# Explainable XRF Element Detection

## with Embedding-Space k-NN and Physics-Based Masks for the HARLOCK Lunar Rover


Alberto Luise,[1] Michele Lombardi,[1] Nicolò Parmiggiani,[2] Andrea Bulgarelli,[2] Samuele Salti,[1] Sara Trevisan,[3] Federico Dogo,[3]
Luca Castaldini,[2] Ambra Di Piano,[2] Francesca Esposito,[4] Antonio Femia,[2] Fabrizio Fiore,[3] Eduardo Medinaceli,[2] Ghulam Mustafa,[1] Alessio Trois [5]

**1**: University of Bologna **2**: INAF/OAS Bologna **3**: INAF O.A. Trieste
**4**: INAF O.A. Cagliari **5**: INAF O.A. Napoli



## ABSTRACT

Deep neural networks are often considered "black boxes", and the lack of transparency can limit their adoption in scientific contexts; in this work, we propose a framework that converts any Convolutional spectral regressor into an explainable predictor without hindering its performance. This is a contribution for **RHEA**, the *Rare-earth & Heavy Elements Analyzer* of the **HARLOCK** project within the **PRORIS** program. It combines multiple analytical techniques to identify Rare-Earth Elements (REE) deposits and associated mineralogical signatures across different spatial scales.

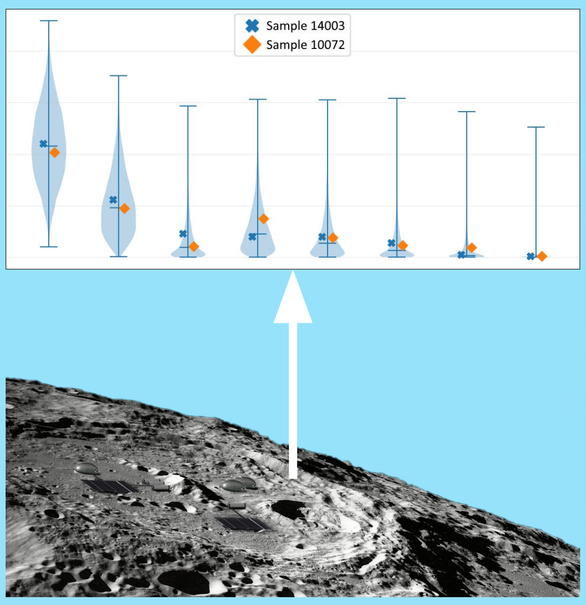
Our approach exploits the CNN's latent representation and preservation of spatial features: after training, the final regression layer of the backbone model is replaced by a **K-Nearest Neighbors (k-NN) regressor** operating directly on the learned embedding space; as a result, every prediction is accompanied by an explanation in the form of the most similar spectra observed during training. This works particularly well thanks to a **custom loss function** designed to optimize the quality of the learned embedding space, without hindering performances. Furthermore, an attention mask is learned for each predicted element starting from a physics-based initialization, providing easily readable heatmaps as well in the form of a mask-by-weights product. Experimental results show that, in a regression task estimating the composition of soil samples, the average R2 score over the predicted elements stays around 0.9 for both the original approach and the explainable Deep k-NN model.


## DATASET

Because a dataset of lunar soil compositions bombarded with X-Rays is not readily available, we developed a simulation pipeline for our training data. First we sample data from a Dirichlet distribution centered around two of the few analyzed lunar rocks present in literature, then we generate the secondary X-ray spectra using the **Fundamental Parameters method**.
It generates spectra based on

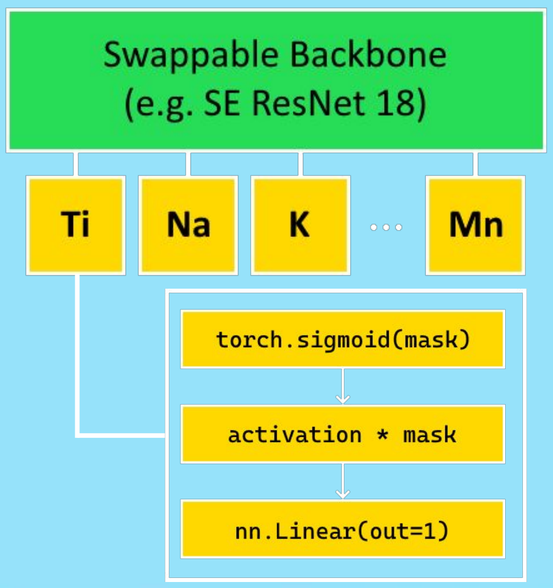
- elemental composition
- geometry of sample, emitter and detector (angles of incidence)
- detector's material and specifics

and convolves the result to match the actual response of the instrument.



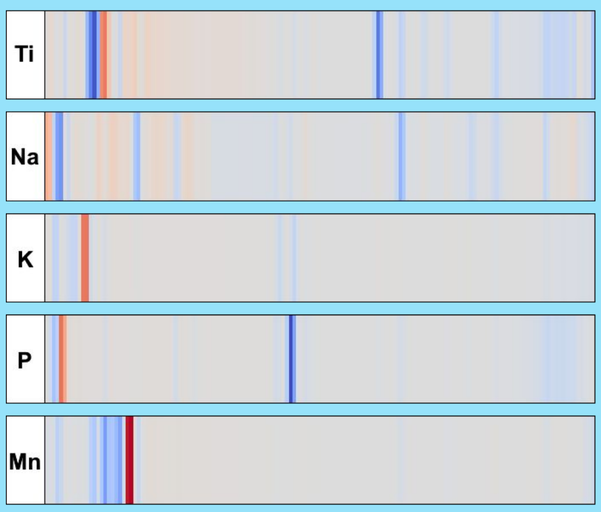

## ARCHITECTURE

Our framework is centered around a **backbone**, which can be any Convolutional Neural Network, and a final layer composed of **multiple Regression Heads**, each predicting the concentration of a different element. Each Regression Head has its own associated **Mask**, which is multiplied element-wise by the last activation of the Backbone during a complete forward pass. Finally, we can replace the last layer of the architecture with a **k-NN Regressor**, which predicts concentrations by local interpolation of embedding-space spectra in the training set.



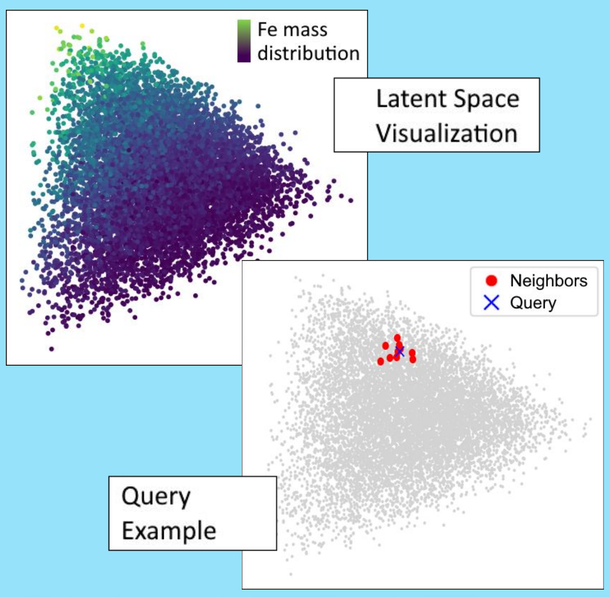

## EXPLAINABILITY

### Physics-Informed Masks




Each mask is multiplied element-wise with the activations of the final convolutional layer, increasing focus on relevant parts of the input with **positive** or **negative** weights.
They are learned at training time and initialized w.r.t. the emission lines of their respective element, also accounting for "interfering" elements with a similar XRF response. This approach is effective thanks to the CNN preserving the spatial features of the input spectra.
The masks help to stabilize training by guiding Gradient Descent towards convergence.

### k-Nearest Neighbors

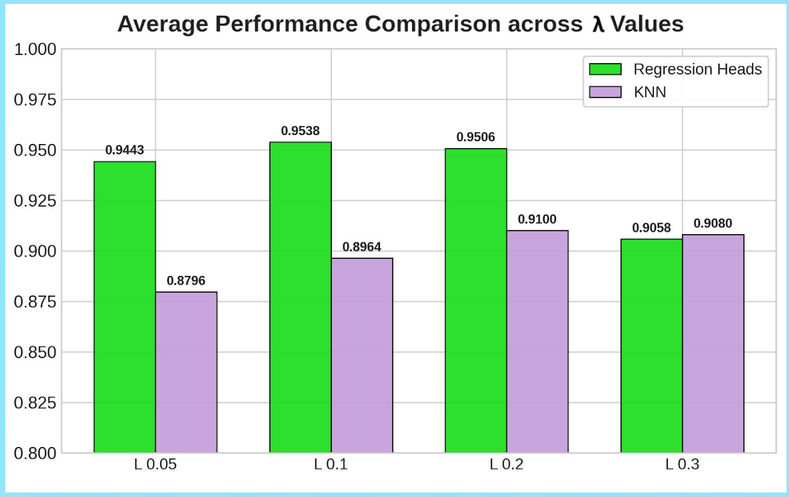



By replacing the regression heads with an inherently interpretable model, a **k-Nearest Neighbor regressor** (which ineterpolates predictions from the training data), we propose an instance based explainability mechanism. The algorithm operates on the latent space of the trained CNN, which preserves spatial features and maintains an optimal ordering thanks to a **custom loss function** integrating MSE and a Triplet Loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MSE}} + \lambda \cdot \mathcal{L}_{\text{Triplet}}$$

Keeping a low λ value maintains a well-ordered latent space without hindering performance.

The **λ** parameter of the loss balances the two metrics; a value for λ too low does not order the latent space enough for k-NN to function, while a value too high hinders the regressor's performance. The sweet spot is around 0.2, where the drop in performance can be considered negligible as the k-NN is mainly employed as an explainability mechanism.




## EXPERIMENTS SUMMARY & FUTURE DEVELOPMENTS

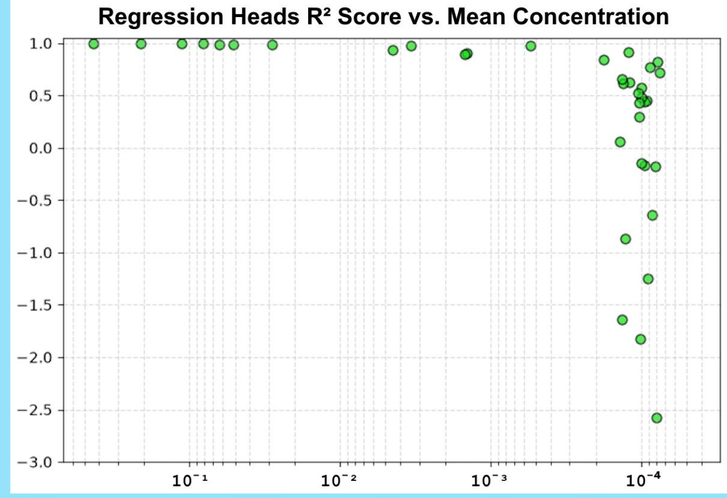
The performance metric employed is the $R^2$ score:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

which shows how much better the model predicts targets than always guessing their average.
The dataset can be built with a minimum threshold for elemental concentration; lowering the threshold below $10^{-4}$ makes the $R^2$ drop below 0, meaning the results on components less than 100 parts per million are worse than always predicting the average.
This is not necessarily a bad result: with a limited exposure time, some elements like Thorium (typically present in 15 ppm) seem to not detectable by our instrument; therefore the graph actually shows that **our algorithm is aligned with actual lab results**.

The next steps to be taken in this research include

- applying TinyML principles to make the algorithm run on the Edge computers (quantization, pruning)
- improve the dataset simulation pipeline
- focus on priority elements
- perform extensive robustness testing





36th annual Astronomical Data Analysis Software and Systems (**ADASS**) conference