# **Electronic supplementary information for:**

# **Computational enrichment of physico-chemical data for a nanomaterial library and development of a ζ-potential read-across model with Isalos Analytics**

Anastasios G. Papadiamantis,a,b Antreas Afantitis,a Andreas Tsoumanis,a Eugenia Valsami-Jones,a Iseult Lynch,\*b Georgia Melagraki\*a

a Nanoinformatics Departments, NovaMechanics Ltd, 1065, Nicosia, Cyprus

b School of Geography, Earth and Environmental Sciences, University of Birmingham, B15 2TT, Birmingham, United Kingdom



**Figure S1.** Scatterplot of the correlation between the normalised experimentally measured ζ-potential and the normalised predicted ζ-potential values for the test set following model development. The red line corresponds to the experimental and predicted values being equal.

## S1: QSAR model report form (QMRF) for the ζ-prediction model, following the OECD template

Model for prediction of the ζ-potential of engineered inorganic nanomaterials in water, developed using the Isalos Analytics Platform, Enalos+ nodes.

In order to demonstrate that the developed model is fully consistent with the OECD principles for predictive model validation for regulatory purposes (OECD, 2004), we summarise here all available information about the development and evaluation of the read-across model in a concise manner. For this purpose, the guidance of the JRC QSAR Model Database is followed, making the necessary alterations for nanoinformatics data (Joint Research Centre, 2017).

### Principle 1–A defined endpoint.

|  |  |
| --- | --- |
| Species | ζ-potential of inorganic engineered nanomaterials (ENM) |
| Endpoint | ζ-potential of inorganic ENM in water |
| Endpoint comments | ζ-potential |
| Endpoint units | Millivolts (mV) |
| Dependent variable | For inorganic (metal-based) ENM the ζ-potential of the ENM in water was predicted. |
| Experimental protocol | Full experimental description can be found in Joosens et al., Scientific Data 2019, 6, (1), 46, **DOI**: <https://doi.org/10.1038/s41597-019-0053-2>  Short description: The dataset consists of pristine Ag (n = 3), AgO (n = 1), Ag2S (n = 1), AlOOH (n = 1), Au (n = 5), BaTiO3 (n = 1), CeO2 (n = 13), Zr-doped CeO2 (n = 5) and their (n = 5) aged equivalents, CePO4 (n = 1), CuO (n = 1), Fe2O3 (n = 1), hydroxyapatite (n = 1), SiO2 (n = 14), TiO2 (n = 8), ZnO (n = 7) and ZrO2 (n = 1) ENMs. The dataset contains information on the core chemistry, where applicable coating type and coating charge (positive, negative, neutral), morphology (identified using TEM) and whether the particles were aged or not (no ageing protocol information was provided). Furthermore, the physicochemical characterisation of the ENM in MilliQ water includes: core size (measured using TEM or STEM), hydrodynamic diameter, ζ-potential, specific and geometric surface area (calculated using primary particle size and morphology) and energy band gap (calculated from UV-Vis measurements using Tauc plots). |
| Endpoint data quality and variability | The ζ-potential in milliQ water was measured experimentally for 69 different ENM. No data gaps existed in the dataset. |

### Principle 2–An unambiguous algorithm.

|  |  |
| --- | --- |
| Type of model | Machine learning, *k*-nearest neighbour (*k*NN) algorithm |
| Explicit algorithm | Use *k*NN with *k* value equal to 7. |
| Descriptors in the model | Statistically significant descriptors used for ζ-potential prediction: ENM core size, type of coating, metal ionic radius, absolute electronegativity, sum of metal electronegativity divided by the number of oxygen atoms present in a particular metal oxide. |
| Descriptor selection | Number and type of descriptors initially screened: 20 descriptors:  Physicochemical: chemical formula, type of coating, coating charge, ageing (Yes/No), core size, morphology, geometric surface area  Molecular descriptors: metal atomic and ionic radii, metal electronegativity, number of metal atoms, number of oxygen atoms, sum of metal electronegativity, sum of metal electronegativity divided by the number of oxygen atoms present in a particular metal oxide, metal charge, metal periodic table period and group, absolute electronegativity, energy band gap.  Method used to select the descriptors: *BestFirst* variable selection along with *CfsSubsetEval* evaluator. |
| Algorithm and descriptor generation | Experimental measurements: See Joosens et al., Scientific Data 2019, 6, (1), 46, **DOI**: <https://doi.org/10.1038/s41597-019-0053-2> |
| Software name and version for descriptor generation | Enalos+ nodes version 1.0 (see Afantitis et al. (2020), **DOI**: 10.2174/0929867327666200727114410) |
| ENMs/Descriptors ratio | 7.14 (50:7, number of data rows divided by the number of significant descriptors in the training set) |

### Principle 3–A defined domain of applicability.

|  |  |
| --- | --- |
| Description of the applicability domain of the model | The applicability domain (APD) is defined by fixed boundaries (threshold). The threshold is calculated by considering Euclidean distances between all training set ENMs. |
| Method used to assess the applicability domain | Euclidean distance method among all training and test ENMs. The distance of each test ENM or each external ENM to each nearest neighbour of the training ENMs set is compared to a predefined APD threshold; if this distance is lower than the threshold then its endpoint prediction can be considered reliable. |
| Software name and version for applicability domain assessment | Enalos+ nodes version 1.0 (see Afantitis et al. (2020), **DOI**: <https://doi.org/10.2174/0929867327666200727114410>) |
| Limits of applicability | APD threshold: 1.947  Calculated using the average and standard deviation of all Euclidian distances in the training set. Predictions outside this threshold are considered unreliable. |

### Principle 4–Appropriate measures of goodness-of-fit, robustness and predictivity.

|  |  |
| --- | --- |
| Availability of the training set | Data available via NanoPharos at https://db.nanopharos.eu/  Dataset name: Papadiamantis et al. (2020), Enrichment, meta-analysis and development of a ζ-potential read-across and safe by design nanoinformatics model for engineered nanomaterials with Isalos Analytics  Information: Joosens et al., Scientific Data 2019, 6, (1), 46, DOI: 10.1038/s41597-019-0053-2 |
| Available information for the training set | 50 tested ENM: Ag, AgO, Ag2S, AlOOH, Au, Ca10(PO4)6(OH)2, CeO2, CexZryO2 (x + y = 1), CePO4, CuO, Fe2O3, Fe3O4, SiO2, TiO2, ZnO  Analytical information on the experimental process can be found in Joosens et al., Scientific Data 2019, 6, (1), 46, **DOI**: <https://doi.org/10.1038/s41597-019-0053-2>  Information on the molecular descriptors calculation can be found at: Kar et al. Ecotoxicology and Environmental Safety, 2014, 107, 162-169, **DOI:** <https://doi.org/10.1016/j.ecoenv.2014.05.026>and Zhang et al., ACS Nano, 2012, 6, 4349-4368, **DOI:** <http://dx.doi.org/10.1021/nn3010087> |
| Data for each descriptor variable for the training set | Yes |
| Data for the dependent variable (response) for the training set | Yes |
| Other information about the training set | Total of 50 data points for the dependent variable (ζ-potential) and 1,050 for the independent variables from 50 ENM. |
| Pre-processing of data before modelling | Gaussian normalisation of descriptors |
| Statistics for goodness-of-fit | Tropsha’s tests, i.e. the coefficient of determination between experimental values and model predictions (R2), validation through an external test set, leave-many-out cross validation procedure and Quality of Fit and Predictive Ability of a continuous QSAR Model, as per Tropsha, Best Practices for QSAR Model Development, Validation, and Exploitation. Molecular Informatics 2010, 29, (6‐7), 476-488 https://doi.org/10.1002/minf.201000061 |
| R2 > 0.6 | 0.90 |
| Rcvext > 0.5; Result: 0.904 | 0.913 |
| (R2-R02)/R2 < 0.1 | 0.005 |
| (R2-R0'2)/R2 < 0.1 | 0.001 |
| |R2-R0'2 |<0.3 | 0.004 |
| 0.85 < k < 1.15 | 1.053 |
| 0.85 < k’ < 1.15 | 0.869 |
| Robustness – Statistics obtained by Y-scrambling | Accuracy in test = 0.00 (10 iterations) |
| Robustness – Statistics obtained by bootstrap | Yes, see above |
| Robustness – Statistics obtained by other methods | Yes, see above |
| Availability of the external validation set | Yes |
| Available information for the external validation set | Yes |
| Data for each descriptor variable for the external validation set | Yes |
| Data for the dependent variable for the external validation set | Yes |
| Other information about the external validation set | Total of 20 data points from 20 ENM for the dependent variable (ζ-potential).  420 data points for the independent variables. |
| Experimental design of test set | Partitioning of the initial dataset using random, stratified sampling (75:25 training : test sets) |
| Predictivity – Statistics obtained by external validation | R2 > 0.6; Result: 0.90  Rcvext > 0.5; Result: 0.913  (R2-R02)/R2 < 0.1; Result: 0.005  (R2-R0'2)/R2 < 0.1; Result: 0.001  |R2-R0'2 |<0.3; Result: 0.004  0.85 < k < 1.15; Result: 1.053  0.85 < k’ < 1.15; Result: 0.869 |
| Predictivity – Assessment of the external validation set | The external validation set is 25% of the initial dataset and all predictions for the validation set fall within the domain of applicability |
| Comments on the external validation of the model | Ν/Α |

### Principle 5–Α mechanistic interpretation.

|  |  |
| --- | --- |
| Mechanistic basis of the model | Absolute electronegativity, metal ionic radius and sum of metal electronegativity divided by the number of oxygen atoms present will regulate the surface charge of the bare ENM. ENM core size will determine the Stern and diffusion layers and the distance from the surface of the bare ENM over which its charge will persist. ENM coating will affect the resulting observed ζ-potential. If sufficiently thick, the coating may mask the surface charge of the bare ENM and the observed ζ-potential will depend on the coating charge. If not thick enough, the observed ζ-potential will be the result of the interaction of the coating charge and thickness with the base ENM surface charge. |
| A priori or a posteriori mechanistic interpretation | The ζ-potential of an ENM can be predicted using a combination of molecular and physicochemical descriptors. These descriptors play a critical role in the resulting surface charge and thus ζ-potential of the ENM. The molecular descriptors will directly affect the surface charge of the pristine ENM. The size of the ENM will affect the size of the Stern and diffusion layers and the amount by which the coating can mask this surface charge. The thickness and surface of the coating will affect the resulting observed ζ-potential. |
| Other information about the mechanistic interpretation | No other information available. |