Project Title	European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructure
Project Acronym	ESCAPE
Grant Agreement No	824064
Instrument	Research and Innovation Action (RIA)
Topic	Connecting ESFRI infrastructures through Cluster projects (INFRA-EOSC-4-2018)
Start Date of Project	04.02.2019
Duration of Project	48 Months
Project Website	https://projectescape.eu

D4.5 - RELEASE OF PROTOTYPE MACHINE LEARNING-ENABLED ARCHIVE SERVICES PROVIDING VALUE-ADDED CONTENT TO ARCHIVES

Work Package	WP4, Connecting ESFRI projects to EOSC through VO framework
Lead Author (Org)	Martino Romaniello (ESO) & Mark Allen (CNRS-ObAS)
Contributing Author(s) (Org)	Felix Stoehr (ESO), Kai Polsterer (HITS)
Due Date	31.01.2022, M36
Date	28.01.2022
Version	2.0

Dissemination Level

X	PU: Public
	PP: Restricted to other programme participants (including the Commission)
	RE: Restricted to a group specified by the consortium (including the Commission)
	CO: Confidential, only for members of the consortium (including the Commission)

Versioning and contribution history

Version	Date	Authors	Notes
0.1	21.12.2021	Mark Allen (CNRS-ObAS), Martino Romaniello (ESO), Felix Stoehr (ESO)	First draft and definition of content
0.2	22.12.2021	Mark Allen (CNRS-ObAS), Martino Romaniello (ESO), Felix Stoehr (ESO)	Complete draft of sections and content.
0.3	11.01.2022	Mark Allen (CNRS-ObAS), Martino Romaniello (ESO), Felix Stoehr (ESO)	Revised complete draft.
0.4	23.01.2022	Mark Allen (CNRS-ObAS), Martino Romaniello (ESO), Felix Stoehr (ESO), Kai Polsterer (HITS)	Updated complete draft.
1.0	25.01.2022	Mark Allen (CNRS-ObAS), Martino Romaniello (ESO), Felix Stoehr (ESO), Kai Polsterer (HITS)	Version submitted to Project Manager for review.
2.0	28.01.2022	Mark Allen (CNRS-ObAS), Martino Romaniello (ESO), Felix Stoehr (ESO), Kai Polsterer (HITS)	Version following comments from the internal project review of the report by Kay Graf.

Disclaimer

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 210506816.

Release of prototype machine learning-enabled archive services providing value-added content to archives

Table of Contents

Executive Summary

This report constitutes deliverable D4.5 and details the activities carried out in Task 4.3 to explore the application of techniques from the field of Machine Learning to astronomical archive search capabilities. They were carried out in a collaboration between ESCAPE partners CDS, ESO and HITS. Two complementary Deep Learning approaches were followed. One part of the team put a special emphasis on exploring how to make the output of the Deep Learning (the identification of similar data) interpretable on a physical basis with high precision. As this places time consuming demands on precomputing and training the machine learning models, the other part of the team was dedicated to experimenting with models that allow for an extremely quick on-the-fly training. Ultimately, we were able to prove the viability of enabling archival searches based on data similarity, in which users upload a data item they are interested in, and the system returns similar assets. Our results indicate that, at least to some extent, the similarity can be interpreted in terms of physical properties of the celestial objects. In addition to this, the chosen architecture has also serendipitously shown a tendency to distinguish intrinsic features from those superimposed by the Earth atmosphere. If confirmed, this finding opens an interesting window of opportunity into improving observatory operations and efficiency, in that the science data itself can be used to correct for telluric features, thus sparing precious telescope time that would otherwise have to be devoted to collecting calibration data.

The results and associated prototypes have been presented to various communities over the course of the ESCAPE project. These presentations have included demonstrations of the prototype tools, and we have used the scientific results obtained to highlight the potential of these new kinds of services.

The ESCAPE project, and the wide range of expertise of the partners, has created the ideal interdisciplinary connection needed for such an exploratory work. The further developments and explorations that are needed to bring these early results to production systems are out of its scope. ESCAPE was instrumental in showing the basic feasibility of the stated goals, thus taking a first, very significant step towards eventually reaching them.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

1. Introduction

ESO hosts and operates science archives for the ALMA and La Silla Paranal Observatories[1]. These archives form an integral part of the respective science operations and are fundamental contributors to the science output of the facilities, with well over 30% of the resulting refereed papers making use of archival data. This is illustrated in Figure 1, where the refereed papers published in 2021 that made use of ESO data are broken down according to the source of the data: proprietary only (PI, Principal Investigator), archive only and a combination of the two.

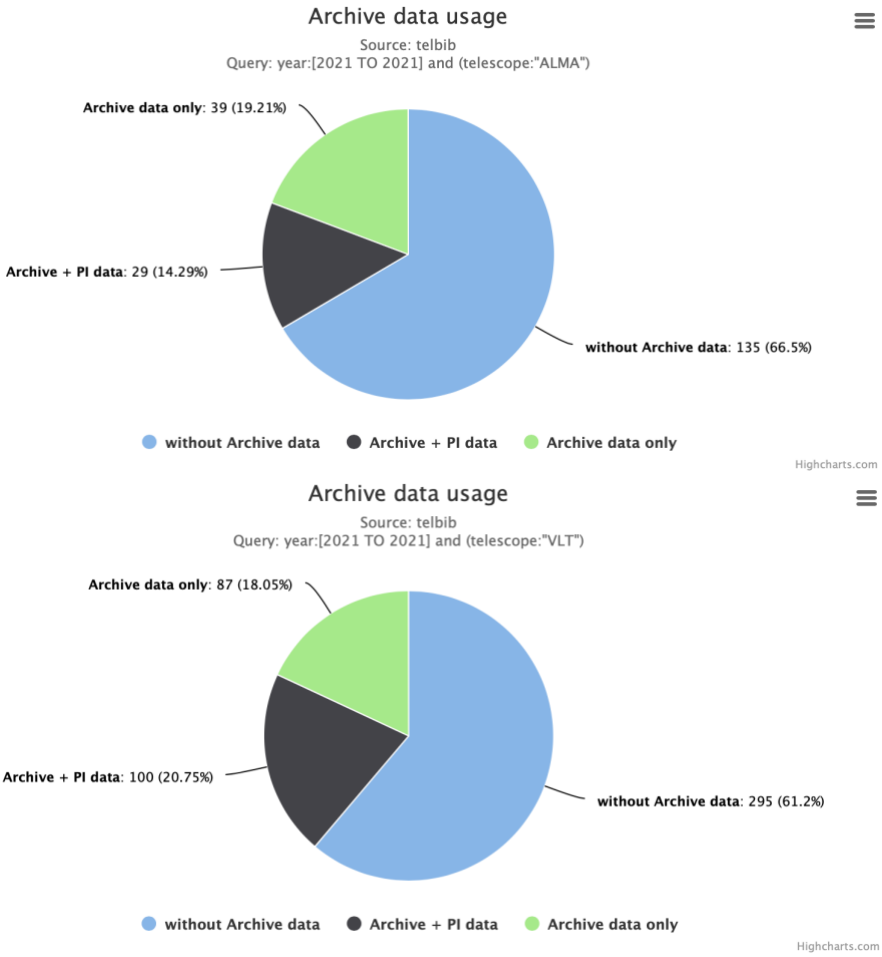

Figure 1. Breakdown of the refereed papers published in 2021 that made use of ESO data. Top panel: ALMA. Bottom panel: La Silla Paranal Observatory. Source: ESO Telescope Bibliography, telbib.eso.org.

[1] Useful links: ESO La Silla Paranal Observatory Science Archive, ALMA Observatory Science Archive, ESO home page, ALMA Observatory home page.

This is part of a sustained upwards trend that has continued throughout the years, as exemplified in Figure 2 for the La Silla Paranal Observatory, which firmly establishes the role of science archives as an indispensable tool in the contemporary landscape of forefront research in Astronomy. This will be extend also to the Extremely Large Telescope (ELT), the Landmark ESFRI that is currently in construction with operations to begin as of 2027.

With millions of data files being made available to the science community, it is crucial to the success of the archives that they provide advanced query capabilities to guide the researchers to identify which of those data are of interest for their specific use. Furthermore, lowering the barriers to data access is an integral part of the ambitions of Open Science initiatives such as EOSC, to broaden the use of the data itself. Our present quest has the potential to further open the use of data by making them findable without requiring a specific knowledge of the source, thus allowing researchers to focus fully on pursuing their scientific goals.

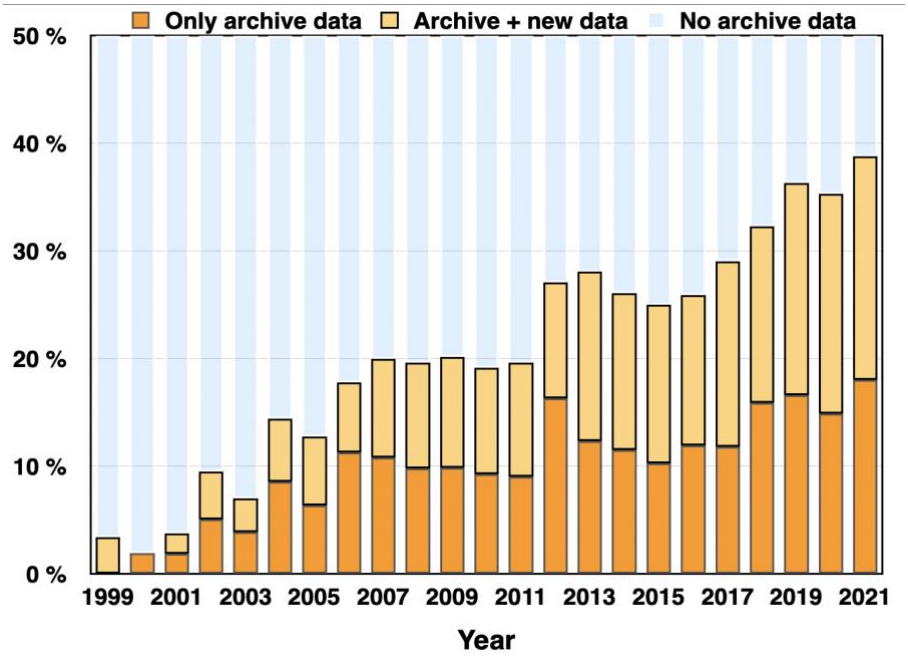

Figure 2. Evolution with time of the contribution of the La Silla Paranal Observatory Science Archive to the referred publications making use of data from the facility. Source: ESO Telescope Bibliography, telbib.eso.org.

A considerable effort was made in recent years at ESO, and elsewhere, to evolve the data search capabilities of archives, from the mere technical description of the instrumental setup that was used in taking the data, towards a science-user oriented approach, where the properties of the data are used to constrain the archive search. As an example, we show in Figure 3 the set of query metadata that are used in the ESO Archive Science Portal, which includes items that relate directly to properties of the data, like spectral range and resolution, signal-to-noise ratio, image sensitivity, and on-sky resolution.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

Figure 3. Screenshot of the ESO Archive Science Portal to highlight the query parameters exposed to users. They include a description of the data based on its properties, like items like spectra range and resolution, signal-to-noise ratio, sensitivity, and on-sky resolution.

2. Machine learning for archival data

The scope of ESCAPE Task 4.3 is to advance further along this trajectory, exploring if, and to what extent, it is possible to incorporate novel techniques from the general field of Artificial Intelligence into the archive search capabilities. State-of-the-art archival queries are based on the explicit specification of criteria, therefore a usual approach of applying machine learning methods would be in the direction of extracting a set of relevant physical properties and attaching it to the relevant archive entry.

Supervised learning methods are typically trained to solve specific tasks like classifying astrophysical sources or extracting properties via a regression approach, both of which would be well suited to the current schema of database access. Even though supervised learning has often shown excellent performance, it is limited by its need for high quality-labelled training data, as well as the limitation that each model is trained to only solve a very specific task. Given the large number of archive assets at ESO and their broad variety, covering essentially all types of celestial sources, together with the multitude of use-cases that scientists are interested in, we decided to focus on explorative, rather than supervised learning methods. Our goal is to at least minimise, or ideally make obsolete, the demand of pre-defined categories. We focused on unsupervised methods, namely dimensionality reduction, to explore, to what extent the structuring of the data could be self-consistently provided as outcome of the analysis.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

We deliberately decided to follow two complementary Deep Learning approaches. Since our intended audience is astronomers (who query the science archives), one part of the team put a special emphasis on exploring how to make the results interpretable on a physical basis with high precision. As this places time consuming demands on precomputing and training the machine learning models, the other part of the team was dedicated to experimenting with models that allow for an extremely quick on-the-fly training. Even though this is by construction more limited with respect to precision of the representation of the data items, it provides to the user the flexibility to retrain the model to their specific needs. **Ultimately, we were able to prove the viability of enabling archival searches based on data similarity, in which users upload a data item they are interested in, and the system returns similar assets.** Furthermore, our results indicate that, at least to some extent, the similarity can be interpreted in terms of physical properties of the celestial objects.

The participation of ESO in the ESCAPE project has enabled the deployment of resources at ESO to pursue this exploration, which would have otherwise not been possible. This is not only because of limitations in the resources themselves available at ESO. Perhaps more importantly, the ESCAPE project, and the wide range of expertise of the partners, has created the ideal interdisciplinary connection needed for such an exploratory work. This task directly involves the ESCAPE partners of ESO, HITS and CNRS-ObAS, and is a linking activity between WP4 and WP3, with activities integrated into the work plans of both work packages.

3. System setup

Hardware setup

Deep Learning requires specialised hardware. In particular, Graphical Processing Units (GPU) that accelerate the training of a machine-learning model by at least one order of magnitude. Optimising the hardware performance is very important, because finding a good Deep Learning model is an iterative process where the free parameters, like the learning rate, the exact network layout, the number of elements that are sent for learning in one batch etc., have to be explored. This process is called hyper-parameter search. In a typical exploration workflow, therefore, several models are set up for training in parallel at any given time.

ESO

To support an efficient exploration, two dedicated, performant servers were acquired. In addition to a very powerful 64 core CPU as well as 512GB of RAM, each server contains 4 high-end Nvidia Titan RTX GPUs with 24GB of RAM per GPU. This allowed us to execute eight model training runs in parallel and the machines were used extensively and exclusively for the ESCAPE project.
A high-speed nVMe SSD is used for the operating system and several standard harddisks provide 16TB of storage capacity. The servers share the same disk-space including the user's home directories so that it is easy to run on both machines simultaneously.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

HITS

HITS operates a GPU cluster equipped with ≈300 GPUs of the latest Nvidia generation (A40 K40 and RTX3080), which is directly connected to a storage network. Access was provided to this cluster for the project partners. As HITS was working on rapidly trainable models for an interactive retraining, most of the actual training happened on laptops or desktop computers, all equipped with user-grade GPUs.

Software setup

There are two main standard Deep Learning frameworks available. *Tensorflow*[2] (by Google) and *PyTorch*[3] (by Facebook). Both frameworks are released as open-source and are very similar in their features. We have decided to use *PyTorch* for the deep convolutional architecture and *Tensorflow* for the plain autoencoder. The installation was straight-forward using the standard instructions.

Deep Learning exploration is typically done in Python which we also have used here. The standard plotting library matplotlib was used to display the results of the learning, e.g. the input and predicted spectra.

The convergence of the models was checked in the browser using the *tensorboard* software which is available for *PyTorch* through the *torchvision* package. In regular intervals, our code was set to add data-points and images to our *tensorboard*.

We have explored two different approaches for the prototypes to provide interactivity. Whereas the application at HITS made use of *matplotlibs* widgets, the prototypes at ESO were built around Jupyter Notebooks. The latter were then converted into a web-application using the *voila* software package and server.

Deep Learning network setup

The exploration of the ESCAPE work was split between ESO and HITS using two different approaches. While at HITS the resolution of the spectra was purposefully reduced substantially before the learning phase and simple architecture like fully-connected auto-encoders were trained, at ESO the full information of the spectra was used for training. Again, an auto-encoder was trained, but this time using a deep convolutional network.

In both our cases, and in general, the definition of the actual network architecture is the easy part. The hard work is to prepare the data so that it can be used for training, to explore the parameter space, to find ways to allow the model to converge and to analyse and interpret the results.

An auto-encoder is a widely adopted network architecture for unsupervised learning, i.e. Deep Learning where no human has pre-classified a sub-sample of the training set before. Instead, the

[2] https://www.tensorflow.org
[3] https://pytorch.org

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

network is given input data (e.g. a spectrum) and is requested to try to reproduce the input as accurately as possible in the output. In the centre of the network architecture, however, only a very small amount of information can pass through. In our case, typically between 4 and 128 floating point values. This “information bottleneck”, also called latent space, requires the network to extract the most useful and smallest amount of information, allowing it to then expand that information again into the full spectrum. It turns out that two input entities (spectra) are similar, if they are also similar in the latent space. As such, the compressing and encoding part, as well as the reconstructing and decoding part do not necessarily need to be symmetric.
Like any other neural network, auto-encoders will try to fulfil the given task by taking the shortest route when optimizing the weights and biases. Therefore, extreme care has to be taken - by constantly accompanying the analysis - to ensure that what is learned is really a generalized representation of the data and that it is not just learning artifacts. For example, the HARPS[4] data are combined from two “arms” of the same spectrograph, resulting in a small gap in the middle of each of the spectra. Careful analysis showed that the auto-encoder architectures initially learned to reproduce the gap, rather than focussing on the scientific content. As a result of this, masking had to be applied to make sure that no artefacts were learned.

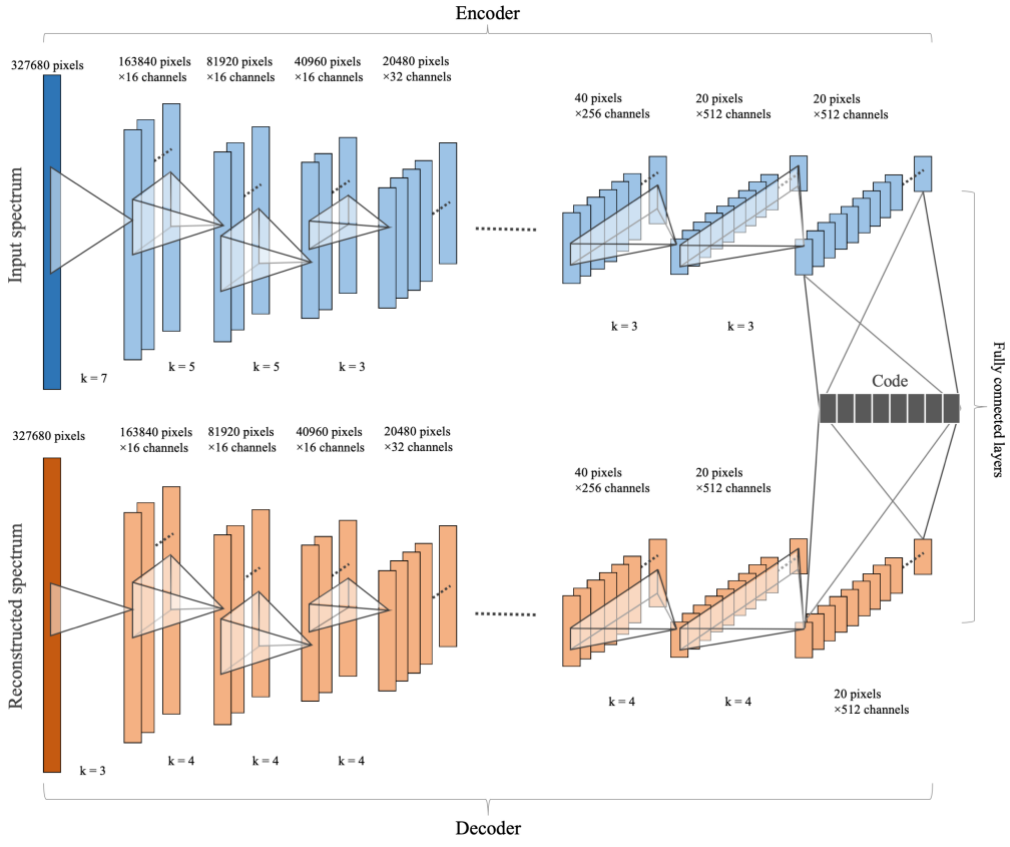

Figure 4. Detailed architecture of the deterministic autoencoder as published in section A.1 of our paper Sedaghat et al. (2021, MNRAS,501,6026; available in Open Access: arXiv).

[4] HARPS is a high-resolution spectrograph operated by ESO at its La Silla site in Northern Chile. It is mainly devoted to the study of stars in the solar neighbourdood. More information can be found at: https://www.eso.org/sci/facilities/lasilla/instruments/harps.html.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union’s Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

In the deep convolution model, an additional constraint was posed onto the latent space. The idea was to impose constraints through a loss function in such a way, that the different dimensions of the latent space are mostly independent of each other, ideally orthogonal. This technique, known as disentanglement, was first proposed by Kingma & Welling (2014, arXiv). The independence of the information content in the disentangled dimensions makes the physical interpretation of the dimensions much easier - or possible in the first place.

Adding disentanglement constraints to the latent space, however, comes with the drawback that the original number of dimensions in the latent space that carry information need to be increased very substantially, e.g. from 6 to 128. This is a phenomenon which is widely discussed in the literature.

In order to identify the dimensions that really carry the information, two independent statistical and information-theoretical methods for finding the number of learned informative features have been developed. We have also measured their true correlation with astrophysical validation labels for a selected subset of the data, involving astronomers with very intimate knowledge of HARPs data. For further information, see below.

For further details about the methods chosen and the results, please refer to the publication N. Sedaghat et al. (2021, MNRAS, 501, 6026; available in Open Access: arXiv). Our implementation is based on autoencoders (Vincent et al 2010, Journal of machine learning research, 11), the de-facto framework for unsupervised approaches in deep learning.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

4. Science Results

Prototypes at HITS

The HITS partner worked on the development of a prototype based on on-the-fly dimensionality-reduction methods for an interactive compressed visualisation, inspection, and interaction with spectral observations. The latent space was set to two dimensions only, to allow a presentation on a screen. This enables scientists to quickly and easily browse massive datasets that are ordered by structural similarities to find classes, outliers/anomalies, and scientifically relevant objects in the dataset. The software shows in real time the lower-dimensional projections and the corresponding reconstructed spectra obtained from the autoencoder. The prototype has been developed in Python and adopts Tensorflow for the neural network model and Matplotlib to generate the main interfaces (see Fig. 5). The autoencoder model was trained with the entire dataset of HARPS spectra by using the HITS GPU cluster, whereas the catalogue containing the unique spectra was only used for visualising the results (i.e., the projections) and for an interactive update when retraining the model. This method enabled us to build an agile interface which allows for visualisation and retraining to be performed on a standard laptop.

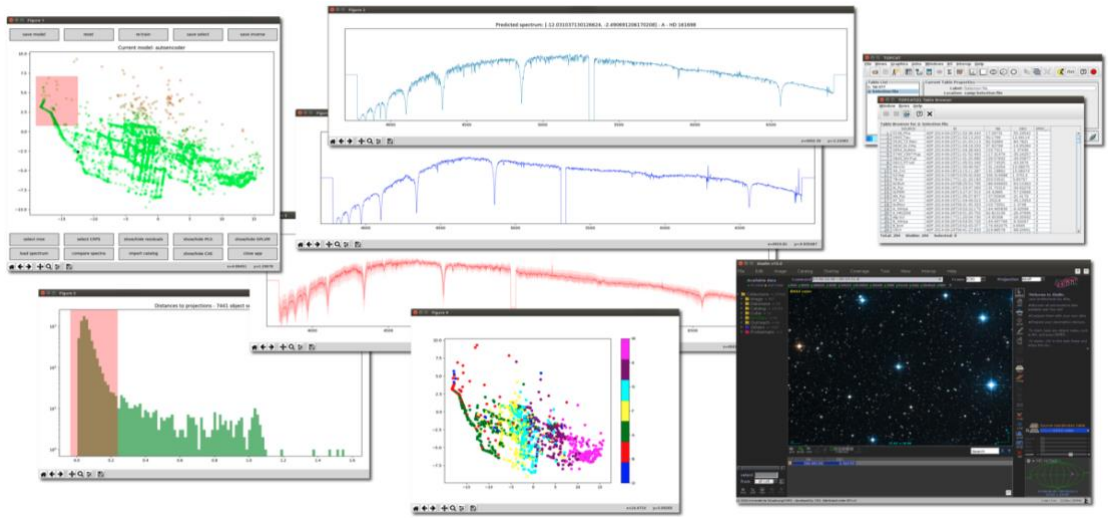

Figure 5. MEGAVIS The prototype developed at HITS to explore the spectra in the HARPS archive. The overview panel allows for interaction with the ML models and for selecting data (upper left). The characteristics of the spectra at the selected coordinate and in the selected area are shown in the three corresponding spectral plots (center). Histogram functions for subset selection as well as an overplot of spectral classes are shown at the bottom. VO tools – such as Topcat (for inspecting the selected sources as tables) and Aladin (for inspecting the corresponding images) – are integrated through VO standards (right)

When inspecting the projections, a clear sequence in structural similarities is visible. By overplotting in different colours the spectral classes of stars, as taken from the CDS SIMBAD astronomical object database, a correlation between the position in the 2D latent space plane and spectral classes, i.e.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

the stellar effective temperatures, can be found (see Fig. 6). This is very interesting, because it relates the results of the network learning to a physical property of the objects that users are potentially interested in as constraint in archive searches.

The prototype is connected to VO tools (Aladin, Topcat, Splat), in order to interact with the original data and to get additional information about the spectral sources. Additional features are available, such as the possibility of visualising projections for all spectra of a particular source in order to retrain and update the model in real time, to select particular subsets for export or exchange, and to interactively further refine the selection. It is also possible to import additional spectra, to calculate their projections and their reconstruction, and to generate catalogues with the spectra's closest neighbours in the projected space. In addition to the autoencoder model, the prototype provides various other dimensionality-reduction methods, such as Principal Component Analysis (PCA), Convolutional Autoencoder (CAE) and Gaussian Process Latent Variable Model (GPLVM).

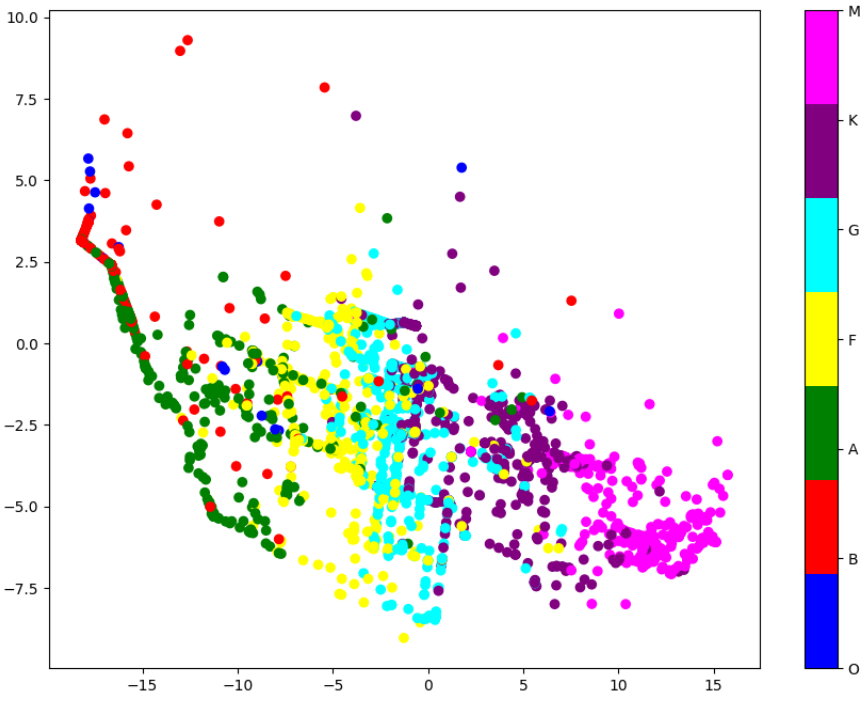

Figure 6. 2D latent space plane and spectral classes

Prototypes at ESO

Latent space similarity

The full details of our results are presented in Sedaghat, Romaniello, Carrick and Pineau (2021, MNRAS, 501, 6026; journal, arXiv). Here, we summarise the most important aspects.

As detailed above, the dimensionality of the latent space of the model was set to 128 to reach a satisfactory level of both reconstruction of the input signal and of disentanglement. This is, then, the maximum amount of information that the network is made capable of preserving, i.e. learning. The first notable result is that the network concentrates the meaningful information on only few of the latent space dimensions. How many exactly is mostly driven by one single parameter, namely how much disentanglement is enforced. Lower degrees of disentanglement result in many significant dimensions. They are, however, not informative, as disentanglement is not effective, leaving significant crosstalk among different dimensions. Increasing the level of disentanglement, on the other hand, results in fewer significant dimensions, but at the expense of losing reconstruction quality to the point that only the overall shape of the spectra is eventually learned. This trade-off is a well-studied characteristic of unsupervised disentanglement methods (e.g. Burgess et al 2018, arXiv:1804.03599). In our case, we find that a disentanglement parameter $\lambda = 0.3$ (see equation 5 of Sedaghat et al. 2021) provides such good trade-off, yielding to 6 significant dimensions, with no significant mutual correlation, i.e. good disentanglement.

In order to make the results useful to our intended audience of astronomers querying science archives, we proceeded to explore to what extent the significant dimensions can be interpreted in terms of physical parameters of the source (radial velocity, effective temperature, surface gravity and metallicity), or of the circumstances under which they were observed (airmass and signal-to-noise ratio, SNR). This is shown in Figure 7, where we show the correlation of these parameters with latent space dimensions for the 128-dimension network.

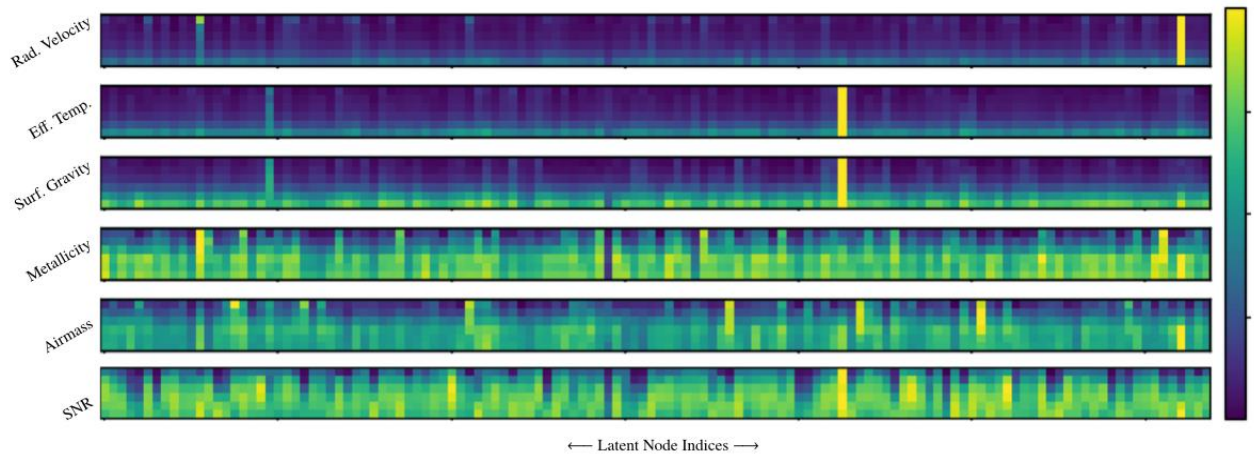

Figure 7. Correlation of physical and observational parameters with latent space dimensions for a 128-dimension network. The correlations are identified by seeking mutual information between the latent nodes and astrophysical validation labels. For radial velocity, effective temperature, and surface gravity, individual nodes stand out, while for metallicity, airmass, and SNR, that is not the case.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

As it can be seen, the information of radial velocity, effective temperature and surface gravity is effectively concentrated on well-defined individual nodes. In this sense, the network has learned by itself that these are significant parameters that, in addition and crucially, are of interest for an astrophysical interpretation of the output of the network. Information on surface gravity and temperature is mostly concentrated in the same node, indicating incomplete disentaglement. No such a correlation is found in the case of metallicity, airmass and SNR. This is especially puzzling in the case of metallicity, which leaves a clear imprint on the spectra in terms of absorption lines. It could, therefore, reasonably be expected to be picked up by the network at a similar level of significance as the effective temperature. One possibility may be that the labels we have used to tag the spectra in terms of metallicity are not accurate enough. Improving on such labels, which we took from the compilation in the SIMBAD database, can be attempted, e.g., by deriving them directly through a thorough analysis of the spectra themselves, or by extensively simulating the data over a wide range of parameters. In both cases, the amount of effort required exceeds the scope of our ESCAPE resources, so they could not be pursued at this stage.

In order to highlight the significance of our Deep Learning results, we have conducted a Principal Component Analysis (PCA) on the same dataset and with the same number of dimensions (128). The results are summarised in Figure 8. As it can be seen, in this case, and as opposed to the case of the Deep Learning network, no clear traces of individual parameters can be seen. In other words, the information about each physical parameter is spread over many dimensions. In this sense, even though the PCA is effective in reconstructing the input signal, it does so without learning any of its features. It, then, cannot be used for our stated goal of distilling information to be used in archive queries.

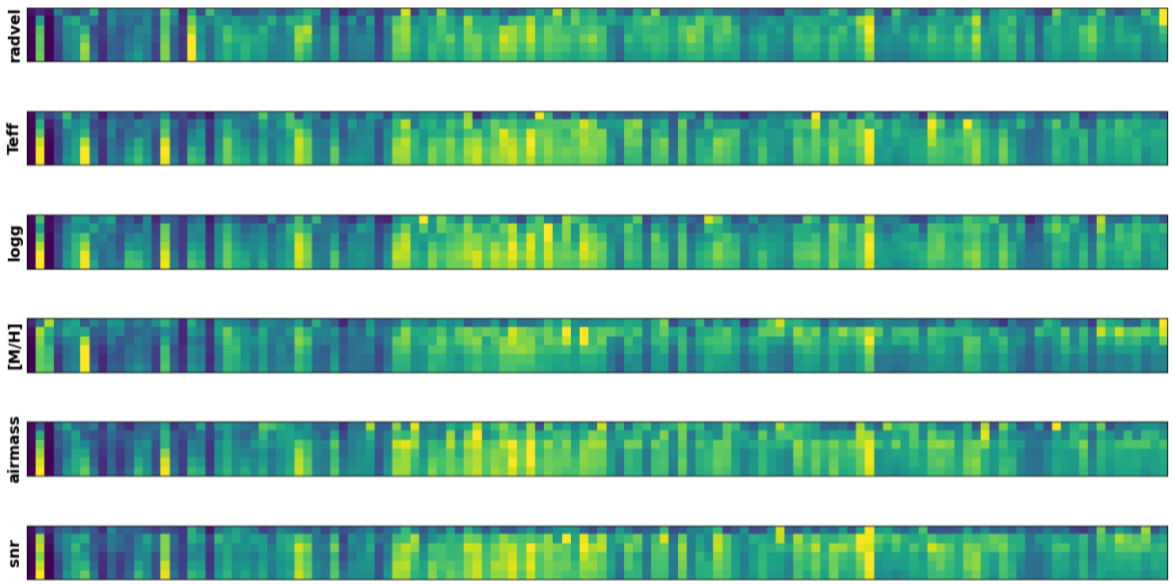

Figure 8. Same as Figure 7, but for a PCA analysis. In this case, no significant concentration of information along any of the significant dimensions is observed.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

Rejection of atmospheric telluric lines

While designed for exploring similarity searches, the chosen network architecture has also serendipitously shown a tendency to reject telluric lines, while retaining the stellar features in the spectra. These are absorption lines imprinted onto the incoming spectrum by molecules in the atmosphere of the Earth, like water vapour, oxygen, ozone and carbon dioxide. While they are quite stable in wavelength, their relative intensities are highly variable in time, depending on the prevailing local conditions (atmospheric temperature and pressure profiles, abundance of the various species, etc.). Telluric lines are a nuisance to extracting the science content from the spectra and they need to be calibrated out. The traditional method to do so is to observe stars with a known intrinsic spectrum as close as possible in conditions to the science target. This is, however, rather intensive in terms of telescope time, a precious commodity. Also, its accuracy is limited by how close the conditions of the science spectra can be reproduced by the calibrating ones. To remedy these, an alternative method was developed based on an accurate modelling of the Earth atmosphere (*molecfit*, Smette et al 2059, A&A 576, A77; see also the page at ESO here). The method shows very good results, and is mainly limited by how accurately the modelling can be performed.

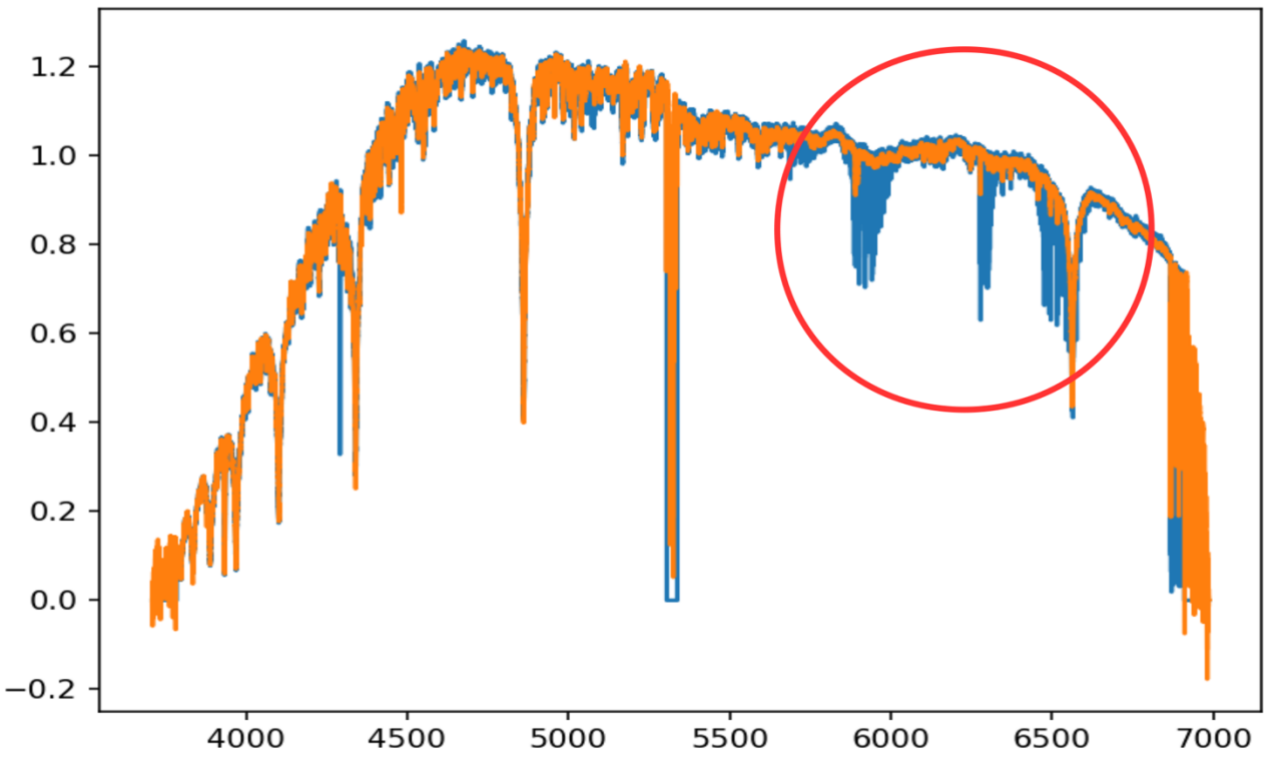

Figure 9. An exemplar of an input spectrum (blue line) and its reconstruction by our Deep Learning network architecture (orange line). The network was not specifically instructed to not reconstruct the telluric feature, which it had no previous knowledge of, but did so of its own accord. The wavelength in Angstroms in represented on the horizontal axis, and the flux in arbitrary units on the vertical one (the absolute flux scale in physical units is not meaningful for instruments like HARPS, i.e. fibre-fed spectrographs).

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

Within ESCAPE, we have started to pursue a completely different approach, based on noticing that the Deep Learning network have a tendency to first reconstruct the stellar features, while they have more difficulties to do the same with the telluric ones. They were not instructed to do so and, in fact, had no previous knowledge of the distinction between telluric and intrinsic stellar features. Still, under appropriate conditions, this effectively provides a way to reject them and provide clean intrinsic spectra (Figure 9).

In order to quantify this, we have computed with *molecfit* the best telluric spectrum for each of the observed spectra. We, then, used them as (pseudo-)ground truth, together with an appropriate metric, to characterize the quality of the rejection and correlate it with the characteristics of the network. Again because of the resource constraints within ESCAPE, we could not complete the analysis. Still, the early results are promising enough to try to proceed further under a different resourcing scheme. One of the aspects to be investigated is whether the learning on a specific dataset can be adapted to a different one, chiefly from a different instrument. If this were the case, existing historical data could be readily applied to new instruments, greatly enhancing the impact of the technique. This would be, for example, of great importance for ELT data, when they will be available, resulting in great savings in the exceedingly precious night-time observations.

Prototype visualisers

We have developed two prototype user services to visualise and interact with the results of our Deep Learning networks. They are both based on python's *matplotlib* plotting library and are exposed as web services via *voila*. One allows to graphically explore the latent space dimensions and how they affect the reconstructed spectra, so that users can form for themselves a first-hand direct impression as to what they mean. The other can be used to select spectra by similarity. Here, the user starts by providing an input spectrum they are interested in and along which latent space dimensions the similarity is to be assessed. The system then returns the closest spectra along the chosen dimensions, as well as a graphical representation of relevant statistical quantities. They are illustrated in Figures 10 and 11 below, respectively.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

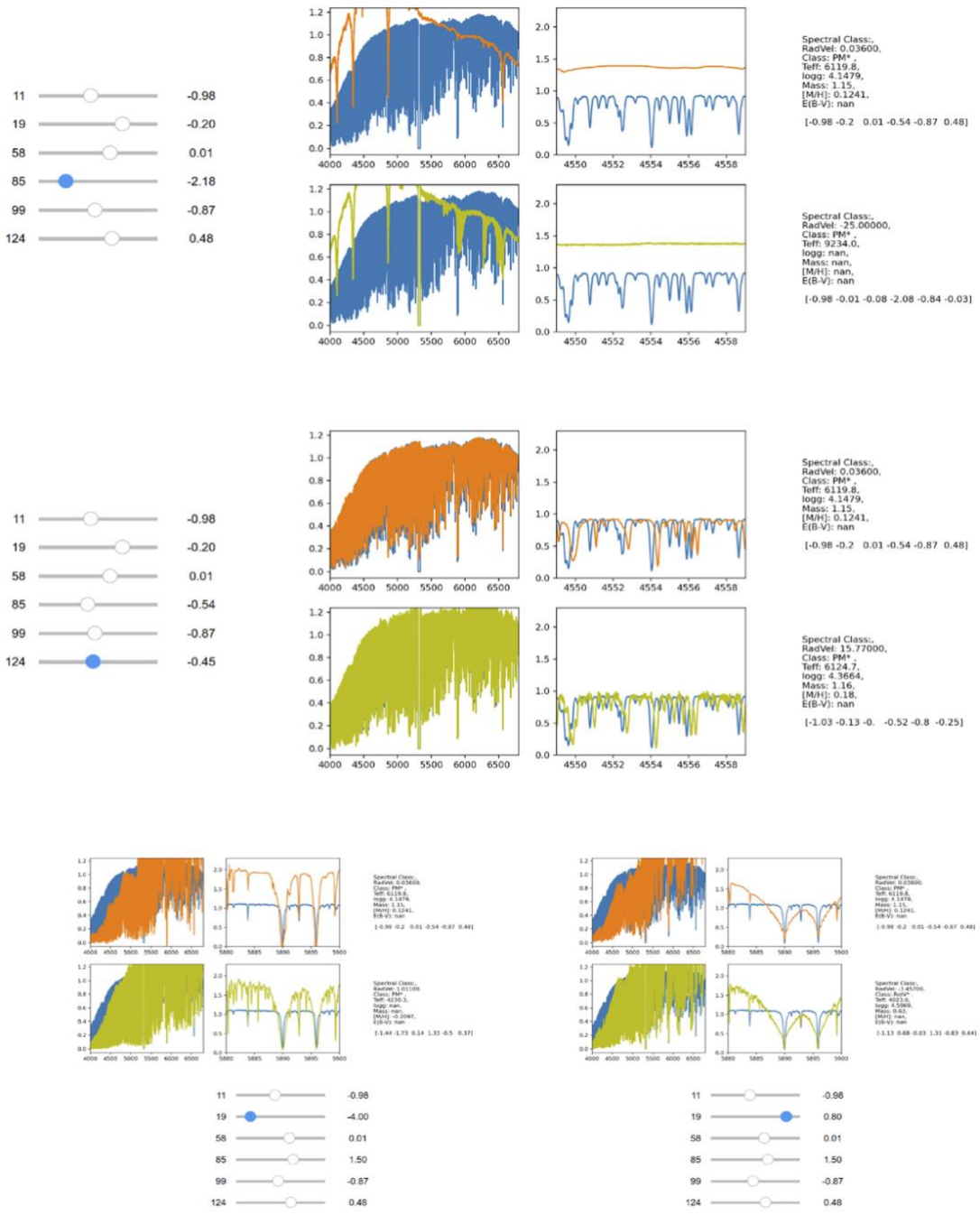

Figure 10. Screenshot of the prototype user service to explore the latent space defined by the Deep Learning network. By adjusting the sliders on the left-hand side, the input spectrum (blue lines in the different panels) is rendered according to the chosen values of the corresponding projections in the latent space. In this way, users can get a hands-on, direct impression of the meaning and effects on the different laten space dimensions. For example, latent dimension 85 controls the overall shape of the spectra continuum (top panel). Dimension 124, on the other hand, is sensitive to the position of the spectra features, i.e. the radial velocity of the star (middle panel). As illustrated in the bottom panel, latent dimension 19 encapsulates the width of the lines.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

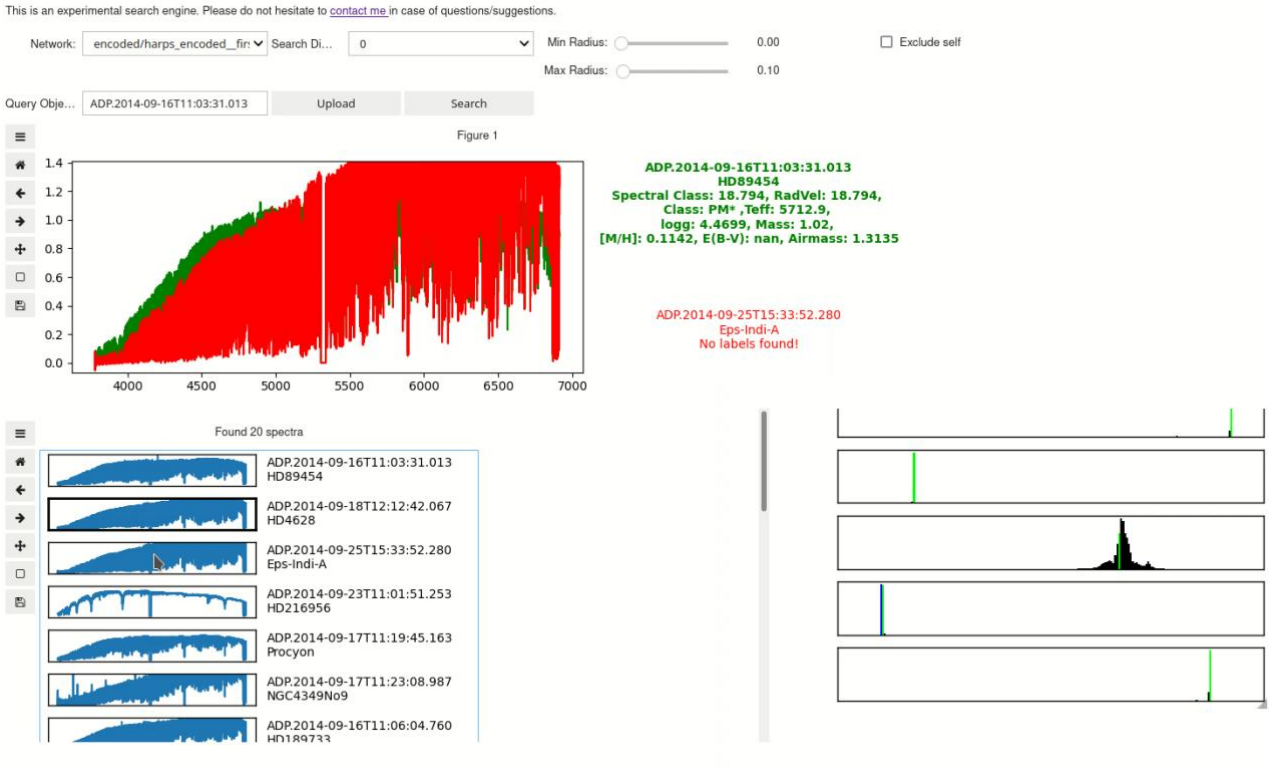

Figure 11. User front-end of the similarity search engine. Users input a spectrum as query template and the system returns the ones that are most similar to it along a user-specified latent space dimension. The meaning of these can, in turn, be explored with the tool illustrated in Figure 10.

5. Demonstrations

The results and associated prototypes have been presented to various communities over the course of the ESCAPE project. The goal has been to show the progress of the developments and to raise awareness of the new approaches being taken for the application of unsupervised Deep Learning in the context of archival data services. These presentations have included demonstrations of the prototype tools, and we have used the scientific results obtained to highlight the potential of these new kinds of services.

We have been able to reach a range of audiences including astrophysics researchers, data scientists, science archive developers, scientific software developers at Astronomy community conferences such as the international 'Astronomical Data Analysis Software and Systems' (ADASS) conference in 2021 and the 2020 conference of the European Astronomical Society. A presentation made at the 'EIROforum workshop on Big Data' in 2020 reached a wider audience of research infrastructures, space agencies and computing related projects in the areas of physics, astronomy and biology.

Demonstrations have been made within ESCAPE events organised by WP4 and WP3. These have included participation by all of the ESFRI that are involved in WP4, and these events have been used to disseminate the results to other ESCAPE partners working on machine learning related activities. These events have been used to discuss and align the work with other ESCAPE partners, so that it benefits from a community-coordination effort. The table below lists the demonstrations that have been made. For each event we include links to the web pages and schedule, and where possible we provide links to the recordings and presentation materials. We also note the number of registered participants.

The content of the different presentations and demonstrations have been adapted to suit the various audiences. Since Deep Learning is a relatively new, and rapidly evolving field, we have provided introductory material to explain the context of the ESCAPE work. In particular, we explain that the prototypes are designed for use within an archive environment where the data products of a research infrastructure are made available to a wide range of science users. Brief introductory explanations of Deep Learning are provided, followed by a description of the ESO archive data sets chosen for the demonstrations. Details of the Deep Learning networks, and their implementations and operations are adapted for the different audiences, with more or less scientific and technical specification provided as needed.

The emphasis of many presentations is on the demonstrations of the interactive interfaces that permit exploration of the results of the Deep Learning. One example is shown in Figure 5 where the MEGAVIS prototype is shown as a suite of interoperable desktop applications.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

Demonstration Title	Event (audience)
The universe speaks for itself: from unsupervised physics to semantic source separation. Martino Romaniello (ESO)	ADASS 2021 (https://www.adass2021.ac.za/) Hybrid (on-line/in-person) meeting hosted in Cape Town, South Africa 24-28 October 2021 Abstract, recording 318 participants
MEGAVIS: real-time spectra analysis and visualization with autoencoders. Antonio D'Isanto (HITS)	ESCAPE - Innovative Workflows in Astronomy & Particle Physics workshop (IWAPP). (https://indico.in2p3.fr/event/20424/) On-line, 9 March 2021 Recording: https://youtu.be/_n1gO_vNU68 72 participants
Letting the data speak for itself. Nima Sedaghat (ESO)	EIROforum workshop on Big Data https://indico.cern.ch/event/881752/ Online, 26-29 October 2020 Presentation: Link 310 participants
MEGAVIS - Real-time spectra analysis and visualization with autoencoders Antonio D'Isanto (HITS)	Astronomische Gesselschaft (German Astronomical Society) meeting, On-line, 22 September, 2020
Retr-Spect: A Deep Information Retrieval Tool for Spectra based on Convolutional Auto-Encoders Nima Sedaghat (ESO)	European Astronomical Society Annual Meeting https://eas.unige.ch/EAS2020/ Online, June 29-July 3, 2020 Abstract
Real-time spectra analysis and visualization with autoencoders. Antonio D'Isanto (HITS) ***Deep Spectral Exploration.*** Nima Sedaghat (ESO)	ESCAPE - Technology Forum 1 https://indico.in2p3.fr/event/20005/ Strasbourg, 4-6 February 2020 Presentations: 1- Link, 2 - Link 38 participants

Data-driven astronomy, Machine Learning versus Deep Learning Nima Sedaghat (ESO)	ESO Knowledge Exchange, September 2019
ESCAPE to victory: building the infrastructure for next generation astronomy Antonio D'Isanto (HITS)	AstroInformatics conference (http://astroinformatics2019.org) Presentation Pasadena, USA, 24-27 June 2019
Is the VO Ready for Machine Learning? Antonio D'Isanto (HITS) (Comment*: An early presentation about the challenges at the beginning of the project*)	IVOA Interoperability Meeting (ESCAPE Milestone MS20) (https://wiki.ivoa.net/twiki/bin/view/IVOA/InterOpMay2019) Paris, 12-17 May 2019 Presentation

6. Discussion

The exploratory work we have carried out has shown that there is considerable potential of unsupervised machine learning techniques being applied to the science archives of ESFRI and other infrastructures that serve a broad user base. The network architectures we have tested showed the ability to reconstruct the input spectra through encoding/decoding, both at the coarse level and in fine details. More importantly, in doing so it condenses relevant information in just a few dimensions in the latent space, several of which can be readily interpreted in terms of physical parameters of the celestial sources. Pushing for disentanglement enhances the physical interpretability of the results, while utilizing oversimplified models enabled an on-the-fly model training and adaptation. These are very important findings in order to be able to present the results and build services on top of them in terms of concepts that are familiar to the intended audience of professional astronomers.

Nevertheless, turning these early results and prototypes into production-grade user services will require significant resources, both at the conceptual and implementation levels. For example, a more thorough exploration is needed to identify the meaning of the latent dimensions that escaped classification in this early phase. Also, this far, we have only worked with spectra of stellar objects. Exposing the networks to different celestial objects, e.g. galaxies of different types, and data from different instruments will be needed to understand and characterise how they behave across the entire variety of data that are found in the ESO archives. Experimenting with domain adaptation, whereby results are transferred from one dataset to a different one, is also required.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

Interaction with users in terms of interfaces will also need to be further explored. We have demonstrated the basic functionalities, but could not devote specific effort to real usability, feature completeness and ergonomic development. All of these are, of course, fundamental to ensure significant and long-lasting user acceptance. This is even more needed in this case that deals with the exploration of complex multi-dimensional manifolds in which the similarities are defined.

In addition to exploring the similarity of spectra, we also started to explore the similarity of images. While first promising results have been obtained, a lot of exploration work will be needed. One challenge is to verify that the network still finds two images being similar. Likewise to spectra where invariances against redshift or object intrinsic luminosities might be important, rotation and flipping invariance are important for images. Standard variational autoencoders may not very well suited for this particular task.

Specific challenges

The technical execution of this task within the ESCAPE project involved dealing with the COVID-19 pandemic, which prevented in-person interactions of the partners involved in the tasks, and it also heavily limited interaction of people at the same institute. The nature of the project involved many exploratory investigations and uses of new techniques and technologies, which are much more easily accomplished with in-person interactions that can favour creativity and fast turn-around of ideas and experiments. This was managed via the use of videoconferencing, but we must acknowledge that the limitations of this way of interacting have had an impact on the task.

7. Conclusions

The work carried out within ESCAPE was intentionally focussed on very specific goals on how to enhance the discoverability of data in ESFRI science archives. The cooperation between CDS, ESO and HITS produced very encouraging results. As a consequence, it was instrumental in raising the awareness at an institutional level on the potentials of Artificial Intelligence approaches in the relevant domains, paving the way for what will be coming next. It was the first time that these advanced data processing techniques were studied and applied at ESO for the benefit of the Science Archive. Our findings on telluric line removal open an interesting window of opportunity into improving observatory operations and efficiency.

These developments and explorations have been well suited to the scope of the ESCAPE project. ESCAPE was instrumental in showing the basic feasibility of the stated goals, thus taking a first, very significant step towards eventually reaching them. It would be of great benefit to pursue further developments based on the results obtained in ESCAPE and accelerate the application of new techniques into ESFRI archives to enable deeper levels of Open Science.

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.