toxFlow v.0.1 Tutorial

Dimitra Danai Varsou
School of Chemical Engineering, National Technical University of Athens

September 13, 2016

Introduction

toxFlow is an application of web tools for Gene Set Variation Analysis (GSVA) and toxicity prediction using read across technique and it is released under GNU General Public Licesce. The application consists of three main parts, in three different tabs (as can be seen in Fig.1). The first two parts can be used independently, while the third part depends on model training, thus it cannot be performed prior to the second part:

1. **GSVA**: In this part, the user can employ the provided tools in order to perform Gene Set Variation Analysis [1] through the samples of an expression data set. Different omics data types could be analyzed (genomics, proteomis). Using GSVA analysis tools a group of all statistically significant gene set are presented in a table along with corresponding acyclic graphs and heatmaps, also providing links to pathway databases whenever possible.
2. **Read across training**: In this part training of a toxicity-prediction model is performed, via read across techniques [2] and leave-one-out cross-validation method [3]. Using read across technique a table that contains all NPs with a successful prediction for the toxicity index is presented, as well as correlation coefficient R^2 (as an index of successful prediction) and a diagram of NPs with their neighbours (NPs' universe).
3. **Read across prediction**: After model training, the user can predict the unknown value of toxicity indices of a nanoparticles' (NPs) data set. After prediction a table that contains the predicted value of toxicity index for all the NPs is presented along with the NPs' universe diagram.

1 GSVA

1.1 Import data

Before running the GSVA analysis, the user should import two files. The first file (**Biological data**) must contain the samples of an expression data set. The file must have a specific form, in order to be read properly: it must be a .csv file where the columns contain samples and the rows contain genes or proteins. Additionally, the first column must contain the names of genes or proteins and the first row must contain the names of the samples. The second file (**Data classification**) should be a .csv file with two columns. The first column (named «ID») must contain the names of the samples, while the second column(named «classification») the classification of the samples into categories, which will be used for the creation of the design matrix (Fig. 3A).

1.2 Adjust parameters of analysis

The supplied data set could be normalized (**Scaling of raw data**) according to the following equation:

$$c_{\mathrm{sc}} = \frac{c_{\mathrm{in}} - min}{max - min} \tag{1}$$

Where c_{in}, the value of the parameter before normalization,
min, the minimum value of the parameter in the set,
max, the maximum value of the parameter in the set and
c_{sc}, the normalized value of the parameter.

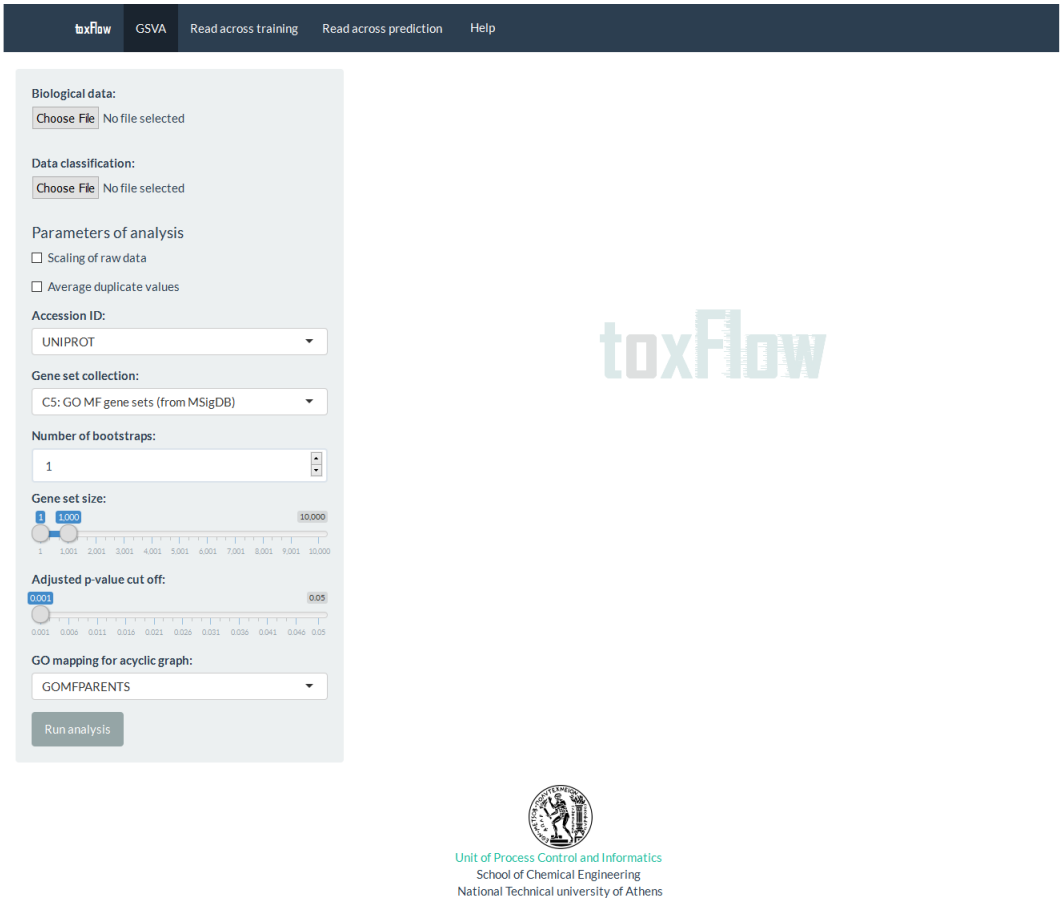

Figure 1: User intreface of toxFlow application. On the top ribbon are shown the three tabs (each for every main part).

If duplicate expression values (for the same gene or protein) exist, the user can choose to replace them by their mean value (by clicking on **Average duplicate values**). Furthermore, the user can select from a dropdown list the **Accession ID** of the gene or protein names (Uniprot, EntrezID, RefSeq or Symbol), as it can be seen in Fig.2.

In the section **Gene set collection** the user can select between the *C5: GO gene sets, MF: GO molecular function* (*GO-MF*) and *CTD Disease-GO molecular function associations* (*CTD-MF*) gene set collections. *GO-MF* is taken from MSigDB v5.1 and contains 396 gene sets in GO terms and *CTD-MF* is taken from Comparative Toxicogenomics Database and contains 3992 gene sets in GO IDs. The user can also import another gene set collection in order to perform the analysis. In this case, the file must be in .csv format with two columns. The first column must contain GOterms and the second the corresponding EntrezIDs (Fig. 3B).

Additionally, in **Number of bootstraps** field the user can choose the number of bootstrap iterations to be performed in GSVA function (default value is 1 bootstrap) and in **Gene set size** can control the minimum and maximum size of the resulting gene sets (default value is 1 and 1000 respectively). In **Adjusted p-value cut off** the user can control the threshold of adjusted p-value in order to select the significant of the gene sets that result from the linear analysis of the GSVA enrichment scores (Fig. 2). Finally the user in **GO mapping for acyclic**

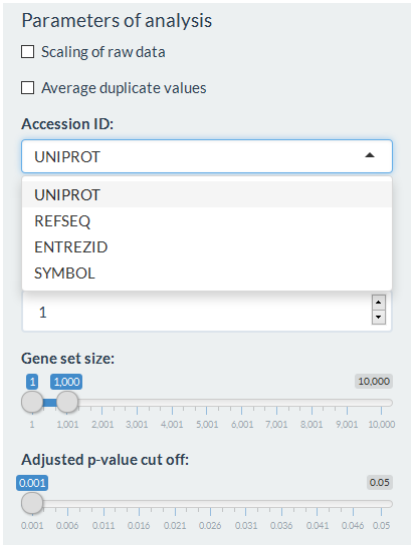

Figure 2: Parameters of analysis and Accession ID selection.

graph can choose between GOMFPARENTS and GOMFCHILDREN, as the main parameter of the acyclic graph that depicts the significant gene sets and the hierarchical relations with other gene sets of Gene Ontology (Fig. 3C).

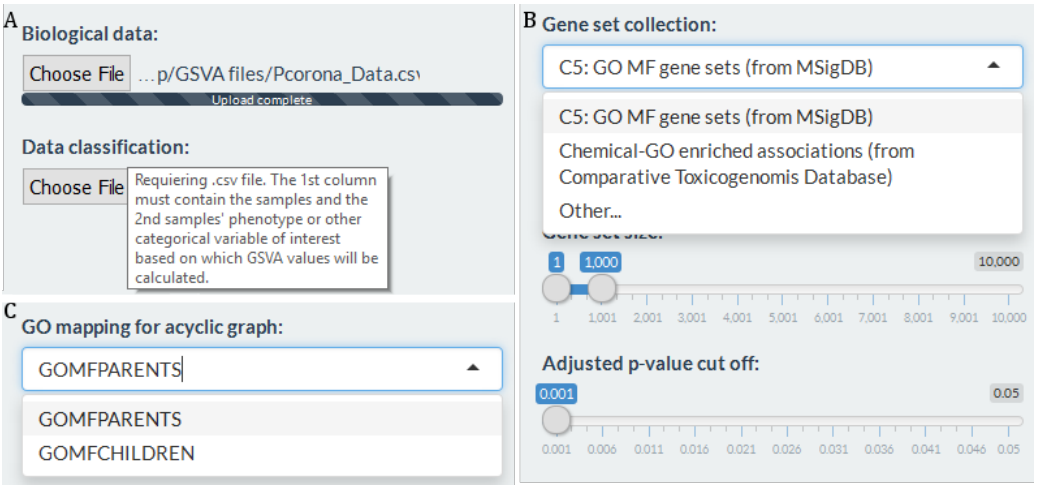

Figure 3: A: Import data, B: Gene set collection selection, C: GO mapping selection.

1.3 Results

By clicking **Run analysis** the application is performing the analysis, according to the parameters above and then exports a table that contains the significant gene sets, their GO ID, their size, the adjusted p-value based on Benjamini-Hochberg (BH) [4] multiple correction method of the linear model, and the counts (the number of genes in the initial data set that are found in the gene sets of the gene set collection). The user can download the table in the form of a .csv file. In addition the application produces a heatmap with the gene sets that result from GSVA and an acyclic graph (Fig. 6).

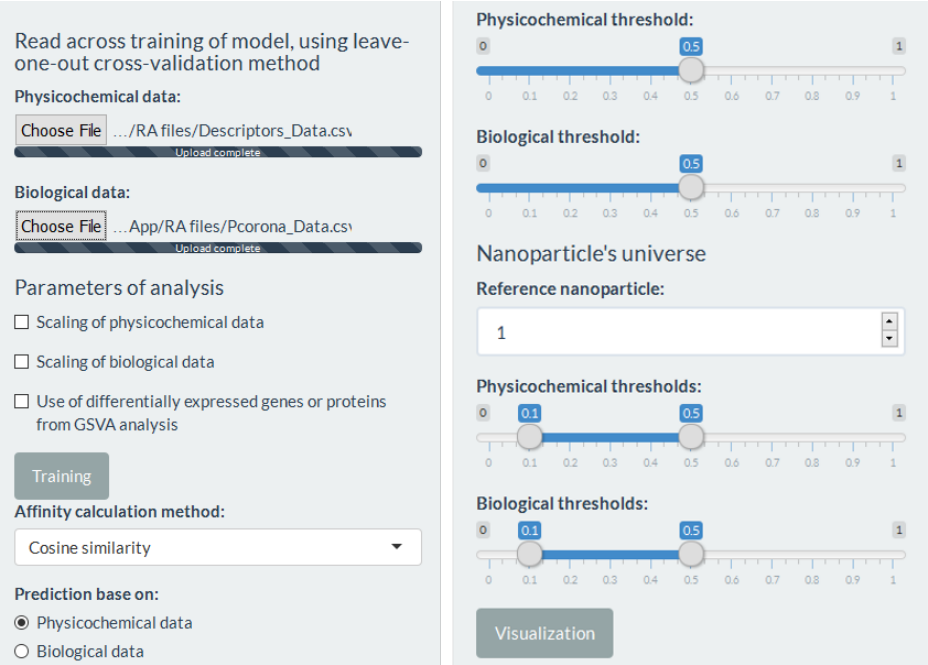

Figure 4: Read across model parameters

2 Read across model

2.1 Import data

The user must upload two .csv files in the application: The first one (**Physicochemical data**) must contain the values of physicochemical descriptors (samples in columns and descriptors in rows). The second one (**Biological data**) must contain the samples of an expression data set (samples in columns and genes or proteins in rows). Both files should include in the first row and in the first column the names of indices and samples. Also, both files must contain the values of the toxicity index, which will be predicted by the model, in the first row.

2.2 Adjust parameters of analysis

Furthermore the user can select if the physicochemical and expression data should be normalized (**Scaling of physicoshemical data**, **Scaling of biological data**) according to equation 1. Also the user can select whether only the significant genes or proteins (according to the data) from the GSVA analysis will be used. It is implied that in the previous section the user will have analysed the same biological data.

For the estimation of affinity between NPs, the user can choose from a dropdown list (**Affinity calculation method**) one of the following options: cosine similarity, Manhattan and Euclidean distance. In **Prediction base** the user can choose if the prediction will be calculated on a physicochemical or on biological base. Finally the user can define the physicoshemical and biological threshold that control the selection of neigbouring NPs from two sliders. By pressing the button **Training** the model training begins. The user can change the affinity method, the calculation base and the two thresholds and the results will be updated automatically.

Also the user can visualize the NPs «universe»: the user can choose a reference NP and observe its neighbors in color code, by adjusting the physicochemical and biological thresholds. Every time the user changes the reference NP and the thresholds, the user should press the button **Visualize**, in order to update the diagram. All parameters are shown in Figure 4.

2.3 Results

The analysis produces a table that contains all NPs with a successful prediction for the toxicity index, with the actual and the predicted value of this index. The user can download this table in the form of a .csv file. In addition the correlation coefficient R^2 is presented, as well as the NPs' universe diagram (Fig. 7).

3 Read across prediction

3.1 Import data

The last section of the application can be used after the model training. In that way, the toxicity index of a data set can be predicted, when all the physicochemical and biological indices that where used in training are known values. The user must upload two .csv files in the application: The first one (**Physicochemical data**) must contain the values of physicochemical descriptors (samples in columns and descriptors in rows). The second one (**Biological data**) must contain the samples of an expression data set (samples in columns and genes or proteins in rows). In both files, the first row and the first column must contain the names of indices and samples.

3.2 Adjust parameters of analysis

Furthermore the user can select if physicochemical and expression data should be normalized (**Scaling of physicoshemical data**, **Scaling of biological data**) according to equation 1. Additionally, the user should indicate whether the data should be filtered based on the significant data points (genes or proteins) given by GSVA analysis, if such an analysis has been performed. It is implied that in previous section user will have analysed the same biological data. The calculation of affinity method and the thresholds are the same with the training section. By clicking on **Prediction** begins the prediction process.

Also the user can visualize the NPs «universe»: the user can choose a reference NP and observe its neighbors in color code, by adjusting the physicochemical and biological thresholds. Every time the user changes the reference NP and the thresholds, the user should press the button **Visualize** in order to update the diagram. All parameters are shown in Figure 5.

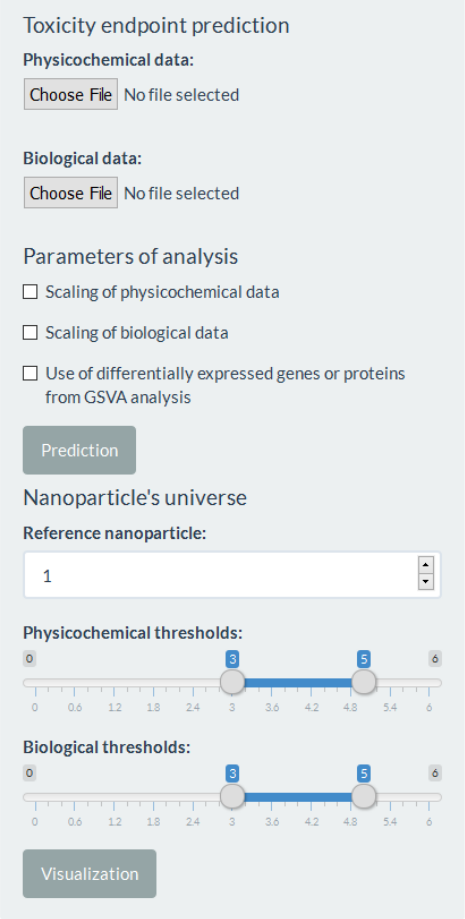

Figure 5: Parameters of prediction.

3.3 Results

The analysis produces a table that contains the predicted value of toxicity index for all the NPs, which the user can download in the form of a .csv file. In addition the NPs' universe diagram is presented (Fig 8).

Figure 6: Sample GSVA results' presentation

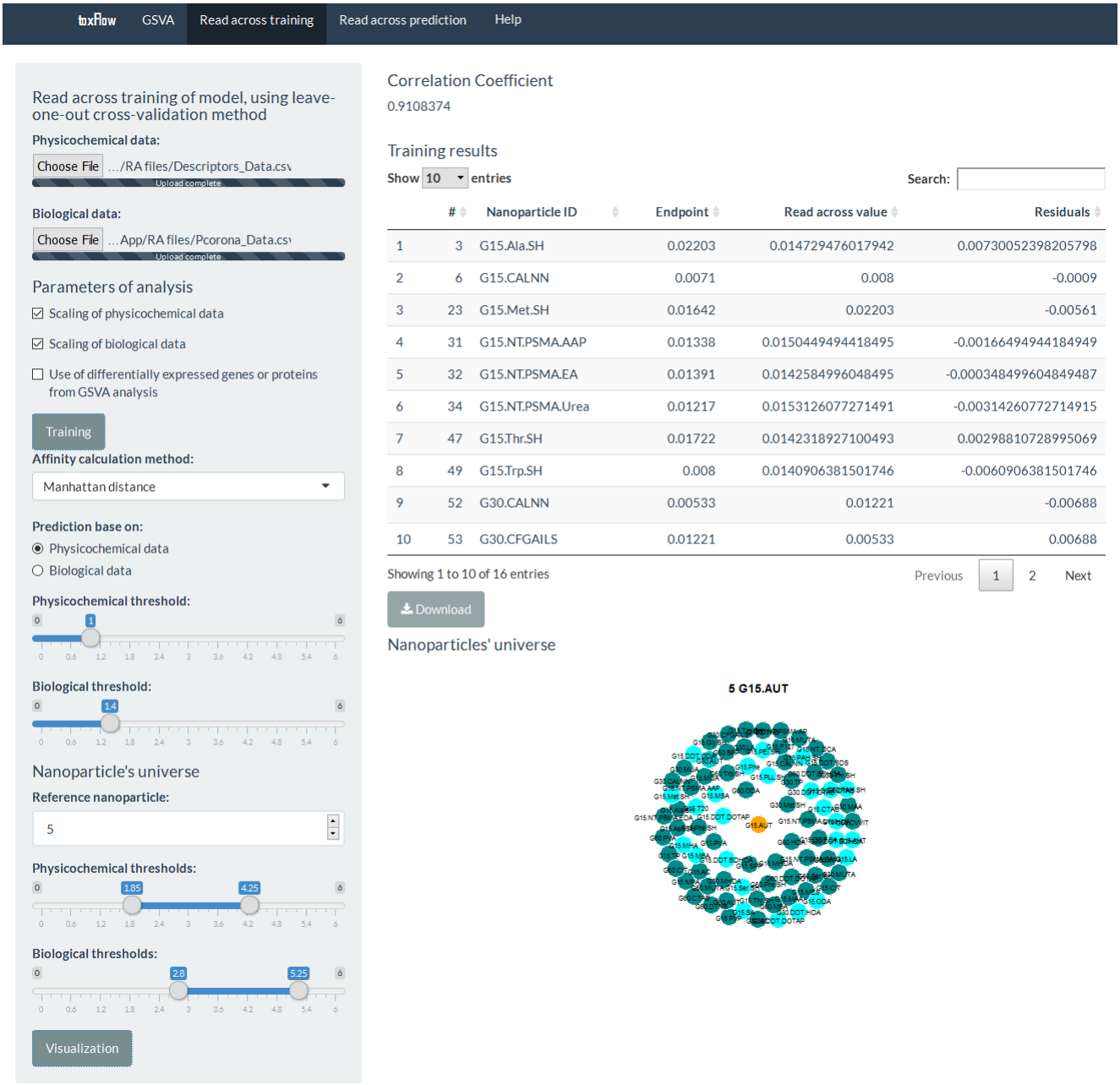

Figure 7: Sample Read across model results' presentation

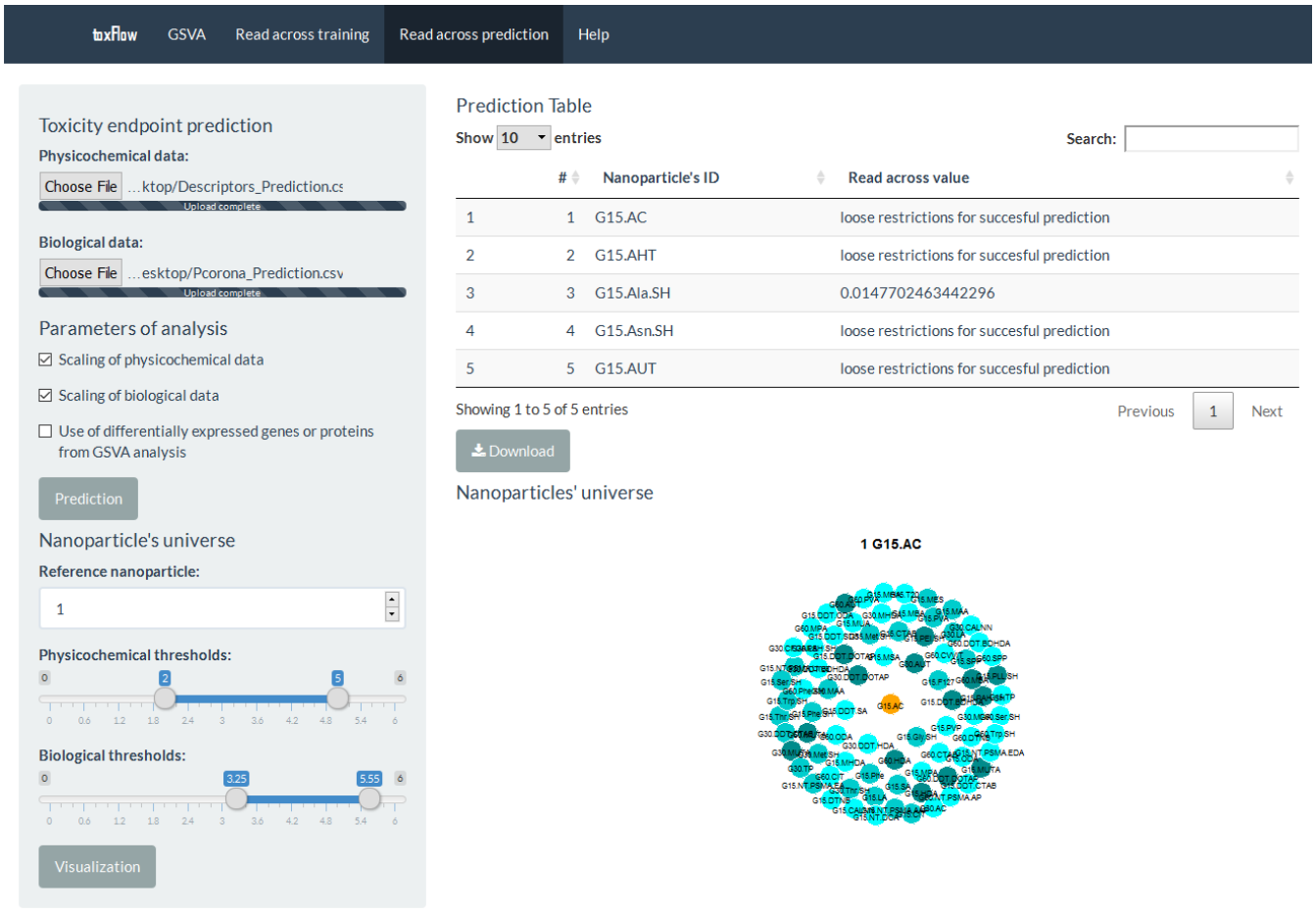

Figure 8: Sample Read across prediction results' presentation

References

[1] S. Hänzelmann, R. Castelo, J. Guinney, *GSVA: gene set variation analysis for microarray and RNA-Seq data*, BMC Bioinformatics, 14:7, 2013, Available online in: `http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-7`

[2] I. Shah, J. Liu, R. S. Judson, R. S. Thomas, G. Patlewicz, *Systematically evaluating read-across prediction and performance using a local validity approach characterized by chemical structure and bioactivity information*, Regulatory Toxicology and Pharmacology, 79: 12-24, 2016

[3] P.N. Tan, M. Steinbach, V. Kumar, *Introduction to Data Mining*, Pearson Addison-Wesley, 2005

[4] J. D. Storey, R. Tibshirani, *Statistical significance for genomewide studies*, PNAS, vol.100, no. 16, 2003