# (Q)SAR model reporting format (QMRF)

| | Element | Explanation |
|---|---|---|
| **1.** | **QSAR identifier** | |
| 1.1. | QSAR identifier (title) | CeresAI-nano: Machine learning-based prediction of plant length response under nanoparticle exposure |
| 1.2 | Other related models | Deng, Peng, et al. "Development potential of nanoenabled agriculture projected using machine learning." *Proceedings of the National Academy of Sciences* 120.25 (2023): e2301885120. https://doi.org/10.1073/pnas.2301885120 |
| 1.3. | Software coding the model | KNIME Analytics Platform v5.1.2<br>Isalos Analytics Platform v1.0.0<br><br>Link to web service:<br>https://enaloscloud.novamechanics.com/chiasma/agrinano/ |
| **2.** | **General information** | |
| 2.0 | Abstract | A machine learning model for the prediction of the length response (positive/negative) of the root, shoot, or whole plant following nanoparticle (NP) treatment. The predictions are based on plant properties, NPs' atomistic descriptors and experimental conditions. |
| 2.1. | Date of QMRF | May 22, 2025 |
| 2.2. | QMRF author(s) and contact details | Dimitra-Danai Varsou: varsou@novamechanics.com<br>Aikaterini Theodori: theodori@novamechanics.com<br>Andreas Tsoumanis: tsoumanis@novamechanics.com<br>Maria Antoniou: antoniou@novamechanics.com<br>Georgia Melagraki: georgiamelagraki@gmail.com<br>Antreas Afantitis: afantitis@novamechanics.com |
| 2.3. | Date of QMRF update(s) | NA |
| 2.4. | QMRF update(s) | NA |

| | | |
|---|---|---|
| 2.5. | Model developer(s) and contact details | Dimitra-Danai Varsou: varsou@novamechanics.com<br>Aikaterini Theodori: theodori@novamechanics.com<br>Andreas Tsoumanis: tsoumanis@novamechanics.com<br>Maria Antoniou: antoniou@novamechanics.com<br>Georgia Melagraki: georgiamelagraki@gmail.com<br>Antreas Afantitis: afantitis@novamechanics.com |
| 2.6. | Date of model development and/or publication | May 10, 2025 |
| 2.7. | Reference(s) to main scientific papers and/or software package | |
| 2.8. | Availability of information about the model | The model is proprietary: the source code is confidential; however, the description of the modelling workflow is presented in the original research article, the model is implemented as a public web service, and the curated and enriched dataset used for model development is available in the nanoPharos database. |
| 2.9. | Availability of another QMRF for exactly the same model | NA |
| **3** | **Defining the endpoint - OECD Principle 1: "A DEFINED ENDPOINT"** | **PRINCIPLE 1: "A DEFINED ENDPOINT". ENDPOINT refers to any physicochemical, biological, or environmental property/activity/effect that can be measured and therefore modelled. The intent of PRINCIPLE 1 (a (Q)SAR should be associated with a defined endpoint) is to ensure clarity in the endpoint being predicted by a given model, since a given endpoint could be determined by different experimental protocols and under different experimental conditions. It is therefore important to identify the experimental system and test conditions that is being modelled by the (Q)SAR.** |
| 3.1. | Species | Crop plants grown for food, including cucumber, bean, wheat, rice, tomato, and maize. |
| 3.2. | Endpoint | Plant length response |
| 3.3 | Comment on endpoint | The length of the root, shoot, or whole plant following NP treatment is classified as positive or negative compared to the control treatment. |
| 3.4. | Endpoint units | NA |
| 3.5. | Dependent variable | When the root, shoot, or whole plant length is greater than that of the control treatment, the response to NP treatment is classified as "positive"; otherwise, it is classified as "negative". |
| 3.6. | Experimental protocol | Modelling was performed upon the dataset compiled by Deng et al. (2023), which was extensively curated and enriched using computationally derived atomistic descriptors for the NPs. Details of the experimental protocol for the NP-plant interactions studies included in the original dataset can be |

<table>
<tr><td></td><td></td><td>found in the supporting information accompanying the work of Deng et al. The protocol of the generation of the atomistic descriptors is described in §4.5.</td></tr>
<tr><td>3.7.</td><td>Endpoint data quality and variability</td><td>Details of the experimental protocol for the plant length response measurement can be found in the studies included in the original dataset of Deng et al. (see supporting information files).<br><br>Following cleansing of the “Length” dataset, the positive values of the label “Length” were re-calculated and normalised based on the control value to extract a more interpretable output (originally positive values were normalised based on the experimental value while negative values were normalised based on the control value). In cases where raw length values were not directly available (e.g., were not provided in textual format), the “WebPlotDigitizer” tool was employed to extract these values from plots presented in the original referenced articles. The systematic error associated with the digitisation process was quantified using the relative absolute error, which was estimated to be less than 1%. Then, for the output variable of length a specific class was assigned, “positive” (for positive responses, i.e., increased length) or “negative” (for negative responses, i.e., reduced length) based on the NP impact compared to the controls. As a result, the 25 control data points were removed as they did not fall under either of the “positive”/“negative” classes.<br><br>Classes distribution between sets after data curation:<br><br>
<table>
<tr><th>Class</th><th>Training set before oversampling</th><th>Training set after oversampling</th><th>Test set</th></tr>
<tr><td>Positive</td><td>0.26</td><td>0.50</td><td>0.25</td></tr>
<tr><td>Negative</td><td>0.74</td><td>0.50</td><td>0.75</td></tr>
</table>
<br>The oversampling is explained in detail in §6.6.</td></tr>
<tr><td>4</td><td>Defining the algorithm - OECD Principle 2: “AN UNAMBIGUOUS ALGORITHM”</td><td>PRINCIPLE 2: “AN UNAMBIGUOUS ALGORITHM”. The (Q)SAR estimate of an endpoint is the result of applying an ALGORITHM to a set of structural parameters which describe the chemical structure. The intent of PRINCIPLE 2 (a (Q)SAR should be associated with an unambiguous algorithm) is to ensure transparency in the model algorithm that generates predictions of an endpoint from information on chemical structure and/or physicochemical properties. In this context, algorithm refers to any mathematical equation, decision rule or output approach.</td></tr>
<tr><td>4.1.</td><td>Type of model</td><td>Decision trees, Extreme Gradient Boosting (XGBoost)</td></tr>
</table>

| | | |
|---|---|---|
| 4.2. | Explicit algorithm | The Extreme Gradient Boosting (XGBoost) open-source library is used to implement the gradient boosting framework, which relies on a class of ensemble machine learning algorithms built from decision tree models. Ensemble learning combines multiple base learners to produce a more accurate final prediction. In XGBoost, trees are added iteratively to the ensemble, with each new tree aiming to reduce the prediction error (loss) of the preceding models. For classification tasks, XGBoost offers distinct loss functions for binary and multiclass problems. |
| 4.3. | Descriptors in the model | 1. Total concentration: Total exposure concentration of NP treatment in mg/L<br>2. Species: Plant species (cucumber, bean, wheat, rice, tomato, maize)<br>3. MT: Measured tissue (root, shoot, plant)<br>4. Cultured: Cultivation method (medium, hydroponic, soil)<br>5. Category: Plant carbon fixation metabolic pathway ($C_3$ or $C_4$ photosynthetic process)<br>6. Duration: NP treatment duration (time elapsed from the exposure of the plant to NPs to the measurement) in days<br>7. Photoperiod: Hours of plant exposure to light per day in hours/day<br>8. D12: The average difference of the coordination parameter between core and shell atoms which can be calculated via the NanoConstruct tool: http://enaloscloud.novamechanics.com/riskgone/nanoconstruct/ |
| 4.4. | Descriptor selection | From the initial pool of descriptors (69 in total), 31 were filtered out using low variance and correlation filtering (see §6.6). The information gain of all remaining descriptors (38) was calculated and descriptors with zero information gain score were excluded from the modelling, as they were not considered critical for establishing a predictive relationship. 20 descriptors were selected from the InfoGain filtering. Finally, to assess the stability and relevance of each of the 20 remaining descriptors, permutation importance was calculated within a 10-fold cross-validation scheme. In each fold, individual descriptors in the validation subset were randomly shuffled, and the resulting drop in the predictive performance of the XGBoost model (compared to the baseline/unshuffled validation subset) was recorded. Normalization was performed within the cross-validation loop to prevent data leakage. A feature was considered important if its mean permutation-induced performance drop across folds was greater than zero. Descriptors with consistently low importance were removed prior to final training. The important features that emerged from the feature permutation process are listed in §4.3. |

<table>
<tr><td>4.5.</td><td>Algorithm and descriptor generation</td><td>Atomistic simulations. To perform the simulations and acquire the computational descriptors, the size, the shape, and the phase of the NPs were needed. The required crystallographic information files (CIF) were obtained from the Crystallography Open Database (COD).<br><br>
<table>
<tr><th>NP core</th><th>Diameter [nm]</th><th>COD DB code of CIF</th><th>Force field using OPENKIM ID</th></tr>
<tr><td>Ag</td><td>13.8, 20</td><td>1509146</td><td>EAM_Dynamo_AcklandTichyVitek_1987v2_Ag__MO_055919219575_000</td></tr>
<tr><td>GQD</td><td>2.5</td><td>1200017</td><td>DUNN_WenTadmor_2019v1_C__MO_584345505904_000</td></tr>
<tr><td>CuO</td><td>30, 40</td><td>1011148</td><td>LJ_ElliottAkerson_2015_Universal__MO_959249795837_003</td></tr>
<tr><td>$Fe_3O_4$</td><td>6.7</td><td>1011032</td><td>EAM_Dynamo_AcklandTichyVitek_1987_Ag__MO_212700056563_005</td></tr>
<tr><td>$SiO_2$</td><td>15</td><td>9011493</td><td>Sim_LAMMPS_Vashishta_BroughtonMeliVashishta_1997_SiO__SM_422553794879_000</td></tr>
<tr><td>$TiO_2$</td><td>6.5, 21</td><td>1010942</td><td>Sim_LAMMPS_MEAM_ZhangTrinkle_2016_TiO__SM_513612626462_000</td></tr>
<tr><td>ZnO</td><td>25</td><td>1011258</td><td>Sim_LAMMPS_ReaxFF_RaymandVanDuinBaudin_2008_ZnOH__SM_449472104549_001</td></tr>
</table>
</td></tr>
<tr><td>4.6.</td><td>Software name and version for descriptor generation</td><td>NanoConstruct: Nanoparticle Construction Tool Powered by Enalos RiskGONE Cloud Platform, http://enaloscloud.novamechanics.com/riskgone/nanoconstruct/</td></tr>
<tr><td>4.7.</td><td>Chemicals/Descriptors ratio</td><td>70 observations (training set after oversampling) / 8 features (see §4.3)</td></tr>
<tr><td>5</td><td>Defining the applicability domain - OECD Principle 3: “A DEFINED DOMAIN OF APPLICABILITY”</td><td>PRINCIPLE 3: “A DEFINED DOMAIN OF APPLICABILITY”. APPLICABILITY DOMAIN refers to the response and chemical structure space in which the model makes predictions with a given reliability. Ideally the applicability domain should express the structural, physicochemical and response space of the model. The CHEMICAL STRUCTURE (x variable) space can be expressed by information on physicochemical properties and/or structural fragments. The RESPONSE (y variable) can be any physicochemical, biological or environmental effect that is being predicted. According to PRINCIPLE 3 a (Q)SAR should be associated with a defined domain of applicability. Section 5 can be repeated (e.g., 5.a, 5.b, 5.c, etc) as many times as necessary if more than one method</td></tr>
</table>

| | | |
|---|---|---|
| | | **has been used to assess the applicability domain.** |
| 5.1. | Description of the applicability domain of the model | The applicability domain (APD) is defined by fixed boundaries, the APD threshold (§5.2), calculated by considering Euclidean distances between all molecules in the training set. The distance of a test compound to its nearest neighbour in the training set is compared to the predefined applicability domain threshold. If the distance is beyond this threshold, then the prediction is considered unreliable. |
| 5.2. | Method used to assess the applicability domain | The distance of a test molecule to its nearest neighbour in the training set is compared to the pre-defined threshold, $thr = \langle d \rangle + Z\sigma$. First, the average Euclidean distances between all pairs of training data are calculated and then the set of distances that were lower than the average is formulated. The $\langle d \rangle$ and $\sigma$ values are finally determined as the average and standard deviation of all distances included in the remaining set. $Z$ is an empirical parameter with a value of 0.5. |
| 5.3. | Software name and version for applicability domain assessment | Isalos Analytics Platform v1.0.0,<br>Function: Statistics → Applicability Domain → APD |
| 5.4. | Limits of applicability | APD threshold = 5.071 |
| **6** | **Defining goodness-of-fit and robustness (internal validation) – OECD Principle 4: "APPROPRIATE MEASURES OF GOODNESS-OF-FIT, ROBUSTENESS AND PREDICTIVITY"** | **PRINCIPLE 4: "APPROPRIATE MEASURES OF GOODNESS-OF-FIT, ROBUSTENESS AND PREDICTIVITY". PRINCIPLE 4 expresses the need to perform validation to establish the performance of the model. GOODNESS-OF-FIT and ROBUSTNESS refer to the internal model performance.** |
| 6.1. | Availability of the training set | The training set is available in the nanoPharos database: https://db.nanopharos.eu/Queries/Datasets.zul?datasetID=np31 |
| 6.2. | Available information for the training set | NP-plant interactions dataset from Deng et al., 2023, cleaned and computationally enriched. |
| 6.3. | Data for each descriptor variable for the training set | The training set is available in the nanoPharos database: https://db.nanopharos.eu/Queries/Datasets.zul?datasetID=np31 |
| 6.4. | Data for the dependent | The training set is available in the nanoPharos database: https://db.nanopharos.eu/Queries/Datasets.zul?datasetID=np3 |

<table>
<tr><td></td><td>variable for the training set</td><td>1</td></tr>
<tr><td>6.5.</td><td>Other information about the training set</td><td>47 (12 “positives” and 35 “negatives”) out of the 113 NP treatments of the Deng et al. (2023) curated dataset were selected as the training set through stratified random partitioning, i.e., the distribution of “positive” and “negative” classes was (approximately) retained in the training and test sets.</td></tr>
<tr><td>6.6.</td><td>Pre-processing of data before modelling</td><td>To correct class imbalance (see §3.7) the training set was oversampled using the Synthetic Minority Over-sampling Technique (SMOTE, with k=5) to a final count of 70 treatments (35 “positives” and 35 “negatives”). Data preprocessing also included: Low variance filter (cutoff limit of 20% to filter out columns), two-sided Spearman’s correlation analysis (upper threshold of 0.99 to filter out highly correlated columns), Gaussian normalization of descriptors (z-score). Variable selection was performed according to §4.4.</td></tr>
<tr><td>6.7.</td><td>Statistics for goodness-of-fit</td><td>Training set:<br><table><tr><th>Metrics</th><th>Values</th></tr><tr><td>Accuracy</td><td>0.986</td></tr><tr><td>Balanced accuracy</td><td>0.986</td></tr><tr><td>Sensitivity</td><td>1.000</td></tr><tr><td>Precision</td><td>0.972</td></tr><tr><td>Specificity</td><td>0.971</td></tr><tr><td>F1-score</td><td>0.986</td></tr><tr><td>Matthews Correlation Coefficient (MCC)</td><td>0.972</td></tr></table></td></tr>
<tr><td>6.8.</td><td>Robustness - Statistics obtained by leave-one-out cross-validation</td><td>Robustness – Statistics obtained by leave-one-out (LOO) cross-validation (training set):<br><table><tr><th>Metrics</th><th>LOO</th></tr><tr><td>Accuracy</td><td>0.843</td></tr><tr><td>Balanced accuracy</td><td>0.843</td></tr><tr><td>MCC</td><td>0.686</td></tr></table></td></tr>
<tr><td>6.9.</td><td>Robustness - Statistics obtained by leave-many-out cross-validation</td><td>Robustness – Statistics obtained by ten-fold cross-validation (stratified selection based on class, training set):<br><table><tr><th>Metrics</th><th>10-fold</th></tr><tr><td>Accuracy</td><td>0.886</td></tr><tr><td>Balanced accuracy</td><td>0.886</td></tr><tr><td>MCC</td><td>0.777</td></tr></table></td></tr>
<tr><td>6.10.</td><td>Robustness - Statistics obtained by Y-scrambling</td><td>Statistics (of the test set) for 10 iterations with scrambled endpoint values:<br><table><tr><th>Randomisation</th><th>Accuracy</th><th>Balanced accuracy</th><th>MCC</th></tr><tr><td>1</td><td>0.550</td><td>0.467</td><td>-0.061</td></tr></table></td></tr>
</table>

<table>
<tr><td></td><td></td><td>
<table>
<tr><td>2</td><td>0.550</td><td>0.467</td><td>-0.061</td></tr>
<tr><td>3</td><td>0.400</td><td>0.467</td><td>-0.061</td></tr>
<tr><td>4</td><td>0.650</td><td>0.533</td><td>0.067</td></tr>
<tr><td>5</td><td>0.450</td><td>0.467</td><td>-0.058</td></tr>
<tr><td>6</td><td>0.400</td><td>0.533</td><td>0.067</td></tr>
<tr><td>7</td><td>0.475</td><td>0.417</td><td>-0.146</td></tr>
<tr><td>8</td><td>0.375</td><td>0.383</td><td>-0.204</td></tr>
<tr><td>9</td><td>0.525</td><td>0.550</td><td>0.087</td></tr>
<tr><td>10</td><td>0.500</td><td>0.567</td><td>0.118</td></tr>
</table>
</td></tr>
<tr><td>6.11.</td><td>Robustness - Statistics obtained by bootstrap</td><td>Resampling of the test set 1000 times with replacement:
<table>
<tr><th>Metrics</th><th>Accuracy</th><th>Balanced accuracy</th><th>MCC</th></tr>
<tr><td>Mean</td><td>0.847</td><td>0.830</td><td>0.619</td></tr>
<tr><td>Median</td><td>0.850</td><td>0.833</td><td>0.629</td></tr>
<tr><td>Standard deviation</td><td>0.056</td><td>0.073</td><td>0.142</td></tr>
<tr><td>0.25-quantile</td><td>0.784</td><td>0.800</td><td>0.529</td></tr>
<tr><td>0.975-quantile</td><td>0.955</td><td>0.950</td><td>0.875</td></tr>
</table>
</td></tr>
<tr><td>6.12.</td><td>Robustness - Statistics obtained by other methods</td><td>NA</td></tr>
<tr><td>7</td><td>Defining predictivity (external validation) – OECD Principle 4: “APPROPRIATE MEASURES OF GOODNESS-OF-FIT, ROBUSTENESS AND PREDICTIVITY”</td><td>PRINCIPLE 4: “APPROPRIATE MEASURES OF GOODNESS-OF-FIT, ROBUSTENESS AND PREDICTIVITY”. PRINCIPLE 4 expresses the need to perform validation to establish the performance of the model. PREDICTIVITY refers to the external model validation. Section 7 can be repeated (e.g., 7.a, 7.b, 7.c, etc) as many times as necessary if more validation studies need to be reported in the QMRF.</td></tr>
<tr><td>7.1.</td><td>Availability of the external validation set</td><td>The (external validation) test set is available in the nanoPharos database: https://db.nanopharos.eu/Queries/Datasets.zul?datasetID=np31</td></tr>
<tr><td>7.2.</td><td>Available information for the external validation set</td><td>NP-plant interactions dataset</td></tr>
<tr><td>7.3.</td><td>Data for each descriptor variable for the external validation set</td><td>The (external validation) test set is available in the nanoPharos database: https://db.nanopharos.eu/Queries/Datasets.zul?datasetID=np31</td></tr>
<tr><td>7.4.</td><td>Data for the dependent variable for the external validation set</td><td>The (external validation) test set is available in the nanoPharos database: https://db.nanopharos.eu/Queries/Datasets.zul?datasetID=np31</td></tr>
</table>

<table>
<tr><td>7.5.</td><td>Other information about the external validation set</td><td>40 treatments (10 positives and 30 negatives) were included in the (external validation) test set, which were not involved in model development, but were rather used solely for validation purposes.</td></tr>
<tr><td>7.6.</td><td>Experimental design of test set</td><td>Selected through random stratified sampling from the Deng et al. (2023) curated dataset (the distribution of “positive” and “negative” classes was (approximately) retained in the training and test sets).</td></tr>
<tr><td>7.7.</td><td>Predictivity - Statistics obtained by external validation</td><td>Test set:<table><tr><th>Metrics</th><th>Values</th></tr><tr><td>Accuracy</td><td>0.850</td></tr><tr><td>Balanced accuracy</td><td>0.833</td></tr><tr><td>Sensitivity</td><td>0.867</td></tr><tr><td>Precision</td><td>0.929</td></tr><tr><td>Specificity</td><td>0.800</td></tr><tr><td>F1-score</td><td>0.897</td></tr><tr><td>MCC</td><td>0.630</td></tr></table></td></tr>
<tr><td>7.8.</td><td>Predictivity - Assessment of the external validation set</td><td>The test set is 35% of the initial dataset and retains the classes distribution, so that reliable conclusions on the model generalisability can be derived. 100% of the test set predictions fall within the domain of applicability limits of the model.</td></tr>
<tr><td>7.9.</td><td>Comments on the external validation of the model</td><td>The test set was normalized using the Gaussian normalization function, which was also applied on the training set.</td></tr>
<tr><td>8</td><td>Providing a mechanistic interpretation - OECD Principle 5: “A MECHANISTIC INTERPRETATION, IF POSSIBLE”</td><td>PRINCIPLE 5: “A MECHANISTIC INTERPRETATION, IF POSSIBLE”. According to PRINCIPLE 5, a (Q)SAR should be associated with a mechanistic interpretation, if possible.</td></tr>
<tr><td>8.1.</td><td>Mechanistic basis of the model</td><td>The parameters studied to unravel the NP-plant interactions usually include (but are not limited to) NP composition, shape, and size, NP surface charge and modification, plant species, treatment concentration, exposure duration, and the application method or exposure route. The factors influencing plant length -and plant growth in general- are complex and sometimes not entirely known, and can be either beneficial, promoting plant development (e.g., increasing root and shoot length) or detrimental, such as inhibiting root elongation. These effects depend on NP type and size, treatment concentration, and plant species.<br><br>The type of NPs, their composition, size, and the physicochemical properties derived from their type can either positively or negatively regulate root length. In our case, the used atomistic descriptor (D12) encodes information about NP composition, shape, size, and crystallinity into a single value. NP exposure concentration has a varying impact on root development: some NPs promote root and shoot elongation at low concentrations but reduce root length at higher</td></tr>
</table>

| | | |
|---|---|---|
| | | concentrations. In the final model, the concentration factor is represented by the total concentration variable. Finally, the NP effect varies across plant species and can be both positive and negative. This is the case with Ag NPs, which, within the same concentration range, were found to promote root length in the case of barley and reduce it in the case of lettuce (due to seed treatments in the latter). The plant species and category are also incorporated into our final model, as well as other features encoding experimental conditions (i.e., exposure duration, photoperiod, MT, and cultivation method – see §4.3). It can be concluded that the highly influential factors for assessing plant length response are included in our model and, as they encode controllable experimental conditions (e.g., laboratory, mesocosms, greenhouse, or hydroponic setups, as opposed to *in situ* experiments and field-based data), the model can serve as a primary tool for evaluating plant length promotion or inhibition and therefore, save time from time intensive experimental cycles. |
| 8.2. | A priori or a posteriori mechanistic interpretation | A posteriori mechanistic interpretation. |
| 8.3. | Other information about the mechanistic interpretation | NA |
| **9** | **Miscellaneous information** | |
| 9.1. | Comments | The model development was performed within an autoML scheme. In this report only the final model is documented. The final model was selected between a group of ML methodologies (gradient boosted trees, naïve Bayes, logistic regression, decision tree, random forest, neural network, XGBoost trees) which were optimised through a five-fold cross validation process. Once the optimal hyperparameters were identified for each algorithm, the models were re-trained using the full training set. Their generalisability was then assessed on the validation set (comprising only real observations) to select the final model. The validation set comprised of 26 NP observations from the original set that retained the initial distribution of the classes. |
| 9.2. | Bibliography | 1. The original data were retrieved from the study of Deng et al. "Development potential of nanoenabled agriculture projected using machine learning." *Proceedings of the National Academy of Sciences* 120.25 (2023): e2301885120. https://doi.org/10.1073/pnas.2301885120<br>2. Read more on atomistic descriptors calculation in: Kolokathis et al. "NanoConstruct: A web application builder of ellipsoidal nanoparticles for the investigation of their crystal growth, stability, and the calculation of atomistic descriptors." *Computational and Structural Biotechnology Journal* 25 (2024): 81-90. https://doi.org/10.1016/j.csbj.2024.05.039<br>3. The XGBoost algorithm is implemented in the Isalos Analytics Platform: https://www.docs.isalos.novamechanics.com/classification. |

| | | |
|---|---|---|
| | | html#xgboost |
| 9.3 | Supporting information | The curated and enriched dataset is available in the nanoPharos database, where the training, validation, and test observations are clearly indicated: https://db.nanopharos.eu/Queries/Datasets.zul?datasetID=np31<br><br>The data curation process is documented in detail in the supporting information files of the publication: |