Supporting Information

NanoTox: Development of a parsimonious *in silico* model for toxicity assessment of metal-oxide nanoparticles using physicochemical features

Nilesh Anantha Subramanian[1,2], Ashok Palaniappan[3*]

[1]Department of Medical Nanotechnology, School of Chemical and BioTechnology, SASTRA Deemed University, [2]Department of Computer Science and Engineering, IIT Madras, Chennai, [3]Department of Bioinformatics, School of Chemical and BioTechnology, SASTRA Deemed University, Thanjavur 613401. India

[*]Corresponding author: apalania@scbt.sastra.edu

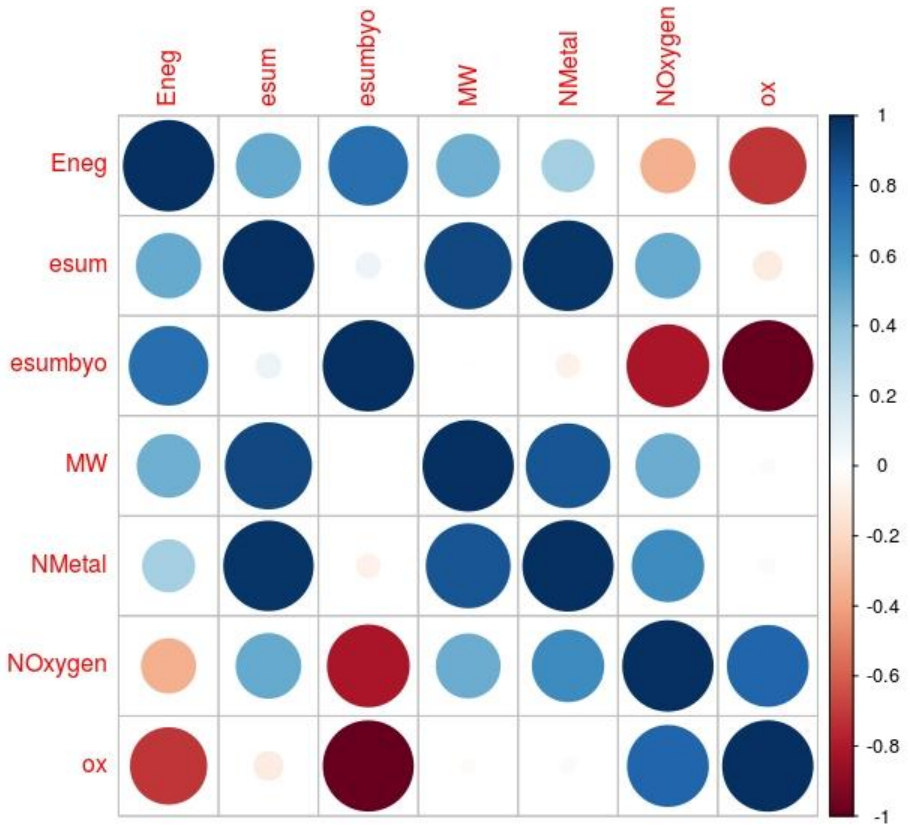

Figure S1. Correlogram of periodic table descriptors. The high absolute correlation between many pairs of features is evident, and would be detrimental to effective learning.

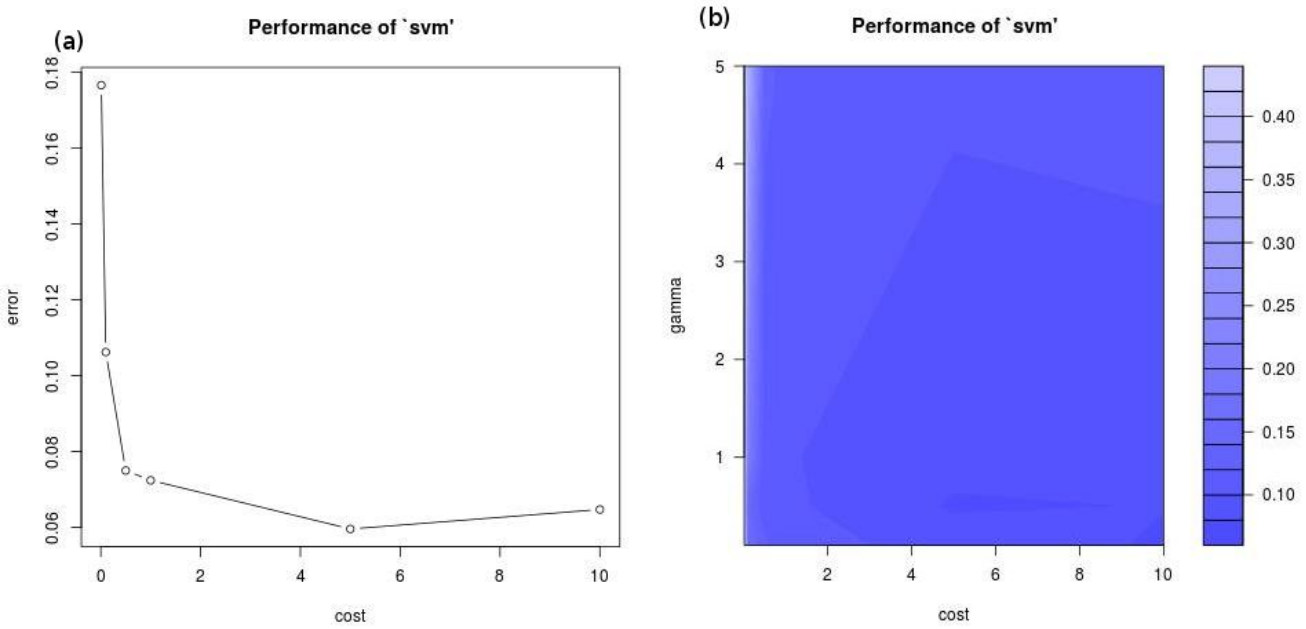

Figure S2. Hyperparameter optimization of the SVM. (a) Linear kernel; notice the error hits a minimum at cost =5. (b) Radial kernel; notice the dark blue patch at the bottom right corner.

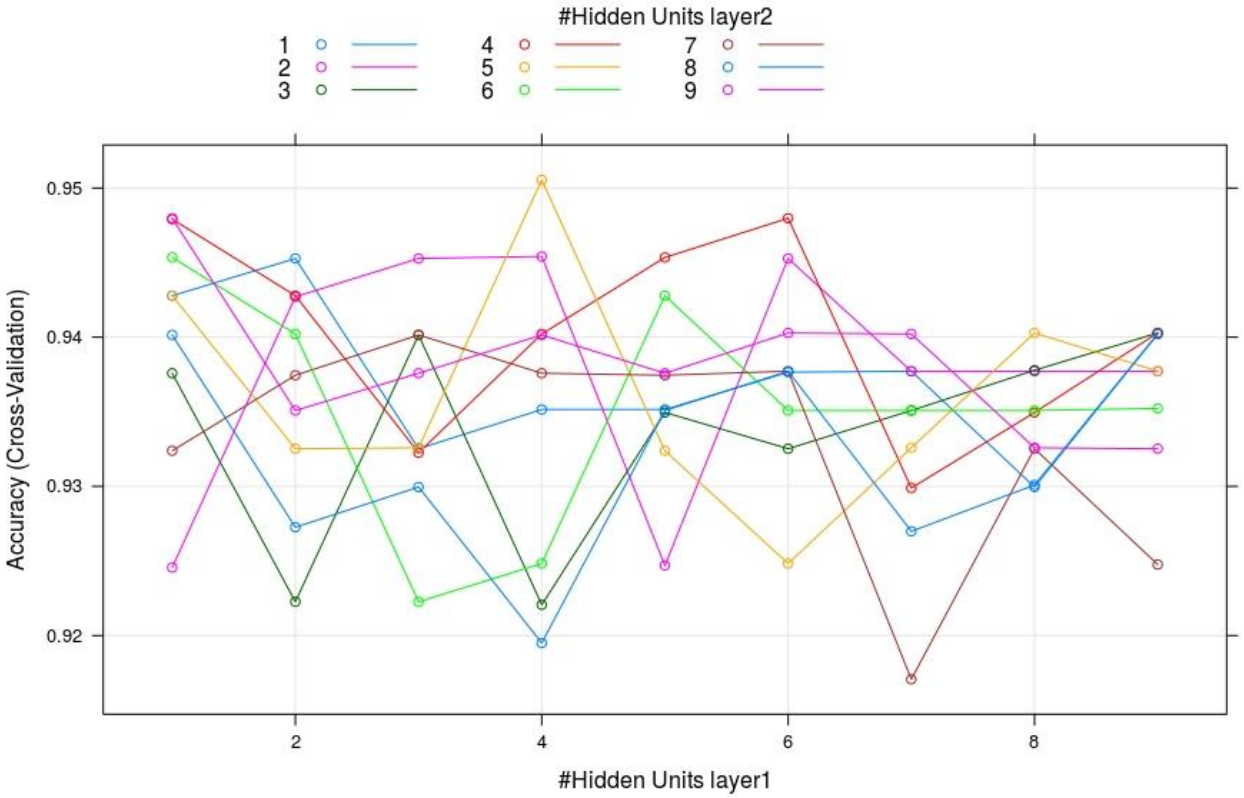

Figure S3. Hyper-parameter optimization of the neural network - 2 layer. At #HiddenUnits_Layer1 = 4, and #HiddenUnits_layer2 = 5 (yellow), the cross-validated accuracy hits a maximum.

Table S1. Hyperparameter optimization for the models in our study. For logistic regression, the default threshold was used (0.5). The default value for the number of trees (=500) in Random forest was used; increasing beyond this value is unlikely to yield significant gains.

No.	Classifier	Hyperparameters	Optimisation	Error
1	Logistic regression	Threshold	(0.5)	n/a
2	Random forest	1. #trees 2. m_{try}	(500) 2	0.029
3	Support vector machine	Linear Kernel 1. Cost	5	0.060
		Radial Kernel 1. Cost 2. Gamma	10 0.1	0.078
		Poly Kernel 1. Cost 2. Gamma 3. Degree	5 5 4	0.065
4	Neural networks	1. Size of hidden layer 1 2. Decay rate	5 0.1	0.057
		1. Size of hidden layer 1 2. Size of hidden layer 2	4 5	0.050

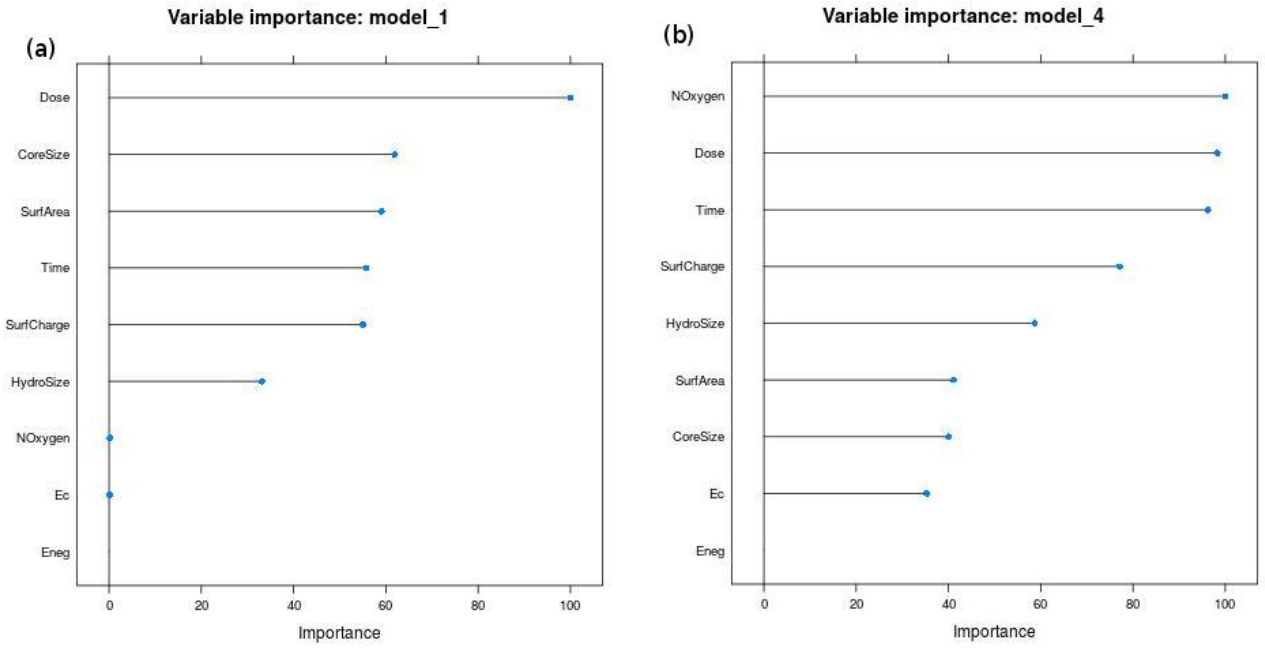

Figure S4. Variable importance plots using the `varimp` function of the `R caret` package. (a) Logistic regression; Dose is by far the most important, but NOxygen does not emerge important at all. (b) Neural Network 1-layer; NOxygen, Dose, and Time are ranked nearly equally important.

Table S2. Summary of the logistic regression model. The size of the coefficient and its significance are shown. Dose is seen to be extremely significant, while CoreSize, SurfArea, Time and SurfCharge emerge as highly significant. No significance is attached to NOxygen, Eneg, and Ec in this model.

Coefficient	Estimate	Std.Error	z value	Pr(>|z|)	Significance size
(Intercept)	29.9642	6412.2589	0.005	0.996272	
CoreSize	-4.3904	1.103	-3.98	6.88E-05	***
HydroSize	1.5645	0.7343	2.131	0.033124	*
SurfCharge	-1.7914	0.5067	-3.535	0.000407	***
SurfArea	-4.754	1.2526	-3.795	0.000147	***
Ec	-12.6625	1498.7043	-0.008	0.993259	
Time	1.0105	0.2818	3.585	0.000336	***
Dose	6.2585	0.974	6.425	1.32E-10	***
Eneg	7.9834	8754.8124	0.001	0.999272	
NOxygen	-10.5567	792.911	-0.013	0.989377	

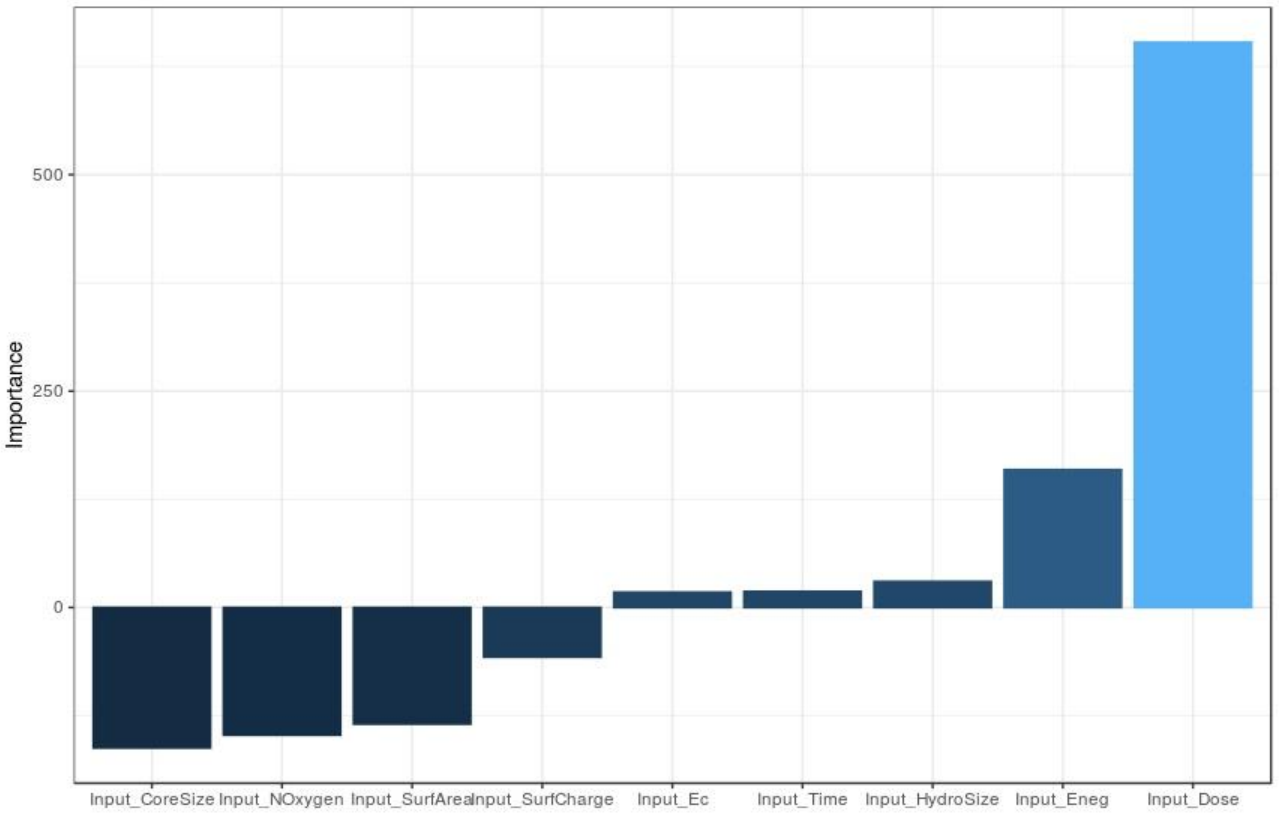

Figure S5. Relative importance plots of the Neural Network -2 layer model using `NeuralNetTools`. Dose is maximally positively correlated with the 'Toxic' class, followed by Eneg. CoreSize and NOxygen are maximally correlated with the 'NonToxic' class, followed by SurfArea.

Table S3. Garson importance analysis of the Neural Network 1L model. NOxygen, Dose and Time emerged almost equally important.

S.No	Feature	Importance
1	CoreSize	0.08220930
2	HydroSize	0.10823040
3	SurfCharge	0.13387495
4	SurfArea	0.08374302
5	Ec	0.07566291
6	Time	0.16056208
7	Dose	0.16340571
8	Eneg	0.02653256
9	NOxygen	0.16577907

Table S4. Relative importance analysis of the Neural network - 2L model. The sign of the variable ('+', '-') indicates favouring the 'nontoxic' or 'toxic' outcome classes, respectively. This table provided the data for Fig. S6.

S.No	Feature	Importance
1	Input_CoreSize	162.23178
2	Input_HydroSize	-30.32187
3	Input_SurfCharge	57.21999
4	Input_SurfArea	134.51806
5	Input_Ec	-17.70011
6	Input_Time	-18.37538
7	Input_Dose	-652.88422
8	Input_Eneg	-159.16791
9	Input_NOxygen	147.33822

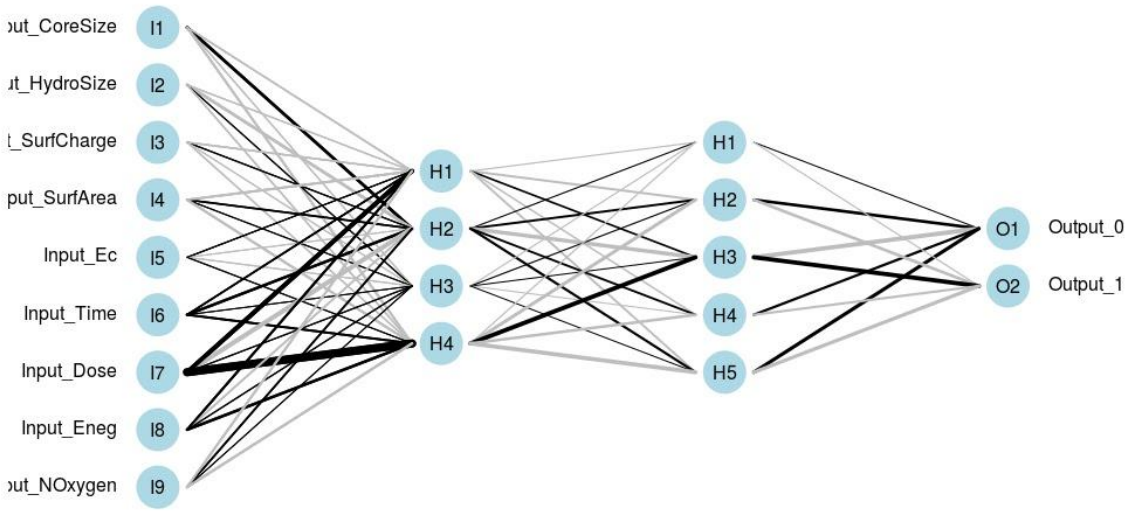

Figure S6. A schematic of the trained neural network- 2 layer model, with the weights of the connections indicated by the linewidth. Black lines indicate positive weights, and gray lines indicate negative weights. Two output units are seen, one for the 'toxic', and other for the 'non-toxic' classes.