



# Innovative Scientific Data Exploration and Exploitation Applications for Space Sciences

---

User manual for SDA S-Disco

(D4.6)





## Info Sheet

Project	Innovative Scientific Data Exploration and Exploitation Applications for Space Sciences
Acronym	EXPLORE
Grant Agreement	101004214
Webpage	explore-platform.eu
Work Package	4
Work Package Leader	UNIMAN
Delivery Title	User Manual for SDA S-Disco
Delivery Number	4.6
Delivery Type	Report
Version	1.0
Date of Issue	28/07/2023
Dissemination Level	Public
Lead Beneficiary	ACRI-ST
Partner Contribution	KNOW

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004214.

The dissemination of results herein reflects only the author's view, and the European Commission is not responsible for any use that may be made of the information it contains.

The information contained in this report is subject to change without notice and should not be construed as a commitment by any members of the EXPLORE Consortium. The information is provided without any warranty of any kind.

© COPYRIGHT 2022-2023 The EXPLORE Consortium. All rights reserved.





## List of Authors

Name	Partner
Nick Cox	ACRI-ST
Manuela Rauch	KNOW
Emmanuel Bernhard	ACRI-ST

## Document Revision History

Version	Author	Reason of change	Section(s)
1.0	ACRI-ST, KNOW	First version	All





## Table of Contents

<b>1. Introduction .....</b>	<b>5</b>
1.1. Purpose and Scope .....	5
1.2. Applicable and Reference Documents.....	5
<b>2. EXPLORE scientific data applications .....</b>	<b>7</b>
<b>3. S-Disco User Manual.....</b>	<b>8</b>
3.1. Introduction .....	8
3.2. User Interface (UI) overview .....	8
3.1. User Interface (UI) details.....	10





# 1. Introduction

## 1.1. Purpose and Scope

This document is the User Manual for the S-Disco scientific data application developed by the EXPLORE project.

The document is structured as follows:

- Chapter 1 – Introduction (this chapter)
- Chapter 2 – EXPLORE scientific data applications
- Chapter 3 – S-Disco User Manual

## 1.2. Applicable and Reference Documents

### 1.2.1. Applicable Documents

Table 1: Applicable documents

Title	Description
[AD-1]	Grant Agreement 101004214 — EXPLORE
[AD-2]	Consortium Agreement – EXPLORE
[AD-3]	Use Case Analysis and User Requirements Study (D2.1) – EXPLORE
[AD-4]	SDA framework definition and EXPLORE-TEP User Manual (D2.2) – EXPLORE
[AD-5]	SDA S-Disco (D4.4) – EXPLORE

### 1.2.2. Reference Documents

Table 2: Reference documents

Title	Description
[RD-1]	Baron & Poznanski, 2017, The weirdest SDSS galaxies: results from an outlier detection algorithm, MNRAS, 465, 4530
[RD-2]	Reis et al. 2018, Detecting outliers and learning complex structures with large spectroscopic surveys - a case study with APOGEE stars, MNRAS, 476, 2117
[RD-3]	McInnes et al., 2018, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, arXiv:1802.03426
[RD-4]	Reis et al., 2021, Effectively using unsupervised machine learning in next generation astronomical surveys, Astronomy and Computing, 34
[RD-5]	Majewski et al. 2016, The Apache Point Observatory Galactic Evolution Experiment (APOGEE) and its successor, APOGEE-2, Astronomische Nachrichten, 337, 863





[RD-6]	Gaia Collaboration. Vallenari et al. 2023, Gaia Data Release 3. Summary of the content and survey properties, A&A, 674, A1 (arXiv:2208.00211)
[RD-7]	HoloViews, version 1.17.0, doi: 10.5281/zenodo.8184918 ( <a href="https://github.com/holoviz/holoviews/tree/v1.17.0">https://github.com/holoviz/holoviews/tree/v1.17.0</a> )
[RD-8]	Gaia Collaboration, Recio-Blanco et al., 2023, Gaia Data Release 3. Analysis of RVS spectra using the General Stellar Parametriser from spectroscopy, A&A, 674, A29 (arXiv:2206.05541)

### 1.2.3. Abbreviations and Acronyms

Table 3: Abbreviations and acronyms list

AI	Artificial Intelligence
DR3	Data Release 3
ESA	European Space Agency
ML	Machine Learning
RVS	Radial Velocity Spectrograph
SDA	Scientific Data Application
UI	User Interface
WP	Work Package





## 2. EXPLORE scientific data applications

EXPLORE's main objective is to deploy machine learning (ML) and advanced visualization tools to achieve efficient, user-friendly, realistic exploitation of scientific data from astrophysics and planetary space missions, as well as from supporting ground-based massive surveys. We focus on six different topics, each chosen for their timely importance and their complementary data structures. This diversity and complementarity is key to a future evolution and growth of the platform that is relevant and applicable to the broadest possible user-base within the research community.

Two of EXPLORE's topics are related to Lunar observation, two to Galactic Science and two to stellar characterization. For each of these topics, the state-of-the-art will be enhanced by introducing ML techniques and advanced visualization tools. For each topic, specific tools are created. These Scientific Data Application (or simply Apps) are developed on a dedicated cloud solution (the EXPLORE platform, <https://explore-platform.eu>). The EXPLORE Apps are also made available on other cloud platforms such as ESA Datalabs and open to the community for direct exploitation-on-demand. The apps are released as open-source software.

The EXPLORE Apps will also be used by the consortium to produce enhanced scientific datasets for space science mission exploitation, which will be stored in appropriate archives for public access. Datasets from Gaia and recent lunar (LRO, Clementine, Chandrayaan, etc) space missions are at the core of the EXPLORE project and are complemented with data from previous space missions as well as ground-based surveys.





## 3. S-Disco User Manual

### 3.1. Introduction

S-Disco provides new insights into the content of large spectral databases through the application of novel machine learning algorithms ([RD-1], [RD-2]). The main purpose of S-Disco is to provide new representations (we use the UMAP dimensionality reduction technique; [RD-3]) of large datasets allowing researchers to explore them effectively. In particular, S-Disco provides “weirdness scores” for all spectra helping to identify those sources which are peculiar from the bulk ([RD-4]). Different visualisations are provided to search for special stellar targets or find similar stars in some of the largest spectral stellar datasets to date (APOGEE, Gaia).

S-Disco is available in two flavours, which exist currently in two version of S-Disco, S-Disco (Apogee) and S-Disco (Gaia), which can be started from [https://explore-platform.eu/sda/s-disco\\_apogee](https://explore-platform.eu/sda/s-disco_apogee) and [https://explore-platform.eu/sda/s-disco\\_gaia](https://explore-platform.eu/sda/s-disco_gaia), respectively (login required). In the future these two flavours will be merged into a new version wherein the user can switch between flavours.

The first flavour uses ~75,000 APOGEE ([RD-5]) spectra complemented with APOGEE stellar parameters and Gaia astrometric data.

The second flavour is using the entire public Gaia DR3 ([RD-6]) RVS spectral database, which contains about 1 million stellar spectra. It is complemented with the set of stellar parameters inferred for (most of) the stars ([RD-8]).

This user manual focusses on the “Gaia” flavour of S-Disco (to which the “Apogee” flavour, a proto-type development, is being integrated).

### 3.2. User Interface (UI) overview

The full S-Disco dashboard is shown in Figure 1. The dashboard consists of six panels (areas):

1. Information pane
2. Control panel
3. Main data visualisation panel
4. Selection panel
5. Spectral plotting panel
6. *Visualiser* data analytics panel

Each panel, and their usage, are explained in more detail in the following sections. That said, the objective of S-Disco is to provide an intuitive interface that users can freely explore. This manual is therefore mostly a reference guide and record of available functionalities.

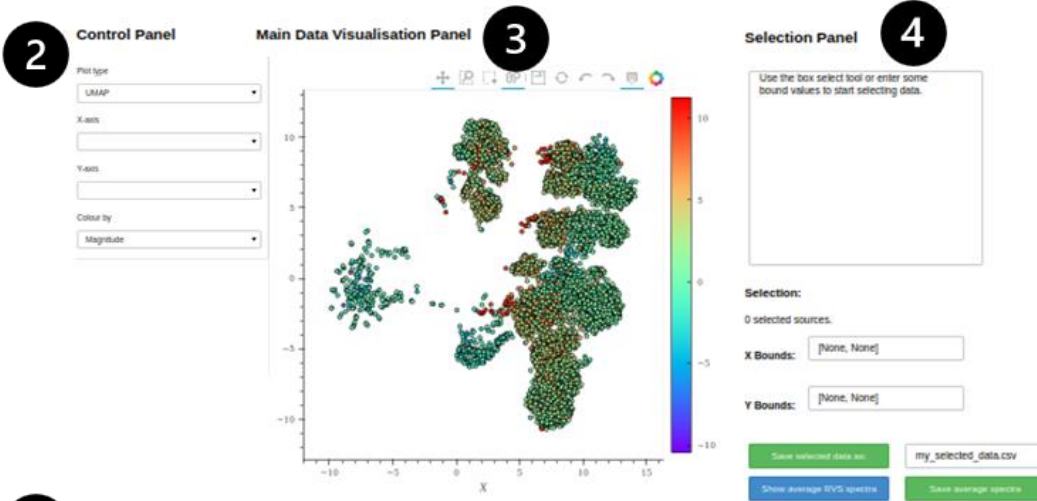




# 1 Dimensionality reduction and similarity mapping for Gaia-RVS spectral data

The Control Panel allows you to visualise the data in several well known parameter spaces using the Plot Type selection menu. You can otherwise select the desired X and Y axis to parameters listed in X-axis and Y-axis which correspond to the data columns. The Colour by list allows you to control the parameter used to colour the data.

The Selection Panel allows you to create a selection within the data using the X bounds and the Y Bounds parameters. You can directly type these into the corresponding text boxes, or use the box select tool available in the Main Data Visualisation Panel. This selection can be saved as a csv file, and the average RVS spectra of the selected data can also be displayed and saved.



## 5 Average RVS spectrum of selected sources

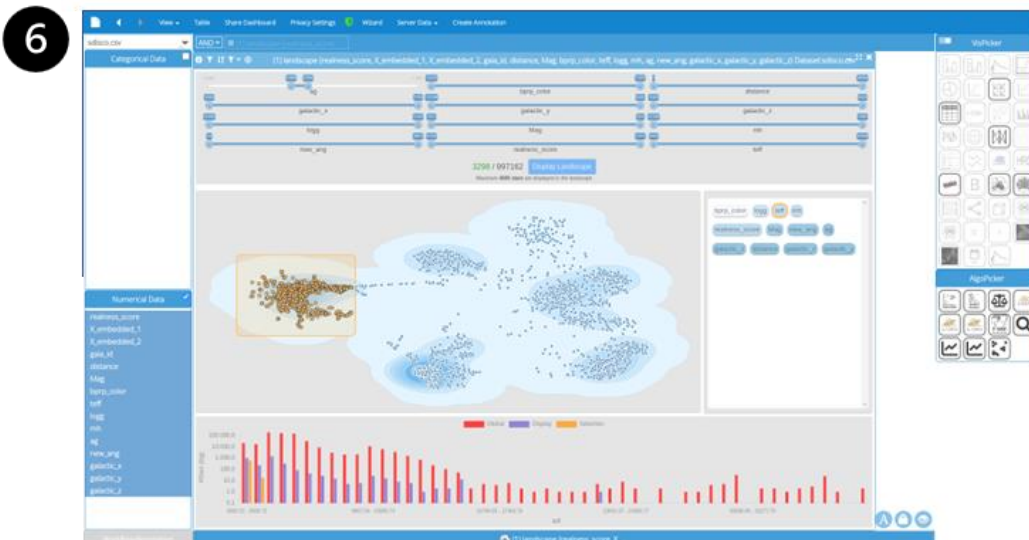
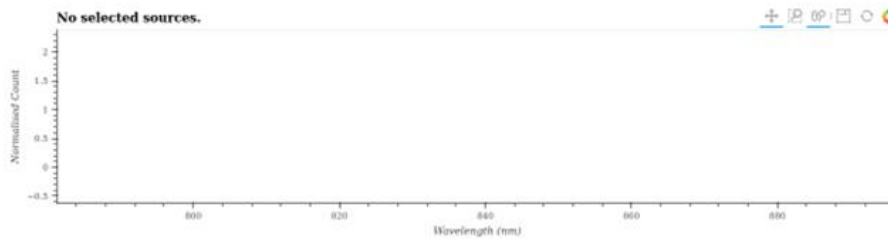


Figure 1. S-Disco full web user interface.



### 3.1. User Interface (UI) details

#### 3.1.1. Information panel

The top panel (❶) provides concise information on how to use different parts of the S-Disco app.

#### 3.1.2. Control panel

The control panel (❷) is the typical starting point for changing and updating the visualisations displayed in the main data visualisation panel (❸).

The user can select one of the pre-defined plot types (UMAP, Galactic side view, Galactic plane, or HR diagram) or create a custom plot selecting for the x-axis and y-axis any of the available data columns (Figure 2):

- X\_umap
- Y\_umap
- galactic\_x
- galactic\_y
- galactic\_z
- distance (derived from Gaia parallax)
- Mag (i.e. Gaia magnitude)
- bprp\_color (Gaia  $B_p-R_p$ )
- galactic\_azimuth
- teff (effective stellar temperature, inferred from Gaia RVS)
- realness\_score
- logg (stellar gravity)
- galactic\_x\_plane
- galactic\_y\_plane

Similarly, displayed data points can be coloured by selecting one of the following options:

- Magnitude
- BP-RP colour
- Distance
- Realness score
- Metallicity
- Effective temperature
- Log g
- None



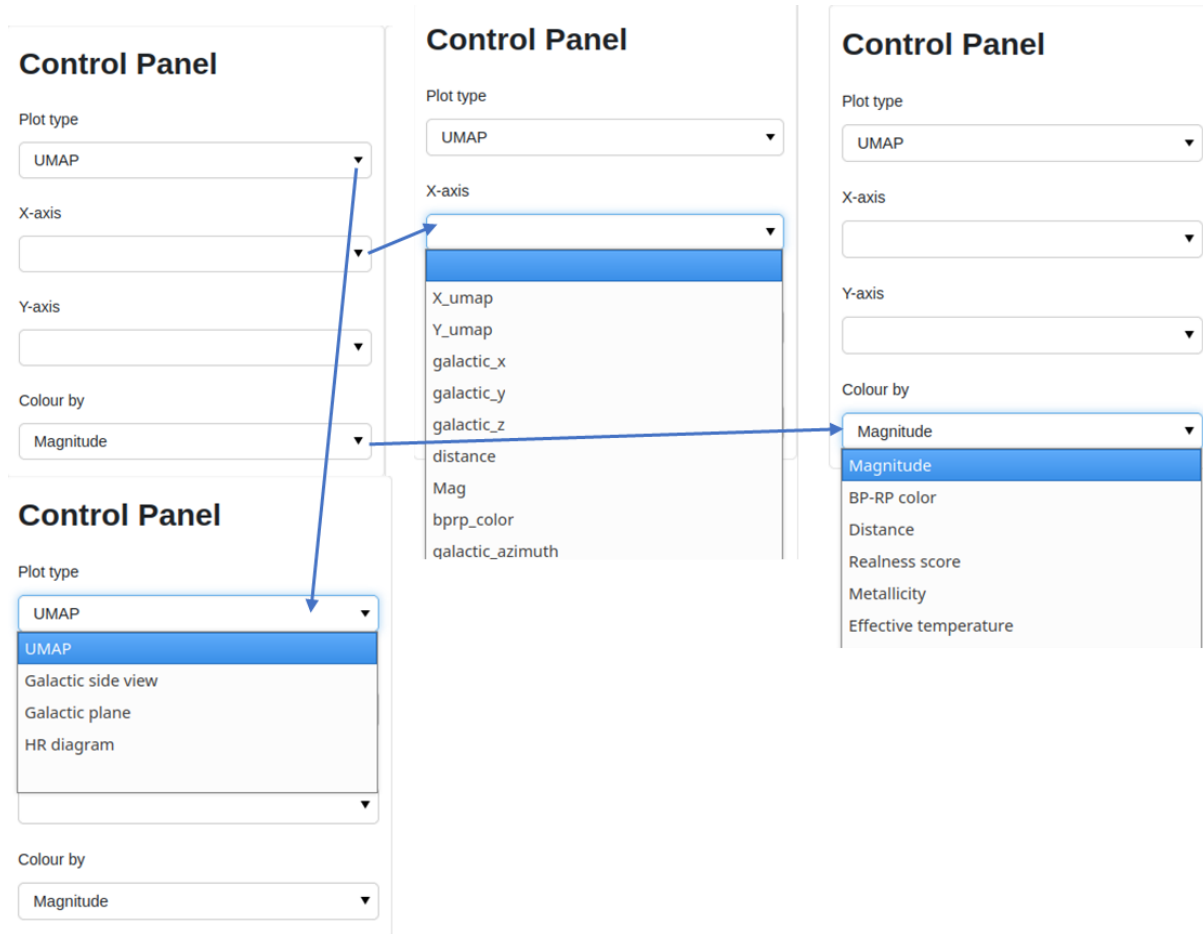



Figure 2. Control panel. The user can select 'plot type', 'x-axis', 'y-axis', and 'colour by' (colour coding of data points).

### 3.1.3. Main data visualisation panel

The main data visualisation panel (3) is the heart of S-Disco allowing to visualise and explore (zoom/select) the full dataset. Plotting 1 million datapoints all at once is difficult and would make the application very slow. We therefore use a data representation technique called Datashader as implemented by the HoloView python library [RD-7]. Using a decimation factor (of 8000 in the case of S-Disco) only a subset of (representative) data points are shown. Figure 3 shows an example of the Datashader visualisation as more data points are revealed upon zooming in using the "Box Zoom" function:  (in the plot tooltip bar).

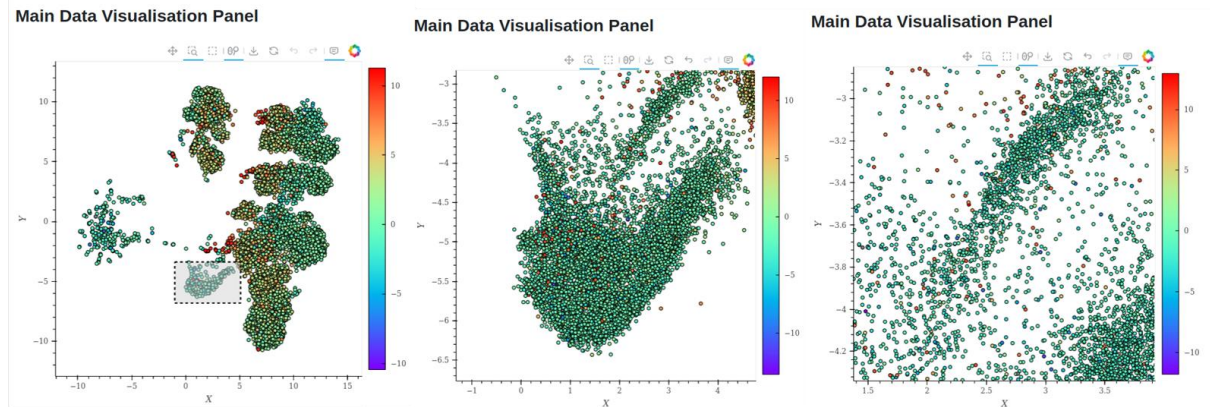


Figure 3. Main data visualisation panel. Three progressive levels of box zoom are shown with increasing levels of detail from left to right.

### 3.1.4. Selection panel

The Selection Panel (4) is tied to the Main data visualisation panel through the “Box

Select” function:  (in the plot tooltip bar). The user can use the Box Select to select any area on the plot. All objects in the selected area are listed in the Selection Panel.

*Warning: Even a small selection area can imply a large number of targets (and hence longer loading time) as not all targets are necessarily shown. It is prudent to use “Box Zoom” first to reveal targets in a smaller area.*

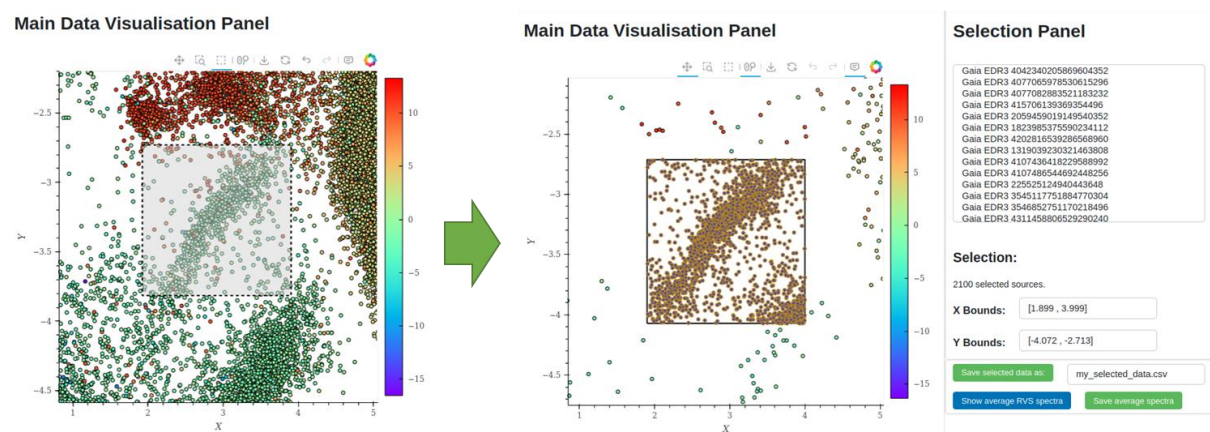


Figure 4. Selection of data points (Box Select). All objects (2100) in the selected region are listed in the Selection panel.

The selected objects can subsequently be saved to a CSV file, and their average RVS spectrum can be plotted (in the spectral plotting panel; 5) and saved. We stress that, by design, the average RVS spectrum needs first to be plotted, and then saved. In fact, plotting the average RVS spectra is also constructing the latter.

### 3.1.5. Spectral plotting panel

The spectral plotting panel “Average RVS spectrum of selected sources” shows the average spectrum (solid blue trace) computed from the sources selected (cf. Selection Panel). The 1-sigma uncertainties are also shown (light coloured trace).

*Warning: the more objects selected the longer it takes to compute the average spectrum. It is not recommended to select more than a few hundred sources.*

*Note: For the average RVS spectrum only spectra with SNR > 20 are used.*

Again, this is an interactive plot allowing the user to zoom in/out.

#### Average RVS spectrum of selected sources

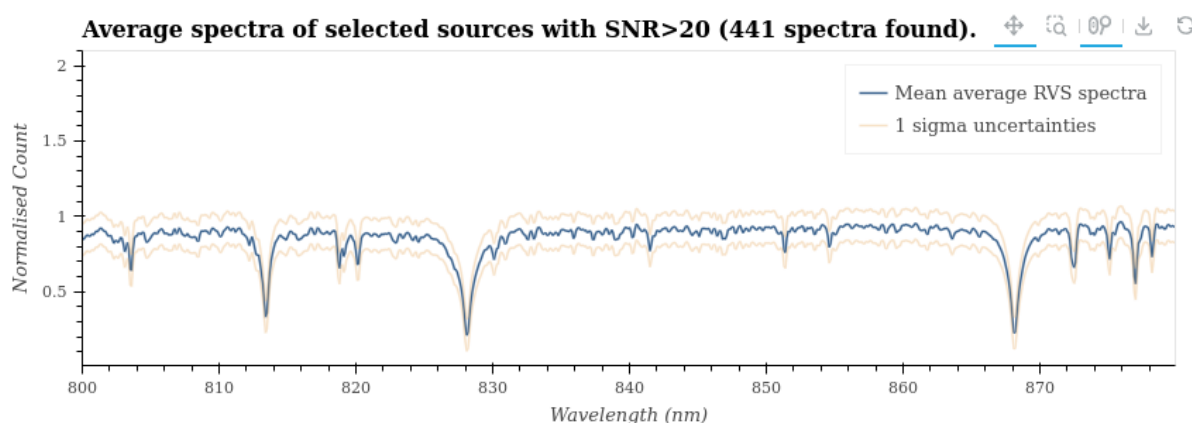


Figure 5. Plot of Average RVS spectrum of selected sources.

### 3.1.6. Visualiser data analytics panel

The Visualiser panel (⑥) provides an additional data analytics tool that can be used to better understand the intricacies of the constructed dataset. There are four main areas to this component (Figure 6):

- ① Data range controls for each of the parameters (max 4000 stars can be displayed).
- ② Visualisation landscape.
- ③ The lower area shows, for the selected parameter (cf. panel ④), the distributions of the selection, displayed and global dataset.
- ④ Data fields sorted according to their correlation coefficient.



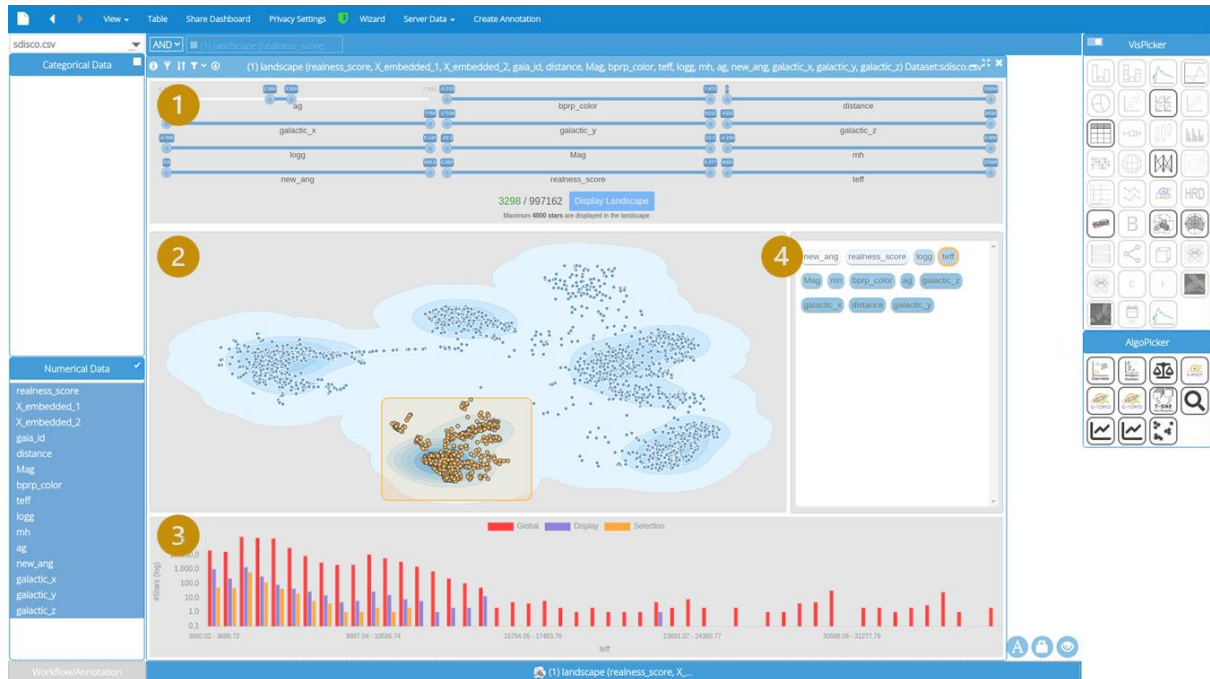


Figure 6. Visualiser data analytics.

The user can freely explore the full parameter space. Data ranges for each parameter can be adjusted. The selection can be updated. Changing the data range and/or the data point selection will automatically update the order of the parameters listed in panel 4. Lighter colours indicate stronger correlation, darker colours a weaker correlation.

Selected stars are shown in orange (middle and lower panel). The global distribution (~1 million sources) is shown as density map in panel 2 and as red bars in panel 3. The stars displayed as individual data points in panel 2 are shown as purple bars in panel 3.

Figure 7 shows the update of the parameter correlation order upon changing the selection of sources.

Figure 8 shows the update of the distribution of parameters for different parameter selections.

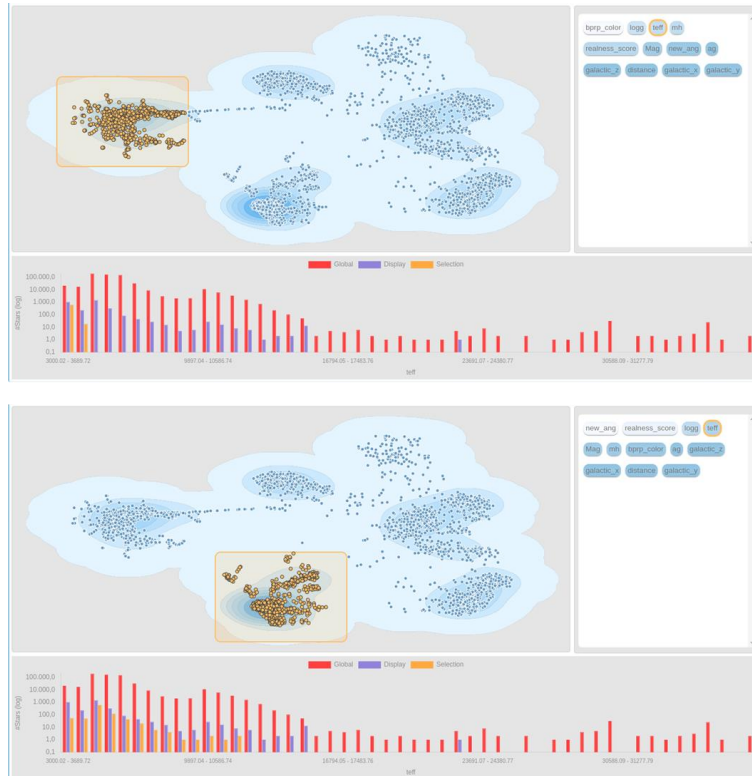


Figure 7. Visualisator analytics. Updating the source selection updates the correlation strength order of the parameters. In the top figure  $\log(g)$  and  $T_{\text{eff}}$  are more important than the “realness score”. For the selection in the bottom figure the latter is more important. Note that the ‘global’ and ‘display’ distributions remain unchanged. Only the ‘selection’ histogram is updated.

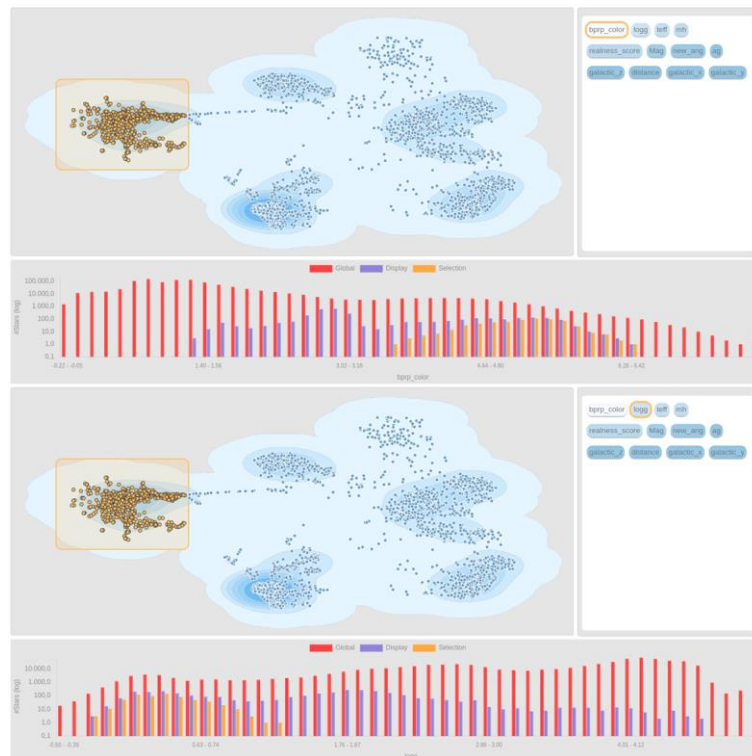


Figure 8. Visualisator analytics. Selecting a different parameter in the middle-right panel updates the source distributions in the lower area. In the top figure ‘bprp color’ is selected, whereas in the lower figure ‘logg’ is.