Representation Learning for Gaia XP DR3

Bernd Doser, Kai Lars Polsterer, Sebastian Trujillo-Gomez
Heidelberg Institute for Theoretical Studies (HITS)

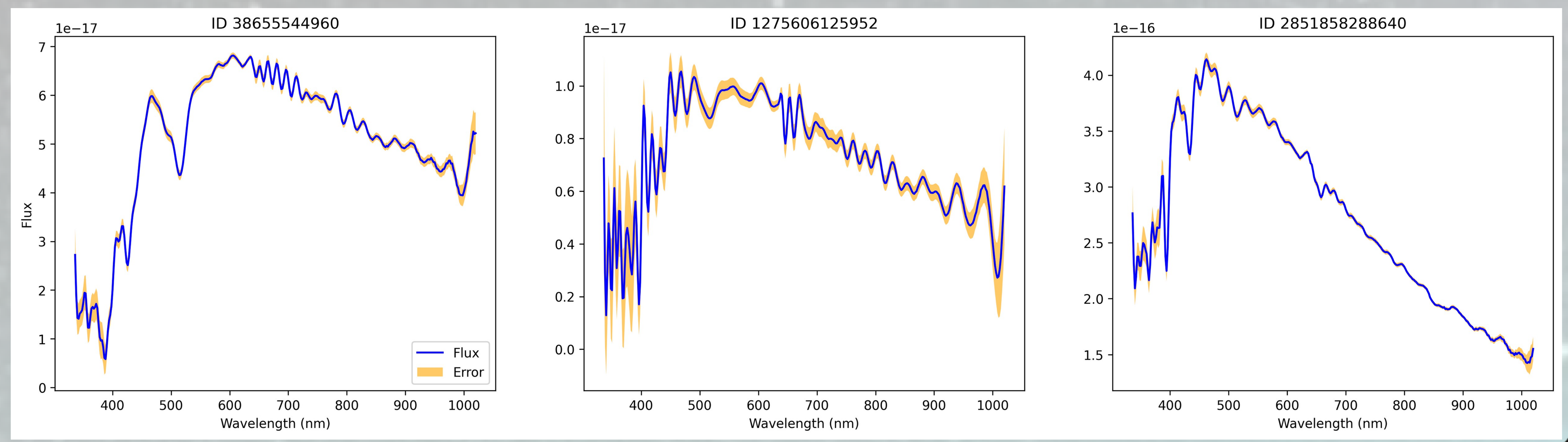
We present a novel representation learning framework for the **Gaia XP DR3 stellar** dataset that leverages two advanced tools: **Spherinator and HiPSter**. Spherinator provides a method for learning compact representations of high-dimensional data, including images, point clouds, data cubes, time series, and spectra. Our training process uses variational autoencoders with **hyperspherical 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 (hierarchical progressive surveys) 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 observational data from Gaia XP DR3 and showcase the effectiveness of **cross-disciplinary** tools developed under the **EU SPACE** initiative to enhance data-driven astronomy.

Data: Gaia XP DR3 Spectroscopic Survey

The Gaia Data Release 3 (DR3) represents the largest spectroscopic survey ever conducted, encompassing approximately **220 million low-resolution spectra**. The survey captures data across two photometric channels: the blue photometer (BP, 330-680 nm) and red photometer (RP, 630-1050 nm), collectively known as XP spectra. Each spectrum in DR3 represents a time-averaged mean spectrum, parameterized using **Hermite polynomial basis functions** with 55 coefficient amplitudes per channel, enabling efficient storage and transmission of spectral information.

Gaia Intro

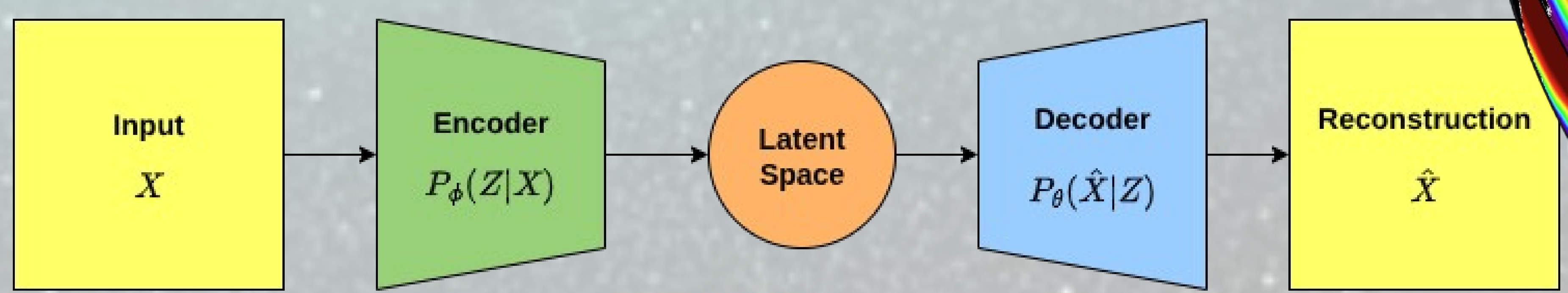


PEST: Data Preparation Pipeline

The **Preprocessing Engine for Spherinator Training (PEST)** serves as a comprehensive data preparation tool designed to handle diverse input formats. PEST integrates all preprocessing steps into a unified workflow, producing standardized data containers using the **Apache Parquet** format for optimal performance and portability.

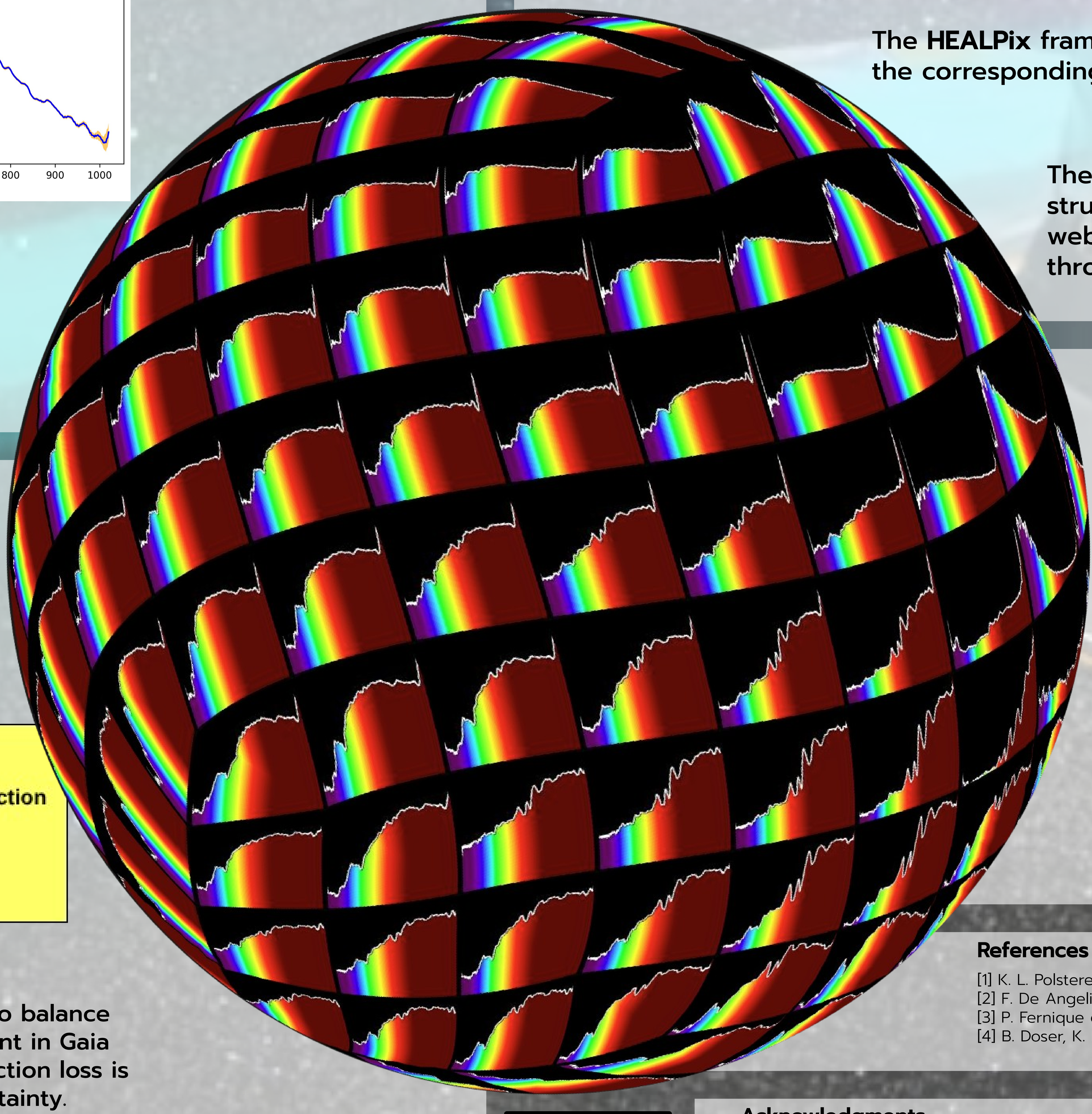
Spherinator: Variational Autoencoder Architecture

The Spherinator implements a novel variational autoencoder (VAE) that maps input data to a **(hyper-)spherical latent space**, built using the **PyTorch Lightning** framework. This architecture offers exceptional flexibility, supporting **multiple data modalities** including images, spectra, data cubes, graphs, and point clouds through modular encoder-decoder designs. For spectroscopic data processing, the model employs a **one-dimensional convolutional neural network (1D-CNN)** architecture optimized for sequential spectral features.



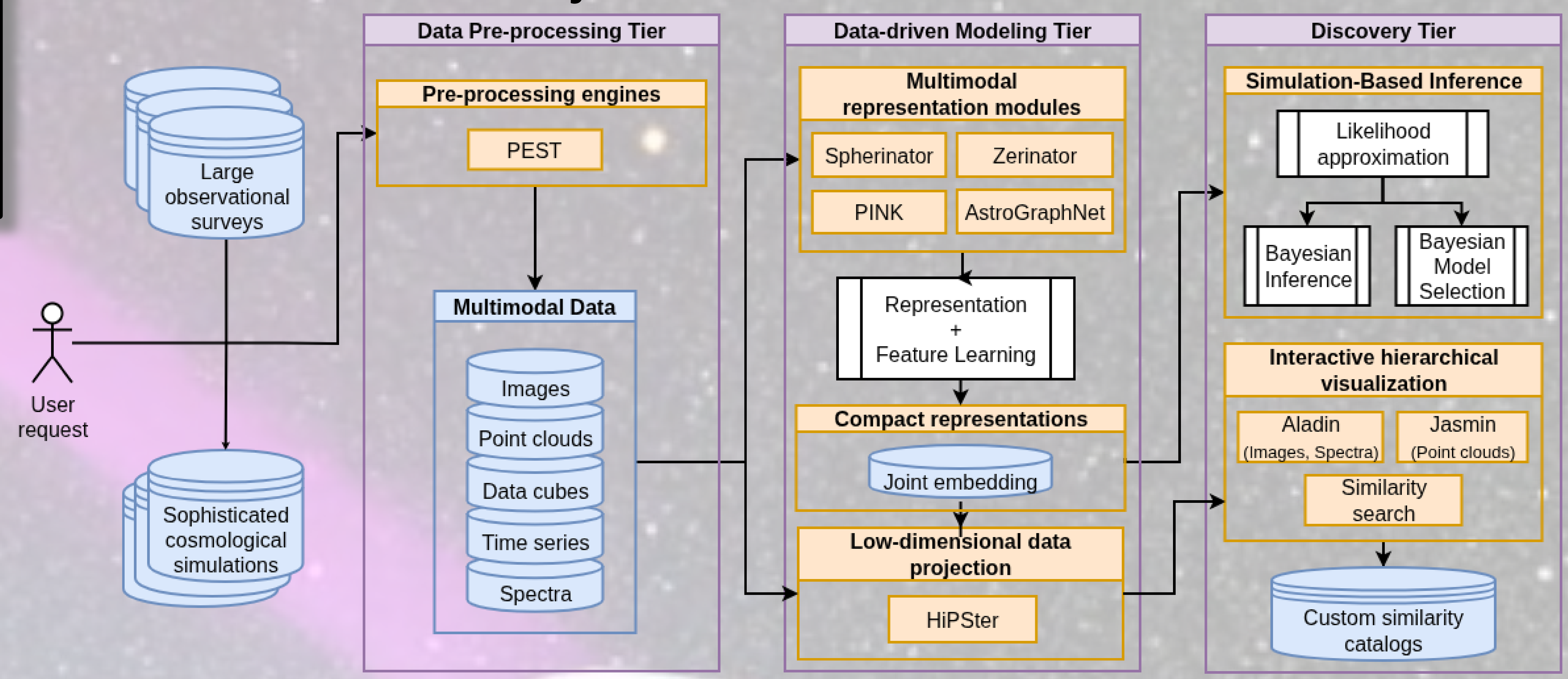
Uncertainty-Aware Loss Function

The VAE loss function combines **Kullback-Leibler (KL) divergence** with a reconstruction term to balance latent space regularization and data fidelity. To incorporate observational uncertainties inherent in Gaia flux measurements, we employ a **negative log-likelihood (NLL)** approach where the reconstruction loss is computed against a normal distribution parameterized by the observed flux and its 1-σ uncertainty.



Read The Docs

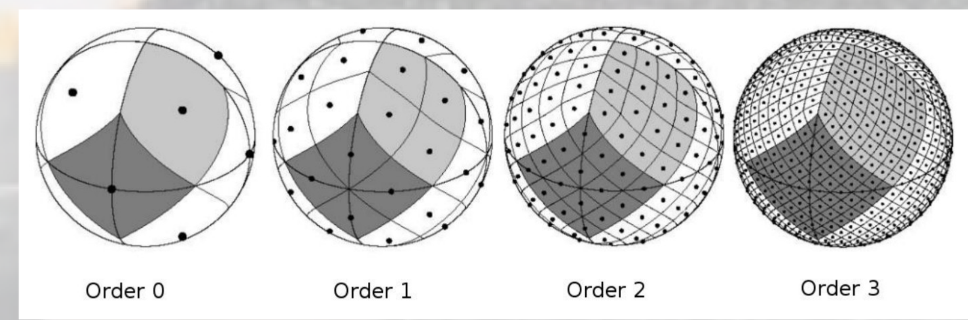
Project X: The Workflow



HIPSter: Inference and Visualization

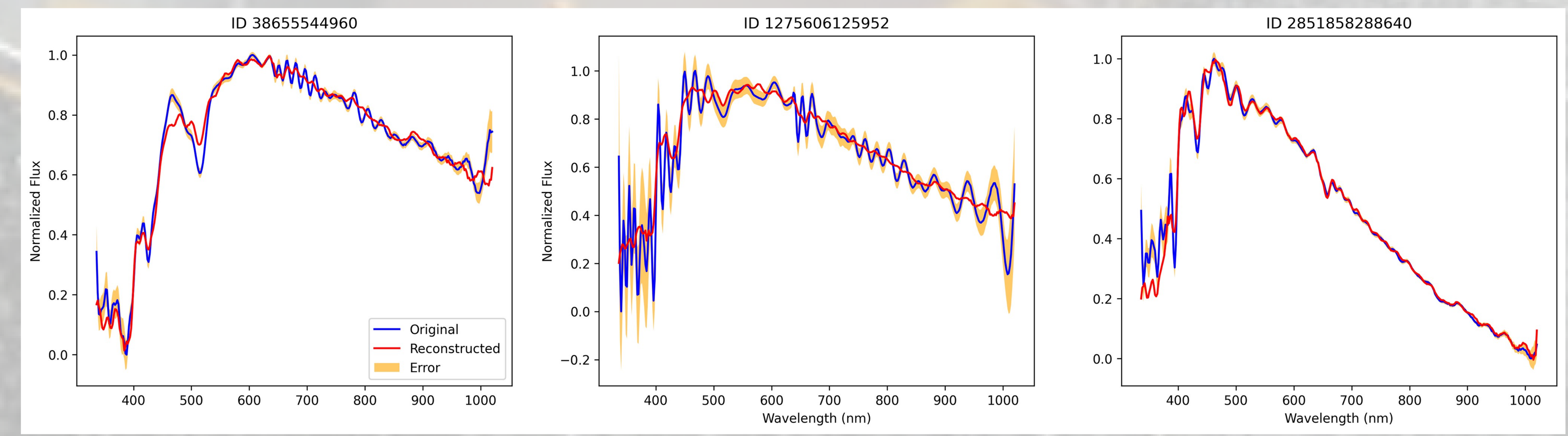
The **HEALPix** framework is used to generate a **Hierarchical Progressive Survey (HiPS)** [3] to map the corresponding spherical latent space positions.

The resulting HiPS data structure is optimized for web-based visualization through **Aladin-Lite**.



Reconstruction Performance

The Spherinator model demonstrates excellent reconstruction capabilities on Gaia XP spectra, 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.



References

[1] K. L. Polsterer, B. Doser, A. Fehlner, and S. Trujillo-Gomez, ADASS XXXIII (2024).
[2] F. De Angeli et al., Astronomy & Astrophysics, 674, A2 (2023).
[3] P. Fernique et al., Astronomy & Astrophysics 578, A114 (2015).
[4] B. Doser, K. L. Polsterer, A. Fehlner, and S. Trujillo-Gomez, ADASS XXXIV (2025).

TRY IT OUT

Acknowledgments

Funded by the European Union. 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. Neither the European Union nor the granting authority can be held responsible for them.
We gratefully acknowledge the generous and invaluable support of the Klaus Tschira Foundation.
This research has made use of "Aladin sky atlas" developed at CDS, Strasbourg Observatory, France. We are grateful to Thomas Boch for help with AladinLite..

Co-funded by the European Union

ADASS XXXV
Görlitz, Germany, 2025