

# toxFlow v.0.1 Tutorial

Dimitra Danai Varsou

School of Chemical Engineering, National Technical University of Athens

Contact: [dimitra.varsou@gmail.com](mailto:dimitra.varsou@gmail.com)

December 12, 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 [3]. 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/toxflow/>, 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 the genes or proteins and the rows contain the samples. Additionally, the first row must contain the names of genes or proteins and the first column 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

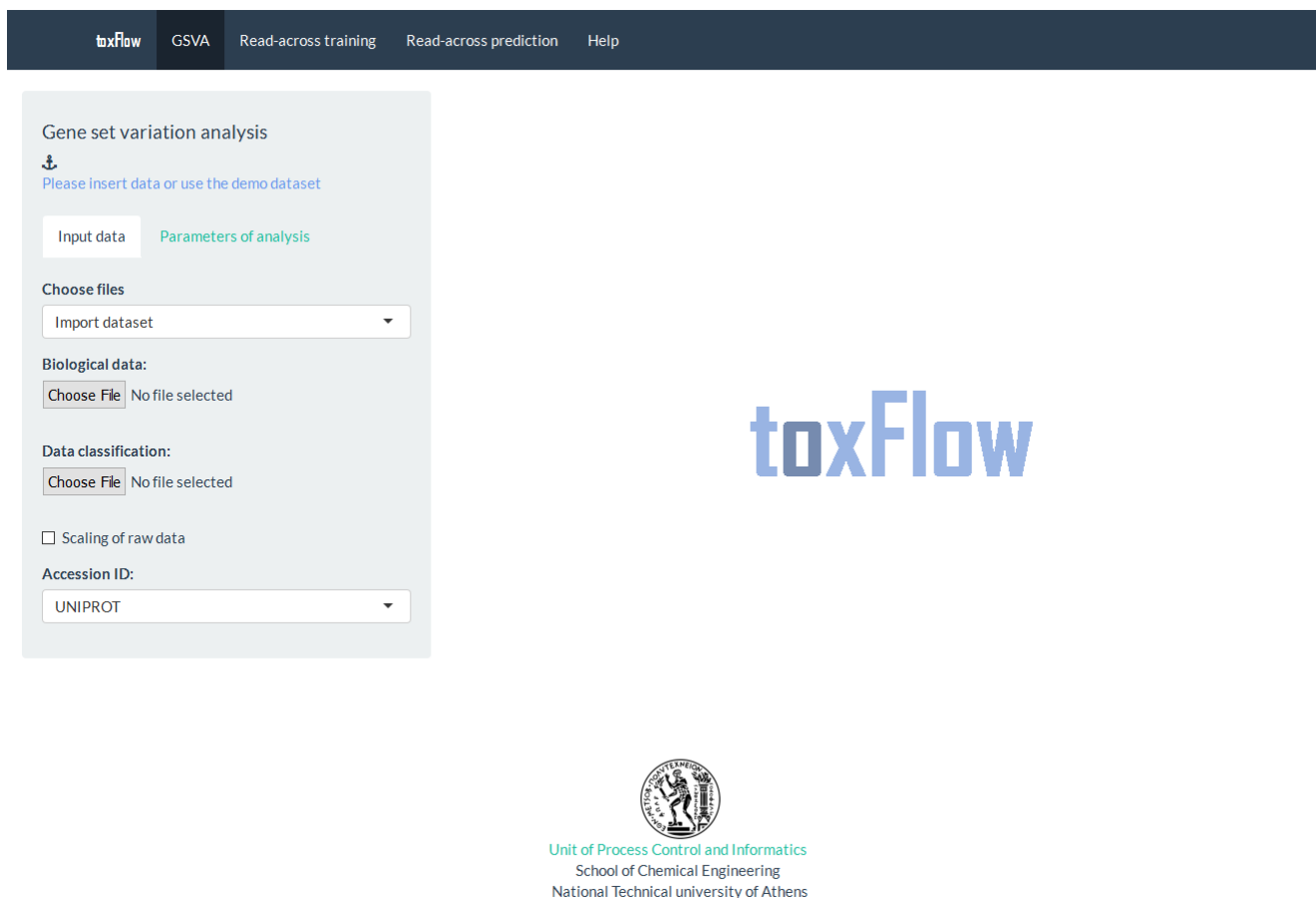


Figure 1: User interface of toxFlow application. On the top ribbon are shown the three tabs (each for every main part).

user can see an example of the analysis for anionic-cationic classification (Fig. 3A). The dataset comes from Walkey *et al.* (2014) published article [4] 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:

$$c_{sc} = \frac{c_{in} - min}{max - min} \quad (1)$$

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.

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*

Gene set variation analysis

A. [Please insert data or use the demo dataset](#)

B. **Input data** Parameters of analysis

C. Choose files

Import dataset | Import dataset | Use demo dataset

Data classification:

Choose File No file selected

D. ☐ Scaling of raw data

E. Accession ID:

UNIPROT | UNIPROT | REFSEQ | ENTREZID | SYMBOL

Gene set variation analysis

F. **Input data** Parameters of analysis

G. Gene set collection:

C5: GO MF gene sets (from MSigDB) | C5: GO MF gene sets (from MSigDB) | Chemical-GO enriched associations (from Comparative Toxicogenomics Database) | Other...

Adjusted p-value cut off:

0.001 | 0.001 1.000 10.000 | 0.05

GO mapping for acyclic graph:

GOMFPARENTS | GOMFPARENTS | GOMFCHILDREN

H. Run analysis

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.

A. Number of bootstraps:

1

B. Gene set size:

1 6000

Adjusted p-value cut off:

C. 0.001 | 0.001 0.006 0.011 0.016 0.021 0.026 0.031 0.036 0.041 0.046 0.05 | 0.05

GO mapping for acyclic graph:

D. GOMFPARENTS | GOMFPARENTS | GOMFCHILDREN

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.

	A	B	C	D	E	Genes/Proteins G		H	I	J	K	L
1	Samples	P01024	P01834	P0C0L4	P02647	P01871	P01857	P02649	B9A064	P10909	P01876	P04004
2	G15.AC	0.073492	0.011657	0.027369	0.004562	0.016726	0.014698	0.085656	0.003548	0.012671	0.008109	0.033452
3	G15.AHT	0.021657	0.015957	0.055091	0.007979	0.018237	0.015578	0.0019	0.00266	0.005319	0.011398	0.021657
4	G15.Ala.SH	0.081059	0.005618	0.017657	0.005618	0.008026	0.00321	0.107544	0.001605	0.008828	0.005618	0.085072
5	G15.Asn.SH	0.084719	0.011876	0.015835	0.007126	0.015044	0.004751	0.102138	0.005542	0.005542	0.006334	0.071259
6	G15.AUT	0.024524	0.014799	0.054968	0.012685	0.01649	0.006342	0.012262	0.002537	0.008034	0.009302	0.09556
7	G15.CALNN	0.043651	0.005952	0.034392	0.008598	0.006614	0.01455	0.010582	0.003307	0.007937	0.007937	0.054233
8	G15.CIT	0.127462	0.028993	0.03884	0.018053	0.011488	0.032276	0.012035	0.013676	0.007659	0.019694	0.037199
9	G15.CTAB	0.091865	0.019635	0.074334	0.047686	0.009116	0.021038	0.015428	0.009116	0.010519	0.021038	0.036466
10	G15.DDT.BDHDA	0.10322	0.039773	0.026515	0.047348	0.03125	0.026515	0.015152	0.005682	0.023674	0.018939	0.018939
11	G15.DDT.CTAB	0.09299	0.022175	0.026466	0.05794	0.017167	0.011445	0.015737	0.005722	0.044349	0.023605	0.030043
12	G15.DDT.DOTAP	0.043282	0.025194	0.033592	0.013566	0.023902	0.016796	0.004522	0.005814	0.010982	0.018088	0.160853
13	G15.DDT.ODA	0.044053	0.013216	0.008811	0.0837	0.013216	0.030837	0.008811	0.013216	0.070485	0.017621	0.017621
14	G15.DDT.SA	0.125887	0.026005	0.018913	0.06383	0.026005	0.026596	0.046099	0.013002	0.088652	0.015957	0.042553
15	G15.DDT.SDS	0.107543	0.018671	0.044063	0.045556	0.014937	0.024645	0.023898	0.005228	0.035848	0.008962	0.023152
16	G15.DTNB	0.036288	0.008328	0.007733	0.005949	0.005949	0.008328	0.033314	0.00238	0.002974	0.006544	0.074361
17	G15.F127	0.085791	0.016086	0.042002	0.02681	0.013852	0.013405	0.018767	0.005809	0.023682	0.025022	0.021895

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. 2). 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.

## 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

	A	B
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

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-(phosphonoxy)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-propanol 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.

contain the values of physicochemical descriptors (samples in rows and descriptors in columns). The second one (**Biological data**) must contain the samples of an expression data set (samples in rows and genes or proteins in columns). Both files should include in the first column and in the first row the names of the samples and the indices respectively and should have the same number of rows (same samples). Also, both files must contain the values of the toxicity index, which will be predicted by the model, in the second column (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 [4]. For defining similarity physicochemical descriptors and protein corona composition data were used, while cell association was used as the end-point.

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

	Toxicity index			Descriptors							
	A	B	C	D	E	F	G	H	I	J	K
1		net.c	class	lspri_synt	lspri_seru	lspri_rel	lspri_diff	lspri_rel_c	zav_synth	zav_serun	pdi_synt
2	G15.AC	-5.1839	1	0.18253	0.454404	2.489473	0.271874	1.489473	22.36	57.53	0.084
3	G15.AHT	-1.00854	0	0.45821	0.525747	1.147394	0.067537	0.147394	30.95	90.06	0.399
4	G15.Ala-S	-5.50439	1	0.223534	0.274761	1.22917	0.051227	0.22917	22.64	44.43	0.147
5	G15.Asn-S	-5.67669	1	0.27362	0.327264	1.196055	0.053645	0.196055	23.09	37.75	0.15
6	G15.AUT	-1.31567	0	0.365436	0.389573	1.066051	0.024138	0.066051	23.8	55.98	0.326
Samples	G15.CALN	-7.13797	1	0.20691	0.265327	1.282332	0.058417	0.282332	25.22	38.8	0.144
	G15.CIT	-5.41982	1	0.210431	0.292836	1.391602	0.082405	0.391602	18.65	54.03	0.138
	G15.CTAB	-5.86229	0	0.326142	0.365811	1.121632	0.039669	0.121632	15.6	59.7	0.465
	G15.DDT@	-7.29449	0	0.266579	0.317134	1.189643	0.050555	0.189643	23.15	47.03	0.187
	G15.DDT@	-7.59005	0	0.275461	0.324751	1.178939	0.049291	0.178939	20.53	48.4	0.191
	G15.DDT@	-1.12756	0	0.276498	0.297223	1.074954	0.020725	0.074954	28.17	47.34	0.193
	G15.DDT@	-6.1218	0	0.309989	0.367331	1.184981	0.057342	0.184981	33.6	58.95	0.268
	G15.DDT@	-6.8039	1	0.395907	0.320779	0.810238	-0.07513	-0.18976	82.41	59.93	0.249
	G15.DDT@	-7.67595	1	0.465011	0.359906	0.773974	-0.1051	-0.22603	27.94	100.13	0.302

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

	Toxicity Index			Genes/Proteins				
	A	B	C	D	E	F	G	H
1		net.c	P01024	P01834	P0C0L4	P02647	P01871	P01857
2	G15.AC	-5.1839	0.073492	0.011657	0.027369	0.004562	0.016726	0.014698
3	G15.AHT	-1.00854	0.021657	0.015957	0.055091	0.007979	0.018237	0.015578
4	G15.Ala.SH	-5.50439	0.081059	0.005618	0.017657	0.005618	0.008026	0.00321
5	G15.Asn.SH	-5.67669	0.084719	0.011876	0.015835	0.007126	0.015044	0.004751
Samples	G15.AUT	-1.31567	0.024524	0.014799	0.054968	0.012685	0.01649	0.006342
	G15.CALNN	-7.13797	0.043651	0.005952	0.034392	0.008598	0.006614	0.01455
	G15.CIT	-5.41982	0.127462	0.028993	0.03884	0.018053	0.011488	0.032276
	G15.CTAB	-5.86229	0.091865	0.019635	0.074334	0.047686	0.009116	0.021038
	G15.DDT.BDHDA	-7.29449	0.10322	0.039773	0.026515	0.047348	0.03125	0.026515
	G15.DDT.CTAB	-7.59005	0.09299	0.022175	0.026466	0.05794	0.017167	0.011445
	G15.DDT.DOTAP	-1.12756	0.043282	0.025194	0.033592	0.013566	0.023902	0.016796
	G15.DDT.ODA	-6.1218	0.044053	0.013216	0.008811	0.0837	0.013216	0.030837
	G15.DDT.SA	-6.8039	0.125887	0.026005	0.018913	0.06383	0.026005	0.026596

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

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 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 two distance metrics,

Read-across training of model, using leave-one-out cross-validation method

Please insert data or use the demo dataset

A. Input data Parameters of analysis

Choose files

B. Import dataset

Import dataset

Use demo dataset

Biological data:

Choose File No file selected

☐ Only one file available

☐ Scaling of physicochemical data

C. ☐ Scaling of biological data

D. ☐ Use of differentially expressed genes or proteins from GSVA analysis

Read-across training of model, using leave-one-out cross-validation method

Please insert data or use the demo dataset

Input data Parameters of analysis

Choose files

Import dataset

Physicochemical data:

Choose File No file selected

E. ☒ Only one file available

Available file:

F. ☒ Physicochemical file

☐ Biological file

☐ Scaling of physicochemical data

☐ Use of differentially expressed genes or proteins from GSVA analysis

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.

are transformed to similarity measures according to the equation 6, in order to take values between 0 and 1. The user can define the physicochemical and biological threshold that control the selection of neighbouring 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 read across values are calculated according to eq 7. The user can change the similarity method, the calculation base and the two thresholds and the results will be updated automatically.

Similarity calculation method:

Cosine similarity

Cosine similarity

Manhattan distance

Euclidean distance

Figure 10: Similarity calculation measure selection

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

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

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

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

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

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

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.

$$similarity_{A,B} = \begin{cases} CS_{A,B}, & \text{for cosine similarity} \\ \frac{1}{1+MD_{A,B}} & \text{for Manhattan distance} \\ \frac{1}{1+ED_{A,B}} & \text{for Euclidean distance} \end{cases} \quad (6)$$

$$RAV = \frac{\sum_{i=1}^{L_{ref}} (rel_{i,ref} \times tox_i)}{\sum_{i=1}^{L_{ref}} rel_{i,ref}} \quad (7)$$

where,  $ref$ , the reference NP,

$L_{ref}$ , the total number of neighbors for the reference NP,

$i$ , the  $i^{th}$  neighbor of the reference NP,

$tox_i$ , the known toxicity end-point value of the  $i^{th}$  neighbor and

$rel_{i,ref}$ , the weighting factor corresponding to the  $i^{th}$  neighbor of the reference NP.

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).



Read-across training of model, using leave-one-out cross-validation method

[Please insert data or use the demo dataset](#)

**A. Input data** Parameters of analysis

Choose physicochemical attribute filtering level

0.05 0.24 0.7

0.05 0.12 0.19 0.26 0.33 0.4 0.47 0.54 0.61 0.68 0.7

**B. Choose biological attribute filtering level**

0.05 0.3 0.7

0.05 0.12 0.19 0.26 0.33 0.4 0.47 0.54 0.61 0.68 0.7

Similarity calculation method:

**C.** Cosine similarity

Physicochemical threshold:

0.5

**D.** Biological threshold:

0.5

**E. Prediction base on:**

☒ Physicochemical data

☐ Biological data

**F.** Training

**Nanoparticle's universe**

**G. Reference nanoparticle:**

1

**H. ☒ Nanoparticles' diameter**

Choose File No file selected

**I. ☒ Nanoparticles' classification**

Choose File No file selected

**J. Physicochemical thresholds:**

0.1 0.5

**K. Biological thresholds:**

0.1 0.5

**L.** Visualization

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

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

### 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 were 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.

The screenshot shows the 'Toxicity endpoint prediction' interface. It is divided into two main sections. The left section, titled 'Toxicity endpoint prediction', contains a message: 'The predictions can be generated after read-across training has been performed'. Below this, there are two sections: 'A. Physicochemical data:' with a 'Choose File' button and 'No file selected' text, and 'B. Biological data:' with a 'Choose File' button and 'No file selected' text. At the bottom of this panel is a 'C. Prediction' button. The right section, titled 'Nanoparticle's universe', contains several options: 'D. Reference nanoparticle:' with a dropdown menu showing '1'; 'E.' with two checkboxes, 'Nanoparticles' diameter' and 'Nanoparticles' classification'; 'Physicochemical thresholds:' with two sliders, 'F.' and 'G.', both set to '0.1' and '0.5'; 'Biological thresholds:' with two sliders, 'F.' and 'G.', both set to '0.1' and '0.5'; and 'H. Visualization' button.

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).

Download

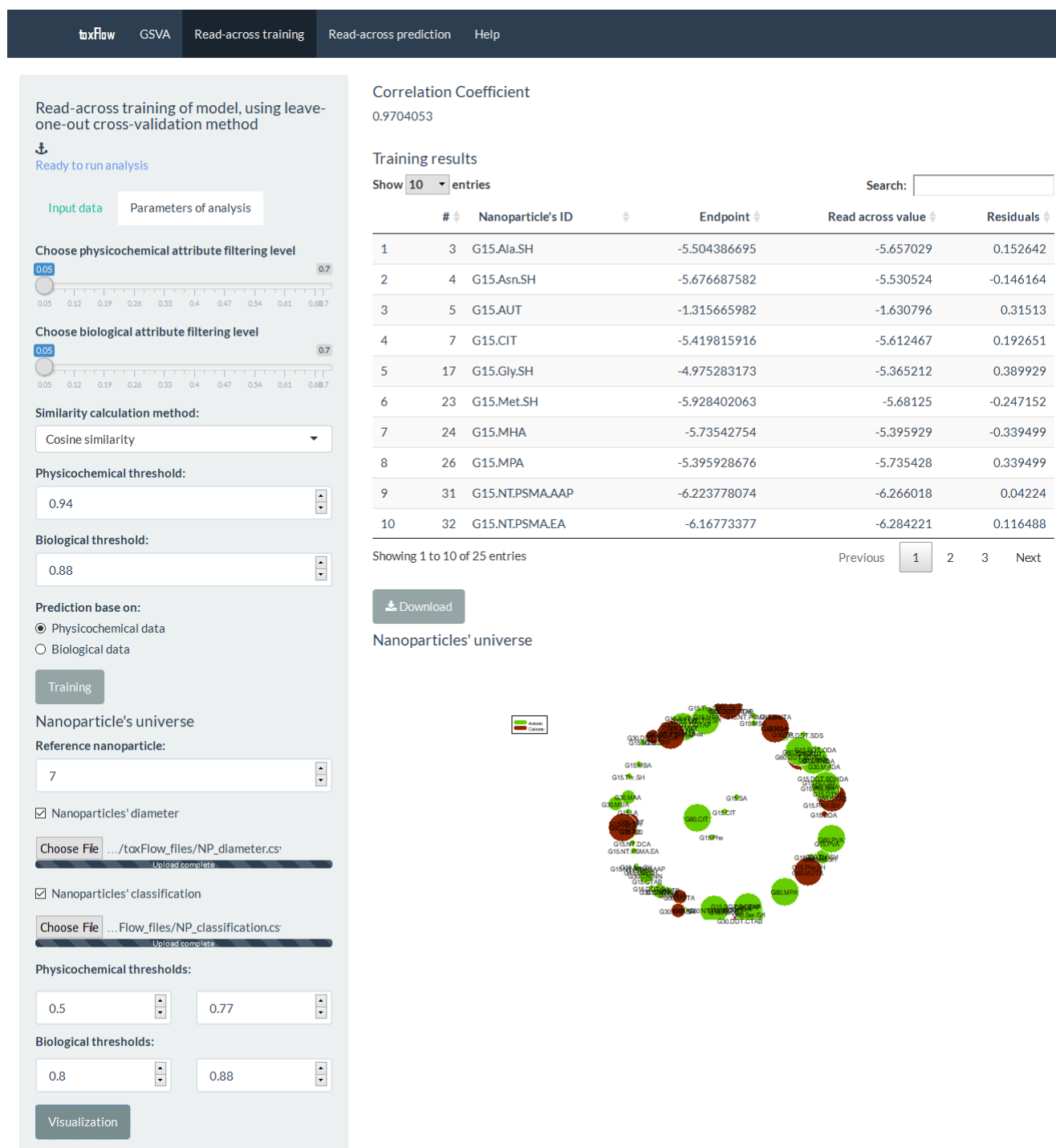
Antonie  
Culic



Download



12





## 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] 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), 24392455, 2014, Available online in: [https://www.researchgate.net/publication/263941898\\_Protein\\_Corona\\_Fingerprinting\\_Predicts\\_the\\_Cellular\\_Interaction\\_of\\_Gold\\_and\\_Silver\\_Nanoparticles](https://www.researchgate.net/publication/263941898_Protein_Corona_Fingerprinting_Predicts_the_Cellular_Interaction_of_Gold_and_Silver_Nanoparticles)
- [5] J. D. Storey, R. Tibshirani, *Statistical significance for genomewide studies*, PNAS, vol.100, no. 16, 2003