**QSAR Model Reporting Format**

**1. QSAR identifier**

**1.1 QSAR identifier (title):** Quantitative Structure Activity Relationships (nano-QSAR) modelling predicting cellular viability of various NMs under diverse experimental conditions. The model is under review in the NanoImpact journal.

**1.2 Other related models:** Diverse models exist targeting cellular viability. However, no other related model has been trained with this specific merged dataset.

**1.3 Software coding the model:** The code has been written in Python. All the Jupiter notebooks can be found here: https://github.com/ammar257ammar/Nanosafety-cell-viability-pycaret.

**2. General information**

**2.1 Date of QMRF**: 10 May 2023.

**2.2 QMRF author(s) and contact details:** Irini Furxhi, Transgero Limited, Cullinagh, Newcastle West, Co. Limerick, Ireland & Dept. of Accounting and Finance, Kemmy Business School, University of Limerick, V94PH93, Ireland. irini.furxhi@transgero.eu. Irini.furxhi@ul.ie

**2.3 Date of QMRF update(s):** Not applicable

**2.4.QMRF update(s):** Not applicable

**2.5 Model developer(s) and contact details:** Ammar Ammar. Department of Bioinformatics BiGCaT, NUTRIM, Maastricht University, The Netherlands. Email: a.ammar@maastrichtuniversity.nl

**2.6 Date of model development and/or publication:** The model is under review in the NanoImpact journal.

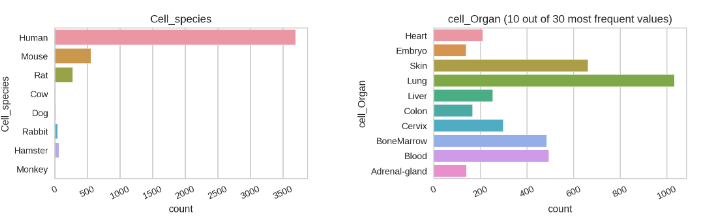
**2.7 Reference(s) to main scientific papers and/or software package:** *Low-code AutoML-augmented Data Pipeline – A Review and Experiments*. Ulla Gain and Virpi Hotti, Journal of Physics: Conference Series, Volume 1828, 2020 International Symposium on Automation, Information and Computing (ISAIC 2020) 2-4 December 2020, Beijing, China, DOI 10.1088/1742-6596/1828/1/012015

**2.8 Availability of information about the model:** All the Jupiter notebooks containing the diverse datasets, the data pre-processing, the validation and execution of the models (e.g., training and external validation sets, source code, and algorithm) can be found here: https://github.com/ammar257ammar/Nanosafety-cell-viability-pycaret. In this QMRF we report solely our best regression universal model.

**2.9 Availability of another QMRF for exactly the same model:** Not applicable

**3. Defining the endpoint – OECD Principle 1**

**3.1 Species:** Various species are taken into account from in vitro experimental settings such as human, mouse, rat, cow, dog, rabbit, hamster, monkey etc. The different species are also followed by the target organ (cell line cultured).



**3.2 Endpoint**: Biological effect, “Other” .

**3.3 Comment on the endpoint:** the hazard evaluation to be predicted based on cellular viability measurements in vitro is formulated either as a numerical output predicted via regression algorithms or expressed as two classes: toxic and not toxic, predicted via classification algorithms. In this QMRF we report solely our best regression universal model.

**3.4 Endpoint units:** Percentage (%) for the regression task or Binary categorical feature for the classification models. In this QMRF we report solely our best regression universal model.

**3.5 Dependent variable**: For the classification algorithms outcome is depended on the corresponding values of cell viability. The lower the viability value, the higher the cytotoxic potential. Thus, toxic were considered the data points with cell viability < 50.0%, and not toxic if when viability is >= 50.0% in a precautionary manner. In this QMRF we report solely our best regression universal model.

**3.6 Experimental protocol:**

**3.7 Endpoint data quality and variability:**.

**4. Defining the algorithm – OECD Principle 2**

**4.1 Type of model:** nano-QSAR models. In this QMRF we report solely our best regression universal model.

**4.2 Explicit algorithm:** Gradient Boosting Regressor

**4.3 Descriptors in the model:**

1. NP\_type object
2. Coating object
3. core\_size\_nm float64
4. hydro\_size\_nm float64
5. Surf\_charge\_mV float64
6. Surface\_area\_m2\_g float64
7. shape object
8. Dose\_microg\_mL float64
9. Duration\_h float64
10. Cell\_name object
11. Cell\_species object
12. cell\_Organ object
13. Cell\_morphology object
14. Cell\_age object
15. cell\_type object
16. sex object
17. Assay object
18. Test\_indicator object
19. nanomaterial\_group object

**4.4 Descriptor selection:** Various pre-processing configurations scenarios.

* ["imputation\_type"])
* ["numeric\_imputation"])
* ["ignore\_features"])
* ["high\_cardinality\_features"])
* ["normalize"])
* ["remove\_outliers"])

**4.5 Algorithm and descriptor generation:** we provide an overview of the eight datasets selected within the supplementary material of the paper. Features were harmonized, for example cell name was found with different annotations such as cell type, cell line, mammalian lines, Cl, C.name/line, Cell-identification etc. The same pattern was observed for cell representative organs (tissue, cell origin, cell-tissue-organ-origin, cell source) and other features such as coating. Exposure dose (μgram/ml) and duration (hour) were converted in the same metrics across the datasets, when feasible. From the available information of the cell name (e.g., BEAS-2B, WI-38, HBE, 16HBE14o) we manually derived information about the cell cultures (species, organ, type, age, gender etc.,) when it was absent. Same principle is followed regarding the assay, where we derived information for the test indicators. Due to the high amount of missing values in the reporting media (cell culture medium or water) where the pchem were measured, we merged the sizes regardless of the medium. In cases where both information was present for example core size measured in water and in culture media, the average values were calculated and taken into account in the final modeling. The same approach was followed for the hydrodynamic size and surface charge.

**4.6 Software name and version for descriptor generation:**

**4.7 Descriptors/Chemicals ratio**: 4,656 data rows with 20 inputs. 3724 rows of training data rows, 20 inputs.

**5. Defining the applicability domain – OECD Principle 3**

**5.1 Description of the applicability domain of the model:** The AD inspections show that 98.4% of the test set predictions fall within the applicability domain of the model.

**5.2 Method used to assess the applicability domain**: the domain of applicability has been defined for the top performing models using the leverage approach by inspecting the percentage of the test set data points falling outside the area defined by standardized residuals (S) and leverage and visualized using Williams plot. A prediction was considered an outlier if it falls outside the area (3 \* S) on the Y-axis and leverage hat value h\* (3P+1/N) on the X-axis where P is the number of independent variables and N is the number of instances in the training set.

**5.3 Software name and version for applicability domain assessment**:

**5.4 Limits of applicability:**

**6. Defining goodness-of-fit and robustness – OECD Principle 4**

**6.1 Availability of the training set:** All the Jupiter notebooks containing the diverse datasets, the data pre-processing, the validation and execution of the models (e.g., training and external validation sets, source code, and algorithm) can be found here: https://github.com/ammar257ammar/Nanosafety-cell-viability-pycaret. In this QMRF we report solely our best regression universal model.

**6.2 Available information for the training set:**

a) Chemical names: Yes

b) CAS numbers; No

c) SMILES; No

d) InChI codes; No

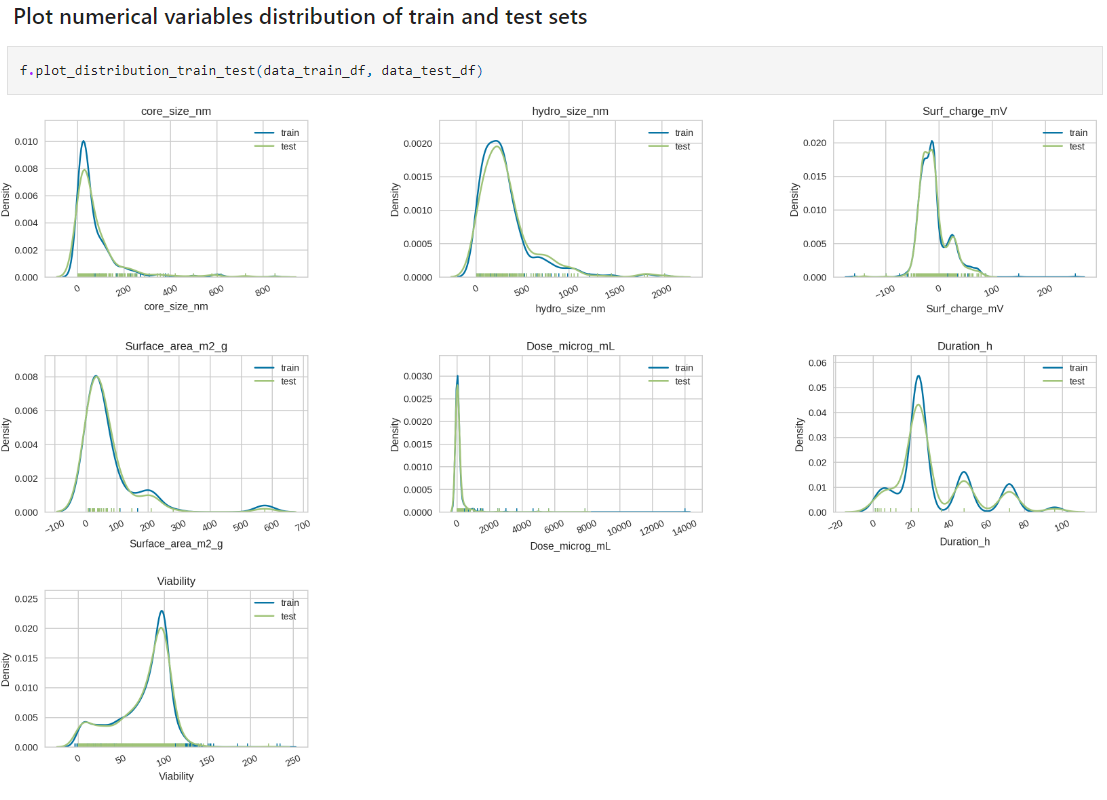
e) MOL files; No

f) Structural formula; No

**6.3 Data for each descriptor variable for the training set:** Yes.

**6.4 Data for the dependent variable (response) for the training set:** Yes.

**6.5 Other information about the training set:**



**6.6 Pre-processing of data before modelling:**

Ignored features: None

High cardinality features: None

Imputation type: simple

Numeric imputation method: mean

Normalization: True

Remove outliers: False

**6.7 Statistics for goodness-of-fit:**

| **Fold** | **MAE** | **MSE** | **RMSE** | **R2** | **RMSLE** | **MAPE** |
| --- | --- | --- | --- | --- | --- | --- |
| **0** | 11.79 | 314.71 | 17.74 | 0.70 | 0.56 | 0.83 |
| **1** | 10.98 | 244.35 | 15.63 | 0.74 | 0.56 | 1.82 |
| **2** | 11.62 | 277.96 | 16.67 | 0.71 | 0.51 | 0.55 |
| **3** | 11.30 | 279.05 | 16.70 | 0.73 | 0.50 | 0.78 |
| **4** | 12.09 | 315.41 | 17.76 | 0.69 | 0.51 | 0.67 |
| **Mean** | 11.56 | 286.30 | 16.90 | 0.71 | 0.53 | 0.93 |

**6.8 Robustness – Statistics obtained by leave-one-out cross-validation:**

**6.9 Robustness – Statistics obtained by leave-many-out cross-validation:**

Outer CV: 10-fold cross-validation for the tuned model

gbr scores: 0.85, 0.81, 0.87, 0.83, 0.9, 0.86, 0.79, 0.87, 0.85, 0.84

gbr mean/std: 0.85 / 0.03

**6.10 Robustness – Statistics obtained by Y-scrambling:**

**6.11 Robustness – Statistics obtained by bootstrap:**

**6.12 Robustness – Statistics obtained by other methods:**

**7. Defining predictivity – OECD Principle 4**

**7.1 Availability of the external validation set:** All the Jupiter notebooks containing the diverse datasets, the data pre-processing, the validation and execution of the models (e.g., training and external validation sets, source code, and algorithm) can be found here: https://github.com/ammar257ammar/Nanosafety-cell-viability-pycaret. In this QMRF we report solely our best regression universal model.

**7.2 Available information for the external validation set:**

a) Chemical names: Yes

b) CAS numbers; No

c) SMILES; No

d) InChI codes; No

e) MOL files; No

f) Structural formula; No

**7.3 Data for each descriptor variable for the external validation set:** Yes.

**7.4 Data for the dependent variable for the external validation set:** Yes

**7.5 Other information about the external validation set:**.

**7.6 Experimental design of test set:** datasetsplit into a training (3724 rows) and test set (932 rows).

**7.7 Predictivity – Statistics obtained by external validation:**.

**7.8 Predictivity – Assessment of the external validation set:**

|  | **Model** | **MAE** | **MSE** | **RMSE** | **R2** | **RMSLE** | **MAPE** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | Gradient Boosting Regressor | 7.60 | 135.26 | 11.63 | 0.87 | 0.36 | 0.58 |

**7.9 Comments on the external validation of the model:**

**8. Providing a mechanistic interpretation – OECD Principle 5**

**8.1 Mechanistic basis of the model:**

**8.2 A priori or a posteriori mechanistic interpretation:**

**8.3 Other information about the mechanistic interpretation:** Feature importance was derived for the top universal performing model as a means of decreasing impurity. The sum of importance is equal to 1 and each feature contributes to the overall prediction with a corresponding value. This analysis allows the identification of the most influential features that allow the accurate prediction of the defined endpoint (biological effect defined as the cellular viability measured in vitro and expressed as % of viable cells under specific exposure conditions).

**9. Miscellaneous information**

**9.1 Comments:**

**9.2 Bibliography:**

**9.3 Supporting information:** The code has been written in Python. All the Jupiter notebooks can be found here: https://github.com/ammar257ammar/Nanosafety-cell-viability-pycaret..

**10. Summary for the ECB Inventory**

**10.1 QMRF number:**

**10.3 Keywords:**

**10.4 Comments**: