  $\beta = 0.1$ . To account for the observational uncertainties present in Gaia flux measurements, we use a negative log-likelihood (NLL) approach. In this method, the reconstruction loss is calculated assuming that the data is represented by a normal distribution defined by the observed flux and its  $1\text{-}\sigma$  uncertainty.

### 3. Conclusions

The Spherinator model demonstrates excellent reconstruction capabilities on the very large Gaia XP spectra dataset, successfully capturing the relevant broad spectral features. The uncertainty-aware training approach results in reconstructions that appropriately reflect the confidence levels in different spectral regions, with higher fidelity in well-constrained wavelength ranges and appropriate uncertainty propagation in noisier regions.

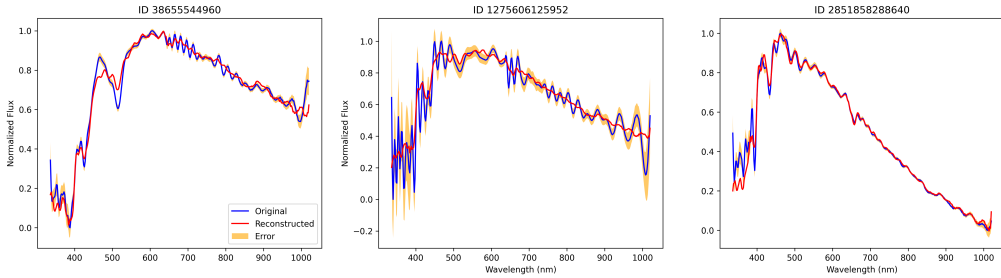


Figure 2. Three examples of Gaia XP spectra showing the original data (blue) and the reconstructions obtained from the learned compressed representations (red). The flux uncertainty is indicated in yellow, and the data has been min-max normalized.

**Acknowledgments.** We gratefully acknowledge the generous and invaluable support of the Klaus Tschira Foundation. This work has received funding from the European High Performance Computing Joint Undertaking (JU) and Belgium, Czech Republic, France, Germany, Greece, Italy, Norway, and Spain under grant agreement No101093441. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European High Performance Computing Joint Undertaking (JU) and Belgium, Czech Republic, France, Germany, Greece, Italy, Norway, and Spain. Code is available at <https://github.com/HITS-AIN/Spherinator>.

### References

- De Angeli, F., Weiler, M., Montegriffo, P., Evans, D. W., Riello, M., Andrae, R., Carrasco, J. M., Busso, G., Burgess, P. W., Cacciari, C., Davidson, M., Harrison, D. L., Hodgkin, S. T., Jordi, C., Osborne, P. J., Pancino, E., Altavilla, G., Barstow, M. A., Bailer-Jones, C. A. L., Bellazzini, M., Brown, A. G. A., Castellani, M., Cowell, S., Delchambre, L., De Luise, F., Diener, C., Fabricius, C., Fouesneau, M., Frémat, Y., Gilmore, G., Giuffrida, G., Hambly, N. C., Hidalgo, S., Holland, G., Kostrzewa-Rutkowska, Z., van Leeuwen, F., Lobel, A., Marinoni, S., Miller, N., Pagani, C., Palaversa, L., Piersimoni,

- A. M., Pulone, L., Ragaini, S., Rainer, M., Richards, P. J., Rixon, G. T., Ruz-Mieres, D., Sanna, N., Sarro, L. M., Rowell, N., Sordo, R., Walton, N. A., & Yoldas, A. 2023, *A&A*, 674, A2. URL <https://doi.org/10.1051/0004-6361/202243680>
- Doser, B., Polsterer, K. L., Fehlner, A., & Trujillo-Gomez, S. 2025, Machine Learning Workflow for Morphological Classification of Galaxies. eprint: 2505.04676, URL <https://arxiv.org/abs/2505.04676>
- Polsterer, K. L., Dosser, B., Fehlner, A., & Trujillo-Gomez, S. 2024, Spherinator and HiPSter: Representation Learning for Unbiased Knowledge Discovery from Simulations. eprint: 2406.03810, URL <https://arxiv.org/abs/2406.03810>