Supplementary Information

Machine learning and materials modelling interpretation of in vivo toxicological response to TiO_2 nanoparticles library (UV and non-UV exposure)

Susana I.L. Gomes[1†], Mónica J.B. Amorim[1†*], Suman Pokhrel[2,3], Lutz Mädler[2,3], Matteo Fasano[4], Eliodoro Chiavazzo[4], Pietro Asinari[4,5], Jaak Jänes[6], Kaido Tämm[6], Jaanus Burk[6] and Janeck J. Scott-Fordsmand[7]

[1]Department of Biology & CESAM, University of Aveiro, 3810-193 Aveiro, Portugal

[2]Department of Production Engineering, University of Bremen, Badgasteiner Str. 1, 28359 Bremen, Germany

[3]Leibniz Institute for Materials Engineering IWT, Badgasteiner Str. 3, 28359 Bremen, Germany

[4]Energy Department, Politecnico di Torino, Corso Duca degli Abruzzi 24, Torino 10129, Italy

[5]INRIM, Istituto Nazionale di Ricerca Metrologica, Strada delle Cacce 91, Torino 10135, Italy

[6]Department of Chemistry, University of Tartu, Ravila 14a, Tartu 50411, Estonia

[7]Department of Bioscience, Aarhus University, Vejlsovej 25, PO BOX 314, DK-8600 Silkeborg, Denmark

† These authors contributed equally to the paper

*Corresponding author:

mjamorim@ua.pt;

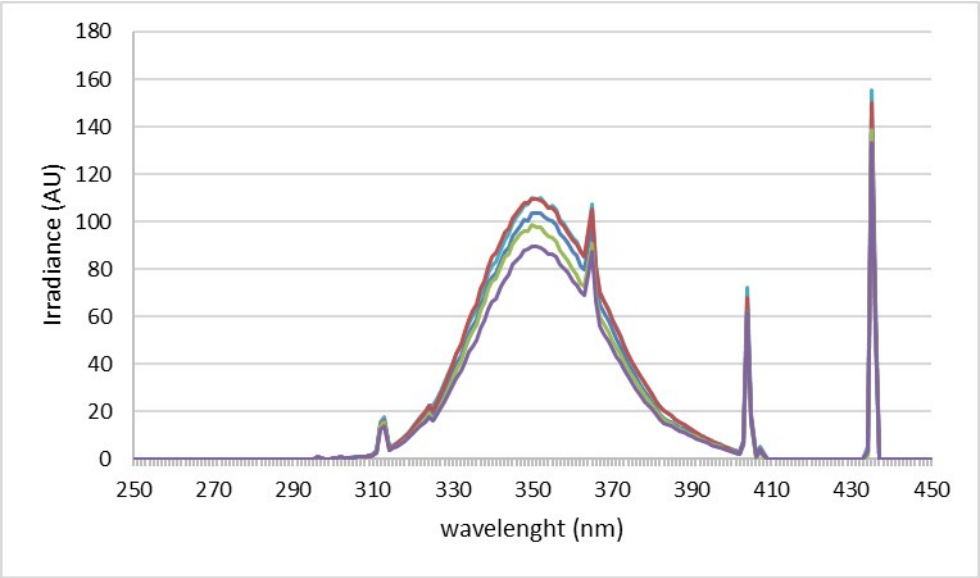

Figure S1. Emission spectra of the lamp used to provide UV radiation (UVP XX-15L Longwave, peak at 365 nm). The different colours correspond to the measurements performed in the different days. AU: arbitrary units

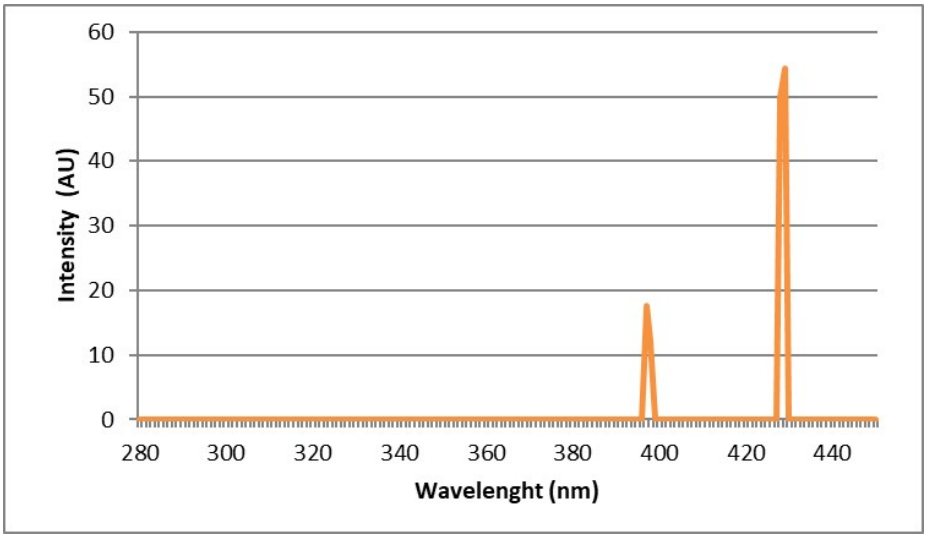

Figure S2. Emission spectra of the fluorescent lamp used for the Non-UV exposure. AU: arbitrary units

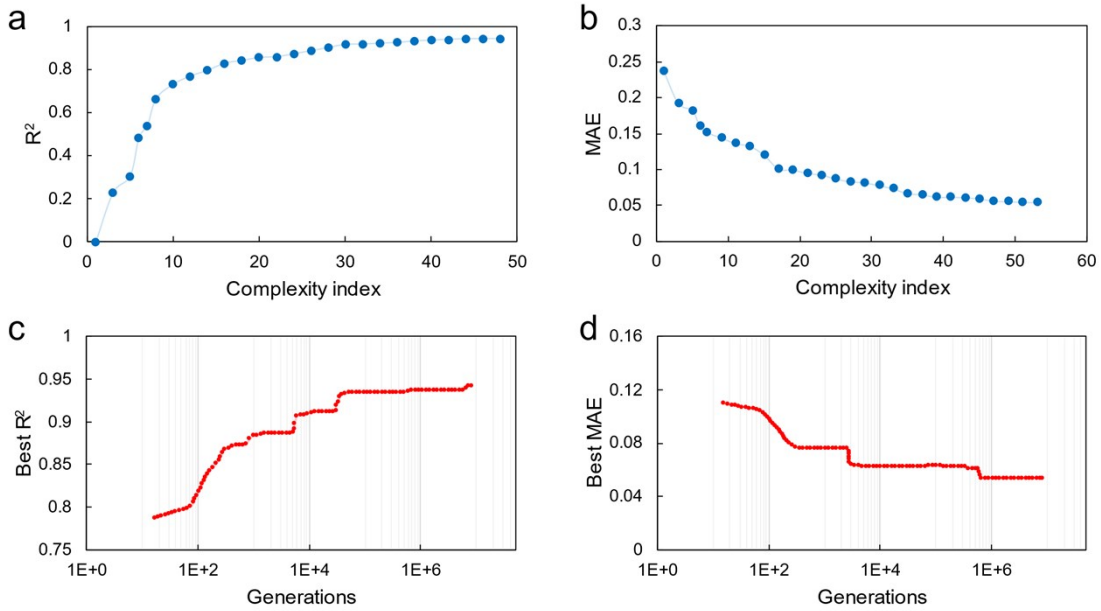

Figure S3. Pareto front of the list of suitable fitting functions *f* identified by the symbolic regressor: **(a)** the R^2 of the function tends to increase with more complex equations, whereas the **(b)** Mean Absolute Error (MAE) to decrease. Clearly, the most complex fitting equation tends to be the most accurate one, while the elbow of Pareto front can be considered as the best compromise between fitting accuracy and complexity of the equation. The following scores for the formula building-blocks are assigned (by default) by the *Eureqa* symbolic regressor to define the complexity index: 1 for constant, addition, subtraction, multiplication; 2 for division; 4 for exponential, natural logarithm, and square root.

Example of model fitting by the symbolic regression algorithm in one step of the pruning process: evolution of the **(c)** R^2 and **(d)** MAE of the best fitting equation during the generations (*i.e.*, iterations) of the genetic algorithm driving the symbolic regressor.

The results depicted in this figure refer to one repetition of the 2nd pruning round of experiments exposed to UV.

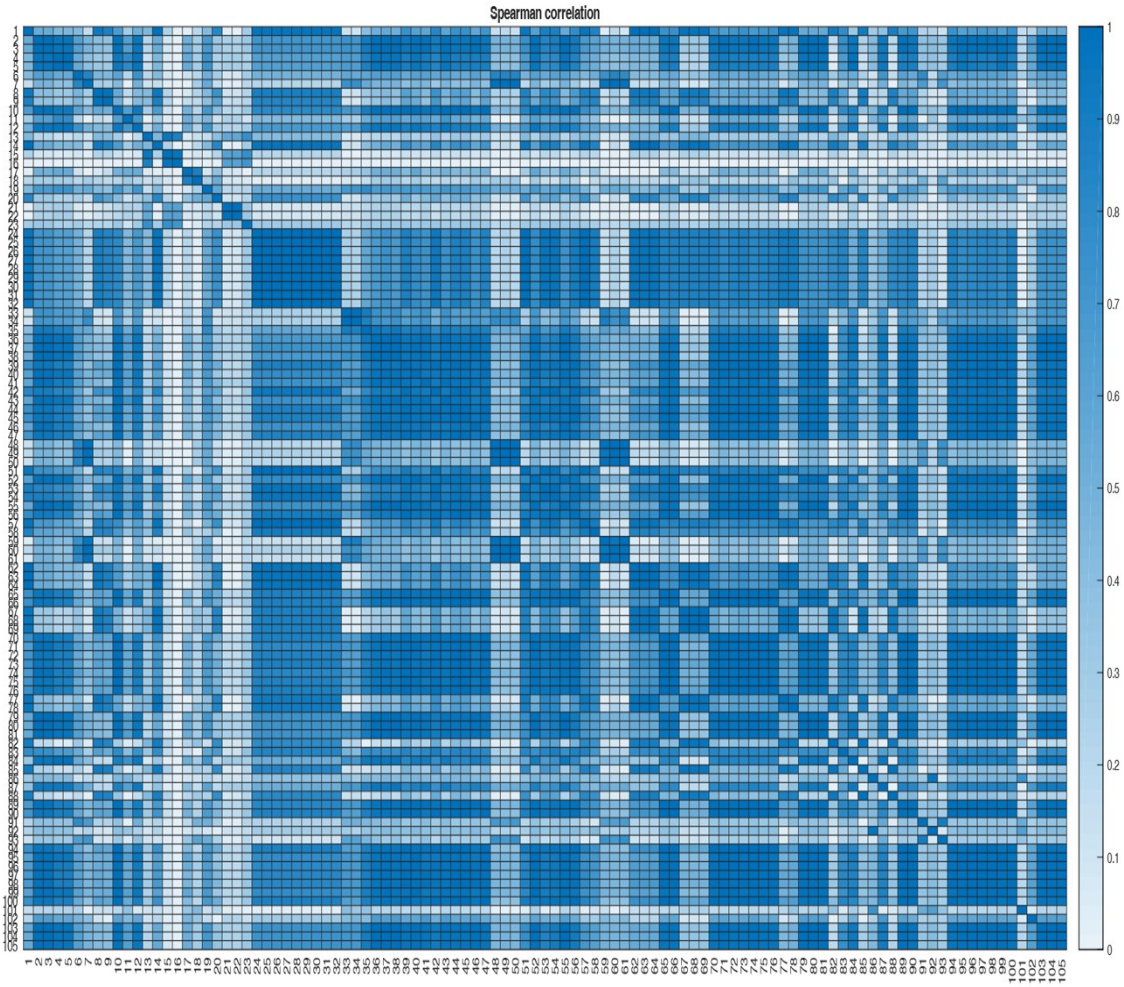

Figure S4. Spearman's correlation coefficient between each pair of TiO_2 toxicity variables (experiments without exposure to UV). The figure reports the 105 variables remaining after the dataset cleaning. The whiter colour tones indicate uncorrelation between each pair of variables, the blue ones indicate correlation. Notice that – given the definition of Spearman's correlation coefficient – the matrix is symmetrical.

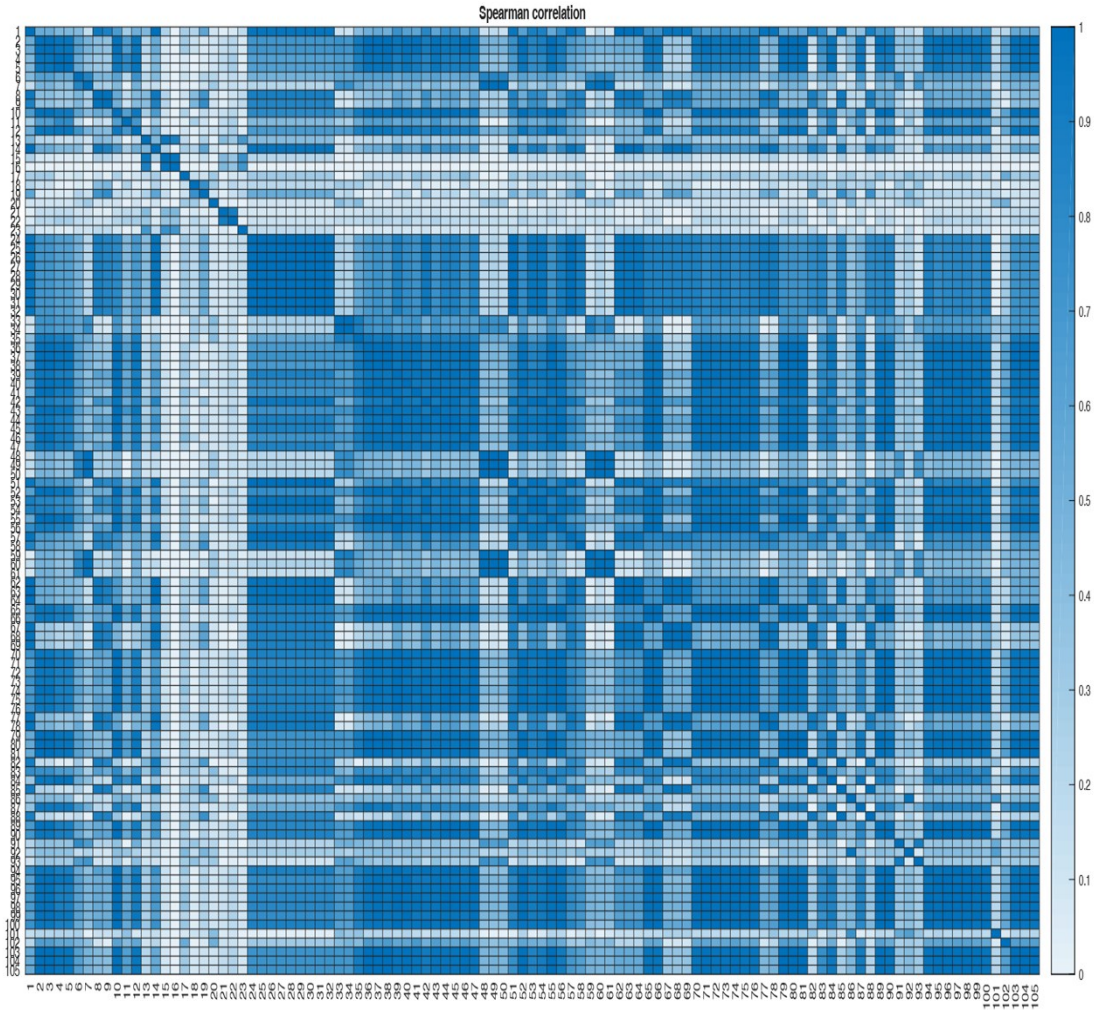

Figure S5. Spearman's correlation coefficient between each pair of TiO_2 toxicity variables (experiments exposed to UV). The figure reports the 105 variables remaining after the dataset cleaning. The whiter colour tones indicate uncorrelation between each pair of variables, the blue ones indicate correlation. Notice that – given the definition of Spearman's correlation coefficient – the matrix is symmetrical.

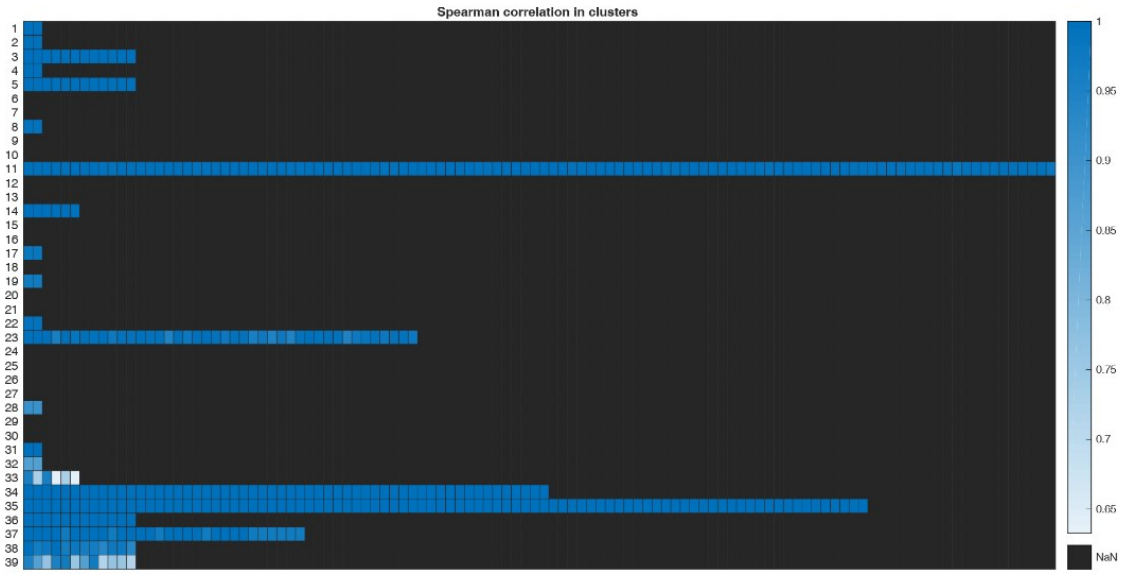

Figure S6. Spearman's correlation coefficient between each pair of variables within the 39 clusters identified by the hierarchical clustering algorithm for experiments without exposure to UV (see Table S2). The whiter colour tones indicate less correlation between each pair of variables, the blue ones more. Note that the black colour simply represents the background of the figure. Clearly, the Spearman's correlation coefficient cannot be computed in clusters made of a sole variable (e.g., cluster #6).

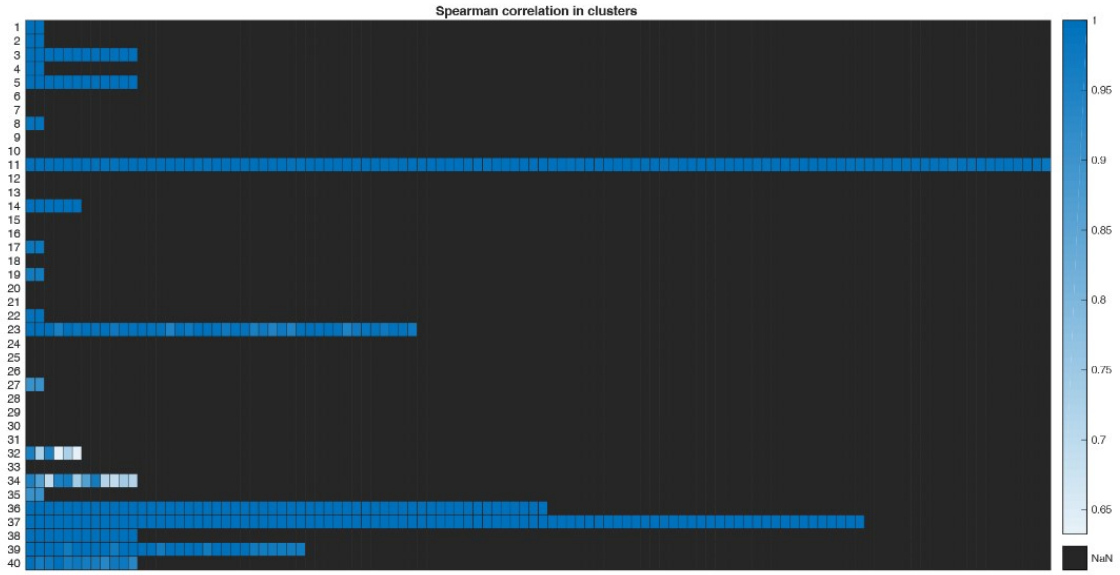

Figure S7. Spearman's correlation coefficient between each pair of variables within the 40 clusters identified by the hierarchical clustering algorithm for experiments exposed to UV (see Table S3). The whiter colour tones indicate less correlation between each pair of variables, the blue ones more. Note that the black colour simply represents the background of the figure. Clearly, the Spearman's correlation coefficient cannot be computed in clusters made of a sole variable (e.g., cluster #6).

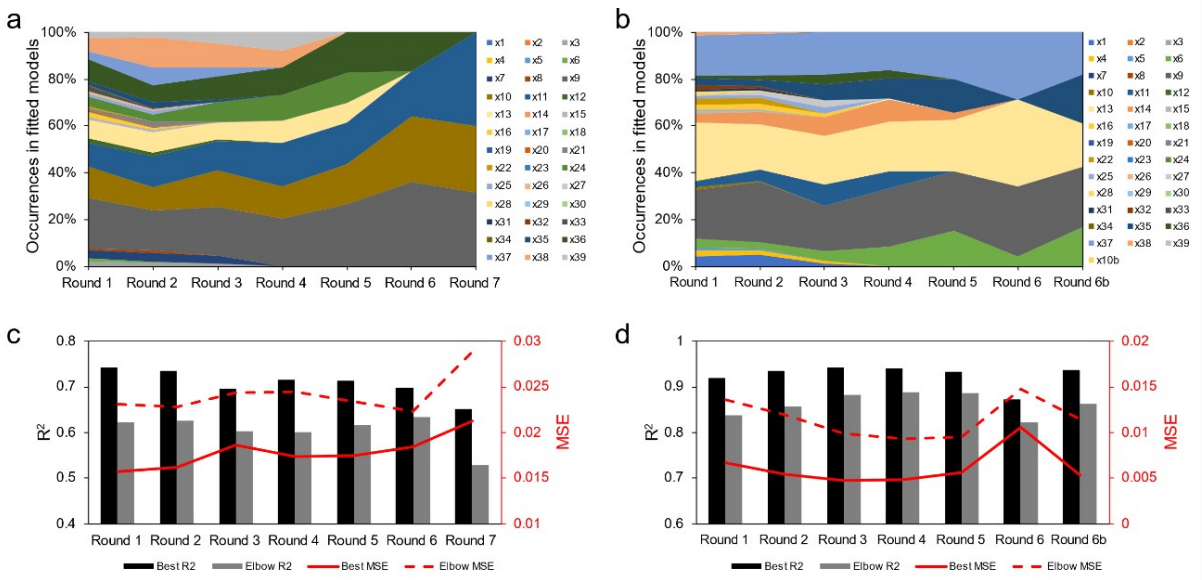

Figure S8. Results of variables pruning. Normalized occurrences of variables x_i in the fitting functions *f* identified by the symbolic regressor for experiments **(a)** without and **(b)** with UV exposure. The definitions of the reported variables $x_1, \ldots, x_{39}$ are reported in the Tables S4 (no exposure to UV) and S5 (exposure to UV). Several rounds of pruning are carried out, in which only the best ranked 40% of variables in terms of occurrence are kept, while the remaining ones pruned. This process is repeated until one of the chosen stopping criteria (based on either a decrease in the coefficient of determination – R^2 or on an increase in the Mean Squared Error – MSE) is met. This is achieved **(c)** at the 7th round for the experiments without UV exposure, **(d)** at the 6th round for the experiments with UV exposure. Notice that, for the UV exposure case, round #5 considers 6 variables (5th pruning iteration, stopping criteria not met), round #6 considers 4 variables (6th pruning iteration, stopping criteria met) and round #6b considers 5 variables (repetition of the 6th pruning iteration with more variables, stopping criteria not met). The symbolic regressor identifies a Pareto front of suitable *f* fitting functions, that is the best compromise between complexity and fitting accuracy of *f*. Here, the error metrics for the most accurate fitting equation ("best", which also has the highest complexity) and the one at the elbow of the Pareto front ("elbow", which shows the best compromise between fitting accuracy and complexity) are reported.

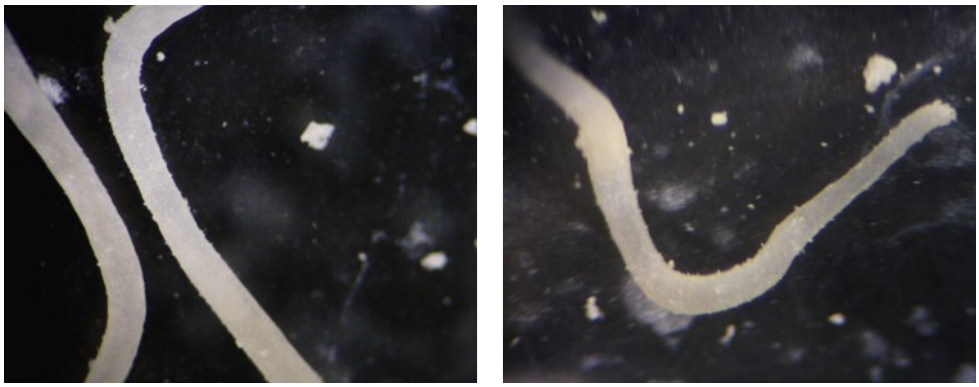

Figure S9. Representative pictures of *E. crypticus* exposed to 100 mg/L of TiO_2 NPs, in ISO water, for 5 days.

Supplementary Movie S1. Bar chart race of the normalized occurrence of variables x_i in the fitting functions f identified by the symbolic regressor for experiments without exposure to UV, per each pruning step. The definition of the reported variables $x_1, \ldots, x_{39}$ and their classification are reported in the Table S4. This movie has been made by Flourish.

Supplementary Movie S2. Bar chart race of the normalized occurrence of variables x_i in the fitting functions f identified by the symbolic regressor for experiments exposed to UV, per each pruning step. The definition of the reported variables $x_1, \ldots, x_{39}$ and their classification are reported in the Table S5. This movie has been made by Flourish.