JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Application Note

Cite This: J. Chem. Inf. Model. 2018, 58, 543−549

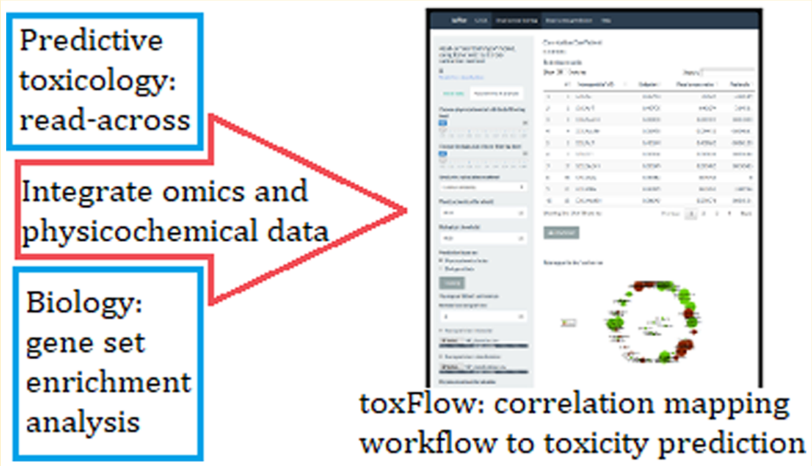
pubs.acs.org/jcim

toxFlow: A Web-Based Application for Read-Across Toxicity Prediction Using Omics and Physicochemical Data

Dimitra-Danai Varsou,[*,†,§] Georgia Tsiliki,[†,§] Penny Nymark,[‡,¶] Pekka Kohonen,[‡,¶] Roland Grafström,[‡,¶] and Haralambos Sarimveis[*,†]

[†]School of Chemical Engineering, National Technical University of Athens, 157 80 Athens, Greece
[‡]Institute of Environmental Medicine, Karolinska Institutet, SE-171 77 Stockholm, Sweden
[¶]Misvik Biology Oy, 20520 Turku, Finland

Supporting Information


Predictive toxicology: read-across
Integrate omics and physicochemical data
Biology: gene set enrichment analysis
toxFlow: correlation mapping workflow to toxicity prediction

ABSTRACT: We present toxFlow, a web application developed for enrichment analysis of omics data coupled with read-across toxicity prediction. A sequential analysis workflow is suggested where users can filter omics data using enrichment scores and incorporate their findings into a correlation-based read-across technique for predicting the toxicity of a substance based on its analogs. Either embedded or in-house gene signature libraries can be used for enrichment analysis. The suggested approach can be used for toxicity prediction of diverse chemical entities; however, this article focuses on the multiperspective characterization of nanoparticles and selects their neighbors based on both physicochemical and biological similarity criteria. In addition, visualization options are offered to interactively explore correlation patterns in the data, whereas results can be exported for further analysis. toxFlow is accessible at http://147.102.86.129:3838/toxflow.

INTRODUCTION

Even though the use of nanomaterials is extended to various applications, current research on nanoparticles (NPs) raises awareness concerning their possible toxic effects on human health.[1] Over the past few years, the popular quantitative structure−activity relationship (QSAR) modeling methodologies for pure chemicals have found successful applications in predicting NPs toxicity. These methodologies, now widely known as nanoQSARs, aim to predict end-points of NPs in a cost- and time-effective way, whilst avoiding in vivo experiments.[2−8] The development of a reliable nanoQSAR model relies on high quality experimental data and also on the availability of sufficiently large and diverse data sets. When the training data set is limited, it is not possible to develop an appropriately validated (nano)QSAR model for predicting quantitative relationships between structures and properties.[9]

In cases of limited data, an alternative nontesting approach for assessing the toxicity of NPs is read-across, a methodology which predicts adverse effects of nontested NPs using information from a single or a group of NPs with similar characteristics. The read-across approach is conceptually similar to connectivity mapping[10] and one that the European Chemicals Agency has recently focused on to clearly and consistently define via the Read-Across Assessment Framework.[11] Unlike conventional chemicals where grouping and read-across are mainly based on structural characteristics, NPs can be characterized using many different aspects such as routes of exposure, material types (e.g., fullerenes, carbon nanotubes, metal oxides, etc.), physicochemical characteristics (e.g., size, shape, surface area, solubility, etc.), biophysical interaction, and biological impact (e.g., protein and lipid corona formation, gene expressions, cellular and organ responses), lifecycle analysis, and biokinetics properties. The plurality of options

Received: March 20, 2017
Published: December 27, 2017

© 2017 American Chemical Society

DOI: 10.1021/acs.jcim.7b00160

however imposes a complexity in the NP characterization, which is the reason why the nanocommunity has only recently begun to rely on grouping and read-across methods for predicting properties of NPs.[9,12−14]

A particular area of interest is how to incorporate biological pathway information to the read-across toxicity prediction and establish the biological similarity of a group of compounds that share structural similarities, have similar toxicological properties etc. For doing so, enrichment computational techniques, already popular in omics analysis, could be employed to regroup the activated functions in biological categories and associate them with already known functional groups, such as biological pathways or gene set collections.[15−18] Gene Set Enrichment Analysis (GSEA)[17] and Gene Set Variation Analysis (GSVA)[18] are popular enrichment computational techniques.

In this work we are proposing a novel read-across approach for defining similar compounds (analogues) to the target NP, whose end-point value is missing. NPs are selected as neighbors to the target NP only if they satisfy two similarity criteria, defined by physicochemical and biological characteristics. Omics enrichment analysis is employed to filter the data and regroup the activated functions in biological categories in order to associate them with already known functional groups, such as biological pathways or gene set collections. Therefore, as a general method, our approach enables model-based connectivity mapping using biological prior information to improve interpretation and increase robustness of results.

The suggested methodology is offered as a web-tool, named toxFlow (http://147.102.86.129:3838/toxflow). toxFlow handles omics and physichochemical data, either jointly or separately, and allows for data scaling, visualization, and exporting of findings. Its applications can be easily extended to toxicity prediction of diverse chemical entities.

■ METHODS

New Approach for the Read-Across Framework. In the present work, the toxicity end-point prediction of the target NP is performed using the weighted average of the corresponding values of the neighbor NPs, i.e. neighboring substances of every target NP are selected by calculating pairwise similarity measures with all available substances separately for physicochemical and biological descriptors and by excluding those substances for which one or both similarity measures do not fulfill predefined thresholds. The proposed workflow assumes that a training data set is available, i.e., a set of NPs for which the toxicity end-point values are known. The complete workflow consists of a number of steps presented analytically below and shown schematically in Figure 1. Raw data could be scaled prior to the analysis, then physicochemical and omics descriptors are filtered based on their correlation to the toxicity end-point. In the resulting data sets, pairwise similarity calculations are used to first select neighboring substances and then predict the unknown toxicity value using read-across.

Data Preprocessing. Input data can be scaled between 0 and 1 in order to be comparable and contribute equally to the read-across predictions. Scaling is optional but is highly recommended when the data exhibit substantially different numerical ranges among attributes. In addition, physicochemical and biological attributes are filtered according to their correlation to the toxicity index. Attributes with correlation less than a certain threshold are excluded from the rest of the analysis.

Gene Set Variation Analysis. With omics data, attributes are often filtered by applying computational techniques that regroup the activated functions in biological categories and associate them with already known functional groups or gene set collections, in order to achieve biological interpretation of findings.[17] In the suggested workflow we are employing GSVA, an enrichment method aiming to estimate over- or underexpressed genes in omics experiments accumulating them into specific functional groups (e.g., biological pathways, Gene Ontology ids). In that way the information of interest is summarized from single gene expression profiles to specific groups of biological categories. GSVA is a special case of enrichment analysis, reported to be quite robust, given that it can detect imperceptible changes in the variation of the pathway activity over a sample population.[18]

As analytically described in the article of Hänzelmann et al.,[18] GSVA can be utilized with a table of normalized expression data and a gene set collection. Gene sets from the collection that do not satisfy a predefined gene set size (number of included genes) are excluded from the analysis. To bring the different expression profiles into the same range, the kernel estimation statistic of the cumulative density function is calculated and the statistics for each sample are normalized in order to gain symmetric values around zero. Finally, the GSVA score is calculated based on the enrichment score (ES), a Kolmogorov−Smirnov like random walk statistic. Bootstrapping and random sampling techniques are employed, by default to the 63.2% of the genes, in order to eliminate technical noise from experimental analysis and produce most reliable results. This procedure is repeated several times until convergence to the statistically significant gene sets is achieved. The latter is evaluated similarly to the GSVA workflow where a linear model is fitted to the GSVA scores data together with a Bayesian empirical model to estimate the significant gene sets based on adjusted Benjamini and Hochberg p-values. The filtered biological data includes only those genes involved in the statistically significant gene sets.

Definition of Neighbors. Similarity measures are calculated between all substances separately for the available physicochemical and filtered biological data. For each NP in the available set, the substances for which both similarity measures satisfy predefined corresponding thresholds are selected as neighbors. Therefore, in order to characterize two NPs as similar they need to satisfy both physicochemical and biological similarity criteria. Particularly, we consider three options for defining the similarity measure, i.e. the cosine similarity (eq 1), the Manhattan distance (eq 2), and the Euclidean distance (eq 3).[19]

$$\mathrm{CS}_{A,B} = \frac{\sum_{i=1}^{N} A_i B_i}{\sqrt{\sum_{i=1}^{N} A_i^{\,2}}\,\sqrt{\sum_{i=1}^{N} B_i^{\,2}}} \tag{1}$$

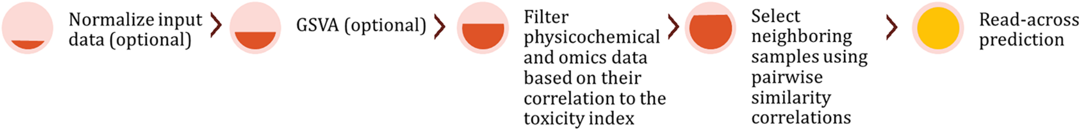

Normalize input data (optional)
GSVA (optional)
Filter physicochemical and omics data based on their correlation to the toxicity index
Select neighboring samples using pairwise similarity correlations
Read-across prediction

Figure 1. Steps of the proposed read-across workflow.

$$\mathrm{MD}_{A,B} = \sum_{i=1}^{N} |A_i - B_i| \tag{2}$$

$$\mathrm{ED}_{A,B} = \sqrt{\sum_{i=1}^{N} (A_i - B_i)^2} \tag{3}$$

where, A and B are vectors containing the physicochemical or biological descriptors of the two NPs on which the similarity measure is computed; A_i, the ith element of vector A; B_i, the ith element of vector B; N, the size of A and B vectors.

The two distance metrics are transformed to similarity measures according to eq 4 shown below, ensuring that all similarity values range between 0 and 1.

$$\mathrm{sim}_{A,B} = \begin{cases} \mathrm{CS}_{A,B} & \text{for cosine similarity} \\ \dfrac{1}{1 + \mathrm{MD}_{A,B}} & \text{for Manhattan distance} \\ \dfrac{1}{1 + \mathrm{ED}_{A,B}} & \text{for Euclidean distance} \end{cases} \tag{4}$$

Read-Across Prediction. The predicted read-aross value (RAV) for a reference NP in the training set is calculated using only its selected neighbors as shown in eq 5, where the weighting factors are based on either physicochemical or biological similarity measures.

$$\mathrm{RAV} = \frac{\sum_{i=1}^{L_{\mathrm{ref}}} (\mathrm{sim}_{i,\mathrm{ref}} \times \mathrm{tox}_i)}{\sum_{i=1}^{L_{\mathrm{ref}}} \mathrm{sim}_{i,\mathrm{ref}}} \tag{5}$$

where, ref is the reference NP; L_{ref}, the total number of neighbors for the reference NP; i, the ith neighbor of the reference NP; tox_i, the known toxicity end-point value of the ith neighbor; $\mathrm{sim}_{i,\mathrm{ref}}$, the weighting factor (similarity) corresponding to the ith neighbor of the reference NP, as defined in eq 5.

Read-Across Validation. In order to validate the performance of the proposed read-across method, an internal Leave-One-Out (LOO) scheme is executed. In each iteration of the LOO method, one example is excluded from the training set and is used as a reference NP, i.e., its end-point value is predicted by applying the proposed read-across method, where neighbors are selected from the rest of the training data. After all iterations have been completed the read-across predictions are compared with actual end-point values using the $R_{\mathrm{LOO}}{}^2$ statistical index to assess the clarity of the calculations.

toxFlow Application. toxFlow has been implemented using R (https://www.r-project.org) and the shiny R package (http://shiny.rstudio.com) for developing web applications. The source code together with a user guide and a video tutorial are available at https://github.com/DemetraDanae/toxFlow (DOI: 10.5281/zenodo.1107622).

toxFlow is a user-friendly tool for performing both GSVA and toxicity prediction using the proposed read-across technique, without any need of prior computational skills. The application consists of three parts. The first two can be used independently, involving data filtering and read-across training respectively (Figure 2), whereas the third part provides read-across toxicity predictions after read-across training has been performed. toxFlow was built for NPs toxicity evaluation and risk assessment, but its use can be extended for the toxicity prediction of other chemical substances and compounds.

For the GSVA part of the analysis, the user first uploads the input data and then defines the parameters for performing the analysis.

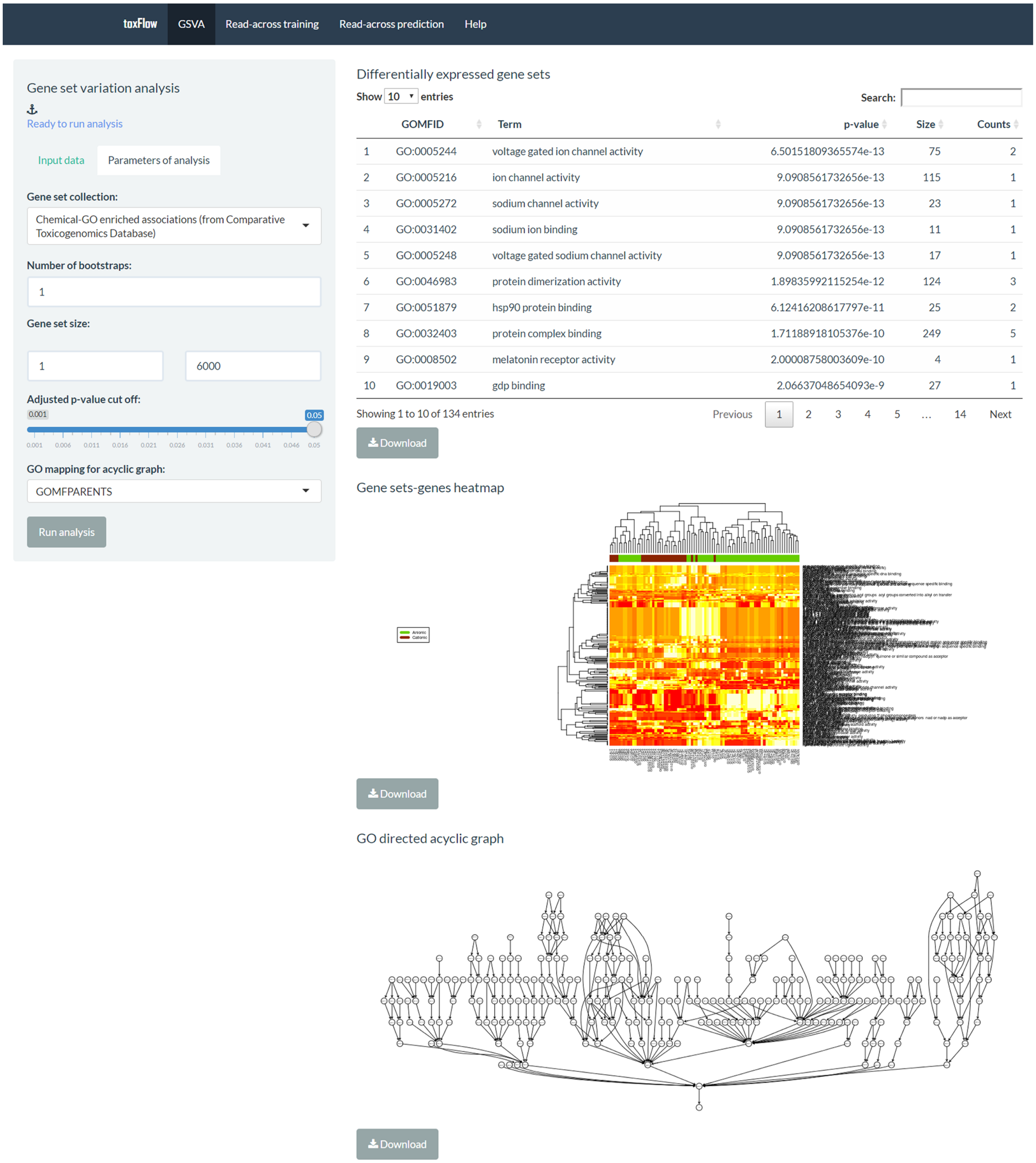
- Input data are uploaded in two csv files: the first file contains omics data for a number of NPs and the second file includes the classification of the same NPs into categories, which is a requirement of the GSVA method.[18] The user should select the Accession ID of the gene or protein names and has the option of scaling the data.
- For the gene set information, the user first chooses from a dropdown menu either a custom-made signature library or one of the two available gene signature libraries associated with toxicity, taken from the Molecular Signature database[20] and the Comparative Toxicogenomics Database,[21] respectively. Additional options are the number of bootstrap iterations to be performed by GSVA, the maximum and minimum size of the resulting gene sets, and the threshold of the adjusted p-value which controls the selection of the significant gene sets.

The application produces a table with all statistically significant gene sets providing links to available information whenever possible. Additionally, a heatmap is produced where the user can compare in color code the calculated ES for each gene set derived from the GSVA function and observe any patterns between the samples and their corresponding classes. Finally, a corresponding acyclic graph is produced to depict the hierarchic relationships between the significant gene sets.

Although the read-across method proposed in this paper requires the availability of both physicochemical and omics data, the toxFlow application gives the option of performing read-across predictions using only physicochemical or biological information. Similarly to GSVA, input data are first uploaded to the application, followed by the definition of the parameters for the read-across analysis.

- Input data are uploaded in one or two csv files depending on the availability of physicochemical/biological information. Both files contain experimental values of the toxicity end-point of interest. When biological data are uploaded, the user has the option to scale them and/or filter them based on the results from GSVA.
- In the "Parameters of analysis" page the user first defines minimum bounds on correlations between individual physicochemical or biological descriptors and toxicity values. If the correlation does not attain this bound, the descriptor is excluded from the rest of the analysis. The user selects the metric for computing pairwise similarity between NPs and sets thresholds on physicochemical and biological similarity measures that define whether two NPs will be considered as neighbors. Finally, the user can choose if the read-across predictions will be calculated using physicochemical or biological similarities (eq 4).

Then the LOO "training" method, as described in the Methods section, is applied. It returns the coefficient of determination ($R_{\mathrm{LOO}}{}^2$) value, a table with all substances on which read-across toxicity end-point prediction is possible, given the thresholds applied, and graphical representations of the substances with their neighbors (NPs' "universe"). To better visualize the NPs' universe, the user could upload two files containing a classification and the sizes of the training NPs. This information defines the color code and the diameter of the circles representing neighboring NPs in the produced figures. The user can optimally tune the parameters of the method (the correlation levels and the physicochemical and biological thresholds) with the objective of



Gene set variation analysis

Ready to run analysis

Gene set collection: Chemical-GO enriched associations (from Comparative Toxicogenomics Database)

Number of bootstraps: 1

Gene set size: 1 6000

Adjusted p-value cut off: 0.001 0.05

GO mapping for acyclic graph: GOMFPARENTS

Differentially expressed gene sets

Show 10 entries Search:

	GOMFID	Term	p-value	Size	Counts
1	GO:0005244	voltage gated ion channel activity	6.50151809365574e-13	75	2
2	GO:0005216	ion channel activity	9.0908561732656e-13	115	1
3	GO:0005272	sodium channel activity	9.0908561732656e-13	23	1
4	GO:0031402	sodium ion binding	9.0908561732656e-13	11	1
5	GO:0005248	voltage gated sodium channel activity	9.0908561732656e-13	17	1
6	GO:0046983	protein dimerization activity	1.89835992115254e-12	124	3
7	GO:0051879	hsp90 protein binding	6.12416208617797e-11	25	2
8	GO:0032403	protein complex binding	1.71188918105376e-10	249	5
9	GO:0008502	melatonin receptor activity	2.00008758003609e-10	4	1
10	GO:0019003	gdp binding	2.06637048654093e-9	27	1

Showing 1 to 10 of 134 entries Previous 1 2 3 4 5 … 14 Next

Gene sets-genes heatmap

GO directed acyclic graph

Figure 2. toxFlow interface for the GSVA analysis: all parameters involved in the analysis can be seen in the left gray area. Results are presented in the right-hand side and include a table with all statistically significant gene sets, along with the corresponding heatmap and acyclic graph. Both the heatmap and the acyclic graph can be downloaded in HQ images for detailed examination.

achieving a high (close to 1) R_{LOO}^2, which means that the model can be used for computing predictions for NPs with unknown end-point toxicity values.

After a satisfactory read-across model has been produced, the user can upload csv files containing physicochemical and/or biological data for NPs with unknown end-point values. The data should be compatible with the training data, i.e., they should contain the same physicochemical or biological descriptors. The application produces a table containing the predicted value of toxicity index for all NPs with unknown toxicity, along with their universe diagram. Additional information is included in the detailed user manual of the toxFlow application.

■ CASE STUDY

The proposed read-across method and the toxFlow application are demonstrated on data derived from the recent publication of Walkey et al.[22] that consists of omics data in the form of 129 protein corona fingerprints (PCF) for 84 culture medium incubated gold anionic and cationic NPs (diameters of 15, 30, and 60 nm), their physicochemical descriptors, and additional measurements of their cell association with human A549 cells. In order to predict NPs toxicity via cellular association, Walkey et al.[22] developed a multivariate linear model that combines the full PCF data set along with the available physicochemical data. As a filtering step, they performed a partial least-squares regression (PLSR) technique to exclude weakly correlated attributes to the cell association. The combined model predicts cell association with ${R_{LOO}}^2 = 0.86$, whereas a follow-up study reported $R^2 = 0.895$ for a 4-fold cross-validation scheme and a limited set of descriptors.[23] Papa et al.[24] and Tsiliki et al.[25] both reported better performances compared to the original study by employing linear and nonlinear regression models to the PCF set. In terms of read-across prediction, Helma et al.[26] applied the lazar read-across framework with their best performance being $R^2 = 0.68$ for random forest models and a 10-fold cross-validation scheme. Nevertheless, their findings are not directly compared with the results reported from others, as they considered an augmented data set containing both gold and silver NPs. Here, we present a novel methodology in the toxicogenomics field; our method integrates biological information to the data, and by that filter the data according to its biological relevance to increase prediction performance.

The application of GSVA analysis utilized NPs classification into anionic/cationic samples, using the *CTD Disease-GO molecular function associations* gene set collection from the Comparative Toxicogenomics Database (*CTD-MF*) and selecting only the gene sets consisting of 1−6000 genes, which resulted in 134 statistically significant genes sets (p-value ≤ 0.05). Figure 2 presents part of the analysis results. (The omics and classification data used for the GSVA analysis are included in the files bio_gsva.csv and NP_classification.csv, respectively.)

We proceeded with performing the proposed read-across method in order to predict the cell association, using the training data included in the files phChem_readAcross.csv (available physicochemical descriptors) and bio_readAcross.csv (containing biological information). Both the physicochemical descriptors and the proteomics data were filtered according to their correlations to the toxicity index, using a common threshold of 0.05. Read-across predictions were calculated using either physicochemical ("PhChem") or the biological ("bio") weighting factors in eq 5 and all the similarity metrics options. The results were optimized in terms of the threshold values on both physicochemical and biological similarities that define the group of neighbors for each query NP. The objective was to maximize the ${R_{LOO}}^2$ statistic, i.e., increase the confidence of accepting predictions, at the expense of obtaining predictions only for a subset of NPs, for which at least one neighbor exists. In order to emphasize how the GSVA filtering influences read-across results, we performed the read-across training with and without GSVA filtering. Table 1 summarizes the optimized read-across validation results without GSVA filtering. For each similarity metric, it depicts the optimal threshold values, the number of NPs on which read-across prediction is possible, that is NPs with at least one neighbor, and the respective ${R_{LOO}}^2$ statistic. Reported values are as high as 0.9704.

Next, we incorporated the GSVA results as described above to the read-across training procedure. Omics data were filtered according to the proteins included in the significant gene sets produced by the previous GSVA, and the results are presented in Table 2. In order to test the sensitivity of the method to varying thresholds, we changed independently the thresholds by ±5% of their optimal value. The results of the sensitivity analysis are also presented in Table 2. It should be emphasized that the resulting ${R_{LOO}}^2$ statistic must be evaluated in relation to the number of NPs for which a read-across prediction is possible. We observed that the results are not very sensitive in terms of the ${R_{LOO}}^2$ statistic. As expected, increasing the values of the thresholds results in having fewer NPs with at least one neighbor for which read-across prediction can be performed. We also observe that consideration of the GSVA findings improved the results of the Manhattan and Euclidean similarity metrics but deteriorated the accuracy of the cosine similarity metric.

Table 1. Results on Read-Across Toxicity Prediction without GSV Analysis[a]

	thresholds			${R_{LOO}}^2$	
method	PhChem	Bio	no. of NPs	PhChem	Bio
cosine sim	0.94	0.88	25	0.9704	0.9703
Manhattan dist	0.2	0.19	11	0.9299	0.9307
Euclidean dist	0.52	0.5	14	0.9393	0.9406

[a] For each similarity measure, the table presents the results obtained after optimally selecting the physicochemical and biological thresholds.

Table 2. Results on Read-Across Toxicity Prediction Based on the Previous GSV Analysis[a]

	thresholds			${R_{LOO}}^2$	
method	PhChem	Bio	no. of NPs	PhChem	Bio
cosine sim	**0.94**	**0.89**	**30**	**0.9394**	**0.9394**
	0.987	0.89	11	0.9282	0.9283
	0.893	0.89	36	0.8691	0.8674
	0.94	0.9345	11	0.9254	0.9254
	0.94	0.8455	40	0.7868	0.7865
Manhattan dist	**0.25**	**0.22**	**19**	**0.9554**	**0.9557**
	0.2625	0.22	19	0.9550	0.9554
	0.2375	0.22	20	0.9509	0.9511
	0.25	0.231	15	0.9355	0.9368
	0.25	0.209	19	0.9554	0.9557
Euclidean dist	**0.5**	**0.54**	**21**	**0.9726**	**0.9725**
	0.525	0.54	19	0.9673	0.9676
	0.475	0.54	24	0.9558	0.9555
	0.5	0.567	11	0.9301	0.9289
	0.5	0.513	26	0.9501	0.9501

[a] For each similarity measure, the best results are depicted first followed by a sensitivity analysis that altered the physicochemical and biological thresholds by ±5%.

Figure S1 presents the selection of the tuning parameters corresponding to the highest ${R_{LOO}}^2$ value and the read-across predictions for the first ten NPs in the training set. The ${R_{LOO}}^2$ value is shown on the top of the figure. A "universe" plot is also shown to visualize the neighbors of each NP, in this case the compound G15.CIT, where its neighbors are defined from the distance metric values (proximity to the central compound). For this reason, we used the files NP_diameter.csv and NP_classification.csv that contain NP diameters and classes (anionic or cationic), respectively. Their color in the figure is defined by their class and the diameter of the circle by their size. For demonstrating the prediction functionalities we used the same model to predict two unknown NPs, P1 and P2, with artificial data which are included in files phChem_prediction.csv and bio_prediction.csv. Figure S2 shows the produced results along with the universe plot for compound P1. All data files are provided in the Supporting Information.

■ CONCLUSIONS

toxFlow is a new web-based tool that implements a novel read-across method to predict toxicity related end-points of NPs. The development of toxFlow was motivated by the multiple-perspective characterization of NPs and there are plans to include additional characterization categories, such as biokinetics or exposure related features. toxFlow integrates physicochemical, omics and biology information data for read-across prediction, which, to the best of our knowledge, is presented for the first time as a methodology as well as a fully implemented workflow. The application provides a variety of chemoinformatics and omics-related services including data standardization, filtering and visualization, validation, and interpretation of predictive models giving links to external biological databases as well as exporting the findings. Most importantly, toxFlow introduces a novel workflow to filter biological data using a model-based connectivity mapping scheme which exploits biological prior information, in order to improve interpretation and increase robustness of results. The user can apply all or use only some of its functionalities depending on the availability of data, which are currently populating public databases in the nanosafety area, due to initiatives such as the NANoREG project (https://search.data.enanomapper.net/nanoreg/). For example, it can produce read-across prediction using physicochemical or biological data alone. Given its input data flexibility, toxFlow can be a useful NP risk assessment tool not only for expert modelers but also for regulators and the industry. Although toxFlow was designed for NPs, its functionalities can potentially be applied to pure organic molecules taking for example into account multiple biological assay tests. Future updates will include additional gene signature libraries from Wikipathways (http://www.wikipathways.org/) or KEGG (http://www.genome.jp/kegg/) databases, as well as enhanced data standardization and cross-validation functionalities.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.7b00160.

Supplementary Figures S1 and S2 (PDF)

The following data files were used in the toxFlow application to produce the results shown in the case study section of this paper. The data included in these files, except from the ones used for read-across prediction, are based on the supplementary information of Walkey et al.[22]

bio_gsva.csv: input file for GSVA analysis. It contains the PCF of 84 gold NPs with diameters 15, 30, and 60 nm, from LC/MS-MS proteomics data.

bio_prediction.csv: input file (biological file) for read-across prediction. It contains artificial PCF data of P1 and P2 samples.

bio_readAcross.csv: input file (biological file) for read-across training. It contains the PCF of 84 gold NPs with diameters of 15, 30, and 60 nm, from LC/MS-MS proteomics data. It also contains the toxicity index of cells association measured with human A549 cells.

NP_classification.csv: input file for GSVA analysis. It contains the classification of the 84 NPs into categories ("anionic"/"cationic"). This file is also used for the visualization (color) of the NP universe diagram.

NP_diameter.csv: input file for the visualization (size) of NPs universe diagram in read-across training. It contains the diameters of the 84 NPs.

phChem_prediction.csv: input file (physicochemical file) for read-across prediction. It contains the values of physicochemical descriptors (artificial data) of two samples P1 and P2.

phChem_readAcross.csv: input file (physicochemical file) for read-across training. It contains the 40 physicochemical descriptors of 84 gold NPs with diameters 15, 30, and 60 nm, as well as their toxicity cell association index of cell association measured with human A549 cells (ZIP)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: hsarimv@central.ntua.gr (H.S.).
*E-mail: dimitra.varsou@gmail.com (D.-D.V.).

ORCID

Georgia Tsiliki: 0000-0001-7209-3670
Haralambos Sarimveis: 0000-0002-8607-9965

Author Contributions

§D.-D.V. and G.T. contributed equally.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Mr. Angelos Valsamis for his valuable help in deploying and maintaining the web app. G.T., P.N., P.K., R.G., and H.S. have been supported in part by the FP7-eNanoMapper project no 604134.

■ LIST OF ABBREVIATIONS

NPs, nanoparticles; GSEA, Gene Set Enrichment Analysis; GSVA, Gene Set Variation Analysis; ES, enrichment score; RAV, read-across value; LOO, leave-one-out; PCF, protein corona fingerprints; PLSR, partial least-squares regression; CTD-MF, *CTD Disease-GO molecular function associations*

■ REFERENCES

(1) Yokel, R. A.; MacPhail, R. C. Engineered Nanomaterials: Exposures, Hazards, and Risk Prevention. *J. Occup. Med. Toxicol.* **2011**, *6*, 7.

(2) Puzyn, T.; Rasulev, B.; Gajewicz, A.; Hu, X.; Dasari, T. P.; Michalkova, A.; Hwang, H.-M.; Toropov, A.; Leszczynska, D.; Leszczynski, J. Using Nano-QSAR to Predict the cytotoxicity of Metal Oxide Nanoparticles. *Nat. Nanotechnol.* **2011**, *6*, 175−178.

(3) Winkler, D. A.; Mombelli, E.; Pietroiusti, A.; Tran, L.; Worth, A.; Fadeel, B.; McCall, M. J. Applying Quantitative Structure-activity Relationship Approaches to Nanotoxicology: Current Status and Future Potential. *Toxicology* **2013**, *313*, 15−23.

(4) Lubinski, L.; Urbaszek, P.; Gajewicz, A.; Cronin, M.; Enoch, S.; Madden, J.; Leszczynska, D.; Leszczynski, J.; Puzyn, T. Evaluation Criteria for the Quality of Published Experimental Data on Nanomaterials and their Usefulness for QSAR Modelling. *SAR QSAR Environ. Res.* **2013**, *24*, 995−1008.

(5) Singh, K. P.; Gupta, S. Nano-QSAR Modeling for Predicting Biological Activity of Diverse Nanomaterials. *RSC Adv.* **2014**, *4*, 13215−13230.

(6) Toropov, A. A.; Toropova, A. P. Quasi-QSAR for Mutagenic Potential of Multi-walled Carbon-nanotubes. *Chemosphere* **2015**, *124*, 40−46.

(7) Gajewicz, A.; Schaeublin, N.; Rasulev, B.; Hussain, S.; Leszczynska, D.; Puzyn, T.; Leszczynski, J. Towards Understanding Mechanisms Governing Cytotoxicity of Metal Oxides Nanoparticles: Hints from Nano-QSAR Studies. *Nanotoxicology* **2015**, *9*, 313−325.

(8) Pan, Y.; Li, T.; Cheng, J.; Telesca, D.; Zink, J. I.; Jiang, J. Nano-QSAR Modeling for Predicting the cytotoxicity of Metal Oxide Nanoparticles using Novel Descriptors. *RSC Adv.* **2016**, *6*, 25766−25775.

(9) Gajewicz, A.; Jagiello, K.; Cronin, M.; Leszczynski, J.; Puzyn, T. Addressing a Bottle Neck for Regulation of Nanomaterials: Quantitative Read-across (Nano-QRA) Algorithm for Cases when only Limited Data is Available. *Environ. Sci.: Nano* **2017**, *4*, 346−358.

(10) Iorio, F.; Bosotti, R.; Scacheri, E.; Belcastro, V.; Mithbaokar, P.; Ferriero, R.; Murino, L.; Tagliaferri, R.; Brunetti-Pierri, N.; Isacchi, A.; et al. Discovery of Drug Mode of Action and Drug Repositioning from Transcriptional Responses. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 14621−14626.

(11) ECHA, Read-across Assessment Framework.2015. https://echa.europa.eu/documents/10162/13628/raaf_en.pdf (accessed Nov 17, 2016).

(12) Gajewicz, A.; Cronin, M. T.; Rasulev, B.; Leszczynski, J.; Puzyn, T. Novel Approach for Efficient Predictions Properties of Large Pool of Nanomaterials Based on Limited Set of Species: Nano-read-across. *Nanotechnology* **2015**, *26*, 015701.

(13) Lynch, I.; Weiss, C.; Valsami-Jones, E. A Strategy for Grouping of Nanomaterials based on Key Physico-Chemical Descriptors as a Basis for Safer-by-design NMs. *Nano Today* **2014**, *9*, 266−270.

(14) Oomen, A. G.; Bleeker, E. A.; Bos, P. M.; van Broekhuizen, F.; Gottardo, S.; Groenewold, M.; Hristozov, D.; Hund-Rinke, K.; Irfan, M.-A.; Marcomini, A.; et al. Grouping and Read-across Approaches for Risk Assessment of Nanomaterials. *Int. J. Environ. Res. Public Health* **2015**, *12*, 13415−13434.

(15) Chen, X.-H.; Ma, L.; Hu, Y.-X.; Wang, D.-X.; Fang, L.; Li, X.-L.; Zhao, J.-C.; Yu, H.-R.; Ying, H.-Z.; Yu, C.-H. Transcriptome Profiling and Pathway Analysis of Hepatotoxicity induced by Tris (2-ethylhexyl) Trimellitate (TOTM) in Mice. *Environ. Toxicol. Pharmacol.* **2016**, *41*, 62−71.

(16) Smalley, J. L.; Gant, T. W.; Zhang, S.-D. Application of Connectivity Mapping in Predictive Toxicology Based on Gene-expression Similarity. *Toxicology* **2010**, *268*, 143−146.

(17) Subramanian, A.; Tamayo, P.; Mootha, V. K.; Mukherjee, S.; Ebert, B. L.; Gillette, M. A.; Paulovich, A.; Pomeroy, S. L.; Golub, T. R.; Lander, E. S.; et al. Gene Set Enrichment Analysis: a Knowledge-based Approach for Interpreting Genome-wide Expression Profiles. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 15545−15550.

(18) Hänzelmann, S.; Castelo, R.; Guinney, J. GSVA: Gene Set Variation Analysis for Microarray and RNA-seq Data. *BMC Bioinf.* **2013**, *14*, 7.

(19) Witten, I. H.; Frank, E.; Hall, M. A.; Pal, C. J. *Data Mining: Practical Machine Learning Tools and Techniques*; Morgan Kaufmann, 2016.

(20) Liberzon, A.; Subramanian, A.; Pinchback, R.; Thorvaldsdóttir, H.; Tamayo, P.; Mesirov, J. P. Molecular Signatures Database (MSigDB) 3.0. *Bioinformatics* **2011**, *27*, 1739−1740.

(21) Davis, A. P.; Murphy, C. G.; Johnson, R.; Lay, J. M.; Lennon-Hopkins, K.; Saraceni-Richards, C.; Sciaky, D.; King, B. L.; Rosenstein, M. C.; Wiegers, T. C.; et al. The Comparative Toxicogenomics Database: update 2013. *Nucleic Acids Res.* **2013**, *41*, D1104.

(22) Walkey, C. D.; Olsen, J. B.; Song, F.; Liu, R.; Guo, H.; Olsen, D. W. H.; Cohen, Y.; Emili, A.; Chan, W. C. Protein Corona Fingerprinting Predicts the Cellular Interaction of Gold and Silver Nanoparticles. *ACS Nano* **2014**, *8*, 2439−2455.

(23) Liu, R.; Jiang, W.; Walkey, C. D.; Chan, W. C.; Cohen, Y. Prediction of Nanoparticles-cell Association Based on Corona Proteins and Physicochemical Properties. *Nanoscale* **2015**, *7*, 9664−9675.

(24) Papa, E.; Doucet, J.; Sangion, A.; Doucet-Panaye, A. Investigation of the Influence of Protein Corona Composition on Gold Nanoparticle Bioactivity Using Machine Learning Approaches. *SAR QSAR Environ. Res.* **2016**, *27*, 521−538.

(25) Tsiliki, G.; Munteanu, C. R.; Seoane, J. A.; Fernandez-Lozano, C.; Sarimveis, H.; Willighagen, E. L. RRegrs: an R Package for Computer-aided Model Selection with Multiple Regression Models. *J. Cheminf.* **2015**, *7*, 46.

(26) Helma, C.; Rautenberg, M.; Gebele, D. nano-lazar: Read Across Predictions for Nanoparticle Toxicities with Calculated and Measured Properties. *Front. Pharmacol* **2017**, DOI: 10.3389/fphar.2017.00377.

■ NOTE ADDED AFTER ASAP PUBLICATION

This paper was published ASAP on February 26, 2018, with an error in the Abstract paragraph. The corrected version reposted March 6, 2018.