Vythos

A small user guide

February 23, 2021

Contents

1 Introduction

Vythos, is a web-application (https://vythos.jaqpot.org/) developed using R and `shiny` package, that hosts four predictive models for the estimation of toxicity indeces. Interested users can upload a set of engineered nanomaterials (ENMs) with known properties and unknown toxicity and acquire reliable predictions. **Vythos** incorporates the concepts of grouping, read-across [1] and optimization through the solution of a Mixed Integer Linear Programming (MILP) problem. More information can be found in the relevant publication (add link).

Vythos is based in a novel grouping methodology for the prediction of toxicity related endpoints of ENMs. Based on the formulation and the solution of a mathematical optimization problem, the method searches over a space of alternative grouping hypotheses [1], in terms of partition feature (group defining feature), breakpoints (group limits) and selected features in each region, and determines the one providing the most accurate read-across predictions[2]. The two available models are based on two different datasets derived from Literature: the *Gold ENMs* dataset from Walkey *et al*. [3] and the *MeOx ENMs* dataset from Gajewicz *et al*. [5]

The *Gold ENMs* dataset consists of 84 culture medium incubated gold anionic and cationic nanoparticles (diameters of 15, 30 and 60 nm). For each ENM there are available 40 physicochemical descriptors, 129 protein corona fingerprints (biological descriptors) and additional measurements of its cell association with human A549 cells (in mL/ g(Mg)). The cell association was transformed into log2 values. The protein corona fingerprints were filtered by GSVA analysis and only 63 were considered as statistically significant proteins. [4] Following the proposed workflow for validation purposes, the above set was split into training and test sets in a ratio of 66:33 using the Kennard-Stones method. [6]

The second dataset in which the MILP workflow is applied, is derived from the publication of Gajewicz *et al*. (2014) [5] (*MeOx ENMs*). This dataset contains a list of 18 nano-metal oxide ENMs with their toxicity index which refer to the concentration [molar] of metal oxide ENMs that caused a 50% reduction of the cells of human keratinocyte (HaCaT) cell line after 24 hours of exposure (LC_{50}). For these ENMs, there are available 18 quantum-mechanical descriptors from quantum-chemical calculations and 11 image descriptors (derived from Transmission Electron Microscopy images). The endpoints were transformed into $\log{(LC_{50})^{-1}}$. For comparison purposes, training and test sets in this case study were the same as in the original publication.

The proposed method is demonstrated on data derived from the publication of Xia *et al*.[7] (*CNTs*). In this publication the surface adsorption energy of ENMs was studied through five nanodescriptors that represent the surface adsorption interactions (hydrophobicity, hydrogen bond -donors and recipients-, polarity/polarizability, and lone-pair electrons). These surface adsorpiton forces are responsible for the interaction of ENMs with various biological molecules, including proteins thus, they can contribute to the interpretation of the ENMs behavior in different biological media.

Under ideal biological conditions, the surface adsorption properties of ENMs can be measured using a set of probes with various physicochemical properties. Therefore in this study, 28 probe compounds were used in order to measure the adsorption coefficient (k) on a given ENM (in this case on 40 nm diameter Multiwalled Carbon Nanotubes (MWCNTs) coated with hydroxyl derivatives - MWCNT40nm-COOH). The k values were converted to logarithmic scale. For external validation purposes, the dataset was partitioned into training and test sets in a ratio of $75{:}25$ using the Kennard-Stone method.

2 Validation

An external validation approach is used to test the proposed read-across methodology, by dividing the full dataset into training and test subsets. This data partitioning can be achieved either by applying a random partition or a partition method (e.g Kennard-Stones).[8] The training set is used in the developed MILP workflow and determines the optimal set of descriptors and group limits. For the test set, predictions are made using the optimized read-across hypothesis. Eventually, the read-across predictions are compared with the experimental endpoint values

using the q^2_{ext} statistic (Eq. 1).[9]

$$q^2_{ext} = 1 - \frac{\sum_{i=1}^{n} (y_i - \widehat{y_i})}{\sum_{i=1}^{n} (y_i - \bar{y}_{tr})^2} \tag{1}$$

where y_i and $\hat{y}_i$ are the experimental and predicted endpoint values over the test set and $\bar{y}_{tr}$ is the averaged value of the endpoint for the training set.

3 Domain of applicability

In this methodology, where multiple linear equations constitute the read-across model, we consider a prediction as reliable only if all the independent descriptors are within the ranges defined by the training samples. Therefore, before using the model for predicting the endpoint of an external ENM, the input variables should be scaled first according to the original dataset's min-max values. Only if all scaled values are within the range [0,1], the read-across prediction can be accepted.

4 *Gold ENMs* model

4.1 1D model

The selected optimal model for the $net.cell$ ($\log_2(n.cellassociation)$) estimation for gold ENMs, classifies the ENMs into regions A or B according to eq. 2 add paper reference. This model is derived from the solution of a 1D MILP problem thus, one partition features is selected; the Apolipoprotein B-100 ($P04114$), which is a biological descriptor. Thirty-nine descriptors are needed for toxicity predictions derived during training, presented in Table 1.

$$net.cell = \begin{cases} 0.625 \cdot lspri.serum + 0.523 \cdot lspri.relative + 0.333 \cdot zav.serum \\ +2.443 \cdot int.serum + 0.716 \cdot hdrel.serum - 0.337 \cdot vol.ch \\ +0.505 \cdot pdi.rel - 2.607 \cdot zp.rel + 1.610 \cdot zp.synt.sign \\ -0.793 \cdot zp.serum.mag - 1.131 \cdot AS.total - 0.508 \cdot P01024 \\ +0.151 \cdot P00734 - 0.009 \cdot P05154 + 1.577 \cdot P19823 + 0.358 \cdot P12259 \\ +0.384 \cdot P10720 - 0.041 \cdot P68871 + 0.756 \cdot O43866 + 2.653 \cdot P02654 \\ -0.718 \cdot P03952 + 0.063 \cdot P18428 - 0.177 \cdot P02655 - 0.356 \cdot P00751 \\ -0.158 \cdot P02790 + 0.251 \cdot P18065 + 0.588 \cdot P08567 + 0.172 \cdot P01019 \\ +0.576 \cdot P02671 + 0.444 \cdot P00451 + 0.095 \cdot P14618 - 0.373 \cdot P23528 \\ +0.467 \cdot Q99467 - 4.814 \quad \text{if } P04114 \leq 0.447 \text{ region A} \\ \\ -2.600 \cdot pdi.serum - 0.820 \cdot int.rel + 0.052 \cdot zp.synth.mag \\ +2.338 \cdot P01009 + 0.934 \cdot P02749 - 3.186 \cdot P02655 + 4.722 \cdot P27169 \\ +0.746 \cdot P01019 - 5.327 \quad \text{if } P04114 > 0.447 \text{ region B} \end{cases} \tag{2}$$

For more information about these descriptors please refer to the original publication and its supplementary material [3].

The training *Gold ENMs* that belong to regions A and B are presented in Table 2. For each unknown ENM belonging to region A or B, the training ENMs of that region are its neighbors.

Table 1: Semantics of *Gold ENMs* models.

Descriptor	Description
$net.cell$	The $\log_2$-transformed value of cell association
$class$	Surface classification: "anionic" (1) or "cationic" (0)
$lspri.serum$	Localized surface plasmon resonance (LSPR) index after serum exposure
$lspri.relative$	(LSPR index after serum exposure)/(LSPR index after synthesis)
$zav.serum$	z-average hydrodynamic diameter (HD) after serum exposure
$pdi.serum$	Polydispersity index after serum exposure
$int.serum$	Intensity mean HD after serum exposure
$hdrel.serum$	(Z-average HD after serum exposure)/(TEM size)
$vol.ch$	(Volume mean HD after serum exposure) - (Volume mean HD after synthesis)
$pdi.rel$	(Polydispersity index after serum exposure)/(Polydispersity index after synthesis)
$int.rel$	(Intensity mean HD after serum exposure) - (Intensity mean HD after synthesis)
$zp.rel$	(zeta potential after serum exposure)/(zeta potential after synthesis)
$zp.synth.sign$	Sign (signum) of zeta potential after synthesis
$zp.synth.mag$	Magnitude of zeta potential after synthesis
$zp.serum.mag$	Magnitude of zeta potential after serum exposure
$AS.total$	Total surface area
$P04114$	Apolipoprotein B-100
$P01024$	Complement C3
$P01009$	Alpha-1-antitrypsin
$P00734$	Prothrombin
$P05154$	Plasma serine protease inhibitor
$P19823$	Inter-alpha-trypsin inhibitor heavy chain H2
$P12259$	Coagulation factor V
$P10720$	Platelet factor 4 variant
$P68871$	Hemoglobin subunit beta
$O43866$	CD5 antigen-like
$P02749$	Beta-2-glycoprotein 1
$P02654$	Apolipoprotein C-I
$P03952$	Plasma kallikrein
$P00742$	Coagulation factor X
$P09871$	Complement C1s subcomponent
$P20851$	C4b-binding protein beta chain
$P18428$	Lipopolysaccharide-binding protein
$P02655$	Apolipoprotein C-II
$P00751$	Complement factor B
$P02790$	Hemopexin
$P27169$	Serum paraoxonase/arylesterase 1
$P18065$	Insulin-like growth factor-binding protein 2
$P08567$	Pleckstrin
$P01019$	Angiotensinogen
$P02671$	Fibrinogen alpha chain
$P00451$	Coagulation factor VIII
$P14618$	Pyruvate kinase isozymes M1/M2
$P23528$	Cofilin-1
$Q99467$	CD180 antigen

Table 2: Training samples groups, when using the 1D *Gold ENMs* model.

	Training *Gold ENMs*
Region A	G15.AC, G15.AHT, G15.CALNN, G15.CTAB, G15.DDT@DOTAP, G15.DTNB, G15.F127, G15.Gly-SH, G15.MES, G15.Met-SH, G15.MHDA, G15.MSA, G15.MUTA, G15.NT@PSMA-EDA, G15.NT@PSMA-Urea, G15.ODA, G15.T20, G15.SA, G15.PAH-SH, G15.PLL-SH, G15.PVA, G15.PVP, G15.SPP, G15.TP, G30.AC, G30.CFGAILS, G30.DDT@DOTAP, G30.LA, G30.MAA, G30.MUA, G30.MUTA, G30.PAH-SH, G30.Thr-SH, G30.TP, G60.AUT, G60.CIT, G60.CTAB, G60.CIT, G60.CVVIT, G60.DDT@DOTAP, G60.DTNB, G60.MBA, G60.MUTA, G60.ODA, G60.PVA, G60.Trp-SH
Region B	G15.DDT@BDHDA, G15.DDT@CTAB, G15.DDT@ODA, G15.DDT@SA, G15.DDT@SDS, G15.HDA, G15.NT@DCA, G30.DDT@BDHDA, G30.DDT@CTAB, G60.DDT@BDHDA

4.2 2D model

The full MILP model produced in 2D for the *Gold ENMs*' case study is presented next (Eq. 3). The partition feature selected from the physicochemical descriptors is the difference between the Intensity mean hydrodynamic diameter (HD) after serum exposure and the Intensity mean HD after synthesis ($int.rel$) while Apolipoprotein B-10 ($P04114$) is selected from the biological descriptors. The MILP model uses 7 physicochemical and 7 biological descriptors (see Table 1), and classifies incoming samples in four regions A, B, C and D.

$$net.cell = \begin{cases} -0.415 \cdot class + 0.113 \cdot lspri.serum \\ +0.094 \cdot zav.serum + 2.178 \cdot int.serum \\ +0.269 \cdot pdi.rel + 2.654 \cdot zp.synth.sign \\ -0.426 \cdot AS.total - 0.140 \cdot P05154 + 1.648 \cdot P19823 \\ -0.364 \cdot P03952 + 0.464 \cdot P00742 + 0.268 \cdot P09871 \\ +0.664 \cdot P20851 - 0.441 \cdot P23528 \\ -5.627 \quad \text{if } int.rel \leq 0.742 \ \& \ P04114 \leq 0.447 \text{ region A} \\ -0.682 \cdot class - 6.122 \quad \text{if } int.rel \leq 0.742 \ \& \ P04114 > 0.447 \text{ region B} \\ 0.113 \quad \text{if } int.rel > 0.742 \ \& \ P04114 \leq 0.447 \text{ region C} \\ -7.590 \quad \text{if } int.rel > 0.742 \ \& \ P04114 > 0.447 \text{ region D} \end{cases} \tag{3}$$

The training *Gold ENMs* that belong to regions A, B, C and D are presented in Table 3. For each unknown ENM belonging to region A, B, C or D the training ENMs of that region are its neighbors.

5 *MeOx ENMs* model

5.1 1D model

The full 1D model for the $\log(LC_{50})^{-1}$ estimation of metal oxide ENMs, classifies the ENMs into regions A, B and C and it is presented next in Eq. 4. Seven descriptors are needed for toxicity predictions (see Table 4) derived

Table 3: Training samples groups, when using the 2D *Gold ENMs* model.

	Training *Gold ENMs*
Region A	G15.TP, G15.Met-SH, G15.DTNB, G15.DDT@DOTAP, G15.CTAB, G15.AC, G15.AHT, G15.CALNN, G15.F127, G15.Gly-SH, G15.MES, G15.MHDA, G15.MSA, G15.NT@PSMA-EDA, G15.NT@PSMA-Urea, G15.ODA, G15.PAH-SH, G15.PLL-SH, G15.PVA, G15.PVP, G15.SA, G15.SPP, G15.T20, G30.AC, G30.CFGAILS, G30.DDT@DOTAP, G30.LA, G30.MAA, G30.MUA, G30.MUTA, G30.PAH-SH, G30.Thr-SH, G30.TP, G60.AUT, G60.CIT, G60.CTAB, G60.CVVIT, G60.DDT@DOTAP, G60.DTNB, G60.MBA, G60.MUTA, G60.ODA, G60.Trp-SH, G60.PVA
Region B	G15.HDA, G15.DDT@ODA, G15.DDT@SA, G15.NT@DCA, G30.DDT@BDHDA, G30.DDT@CTAB, G60.DDT@BDHDA
Region C	G15.DDT@BDHDA, G15.DDT@CTAB, G15.DDT@SDS
Region D	G15.MUTA

during training, 4 quantum-mechanical and 3 image descriptors. The partition feature is the electrochemical potential (μ). For more information about these descriptors please refer to the original publication ([5]).

$$\log\left(LC_{50}\right)^{-1} = \begin{cases} 3.320 & \text{if } \mu < 0.075 \text{ region A} \\ -0.096 \cdot Core + 2.926 & \text{if } 0.075 \leq \mu \leq 0.537 \text{ region B} \\ 0.723 \cdot \Delta H_f^c - 0.330 \cdot TE - 0.044 \cdot S & \\ +0.006 \cdot d_{V/m} + 0.237 \cdot d_{S_{auter}} & \\ +0.055 \cdot P_Y + 1.829 & \text{if } 0.537 < \mu \text{ region C} \end{cases} \tag{4}$$

Table 4: Semantics of the 1D *MeOx ENMs* model.

Descriptor	**Description**
$\log LC_{50}^{-1}$	The log-transformed concentration of ENMs that causes a 50% reduction of cells after 24 hrs of exposure
μ	The electrochemical potential
χ^c	The Mulliken's electronegativity
$Core$	The core–core repulsion energy
ΔH_f^c	The standard enthalpy of formation of metal oxide nanocluster
TE	The total energy
S	The total softness
$LUMO$	The energy of the lowest unoccupied molecular orbital
$Shift$	The Schuurmann MO shift alpha
A_{R_X}	The aspect ratio X
$d_{V/m}$	The volume/mass diameter
d_{Sauter}	The volume/surface diameter
P_Y	The porosity Y

The training *MeOx ENMs* that belong to regions A, B and C are presented in Table 5. For each unknown ENM belonging to region A, B or C the training ENMs of that region are its neighbors.

Table 5: Training samples groups, when using the 1D *MeOx ENMs* model.

	Training *MeOx ENMs*
Region A	ZnO
Region B	In2O3, CoO
Region C	Bl2O3, ZrO2, Mn2O3, Sb2O3, SiO2, TiO2, V2O3

Table 6: Training samples groups, when using the 2D *MeOx ENMs* model.

	Training *MeOx ENMs*
Region A	Sb2O3
Region B	Bl2O3, CoO, In2O3, Mn2O3, SiO2, TiO2, V2O3, ZrO2
Region C	-
Region D	ZnO

5.2 2D model

The selected optimal 2D model for the $\log(LC_{50})^{-1}$ estimation of metal oxide ENMs, classifies the ENMs into regions A, B, C and D according to eq. 5 add paper reference. This model is derived from the solution of a 2D MILP problem, thus two partition features are selected; the Mulliken's electronegativity (χ^c), which is a quantum-mechanical descriptor and the aspect ratio X (A_{R_X}) which is an image descriptor. Seven descriptors are needed for toxicity predictions derived during training 4 quantum-mechanical and 3 image descriptors (see Table 4). For more information about these descriptors please refer to the original publication [5].

$$\log(LC_{50})^{-1} = \begin{cases} 2.310 & \text{if } \chi^c \leq 0.875 \ \& \ A_{R_X} \leq 0.361 \text{ region A} \\ 0.958 \cdot \Delta H_f^c - 0.585 \cdot LUMO \\ +0.158 \cdot S - 0.647 \cdot Shift \\ -0.099 \cdot d_{V/m} - 0.034 \cdot d_{Sauter} \\ -0.158 \cdot P_Y + 2.564 & \text{if } \chi^c \leq 0.875 \ \& \ A_{R_X} > 0.361 \text{ region B} \\ NA & \text{if } \chi^c > 0.875 \ \& \ A_{R_X} \leq 0.361 \text{ region C} \\ 3.320 & \text{if } \chi^c > 0.875 \ \& \ A_{R_X} > 0.361 \text{ region D} \end{cases} \quad (5)$$

The training *MeOx ENMs* that belong to regions A, B and D are presented in Table 6. For each unknown ENM belonging to region A, B or D the training ENMs of that region are its neighbors. Considering that no training ENMs have been allocated in region C, no predictions can be made if an unknown ENM belongs to this region.

6 *CNTs* model

The full 1D model for the $\log k$ estimation of CNTs, classifies the samples into regions A and B and it is presented next in Eq. 6. Four descriptors are needed for toxicity predictions (see Table 7) derived during training. The partition feature is the hydrogen bond acidity (α). For more information about these descriptors please refer to

the original publication ([7]).

$$\log k = \begin{cases} 1.841 \cdot \pi - 0.016 \cdot \alpha - 2.463 \cdot \beta + 3.106 \cdot V + 2.431 & \text{if } \alpha \leq 0.907 \text{ region A} \\ \\ 4.260 & \text{if } \alpha > 0.907 \text{ region B} \end{cases} \tag{6}$$

Table 7: Variables involved in the 1D *CNTs* model.

Role	Symbol	Description[7]
Endpoint	$\log k$	The adsorption coefficients of the probe compounds on MWCNT40nm-COOH
Partition feature	α	Hydrogen-bond acidity
MLR variables	π	Polarizability
	α	Hydrogen-bond acidity
	β	Hydrogen-bond basicity
	V	Lipophilicity interaction

The training probe compounds that belong to regions A and B are presented in Table 8. For each unknown ENM belonging to region A or B the training ENMs of that region are its neighbors.

Table 8: Training samples groups, when using the 1D *CNTs* model.

	Training probe compounds
Region A	4-chloroanisole, phenethyl alcohol, 1-methylnaphthalene, benzyl alcohol, phenol, benzonitrile, 3-methylphenol, chlorobenzene, p-xylene, bromobenzene, acetophenone, 3,5-dimethylphenol, methyl benzoate, iodobenzene, propylbenzene, 4-chlorotoluene, ethyl benzoate, 4-nitrotoluene, 4-chloroacetophenone, naphthalene
Region B	3-bromophenol

7 Instructions

To initiate the prediction process users must select one of the provided datasets from the **Select dataset** dropdown menu: **Metal oxides ENMs toxicity** that corresponds to the *MeOx ENMs* set or **Gold ENMs toxicity** that corresponds to the *Gold ENMs* set. The users must also select the type of model they want to use; the 1D or 2D MILP model from the corresponding radiobuttons. When a model is selected the necessary variables are presented, as well as the endpoint that will be predicted, and the q^2_{ext} statistic from external validation (Fig. 1).

Users must upload one .csv file containing the dataset of interest by clicking on the **Browse** button in the **Select file** field. The file must contain the values of the necessary descriptors (in columns), including the the ENMs names in the 1rst column. Missing values cannot be handled by this approach. Users are advised to download and fill the template input file different for each model that can be downloaded from the app by clicking on **Template** button.

Input data are automatically normalized, according to Eq. 7.

$$c_{\text{sc}} = \frac{c_{\text{in}} - min}{max - min} \tag{7}$$

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

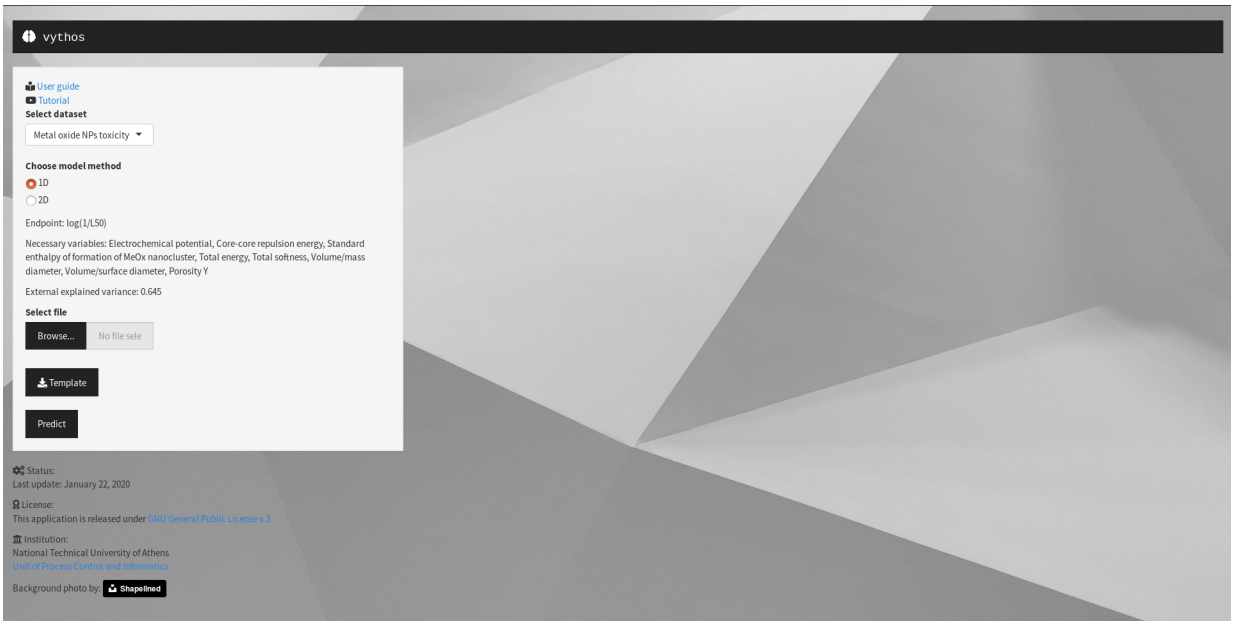

Figure 1: The user interface of **Vythos** application.

If a dataset is uploaded, by pressing the **Predict** button, the prediction process starts, according to the regions where each input sample belongs, otherwise the corresponding button remains disabled till the necessary file is provided.

The analysis produces a table that contains the predicted value of toxicity index for all the provided ENMs and can be downloaded in .csv format by clicking on **Results** button. The table contains also an indication of the reliability of the predictions and the group where each ENM belongs. A graph where the input ENMs are depicted in accordance to the training ENMs is also presented. The "space" is defined by the breakpoints and the toxicity index thus, it can be a 2D graph in the case of one breakpoint (1D MILP model) model (Fig. 2) or a 3D graph (2D MILP model) in the case of two breakpoints (Fig. 3). The diagram can be downloaded in .png format.

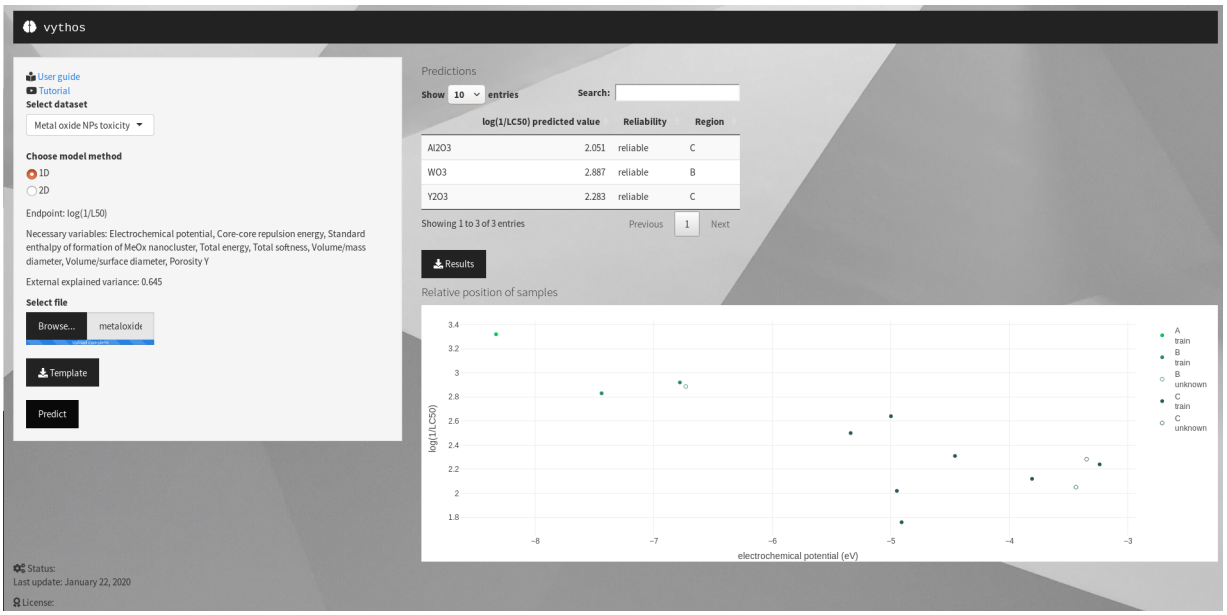

Figure 2: The produced results running 1D-*MeOx ENMs* set. The predictions for $log(LC_{50})^{-1}$ are presented along with the 2D regions plot.

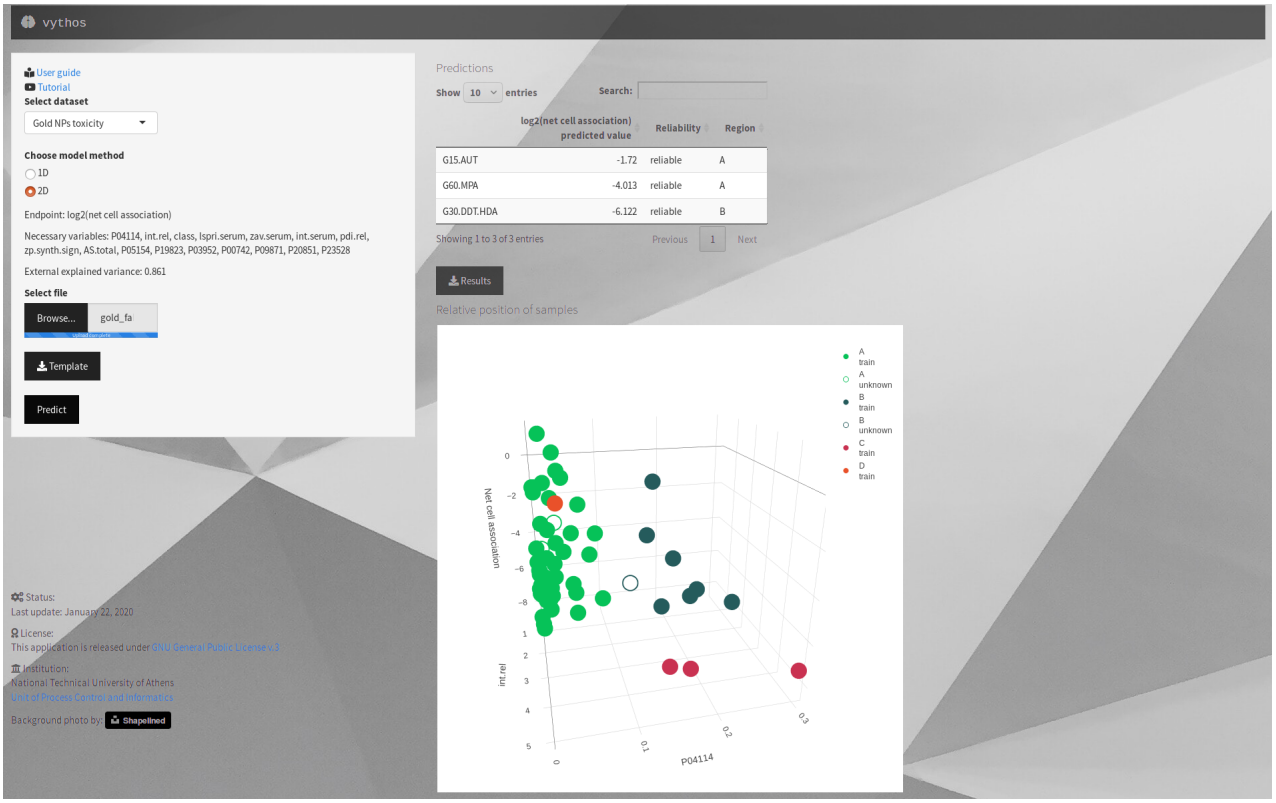

Figure 3: The produced results running 2D-*Gold ENMs* set. The predictions for $net.cell$ are presented along with the 3D regions plot.

8 Abbreviations

- ENM, engineered nanomaterial
- MILP, mixed integer linear programming

9 Acknowledgments

This research work was supported by the Hellenic Foundation for Research and Innovation (HFRI) under the HFRI PhD Fellowship grant (Fellowship Number: 637).

10 Contact information

dimitra.varsou@gmail.com, nikikoutroumpa@gmail.com

NikiKou

Video tutorial

Unit of Process Control and Informatics, School of Chemical Engineering, NTUA

References

[1] ECHA, *Read Across Assessment Framework, Appendix R.6-1 for nanomaterials applicable to the Guidance on QSARs and Grouping of Chemicals. Guidance on information requirements and chemical safety assessment*, 2017, Available online in: https://echa.europa.eu/documents/10162/23036412/appendix_r6_nanomaterials_en.pdf

[2] J. Cardoso-Silva, G. Papadatos, L. Papageorgiou and S. Tsoka, *Optimal Piecewise Linear Regression Algorithm for QSAR Modelling*, Molecular informatics, 38(3): 1800028, 2019

[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://pubs.acs.org/doi/abs/10.1021/nn406018q

[4] D. D. Varsou, G. Tsiliki, P. Nymark, P. Kohonen, R. Grafström, H. Sarimveis, *toxFlow: a web-based application for read-across toxicity prediction using omics and physicochemical data*, 58(3): 543-549, 2017, Available online in: https://pubs.acs.org/doi/10.1021/acs.jcim.7b00160

[5] A. Gajewicz, N. Schaeublin, B. Rasulev, S. Hussain, D. Leszczynska, T. Puzyn and J. Leszczynski, *Towards understanding mechanisms governing cytotoxicity of metal oxides nanoparticles: Hints from nano-QSAR studies*, Nanotechnology, 9(3):313-25, 2015, Available online in: https://www.tandfonline.com/doi/abs/10.3109/17435390.2014.930195?journalCode=inan20

[6] M. Daszykowski, B. Walczak and D. Massart, *Representative subset selection*, Analytica Chimica Acta, 468: 91–103, 2002, Available online in: https://www.sciencedirect.com/science/article/pii/S0003267002006517?via%3Dihub

[7] X. R. Xia, N. A. Monteiro-Riviere, S. Mathur, X. Song, L. Xiao, S. J. Oldenberg, B. Fadeel, J. E. Riviere, *Mapping the surface adsorption forces of nanomaterials in biological systems*, ACS nano, 9074–9081, 2011, 5.

[8] A. Tropsha, *Best practices for QSAR model development, validation, and exploitation*, Molecular Informatics, 29(6-7): 476–488, 2010

[9] A. Tropsha, P. Gramatica, V. K. Gombar, *The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models*,QSAR & Combinatorial Science, 22(1): 69-77, 2003