# toxFlow v.0.1 Tutorial

Dimitra Danai Varsou
School of Chemical Engineering, National Technical University of Athens
Conctact: dimitra.varsou@gmail.com

July 31, 2017

**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 License. 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 analysed (genomics, proteomics). 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 [4]. Using read across technique a table that contains all nanoparticles (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.

The web application is available in: http://147.102.86.129:3838/, the source code and the manual are available at GitHub (https://github.com/DemetraDanae/toxFlow, doi: 10.5281/zenodo.595814) and a video tutorial on YouTube.

# 1 GSVA

## 1.1 Input data

Before running the GSVA analysis, the user should import two files by selecting from the drop-down list **Import files** (Fig. 2C), of the inner tab **Input data** (Fig. 2B). 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 (Fig. 4). 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 (Fig. 5), which will be used for the creation of the design matrix. By selecting from the drop-down list **Use demo dataset** the

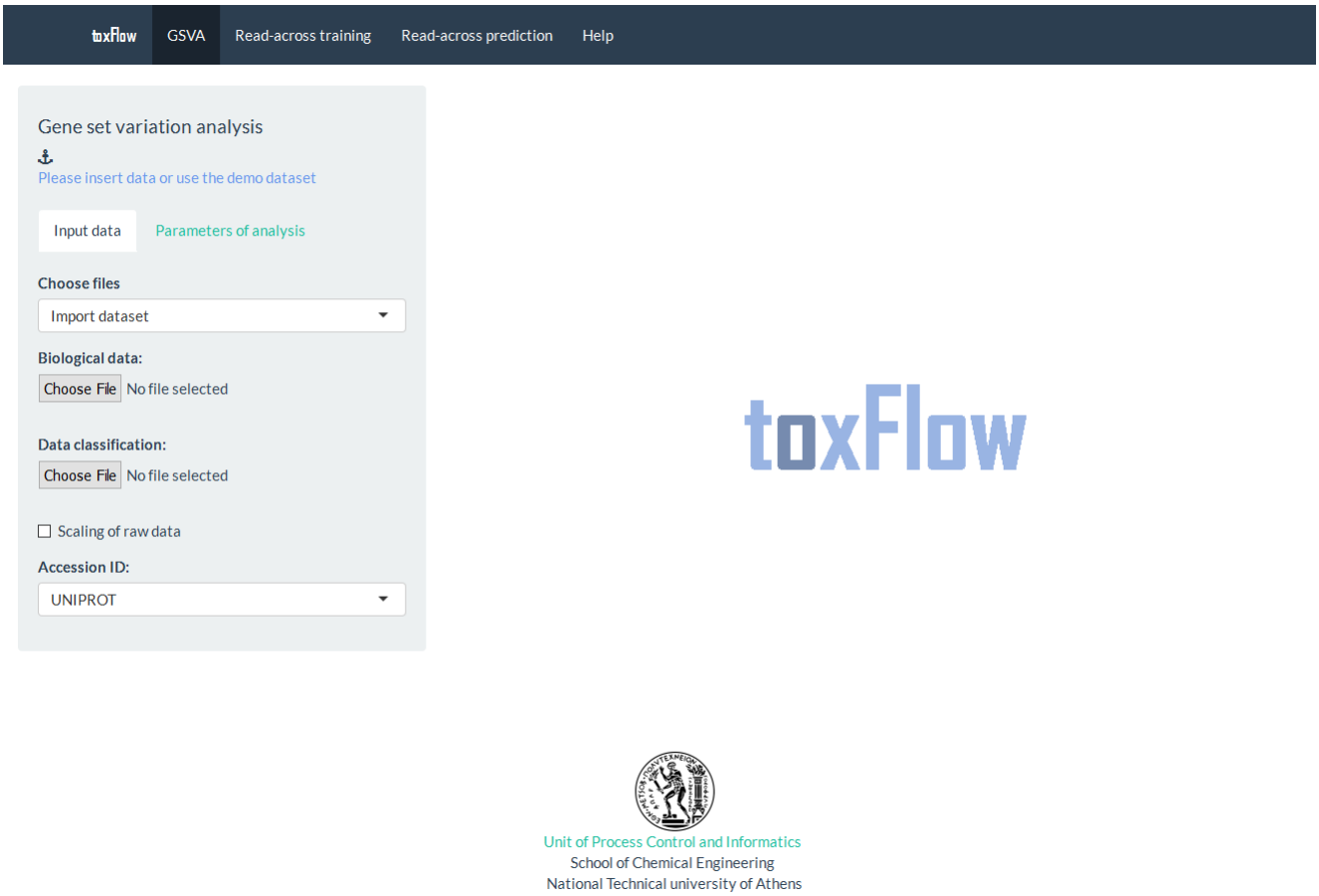



Figure 1: User interface of toxFlow application. On the top ribbon are shown the three tabs (each for every main part).

user can see an example of the analysis for anionic-cationic classification (Fig. 3A). The dataset comes from Walkey *et al.* (2014) published article [3] which consists of protein corona fingerprint for 84 gold NPs with diameter 15, 30 and 60 nm, from LC/MS-MS analysis experiments. These nanoparticles were incubated with cells of A549 cell line (human lung epithelial cancer cells).

The supplied data set could be normalized, by clicking on the check-box (**Scaling of raw data**-(Fig. 2D), according to the following equation:

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

Where $c_{\mathrm{in}}$, 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
$c_{\mathrm{sc}}$, the normalized value of the parameter.

Furthermore, the user can select from a drop-down list the **Accession ID** of the gene or protein names (Uniprot, EntrezID, RefSeq or Symbol), as it can be seen in Fig.2E.

## 1.2 Parameters of analysis

In the section **Gene set collection** of the inner tab **Parameters of analysis** (Fig.2F) 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 (Fig.2G). *GO-MF*

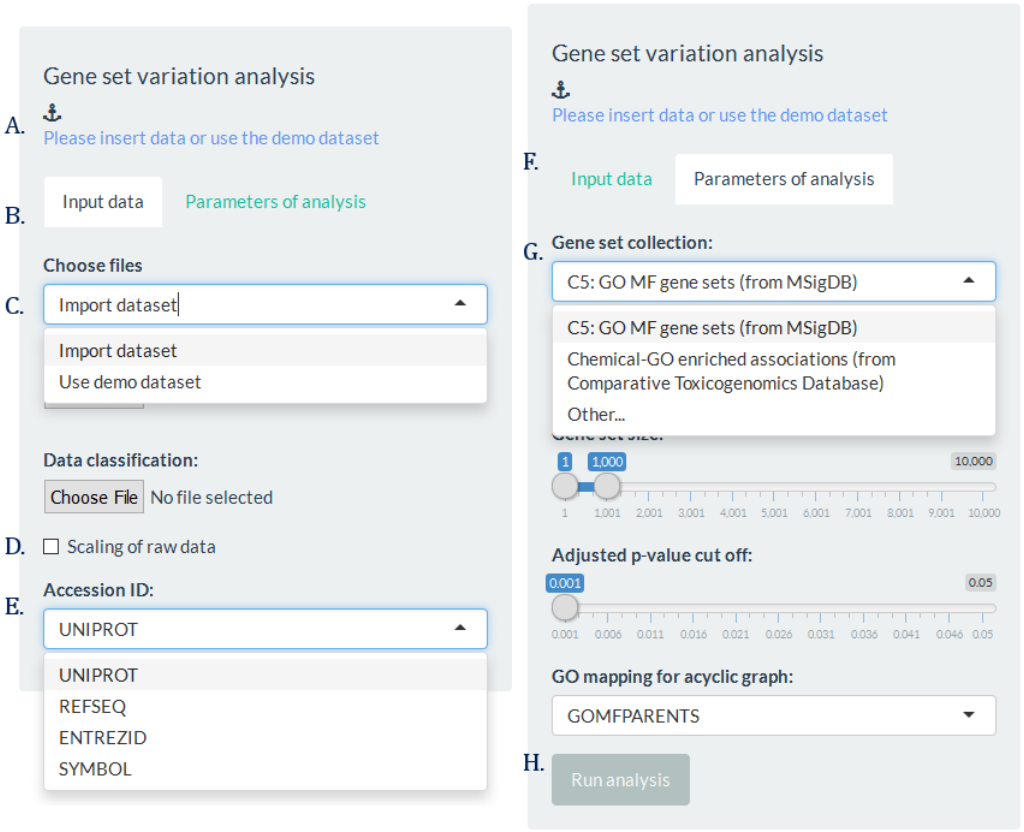



Figure 2: Input data and adjustment of parameters of the GSVA analysis: A. Informative messages, B. Inner tab "Input data", C. Choice of files, D. Scaling of raw data, E. Accession ID selection, F. Inner tab "Parameters of the analysis", G. Gene set collections and H. Run button.

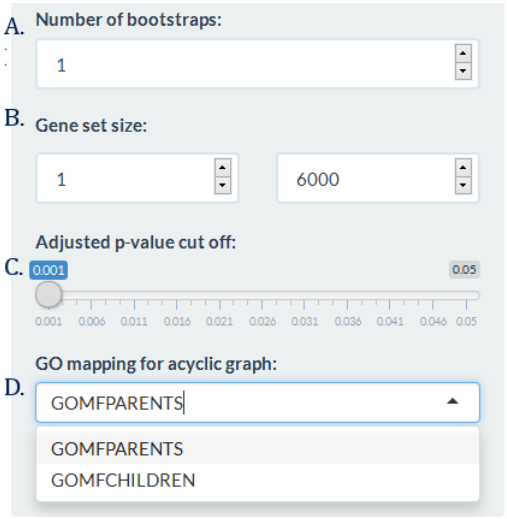



Figure 3: Adjustment of parameters of the GSVA analysis: A. Number of bootstraps, B. Gene set size, C. Adjusted p-value cut off and D. GO mapping.

Samples

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | G15.AC | G15.AHT | G15.Ala.SH | G15.Asn.SH | G15.AUT | G15.CALNN | G15.CIT | G15.CTAB |
| 2 | P01024 | 0.073492144 | 0.021656535 | 0.08105939 | 0.084718923 | 0.024524313 | 0.043650794 | 0.127461707 | 0.091865358 |
| 3 | P01834 | 0.011657375 | 0.015957447 | 0.005617978 | 0.011876485 | 0.014799154 | 0.005952381 | 0.028993435 | 0.019635344 |
| 4 | P0C0L4 | 0.027369488 | 0.055091185 | 0.017656501 | 0.015835313 | 0.054968288 | 0.034391534 | 0.038840263 | 0.074333801 |
| 5 | P02647 | 0.004561581 | 0.007978723 | 0.005617978 | 0.007125891 | 0.012684989 | 0.008597884 | 0.018052516 | 0.047685835 |
| 6 | P01871 | 0.016725798 | 0.018237082 | 0.008025682 | 0.015043547 | 0.016490486 | 0.006613757 | 0.011487965 | 0.00911641 |
| 7 | P01857 | 0.014698429 | 0.015577508 | 0.003210273 | 0.004750594 | 0.006342495 | 0.014550265 | 0.032275711 | 0.021037868 |
| 8 | P02649 | 0.085656361 | 0.001899696 | 0.107544141 | 0.102137767 | 0.012262156 | 0.010582011 | 0.012035011 | 0.01542777 |
| 9 | B9A064 | 0.003547897 | 0.002659574 | 0.001605136 | 0.005542359 | 0.002536998 | 0.003306878 | 0.013676149 | 0.00911641 |
| 10 | P10909 | 0.012671059 | 0.005319149 | 0.00882825 | 0.005542359 | 0.008033827 | 0.007936508 | 0.007658643 | 0.010518934 |
| 11 | P01876 | 0.008109478 | 0.011398176 | 0.005617978 | 0.006334125 | 0.009302326 | 0.007936508 | 0.019693654 | 0.021037868 |
| 12 | P04004 | 0.033451597 | 0.021656535 | 0.085072231 | 0.071258907 | 0.095560254 | 0.054232804 | 0.037199125 | 0.036465638 |
| 13 | Q14624 | 0.002534212 | 0.00493921 | 0.002407705 | 0.003958828 | 0.002536998 | 0.022486772 | 0.064551422 | 0.025245442 |
| 14 | P01009 | 0.006588951 | 0.016717325 | 0.006420546 | 0.003167063 | 0.01987315 | 0.009920635 | 0.006564551 | 0.0371669 |
| 15 | P04114 | 0.079574252 | 0.021276596 | 0.016853933 | 0.026128266 | 0.023678647 | 0 | 0.007658643 | 0.04628331 |
| 16 | P00734 | 0.061834769 | 0.263677812 | 0.123595506 | 0.106096595 | 0.226638478 | 0.141534392 | 0.024617068 | 0.0371669 |
| 17 | P0C0L5 | 0.001520527 | 0.007598784 | 0.001605136 | 0.002375297 | 0.005496829 | 0.001322751 | 0 | 0.003506311 |
| 18 | P01766 | 0.000506842 | 0.000759878 | 0 | 0.000791766 | 0.000422833 | 0.001322751 | 0.002188184 | 0.000701262 |
| 19 | P01008 | 0.112012164 | 0.047112462 | 0.313001605 | 0.288202692 | 0.043128964 | 0.185846561 | 0.085339168 | 0.066619916 |
| 20 | P04196 | 0.044095286 | 0.001899696 | 0.031300161 | 0.040380048 | 0.001691332 | 0.072089947 | 0.019146608 | 0.011921459 |
| 21 | P02652 | 0.003041054 | 0.003039514 | 0.010433387 | 0.006334125 | 0.003382664 | 0 | 0.00273523 | 0.010518934 |
| 22 | P04003 | 0.031931069 | 0.127279635 | 0 | 0.001583531 | 0.101057082 | 0 | 0.010393873 | 0.001402525 |
| 23 | P02766 | 0.002027369 | 0.004179331 | 0.004815409 | 0.003958828 | 0.024524313 | 0.003306878 | 0.001094092 | 0.007713885 |
| 24 | P01860 | 0.005068424 | 0.001139818 | 0.001605136 | 0.000791766 | 0.002114165 | 0.00462963 | 0.007111597 | 0.004207574 |

Proteins

Figure 4: Required format of the .csv file with a sample of input data for GSVA analysis.

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. 6).

Additionally, in **Number of bootstraps** (Fig. 3A) 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** (Fig. 3B) can control the minimum and maximum size of the resulting gene sets (default value is 1 and 6000 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. 3C). Finally the user in **GO mapping for acyclic 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. 3D).

## 1.3 Results

If all the required files are uploaded, by clicking **Run analysis** the application is performing the analysis, according to the parameters above (Fig. 2H). Otherwise the corresponding button remains disabled till all necessary files are provided. The application displays messages in order to help the user to handle the inputs (Fig. 2A). After running the analysis, the application exports a table that contains the significant gene sets, their GO ID, their size, the adjusted p-value based on Benjamini-Hochberg (BH) [5] 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. 14). Both the heatmap and the acyclic graph can be downloaded too.

| | A | B | C |
|---|---|---|---|
| 1 | cnano.ids | Classification | |
| 2 | G15.AC | Anionic | |
| 3 | G15.AHT | Cationic | |
| 4 | G15.Ala.SH | Anionic | |
| 5 | G15.Asn.SH | Anionic | |
| 6 | G15.AUT | Cationic | |
| 7 | G15.CALNN | Anionic | |
| 8 | G15.CIT | Anionic | |
| 9 | G15.CTAB | Cationic | |
| 10 | G15.DDT.BDHDA | Cationic | |
| 11 | G15.DDT.CTAB | Cationic | |
| 12 | G15.DDT.DOTAP | Cationic | |
| 13 | G15.DDT.ODA | Cationic | |
| 14 | G15.DDT.SA | Anionic | |
| 15 | G15.DDT.SDS | Anionic | |
| 16 | G15.DTNB | Anionic | |
| 17 | G15.F127 | Anionic | |
| 18 | G15.Gly.SH | Anionic | |
| 19 | G15.HDA | Cationic | |
| 20 | G15.LA | Anionic | |
| 21 | G15.MAA | Anionic | |
| 22 | G15.MBA | Anionic | |
| 23 | G15.MES | Anionic | |
| 24 | G15.Met.SH | Anionic | |
| 25 | G15.MHA | Anionic | |
| 26 | G15.MHDA | Anionic | |
| 27 | G15.MPA | Anionic | |

Figure 5: Required format of the .csv file with a sample of input data for classification.

| | A | B | C |
|---|---|---|---|
| 1 | GOterm | EntrezID | |
| 2 | 10-hydroxy-9-(phosphonooxy)octadecanoate phosphatase activity | 1914 | |
| 3 | 11-beta-hydroxysteroid dehydrogenase (NADP+) activity | 2800 | |
| 4 | 11-beta-hydroxysteroid dehydrogenase [NAD(P)] activity | 2801 | |
| 5 | 11-beta-hydroxysteroid dehydrogenase [NAD(P)] activity | 2800 | |
| 6 | 11-cis retinal binding | 5068 | |
| 7 | 1,1-dichloro-2-(dihydroxy-4-chlorophenyl)-(4-chlorophenyl)ethene 1,2-dioxygenase activity | 298 | |
| 8 | 1,2-bis(4-hydroxyphenyl)-2-proponol dehydratase activity | 91 | |
| 9 | 1,2-dihydroxy-5,6,7,8-tetrahydronaphthalene extradiol dioxygenase activity | 298 | |
| 10 | 1,2-dihydroxyfluorene 1,1-alpha-dioxygenase activity | 298 | |
| 11 | 1,2-dihydroxynaphthalene-6-sulfonate 1,8a-dioxygenase activity | 298 | |
| 12 | 1,2-dihydroxy-phenanthrene glycosyltransferase activity | 3332 | |
| 13 | 1,2-dihydroxy-phenanthrene glycosyltransferase activity | 591 | |
| 14 | 1,2-dihydroxy-phenanthrene glycosyltransferase activity | 587 | |
| 15 | 1,2-dihydroxy-phenanthrene glycosyltransferase activity | 588 | |
| 16 | 1,2-dihydroxy-phenanthrene glycosyltransferase activity | 2291 | |
| 17 | 1,2-dihydroxy-phenanthrene glycosyltransferase activity | 589 | |
| 18 | 1,2-dihydroxy-phenanthrene glycosyltransferase activity | 592 | |
| 19 | 1,2-dihydroxy-phenanthrene glycosyltransferase activity | 4473 | |
| 20 | 1,2-dihydroxy-phenanthrene glycosyltransferase activity | 4477 | |
| 21 | 1,2-dihydroxy-phenanthrene glycosyltransferase activity | 586 | |
| 22 | 1,2-dihydroxy-phenanthrene glycosyltransferase activity | 2408 | |
| 23 | 1,2-dihydroxy-phenanthrene glycosyltransferase activity | 4605 | |
| 24 | 1,2-dihydroxy-phenanthrene glycosyltransferase activity | 2244 | |
| 25 | 1,2-dihydroxy-phenanthrene glycosyltransferase activity | 3653 | |
| 26 | 1,2-dihydroxy-phenanthrene glycosyltransferase activity | 590 | |
| 27 | 1,2-dihydroxy-phenanthrene glycosyltransferase activity | 2243 | |

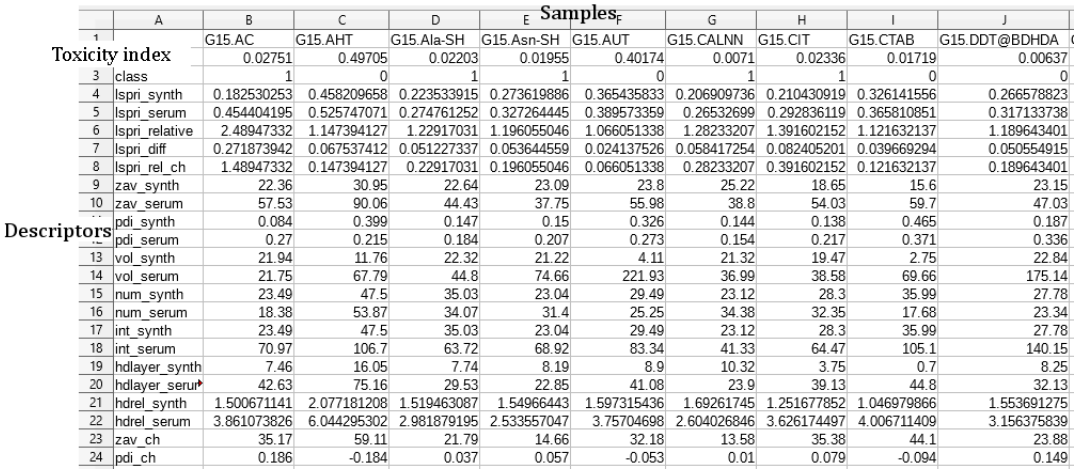
Figure 6: Required format of the .csv file with a sample of a gene set collection.

# 2 Read across training

## 2.1 Import data

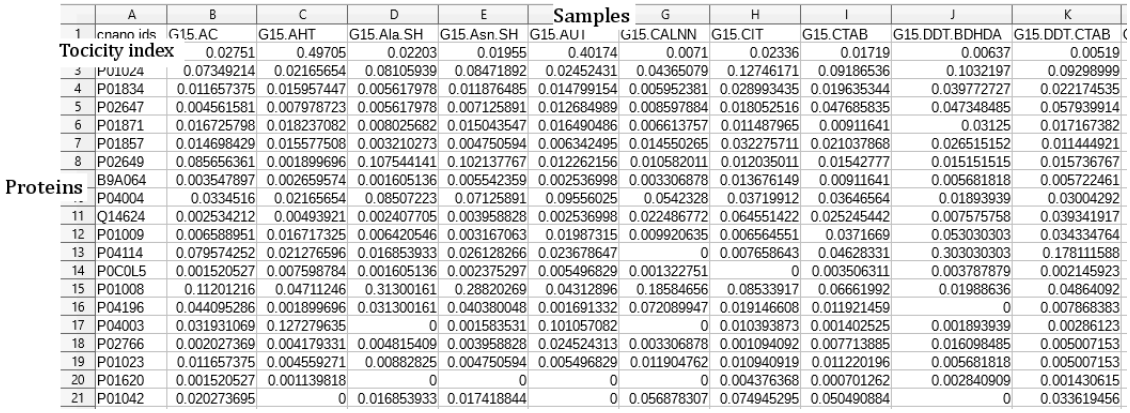
The user must upload two .csv files in the application (from the drop-down list **Import files**, Fig. 9B) of the inner tab **Input data** (Fig. 9A): The first one (**Physicochemical data**) must contain the values of physicochemical descriptors (samples in columns and descriptors in rows). The second 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 the samples and the indeces respectively and should have the same number of columns (same nanoparticles). Also, both files must contain the values of the toxicity

index, which will be predicted by the model, in the second row (Fig. 7 and 8). By selecting from the drop-down list **Use demo dataset** the user can see an example of the analysis. The dataset comes from Walkey *et al.* (2014) published article [3]. For defining similarity physicochemical descriptors and protein corona composition data were used, while cell association was used as the end-point.




| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | G15.AC | G15.AHT | G15.Ala-SH | G15.Asn-SH | G15.AUT | G15.CALNN | G15.CIT | G15.CTAB | G15.DDT@BDHDA |
| 2 | Toxicity index | 0.02751 | 0.49705 | 0.02203 | 0.01955 | 0.40174 | 0.0071 | 0.02336 | 0.01719 | 0.00637 |
| 3 | class | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 |
| 4 | lspri_synth | 0.182530253 | 0.458209658 | 0.223533915 | 0.273619886 | 0.365435833 | 0.206909736 | 0.210430919 | 0.326141556 | 0.266578823 |
| 5 | lspri_serum | 0.454404195 | 0.525747071 | 0.274761252 | 0.327264445 | 0.389573359 | 0.26532699 | 0.292836119 | 0.365810851 | 0.317133738 |
| 6 | lspri_relative | 2.48947332 | 1.147394127 | 1.22917031 | 1.196055046 | 1.066051338 | 1.28233207 | 1.391602152 | 1.121632137 | 1.189643401 |
| 7 | lspri_diff | 0.271873942 | 0.067537412 | 0.051227337 | 0.053644559 | 0.024137526 | 0.058417254 | 0.082405201 | 0.039669294 | 0.050554915 |
| 8 | lspri_rel_ch | 1.48947332 | 0.147394127 | 0.22917031 | 0.196055046 | 0.066051338 | 0.28233207 | 0.391602152 | 0.121632137 | 0.189643401 |
| 9 | zav_synth | 22.36 | 30.95 | 22.64 | 23.09 | 23.8 | 25.22 | 18.65 | 15.6 | 23.15 |
| 10 | zav_serum | 57.53 | 90.06 | 44.43 | 37.75 | 55.98 | 38.8 | 54.03 | 59.7 | 47.03 |
| 11 | pdi_synth | 0.084 | 0.399 | 0.147 | 0.15 | 0.326 | 0.144 | 0.138 | 0.465 | 0.187 |
| 12 | pdi_serum | 0.27 | 0.215 | 0.184 | 0.207 | 0.273 | 0.154 | 0.217 | 0.371 | 0.336 |
| 13 | vol_synth | 21.94 | 11.76 | 22.32 | 21.22 | 4.11 | 21.32 | 19.47 | 2.75 | 22.84 |
| 14 | vol_serum | 21.75 | 67.79 | 44.8 | 74.66 | 221.93 | 36.99 | 38.58 | 69.66 | 175.14 |
| 15 | num_synth | 23.49 | 47.5 | 35.03 | 23.04 | 29.49 | 23.12 | 28.3 | 35.99 | 27.78 |
| 16 | num_serum | 18.38 | 53.87 | 34.07 | 31.4 | 25.25 | 34.38 | 32.35 | 17.68 | 23.34 |
| 17 | int_synth | 23.49 | 47.5 | 35.03 | 23.04 | 29.49 | 23.12 | 28.3 | 35.99 | 27.78 |
| 18 | int_serum | 70.97 | 106.7 | 63.72 | 68.92 | 83.34 | 41.33 | 64.47 | 105.1 | 140.15 |
| 19 | hdlayer_synth | 7.46 | 16.05 | 7.74 | 8.19 | 8.9 | 10.32 | 3.75 | 0.7 | 8.25 |
| 20 | hdlayer_serum | 42.63 | 75.16 | 29.53 | 22.85 | 41.08 | 23.9 | 39.13 | 44.8 | 32.13 |
| 21 | hdrel_synth | 1.500671141 | 2.077181208 | 1.519463087 | 1.54966443 | 1.597315436 | 1.69261745 | 1.251677852 | 1.046979866 | 1.553691275 |
| 22 | hdrel_serum | 3.861073826 | 6.044295302 | 2.981879195 | 2.533557047 | 3.75704698 | 2.604026846 | 3.626174497 | 4.006711409 | 3.156375839 |
| 23 | zav_ch | 35.17 | 59.11 | 21.79 | 14.66 | 32.18 | 13.58 | 35.38 | 44.1 | 23.88 |
| 24 | pdi_ch | 0.186 | -0.184 | 0.037 | 0.057 | -0.053 | 0.01 | 0.079 | -0.094 | 0.149 |

Figure 7: Required format of the .csv file with a sample of physicochemical data.




| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | cnano ids | G15.AC | G15.AHT | G15.Ala.SH | G15.Asn.SH | G15.AUT | G15.CALNN | G15.CIT | G15.CTAB | G15.DDT.BDHDA | G15.DDT.CTAB |
| 2 | Tocicity index | 0.02751 | 0.49705 | 0.02203 | 0.01955 | 0.40174 | 0.0071 | 0.02336 | 0.01719 | 0.00637 | 0.00519 |
| 3 | P01024 | 0.07349214 | 0.02165654 | 0.08105939 | 0.08471892 | 0.02452431 | 0.04365079 | 0.12746171 | 0.09186536 | 0.1032197 | 0.09298999 |
| 4 | P01834 | 0.011657375 | 0.015957447 | 0.005617978 | 0.011876485 | 0.014799154 | 0.005952381 | 0.028993435 | 0.019635344 | 0.039772727 | 0.022174535 |
| 5 | P02647 | 0.004561581 | 0.007978723 | 0.005617978 | 0.007125891 | 0.012684989 | 0.008597884 | 0.018052516 | 0.047685835 | 0.047348485 | 0.057939914 |
| 6 | P01871 | 0.016725798 | 0.018237082 | 0.008025682 | 0.015043547 | 0.016490486 | 0.006613757 | 0.011487965 | 0.00911641 | 0.03125 | 0.017167382 |
| 7 | P01857 | 0.014698429 | 0.015577508 | 0.003210273 | 0.004750594 | 0.006342495 | 0.014550265 | 0.032275711 | 0.021037868 | 0.026515152 | 0.011444921 |
| 8 | P02649 | 0.085656361 | 0.001899696 | 0.107544141 | 0.102137767 | 0.012262156 | 0.010582011 | 0.012035011 | 0.01542777 | 0.015151515 | 0.015736767 |
| 9 | B9A064 | 0.003547897 | 0.002659574 | 0.001605136 | 0.005542359 | 0.002536998 | 0.003306878 | 0.013676149 | 0.00911641 | 0.005681818 | 0.005722461 |
| 10 | P04004 | 0.0334516 | 0.02165654 | 0.08507223 | 0.07125891 | 0.09556025 | 0.0542328 | 0.03719912 | 0.03646564 | 0.01893939 | 0.03004292 |
| 11 | Q14624 | 0.002534212 | 0.00493921 | 0.002407705 | 0.003958828 | 0.002536998 | 0.022486772 | 0.064551422 | 0.025245442 | 0.007575758 | 0.039341917 |
| 12 | P01009 | 0.006588951 | 0.016717325 | 0.006420546 | 0.003167063 | 0.01987315 | 0.009920635 | 0.006564551 | 0.0371669 | 0.053030303 | 0.034334764 |
| 13 | P04114 | 0.079574252 | 0.021276596 | 0.016853933 | 0.026128266 | 0.023678647 | 0 | 0.007658643 | 0.04628331 | 0.303030303 | 0.178111588 |
| 14 | P0C0L5 | 0.001520527 | 0.007598784 | 0.001605136 | 0.002375297 | 0.005496829 | 0.001322751 | 0 | 0.003506311 | 0.003787879 | 0.002145923 |
| 15 | P01008 | 0.11201216 | 0.04711246 | 0.31300161 | 0.28820269 | 0.04312896 | 0.18584656 | 0.08533917 | 0.06661992 | 0.01988636 | 0.04864092 |
| 16 | P04196 | 0.044095286 | 0.001899696 | 0.031300161 | 0.040380048 | 0.001691332 | 0.072089947 | 0.019146608 | 0.011921459 | 0 | 0.007868383 |
| 17 | P04003 | 0.031931069 | 0.127279635 | 0 | 0.001583531 | 0.101057082 | 0 | 0.010393873 | 0.001402525 | 0.001893939 | 0.00286123 |
| 18 | P02766 | 0.002027369 | 0.004179331 | 0.004815409 | 0.003958828 | 0.024524313 | 0.003306878 | 0.001094092 | 0.007713885 | 0.016098485 | 0.005007153 |
| 19 | P01023 | 0.011657375 | 0.004559271 | 0.00882825 | 0.004750594 | 0.005496829 | 0.011904762 | 0.010940919 | 0.011220196 | 0.005681818 | 0.005007153 |
| 20 | P01620 | 0.001520527 | 0.001139818 | 0 | 0 | 0 | 0 | 0.004376368 | 0.000701262 | 0.002840909 | 0.001430615 |
| 21 | P01042 | 0.020273695 | 0 | 0.016853933 | 0.017418844 | 0 | 0.056878307 | 0.074945295 | 0.050490884 | 0 | 0.033619456 |

Figure 8: Required format of the .csv file with a sample of biological data.

Furthermore the user can select if the physicochemical and expression data should be normalized (**Scaling of physicochemical data**, **Scaling of biological data**, Fig. 9C) 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 (Fig. 9D). It is implied that in the previous section the user will have analysed the same biological data. If no GSV Analysis is performed the check-box remains disabled.

In case that only one file of input data (either the physicochemical or the biological) is available, the user can select the appropriate check-box (**Only one file available**) and he is enabled to upload the corresponding file and proceed to the analysis. The file must follow the format presented before.

## 2.2 Adjust parameters of analysis

In order to reduce the noise of the data for the analysis, the user can filter the attributes according to the absolute value of their correlation to the endpoint. For this purpose, two

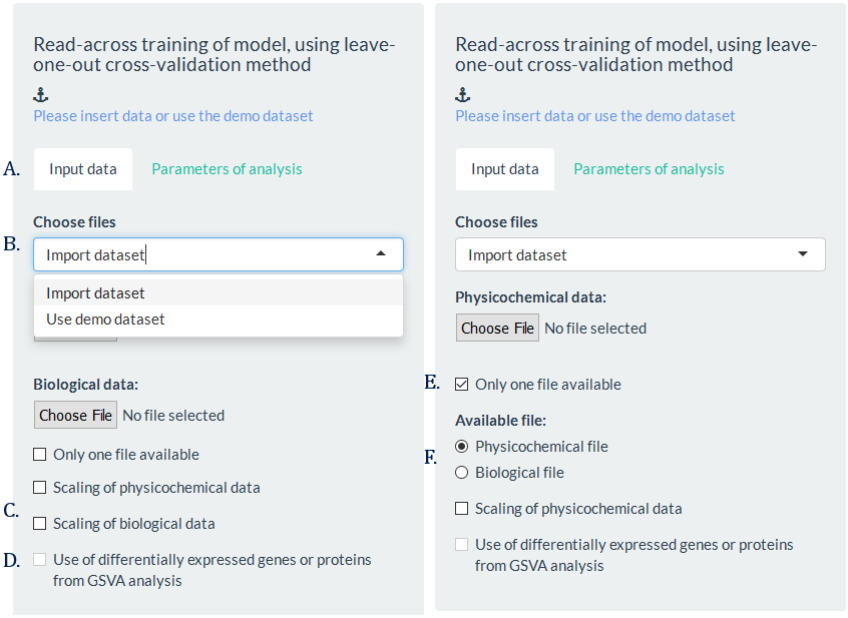



Figure 9: Read across input files: A. Inner tab "Input data", B. Choice of files, C. Scaling of raw data, D. Use of differentially expressed proteins from GSVA, E. Determine whether only one file is available and F. Selection of available file.

sliders are available in order to define the filtering level of the absolute values of the attribute's correlation (2) to the endpoint. For the estimation of similarity between NPs, the user can choose from a drop-down list (**Similarity calculation method**) one of the following options: cosine similarity (3), Manhattan (4) and Euclidean (5) distance (Fig. 10). The user can define the physicoshemical and biological threshold that control the selection of neigbouring NPs from numeric inputs (Fig. 11C). Finally, in **Prediction base** the user can choose if the prediction will be calculated on a physicochemical or on biological base (Fig. 11D). In case that only one file is uploaded the user can control only the corresponding level of attribute filtering and similarity threshold and the prediction base is selected automatically, according to the input file. By pressing the button **Training** (Fig. 11E) the model training begins. This button remains disabled till all necessary files are provided. The user can change the similarity method, the calculation base and the two thresholds and the results will be updated automatically.

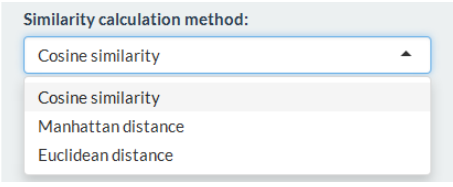



Figure 10: Similarity calculation measure selection

$$r = \frac{\sum_{i=1}^{N}(A_i - \overline{A})(B_i - \overline{B})}{\sqrt{\sum_{i=1}^{N}(A_i - \overline{A})^2}\sqrt{\sum_{i=1}^{N}(B_i - \overline{B})^2}} \tag{2}$$

Where $A_i$, the $i^{th}$ value of the vector $A$,
$B_i$, the $i^{th}$ value of the vector $B$,

$\overline{A}$, the average value of the vector $A$ and
$\overline{B}$, the average value of the vector $B$.

$$cos(\theta) = \frac{\sum_{i=1}^{N} A_i B_i}{\sqrt{\sum_{i=1}^{N} A_i^2}\sqrt{\sum_{i=1}^{N} B_i^2}} \tag{3}$$

$$d_1 = \sum_{i=1}^{N} |A_i - B_i| \tag{4}$$

$$d_2 = \sqrt{\sum_{i=1}^{N} (A_i - B_i)^2} \tag{5}$$

Where $A_i$, the $i^{th}$ value of the vector $A$,
$B_i$, the $i^{th}$ value of the vector $B$ and
$N$, the total number of samples.

Also the user can visualize the NPs «universe»: the user can choose a reference NP (Fig. 11F) and observe its neighbours in color code, by adjusting the physicochemical and biological thresholds (Fig. 11I-J). Also, by selecting **Nanoparticles' diameter** and **Nanoparticles' classification** (Fig. 11G-H) the user can upload two corresponding files and observe the reference's neighbours according to their size and their classification. The file with the NPs size must be a .csv file. The 1st column must contain the samples and the 2nd samples' diameter (Fig. 12). The file with the NPs classification must be a .csv file. The 1st column must contain the samples and the 2nd samples' phenotype or other categorical variable of interest (Fig. 5). Every time the user changes the reference NP, imports the size and/or the classification file, and the thresholds, the user should press the button **Visualize** (Fig. 11K), in order to update the diagram. All parameters are shown in Figure 11.

### 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 prediction value is calculated using the same units as the actual values. 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. 15).

## 3 Read across prediction

### 3.1 Import data

The last section of the application can be used after the read across training. If training is not performed the Prediction button (see below) remains disabled. 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 (Fig. 13A-B): 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. In case that training is performed using only one file, the user is enabled to upload only one file for the prediction.

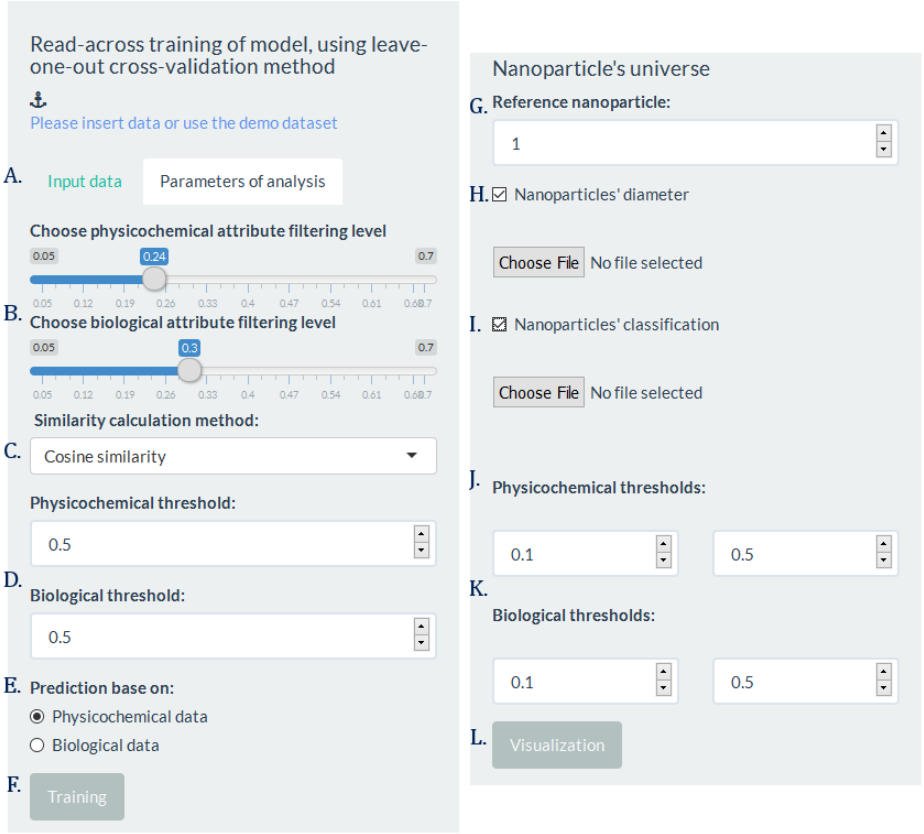



Figure 11: Read across training parameters: A. Inner tab "Parameters of analysis", B. Adjustment of filtering attribute level, C. Similarity measure selection, D. Physicochemical and biological threshold adjustment, E. Prediction base selection, F. Training button, G. Reference NP for visualization, H-I. Additional files for visualization, J-K. Physicochemical and biological thresholds for visualization and L. Visualization button.

| | A | B |
|---|---|---|
| 1 | cnano.ids | Diameter |
| 2 | G15.AC | 15 |
| 3 | G15.AHT | 15 |
| 4 | G15.Ala.SH | 15 |
| 5 | G15.Asn.SH | 15 |
| 6 | G15.AUT | 15 |
| 7 | G15.CALNN | 15 |
| 8 | G15.CIT | 15 |
| 9 | G15.CTAB | 15 |
| 10 | G15.DDT.BDHDA | 15 |
| 11 | G15.DDT.CTAB | 15 |
| 12 | G15.DDT.DOTAP | 15 |
| 13 | G15.DDT.ODA | 15 |
| 14 | G15.DDT.SA | 15 |
| 15 | G15.DDT.SDS | 15 |
| 16 | G15.DTNB | 15 |
| 17 | G15.F127 | 15 |
| 18 | G15.Gly.SH | 15 |
| 19 | G15.HDA | 15 |
| 20 | G15.LA | 15 |
| 21 | G15.MAA | 15 |
| 22 | G15.MBA | 15 |
| 23 | G15.MES | 15 |
| 24 | G15.Met.SH | 15 |
| 25 | G15.MHA | 15 |
| 26 | G15.MHDA | 15 |
| 27 | G15.MPA | 15 |

Figure 12: Required format of the .csv file with a sample of size data.

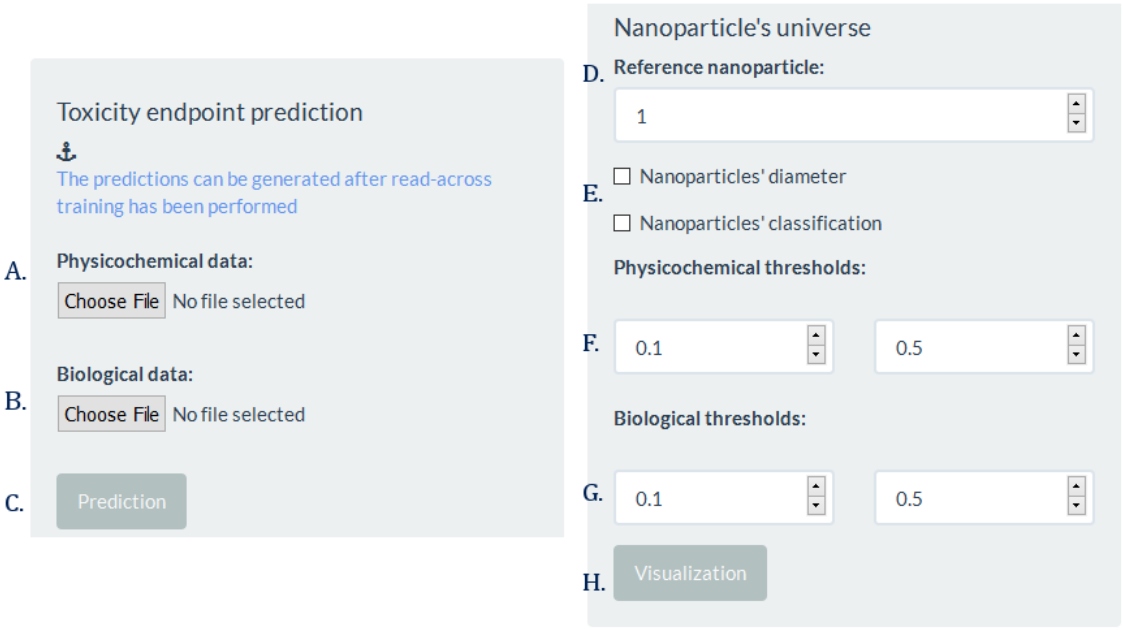



Figure 13: Parameters of prediction: A-B. Choice of input files, C. Prediction button, D. Reference NP for visualization, E. Additional files for visualization, F-G. Physicochemical and biological thresholds for visualization and H. Visualization button.

### 3.2 Adjust parameters of analysis

The normalization of input data, the filtering based on the significant data points (genes or proteins) given by GSVA analysis, the attribute filtering level, the calculation of similarity method and the thresholds are the same as defined at the training section. By clicking on **Prediction** (Fig. 13C) begins the prediction process.

The user can visualize the NPs «universe»: the user can choose a reference NP (Fig. 13D) and observe its neighbours in the training set in color code, by adjusting the physicochemical and biological thresholds and by importing the files (Fig. 13E) with the NPs' size and their classification (if they are available). Every time the user changes the reference NP, imports the size and/or the classification file, and the thresholds (Fig. 13F-G), the user should press the button **Visualize** (Fig. 13H) in order to update the diagram. All parameters are shown in Figure 13.

### 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 16).

toxFlow | GSVA | Read-across training | Read-across prediction | Help

Gene set variation analysis

Ready to run analysis

Input data | Parameters of analysis

Gene set collection:

C5: GO MF gene sets (from MSigDB)

Number of bootstraps:

3

Gene set size:

1 — 6000

Adjusted p-value cut off:

0.001 — 0.02 — 0.05

GO mapping for acyclic graph:

GOMFPARENTS

Run analysis

Differentially expressed gene sets

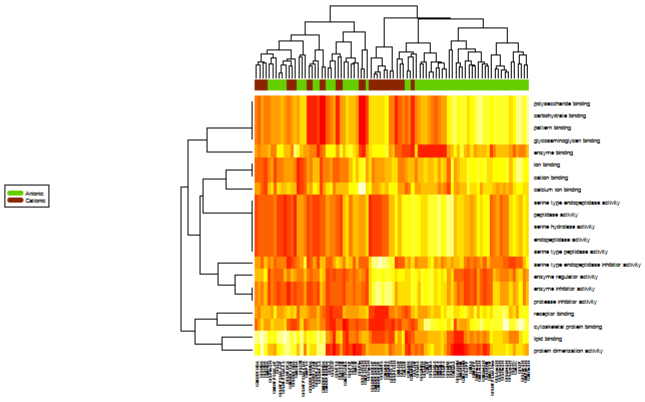
Show 10 entries — Search:

| | GOMFID | Term | p-value | Size | Counts |
|---|---|---|---|---|---|
| 1 | GO:0008092 | cytoskeletal protein binding | 7.47020735529879e-8 | 159 | 4 |
| 2 | GO:0030246 | carbohydrate binding | 0.000005134911582604 | 72 | 4 |
| 3 | GO:0030247 | polysaccharide binding | 0.000005134911582604 | 36 | 4 |
| 4 | GO:0001871 | pattern binding | 0.000005134911582604 | 45 | 4 |
| 5 | GO:0005539 | glycosaminoglycan binding | 0.000005134911582604 | 34 | 4 |
| 6 | GO:0005102 | receptor binding | 0.000010410947657038 2 | 377 | 7 |
| 7 | GO:0008236 | serine type peptidase activity | 0.0000402833799959389 | 45 | 12 |
| 8 | GO:0004175 | endopeptidase activity | 0.0000402833799959389 | 117 | 12 |
| 9 | GO:0017171 | serine hydrolase activity | 0.0000402833799959389 | 46 | 12 |
| 10 | GO:0008233 | peptidase activity | 0.0000402833799959389 | 176 | 12 |

Showing 1 to 10 of 15 entries — Previous 1 2 Next

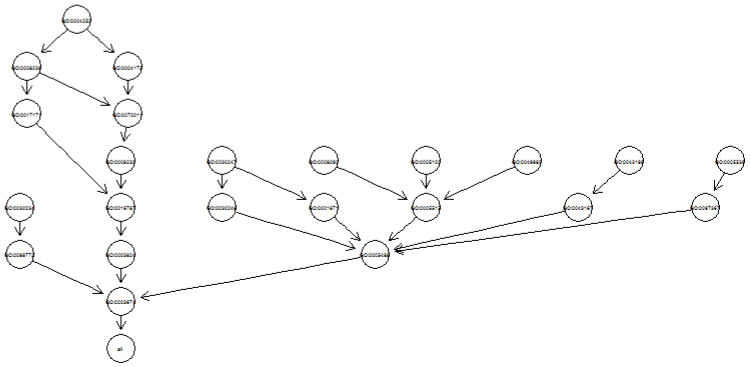
Download

Gene sets-genes heatmap



Download

GO directed acyclic graph



Download

Figure 14: Sample GSVA results' presentation

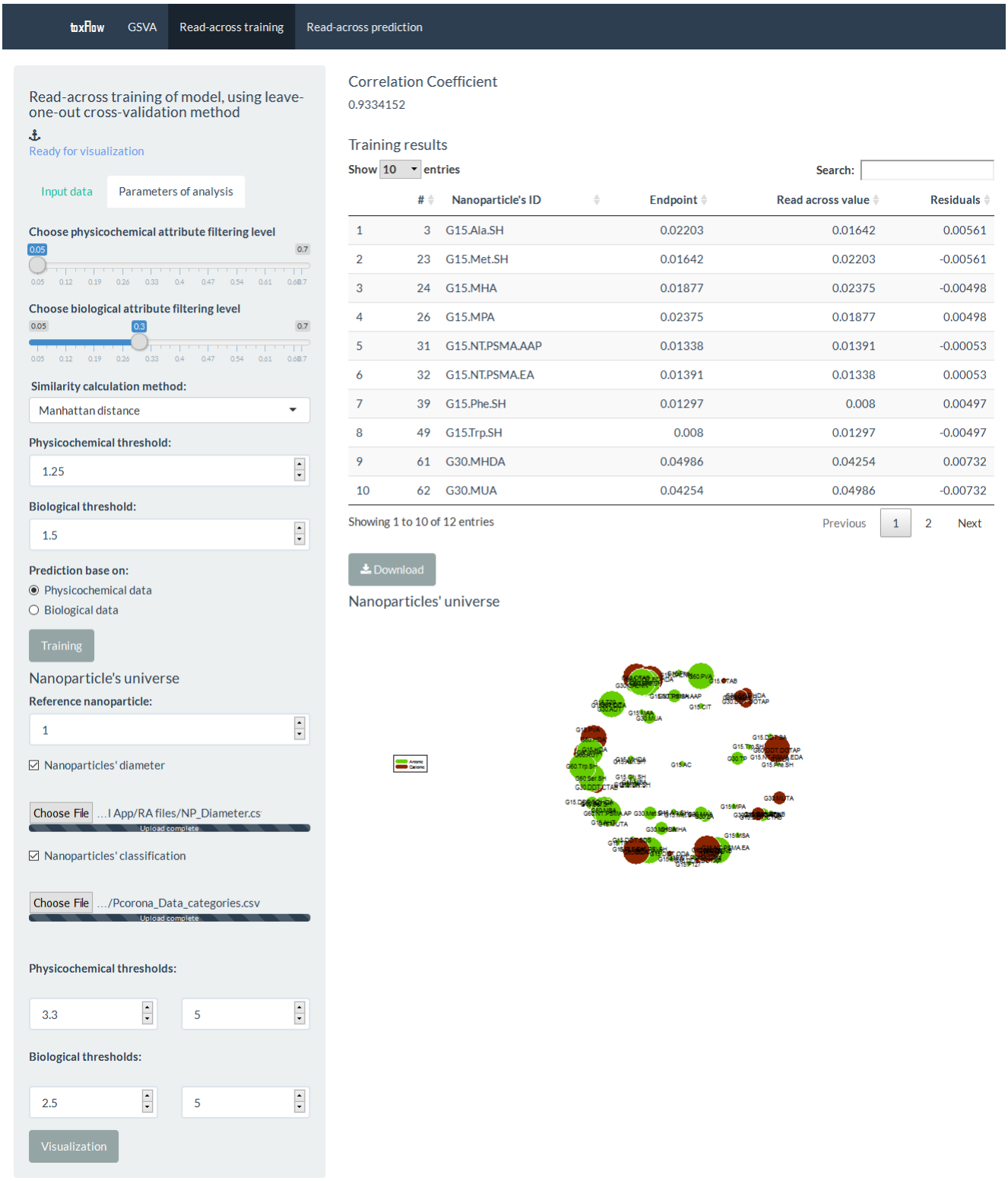



Figure 15: Sample Read across model results' presentation

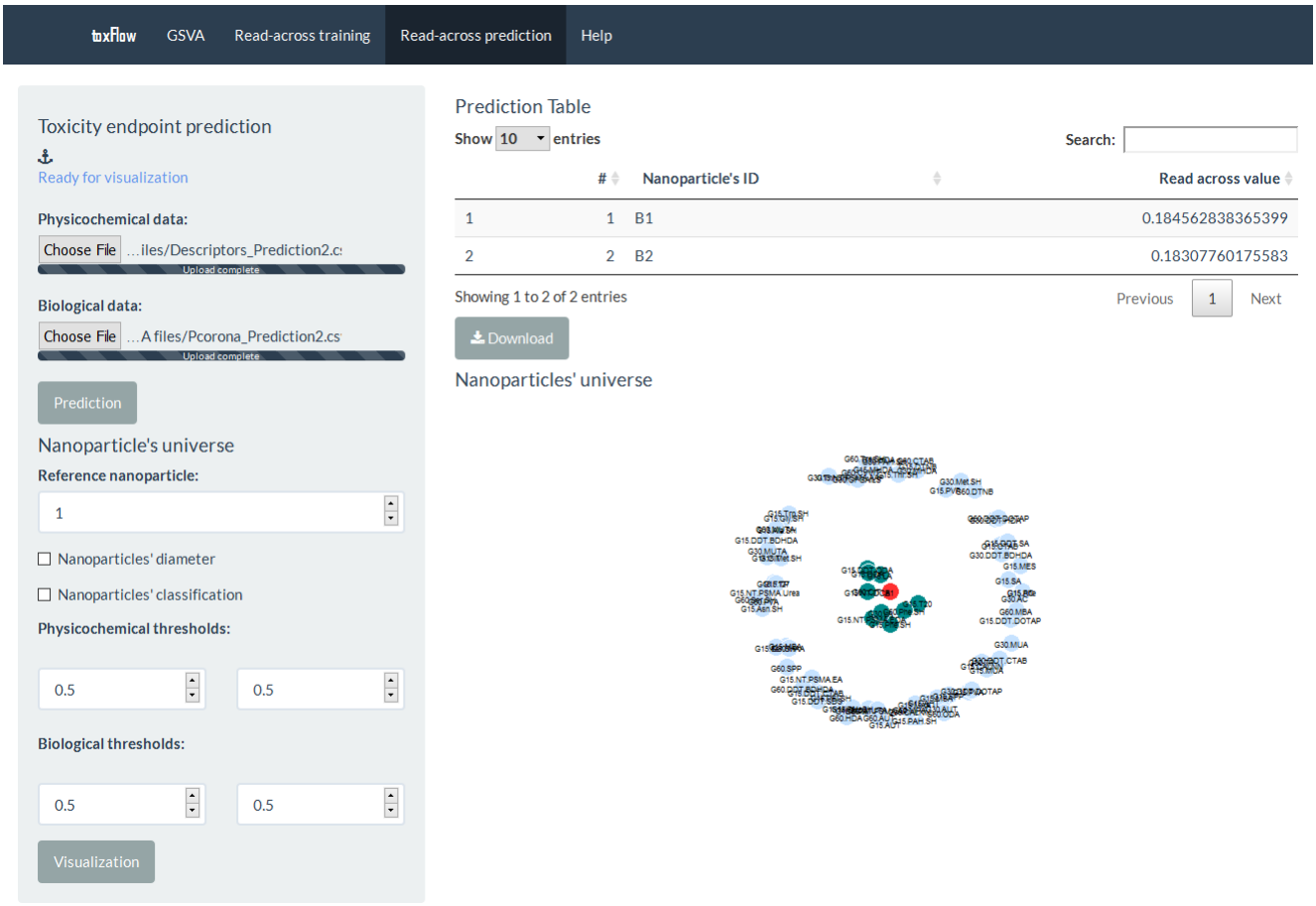



Figure 16: Sample Read across prediction results' presentation

## 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] C. D. Walkey, J. B. Olsen, F. Song, R. Liu, H. Guo, W. Olsen, Y. Cohen, A. Emili, W. C. W. Chan, *Protein Corona Fingerprinting Predicts the Cell Association of Gold Nanoparticles*, ACS Nano, 8 (3), 2439–2455, 2014, Available online in: `https://www.researchgate.net/publication/263941898_Protein_Corona_Fingerprinting_Predicts_the_Cellular_Interaction_of_Gold_and_Silver_Nanoparticles`

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

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