Environmental Science Nano

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Environ. Sci.: Nano*, 2026, 13, 621

Rigorous data curation, enrichment and meta-analysis enable autoML prediction of plant length responses to nanoparticles powered by the Enalos Cloud platform

Dimitra-Danai Varsou, *[ab] Aikaterini Theodori, [a] Anastasios G. Papadiamantis, [bc] Andreas Tsoumanis, [bc] Dimitrios Zouraris, [bc] Maria Antoniou, [bc] Nikoletta-Maria Koutroumpa, [c] Georgia Melagraki, [d] Iseult Lynch [e] and Antreas Afantitis *[abcf]

The application of nanomaterials as fertilizers, biostimulants, and pesticides has been emerging as a promising approach in recent years, aiming to support sustainable and precision agriculture, while simultaneously addressing the challenges of climate change, global population growth, and the search for alternative energy sources (biofuels). In this work, to computationally assess the effects of nanoparticles (NPs) on plant growth (encoded in terms of length of *e.g.*, root, shoot or overall plant length), we performed extensive data curation and enrichment with atomistic descriptors of an existing NP–plant interactions database, ensuring high-quality data for the development of machine learning (ML) models. To address class imbalance, data augmentation techniques were applied. An autoML workflow was developed to optimise and evaluate seven ML algorithms for predicting the plant length response class following NP exposure. The optimised XGBoost model demonstrated superior predictive performance during external validation, achieving an accuracy of 85% and a balanced accuracy of 83%, and its applicability domain was clearly defined. One of the key advantages of the plant length response model is that it requires no experimental input data to generate predictions, thus facilitating virtual screening prior to implementation of controlled experimental setups. The curated dataset has been made findable, accessible, interoperable and reusable (FAIR) *via* the nanoPharos database (https://db.nanopharos.eu/Queries/Datasets.zul?datasetID=np31) and the XGBoost model was documented in a standardized QSAR model report format (QMRF) to enhance its usability and FAIRness and made available as a user-friendly web-application, CeresAI-nano, *via* the Enalos Cloud platform (https://enaloscloud.novamechanics.com/chiasma/agrinano/).

Received 26th September 2025,
Accepted 9th December 2025

DOI: 10.1039/d5en00897b

rsc.li/es-nano

Environmental significance

This study provides a significant advance for the environmental sustainability of nano-enabled agriculture by combining rigorous data curation, atomistic descriptor enrichment, and automated machine learning to predict plant length responses to nanoparticle (NP) exposure. Conventional assessment of NP–plant interactions requires long, resource-intensive experiments that often capture only part of a crop's growth cycle. By contrast, our publicly available curated dataset and predictive model (CeresAI-nano) enable rapid, reproducible, and FAIR (findable, accessible, interoperable, and reusable) virtual screening of nano-agrochemical treatments without the need for experimental input. This capability directly supports the safe- and sustainable-by-design (SSbD) development of new nano-fertilizers, biostimulants, and pesticides, helping stakeholders identify formulations that enhance plant growth while minimising adverse effects on soil health, biodiversity, and human exposure. These efforts accelerate environmentally responsible innovation, and contribute to the long-term resilience of agricultural systems under climate change pressures.

[a] *NovaMechanics MIKE, Piraeus 18545, Greece.*
E-mail: varsou@novamechanics.com, afantitis@novamechanics.com
[b] *Entelos Institute, Nicosia 2102, Cyprus*
[c] *NovaMechanics Ltd, Nicosia 1070, Cyprus*
[d] *Division of Physical Sciences and Applications, Hellenic Military Academy, Vari 16672, Greece*
[e] *School of Geography, Earth and Environmental Sciences, University of Birmingham, B15 2TT Birmingham, UK*
[f] *Department of Pharmacy, Frederick University, Nicosia 1036, Cyprus*

Introduction

Modern agriculture, on which the global food sector[1] and recently an increasing part of fuel production[2] depend, and upon which many Global South countries build their economies,[3] is facing a range of interrelated challenges mainly due to the substantial rise in human-driven

activities.[4] Rapid global population growth is driving increased food demand, placing pressure on agricultural productivity and simultaneously emphasising the need to ensure long-term food security.[1,5–7] At the same time, environmental pollution,[4] declining soil quality and fertility,[5,6] global warming and unprecedented conditions resulting from climate change, are negatively affecting sustainable agrarian development.[3–8] Additional challenges include the ongoing loss of arable land due to urban expansion, the decreasing availability of irrigation water resources, and the low efficiency of agrochemicals primarily caused by the overuse of fertilizers and pesticides[1,5] which further undermine agricultural productivity.

Nanotechnology can play a key role in promoting more precise, sustainable, and resource-efficient, agricultural practices,[5,7] mainly due to the intrinsic properties of nanoparticles (NPs), which stem from their small size, high surface area-to-volume ratio, and tuneable surface charge.[2,9] NPs are widely used in nano-enabled agrochemicals to enhance plant nutrition and resilience, and support biotic and abiotic stress management, all while increasing productivity.[2,6,7,10] For example, NPs are incorporated into fertilizers and pesticides,[6,8,10] as they can serve as nano-carriers for controlled release of nutrients,[3,5,7] target pests efficiently,[5] protect plants against the toxic effects of heavy metals such as Cd,[11] remediate or improve soil quality,[3,8,10,12] and/or facilitate key biochemical processes in plants, such as photosynthesis.[2,5,6,13] It should also be noted that NPs are more effective than conventional fertilizers.[13–15] For example, Nekoukhou *et al.*[16] showed that foliar application of ZnO NPs on dragonhead resulted in increased productivity compared to conventional fertilizer (ZnS) treatments. In addition, nano-sensors are being developed and integrated to monitor and evaluate crop growth, detect infections, assess soil conditions, and track pesticide penetration, thus, contributing to more efficient agricultural practices.[3–5,7,13,17]

However, as is often observed in emerging research areas, nano-agriculture is not free of risks. Apart from the beneficial impact of the nano-enabled agrichemicals, there is an ongoing discussion about the potential undesirable effects of NPs on crop growth, which can sometimes alter plant physiological and biochemical functions in ways that are contrary to expectations (*e.g.*, plant growth inhibition, reduced micronutrient uptake, *etc.*).[10] Nanotoxicity and the potential toxic effects of NPs on human health and the environment, highlight the need for further research, especially in terms of non-monotonic responses.[7] Additionally, it should be considered that beyond the intentional application of NPs to plants and soils, they can also be released into the agroecosystems incidentally during their production, through the use of nano-enabled products, and disposal, through untreated wastewater discharge, landfills, and *via* other pathways.[7,9,10,18] Crops cultured in NPs-contaminated soil or atmosphere, may uptake these through the roots or aerial parts[19] and show reduced growth rates and quality as a result of NP-induced oxidative stress.[20] The interaction between NP properties, soil systems and plant species is highly complex.[5,6,8] The NPs undergo transformation in different systems (*e.g.*, biomolecule corona formation),[5] which poses a barrier in identifying the precise mechanisms through which NPs impact plant growth and development. In this context, the application of machine learning (ML) algorithms may serve as a powerful tool to model, decode, and predict NP-induced effects and shed light on the mechanisms of NPs uptake and translocation in plant tissues, including any transformations and their consequences.

ML and data-driven approaches, including quantitate structure–activity relationship (QSAR) models and read-across methodologies, have been already established in the nanoinformatics field to support the safe and sustainable-by-design (SSbD) development of novel NPs. These approaches aim to optimise NP functionality while eliminating potential adverse effects-even prior to their synthesis.[5,21] Similarly, ML approaches can support SSbD of novel nano-agrichemicals, by predicting specific effects (*e.g.*, oxidative stress and photosynthesis parameters, as well as the NPs uptake and accumulation in different plant tissues), in order to achieve optimal NPs properties that enhance plant growth and resilience, improve soil health, and reduce possible toxicological effects on human health and the environment.[5]

Artificial neural network (ANN) approaches have recently been employed to simulate NP–plant–soil interactions. For instance, Li *et al.*[22] developed eight ANN models to predict the yield and quality responses of rice plant after soil exposure to selenium NPs at four different time intervals (30–120 days). The reported R^2 values for the eight models were in the range of 0.71–0.89 during external validation. Wang *et al.*[23] applied ANN algorithms to predict the transport factor (TF) and root concentration factor (RCF) of different types of NPs in plants grown in soil or hydroponic systems, achieving R^2 values higher than 0.80 in all cases during external validation. ANNs algorithms have also been successfully used to model the uptake of cooccurring CeO_2 NPs and Cd in *Brassica napus* L. cultivated in soil. Rossi *et al.*[24] developed a multilayer perceptron ANN to predict Cd and Ce concentrations in plant roots and leaves based on plant characteristics, achieving R values greater than 0.90 on the test set in all cases.

Min *et al.*[25] explored the effects of metal oxide NPs along with parameters such as crop species (vegetable, legume, fruit, cereal or other) and cultivation conditions on the heavy metal uptake by crops. They developed different ML models, among which the random forest algorithm achieved the highest performance, with an R^2 of 0.62 on the test set for the prediction of the effect of metal oxide NPs on the heavy metal uptake and growth of crops expressed as a weighted response ratio (LnR+). Xu *et al.*[12] investigated *in silico* the potential hazardous effects posed by NPs to soil microbial communities by training random forest models on NP and soil properties to predict microbial diversity using Richness

and Shannon index. These models, evaluated *via* 10-fold cross-validation, achieved R^2 values mostly exceeding 0.70. Li *et al.*[6] also developed several ML models for prediction of the effects of carbon dots on plants (mostly lettuce, tomato, maize, wheat, cucumber, rice, mung beans, and soybean) including nutrient content, quality, growth indicators, photosynthesis, and antioxidant response. These models used features such as carbon dots and plant properties, environmental factors, and experimental conditions. The growth indicator endpoint was predicted with the highest accuracy using a random forest model achieving an R^2 value of 0.62 in external validation. Yu *et al.*[26] applied light gradient boosting machine (LightGBM) models to classify the relative metal/metalloid concentration (RMC) in maize seedlings following seed priming with fourteen different metal oxide NPs at varying concentrations. These models were developed using NP characteristics and two seedling parameters. Using ten different stratified train-test splits, the average accuracy on the test sets reached 0.76.

Great progress has thus been made in the integration of ML methods in nano-agriculture, however key challenges persist in model development. The lack of large-scale experimental studies about the effects of NPs on plant growth, the variations in experimental conditions, methods and protocols, as well as the non-systematic data coding in existing datasets created by different research groups, impede the creation of a uniform dataset that will allow the development of robust ML models. As a result, the earlier work mentioned mainly either addresses very specific nano-agriculture applications (only one type of NP or one crop species) or utilizes small datasets with limited attributes for model development. To overcome this, Deng *et al.*[8] conducted an extensive literature review to identify NP–plant interactions studies and systematically extracted a wide range of relevant data, including NP characterization data, experimental conditions, plant properties, and responses for both treated and control groups. Their effort resulted in the compilation of data from 57 studies into 13 heterogeneous datasets, covering plant growth, photosynthesis, NP uptake, and antioxidant responses for 17 different NPs. This work represents a critical step toward overcoming current data limitations and enabling the development of more generalisable ML models in nano-agriculture. Subsequently, they developed random forest models and obtained predictions about the potential of nano-enabled agriculture across different global regions.[8]

Deng *et al.* made their dataset publicly available on GitHub, where key NP characteristics – such as NP type (*e.g.*, carbide, metal, oxide, macromolecular compound), composition, and shape (*e.g.*, granular, one- and two-dimensional, hollow) – are encoded. However, the absence of explicit information on the core composition of the NPs poses challenges for reproducibility and limits the full exploitation of the data. For instance, calculating atomistic descriptors requires information on the precise NP composition and shape, and deciphering the encoded data from existing datasets such as that compiled by Deng *et al.*,[8] can be time-consuming. Furthermore, the lack of direct links within the dataset to the DOIs of the original publications from which data were extracted impedes quick verification and retrieval of additional details that may be essential for other types of studies, such as mechanistic investigations, meta-analyses, or regulatory risk assessment (*e.g.*, for the calculation of atomistic descriptors the NP crystallographic phase is needed).

In the present study, building upon the dataset compiled by Deng *et al.*,[8] we performed an extensive additional data curation and quality control process, which involved identifying the original publications and cross-checking all details and supplementing with additional information such as the aforementioned core composition and crystal phase information (where available), followed by data enrichment using computationally derived atomistic descriptors based on the elemental composition of the NPs. Our goal was to develop a ML model capable of predicting the effects of NP exposure on plant length (roots, shoots or total). To address data imbalance, we employed a synthetic data generation technique, combined with additional data manipulation strategies aimed at optimising the modelling workflow, which included rigorous data filtering and variable selection through an automated ML (autoML) framework. The resulting optimised models were validated according to the Organisation for Economic Co-operation and Development (OECD) principles for the validation of QSAR models for regulatory purposes, including the generation of a standardised report following the QSAR model reporting format (QMRF),[27] and the final predictive model was implemented as a web-service and distributed *via* the Enalos Cloud platform. To ensure transparency and FAIRness of the data and model, curated data were made available through the nanoPharos database for full exploitation from the stakeholders.

Methods

Our analysis focuses on the "Length" dataset – one of the 13 provided by Deng *et al.*[8] *via* GitHub (accessed on September 15, 2023)[28] – selected for its relatively large size which reduces the potential impact of data curation on the data size. The original dataset consisted of 28 features for 299 NP–plant interaction observations along with the target variable, namely the normalised length response of the root, shoot or plant following the NP uptake. The normalisation of the length target variable in the study of Deng *et al.*[8] was based on the control length for negative responses and on the experimental value for positive responses, which can be considered less intuitive but is ML-friendly, as this approach means that label values were bounded between −1 and 1 with 0 representing the control value. The attributes included plant properties (plant species, carbon fixation category, plant age, growth stage at NP application (germination, seedling, vegetative)) and experimental conditions (exposure pathway (*i.e.*, seed, foliar, root uptake), measured

tissue (root, shoot, overall plant), culture type (medium, hydroponic, soil), total NP content, duration of exposure, photoperiod, illumination intensity, humidity, day and night temperature), NP characteristics (particle size measured by TEM and DLS, Z-potential, purity) and encoded NP type and NP shape variables (for a detailed list of features see Table S1).

Our analysis steps are presented schematically in Fig. 1. Prior to model development, data curation, management, and quality control (QC) of the original dataset took place based on the best practices for curating, cleaning, and processing literature-extracted data for meta-analysis purposes. A detailed workflow was established, with the aim to maximise the extraction of information, correct potential errors from the data curated in the source publication,[8] perform data gap filling where possible, and enrich the dataset with additional key points from the original papers used for data extraction or with computational descriptors calculated in house.[29,30]

The curated dataset was later randomly partitioned into training, validation and test sets using a stratified sampling technique. Training and validation sets were used for model development and hyperparameters optimization, respectively, within the autoML scheme and the test set was used as a blind set for external validation.

Data curation, management and enrichment

Data extraction commenced with identifying and decoding the different materials that were extracted by Deng *et al.*[8] following literature search in the Web of Science, ScienceDirect, and Springer-Link databases. Their literature search led to 57 papers used for meta-analysis. The NP type and morphology were identified based on the abbreviations used by Deng *et al.* (Table 1). These were divided into carbides, metals, and metal oxides for the NP type. Each NP was divided into 2 components (Com1, Com2) and the relative atomic mass (A_r) for each component was extracted separately. Where only one component existed, *e.g.*, graphene, the A_r of Com2 was set to 0.

To QC the data extracted by Deng *et al.*,[8] each dataset row was matched to its corresponding original publication, *e.g.*, rows 1–9 were matched to paper number 5 (ref. 31) in the literature list provided by the writers. All data reported by Deng *et al.*[8] were QCed based on the original publications. Any inconsistencies between the original and the Deng *et al.* publications were highlighted and, where possible, corrected (see Table S2 in the SI for a full list of corrections made to the dataset). Corrections were made in the NP size (sizeTEM), purity, daytime and night temperature (DT and NT respectively), humidity, and illumination descriptors. Similarly, 52 rows from the original dataset were removed, as the size or shape of the NPs were not clearly mentioned in the original papers. Where possible, verification of the data was achieved based on previous studies; *e.g.*, for rows 249–260 (ref. 32) of the original dataset, CuO NP shape and size were identified in a follow-up study by the same group.[33] When included data could not be traced back to the original papers, the respective rows (n = 16, rows 216–231 of the original dataset) were removed. For this reason, the sizeDLS and zeta columns were removed as they contained the most data gaps, *i.e.*, 73% and 66%, respectively.

On top of that, the full text for all the 57 original publications was studied and key points were highlighted or extra data extracted to enrich the dataset further. This included adding 12 rows with additional data from papers number 10 (ref. 34) and 45 (ref. 35). Furthermore, the total concentration (in mg L^{-1}) column was added to include raw data about initial NP concentration. To avoid any bias or artificial enhancement, the original total content (TC (mg)) column was removed to avoid possible assumptions about initial volumes by Deng *et al.*, as well. This led to the removal of 48 rows because the total concentration was not available and could not be calculated from the information in the original publication. For example, for rows 91–108 (ref. 36) concentration was reported in mg kg^{-1} of soil, but the initial soil mass was not provided, preventing conversion to mg L^{-1}.

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.*, provided in textual format), the "WebPlotDigitizer" tool[37] 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% (see SI file for more details). 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. This was done to facilitate the development of a classification model, as preliminary regression models resulted in poor performance ($R^2 < 0.5$). As a result, the 25 control data points were removed as they did not fall under the "positive"/"negative" classes.

The produced dataset, which consisted of 170 NPs and 26 features, was then enriched with a set of 53 atomistic descriptors of NPs in vacuum calculated with the Enalos NanoConstruct tool.[29] The necessary information to perform the calculation of the atomistic descriptors is included in Table 2. The required crystallographic data were obtained from the Crystallography Open Database (COD).[38] Calculations were performed for spherical NPs, as following processing and cleansing only spherical NPs remained in the dataset. For TiO_2 NPs lacking crystallographic phase information, an anatase structure was assumed to prevent the exclusion of the relevant particles and preserve loss of samples from the dataset. Out of the 53 calculated descriptors, 10 provided not a number (NaN) values for some NPs and were excluded. When atomistic descriptor

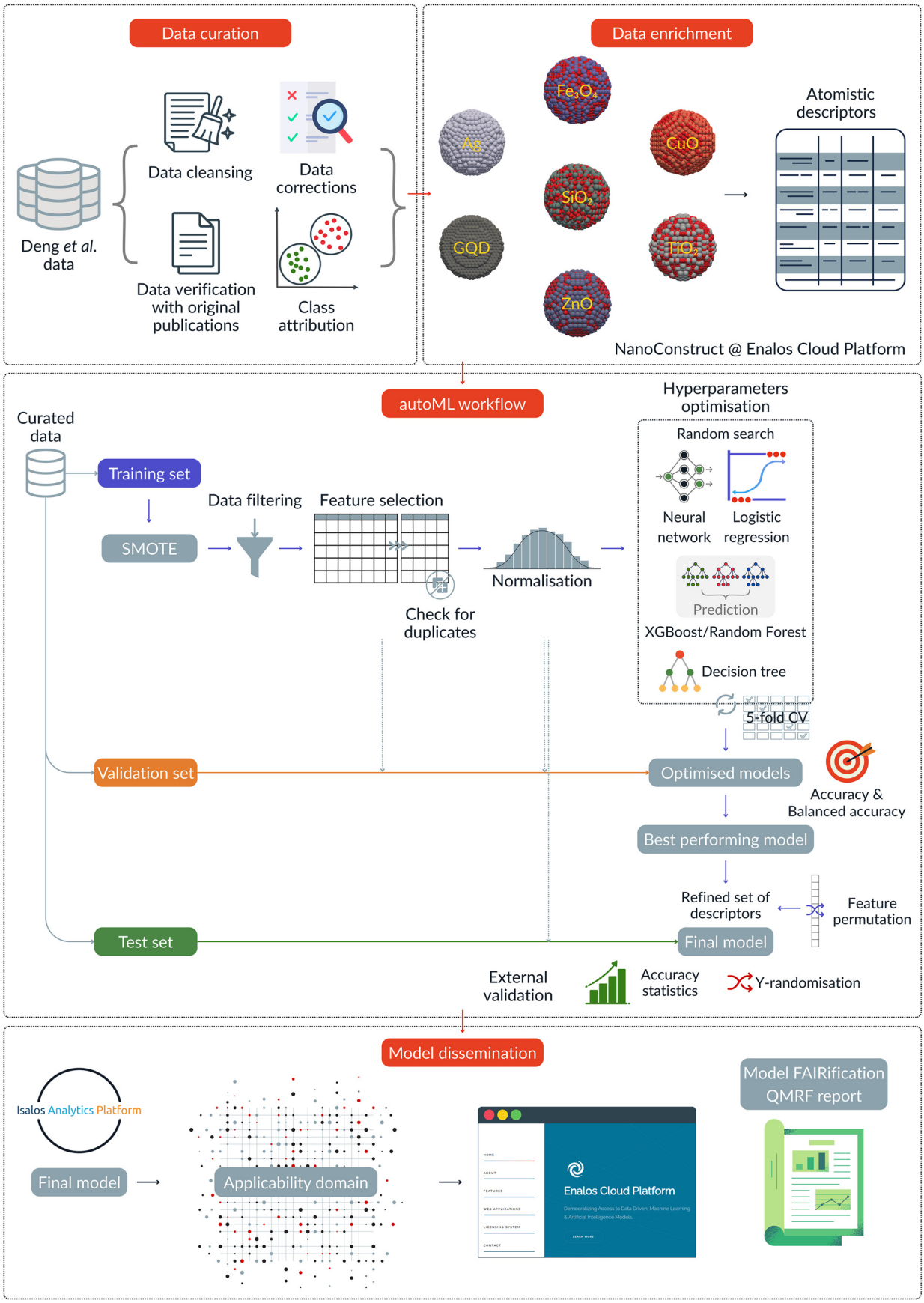

Fig. 1 Schematic workflow of the data curation, analysis and modelling of the NP–plant response data, and deployment of the model and its complete documentation. Note that the curated and enriched dataset has also be made available for re-use. Designed with https://Canva.com.

Table 1 Dataset decoding used for NP identification and data extraction based on the nomenclature used by Deng *et al.*[8]

NP type	
Encoding	Description
Carbide	Number of carbon atoms
Metal	Number of metal atoms
Oxide	Number of oxygen atoms
Component 1 (Com1)	A_r of Com1, *e.g.*, for TiO_2, Com1(Ti) = 47.87
Component 2 (Com2)	A_r of Com2, *e.g.*, for TiO_2, Com2(O) = 16.00
Morphology	
Encoding	Description
Dim0	Granular
Dim1	One-dimensional
Dim2	Two-dimensional
Hollow	Hollow

calculation was not possible, *e.g.*, due to large size of the NP, the respective dataset rows (n = 57) were removed. This included polymeric and large carbon NPs, *e.g.*, graphene sheets and nanotubes (rows 189–210 in the original dataset[39]). The final dataset used for model development, following cleansing and enrichment, contained 69 features and 113 rows, compared to the 28 features and 299 rows of the original one.

Synthetic data generation and class balancing

While data validation and cleaning are essential pre-processing steps when using existing datasets to ensure consistent encoding and high quality of the data, they can inevitably lead to a drastic dataset size reduction, as observed here. Despite being in the "big data" era, data availability remains a challenge in the *in silico* study of materials science and applications.[48–50] Most importantly, reduced dataset size combined with high feature dimensionality is associated with ML model overfitting and classification bias due to potentially imbalanced classes.[50] Considering that the majority of NP observations in the original dataset created by Deng *et al.* resulted in reduced plant length (45.5% "negative" labels, 33.4% "positive" labels, 21.1% "zero" labels-control groups), it is anticipated that the data cleaning process may further exacerbate the phenomenon of class imbalance. Indeed, in the final, revised dataset only 25.7% of labels were "positive" indicating enhanced plant growth relative to the control.

To effectively address both the small dataset size and the class imbalance problem, a synthetic data generation method was applied to populate the minority "positive" class. In this case an oversampling technique that has been successfully applied to similar modelling problems,[51–53] the synthetic minority over-sampling technique (SMOTE),[54] was selected to enhance data balancing without introducing unrealistic data points. In SMOTE, the new data points created correspond to a randomly chosen point along the line segments between a "positive"-labelled observation and one of its k-nearest neighbours (kNN). It should be noted that oversampling was applied exclusively to the training set to ensure that the calculated accuracy metrics are meaningful and express the model performance on real data. Equally significant was the normalisation of the data before applying a distance-based oversampling technique such as SMOTE. To this end, the training data were normalised using the z-score normalisation method prior to the application of SMOTE, and subsequently de-normalised to proceed with the rest of the feature engineering processes.

Data pre-processing

To reduce data dimensionality, enhance model interpretability, and mitigate overfitting, descriptors with low variance ($\leq$0.2) and highly intercorrelated descriptors (Spearman's rank correlation coefficient[53] value $\geq$0.99) were filtered out. Information gain scores were subsequently calculated for the remaining features, and variables with zero information gain were excluded to retain only the most relevant ones for the endpoint prediction. Following all feature selection steps, the training data were examined for duplicate rows to prevent any potential sources of bias in model development. As different descriptors span different

Table 2 NanoConstruct configuration details for the calculation of the atomistic descriptors

NP core	Diameter [nm]	COD code of CIF file	FF using OPENKIM ID[40]
Ag (ref. 31, 34 and 41)	13.8, 20	1509146	EAM_Dynamo_AcklandTichyVitek_1987v2_Ag_ _MO_055919219575_000
Graphene quantum dots (ref. 42)	2.5	1200017	DUNN_WenTadmor_2019v1_C_ _MO_584345505904_000
CuO (ref. 32 and 35)	30, 40	1011148	LJ_ElliottAkerson_2015_Universal_ _MO_959249795837_003
Fe_3O_4 (ref. 43)	6.7	1011032	EAM_Dynamo_AcklandTichyVitek_1987_Ag_ _MO_212700056563_005
SiO_2 (ref. 44)	15	9011493	Sim_LAMMPS_Vashishta_BroughtonMeliVashishta_1997_SiO_ _SM_422553794879_000
TiO_2 (ref. 45 and 46)	6.5, 21	1010942	Sim_LAMMPS_MEAM_ZhangTrinkle_2016_TiO_ _SM_513612626462_000
ZnO (ref. 47)	25	1011258	Sim_LAMMPS_ReaxFF_RaymandVanDuinBaudin_2008_ZnOH_ _SM_449472104549_001

Table 3 Tuned hyperparameters of the seven ML methodologies within the autoML workflow

ML methodologies	Hyperparameters	Search space [min, max; step]
Gradient boosted trees	Number of trees	[50, 100; 20]
Naïve Bayes	Default probability (threshold)	[0.004, 0.01; 0.0001]
Logistic regression	Step size	[0.01, 0.1; 0.01]
Decision tree	Minimum number of records per node	[2, 20; 2]
Random forest	Tree depth	[4, 8; 4]
	Number of trees	[100, 200; 100]
	Minimum child node size	[10, 20; 5]
Neural network	Number of hidden layers	[1, 2; 1]
	Number of hidden neurons per layer	[5, 20; 10]
XGBoost trees	Learning rate (eta)	[0.2, 0.3; 0.1]
	Maximum depth	[5, 10; 5]

numerical scales, the remaining ones were normalised using the *z*-score normalisation method allowing consistent feature contribution to the computational analysis. Finally, one-hot-encoding of the categorical features was performed to be compatible with the employed ML algorithms.

Workflow of model development

The core of the autoML was used for the automated, iterative optimisation of seven ML models adapted from our previous work.[53] Each model underwent an independent optimisation process to identify the best-performing configuration of hyperparameters. For example, the gradient boosted trees algorithm was tuned based on the optimum number of trees and the Naïve Bayes algorithm was tuned based on the default probability (threshold). More complex methodologies required the tuning of multiple hyperparameters such as the random forest method (tree depth, number of trees and minimum child node size), and the extreme gradient boosting (XGBoost) trees methodology (learning rate and maximum depth) (Table 3). In each case, hyperparameter search was conducted using a randomised approach. For each hyperparameter trial configuration, the respective model was trained using a five-fold cross-validation scheme. The quality-of-fit was assessed by averaging the accuracy scores across all folds, producing a single metric to guide selection. The tuning process continued for a maximum of 10 rounds or – if no performance improvement greater than 0.01 was observed over five successive rounds, – early stopping was enabled. 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 determine which model yielded the highest accuracy score and should be selected for final predictions.

Model validation

In order to ensure that the selected model had good generalisation ability for new NPs, an external validation scheme was applied. In this approach, the initial data set was randomly divided into training, validation, and test sets. The training set was oversampled and used to determine the optimal modelling hyperparameters *via* five-fold cross-validation. The validation set was used for model selection by assessing models' performance, and the best-performing model predictive accuracy was assessed using the test set, which did not take part in the model development and selection procedures. The validation and test sets were randomly selected from the original set of NPs prior to any modelling steps using a stratified approach, to retain the distribution of classes (see Table 4).

Accuracy (ACC, eqn (1)) and balanced accuracy (BA, eqn (2)) metrics were calculated during the internal validation of the seven ML models within the autoML scheme. As the BA metric accounts for classes imbalance, it was selected to evaluate model performance, along with the ACC metric.

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \quad (1)$$

$$\mathrm{BA} = \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}\right) \quad (2)$$

where, TP are true positives, TN are true negatives, FP are false positives, and FN are false negatives.

To thoroughly evaluate model predictivity, within the external validation of the selected model from the autoML, in addition to ACC and BA, the sensitivity (SEN, eqn (3)), specificity (SPE, eqn (4)), precision (PRE, eqn (5)), Matthews correlation coefficient (MCC, eqn (6)), and F1-score metrics (eqn (7)) were calculated.

Table 4 Class distribution between training, validation and test sets

Class	Training set before oversampling	Training set after oversampling	Validation set	Test set
Positive	0.26	0.50	0.27	0.25
Negative	0.74	0.50	0.73	0.75

$$\mathrm{SEN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

$$\mathrm{SPE} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{4}$$

$$\mathrm{PRE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{5}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{6}$$

$$\mathrm{F1} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{7}$$

The *y*-randomisation test[53] was performed to ensure that the accuracy of the best-performing model was not a coincidental outcome. In this method, the endpoint values of the training set are randomly shuffled among the NP observations, while the independent descriptors remain unchanged. The entire modelling process is then repeated multiple times (10 in this case) using the original descriptors and the shuffled response values. If the original model is truly robust and reliable, it is expected that the models generated from the randomised data will perform poorly when applied to the validation or test set treatments. Finally, to ensure the robustness of the final model, internal validation tests, leave-one-out (LOO) and ten-fold cross-validation tests were performed.

Applicability domain

The limited size and diversity of NP datasets used for training ML models directly affect their practical applicability under real-life applications. To address this, it is essential to define the model's applicability domain (APD), as recommended by the OECD,[55] in order to enhance stakeholder confidence in the reliability of the model's predictions. Specifically, the APD refers to a subspace of the overall data space within which the model's predictions are based on interpolation rather than extrapolation and are therefore considered reliable.[56] In this study, a distance-based approach was adopted to define the APD limits: An APD threshold value (thr) was defined according to eqn (8); if the distance of a query NP from its nearest training neighbour exceeds the APD threshold, the prediction for this NP is considered unreliable.

$$\mathrm{thr} = \langle d \rangle + Z\sigma \tag{8}$$

To define the APD threshold of eqn (8), Euclidean distances between all NPs of the training set were calculated, and the overall average of these distances is denoted as D_{av}. Next, a subset of training NPs was identified, *i.e.*, those whose distances are lower than D_{av}. Within this subset, a new average distance, $\langle d \rangle$, was calculated along with its standard deviation σ. Z, is an empirical cut-off value, the default value is equal to 0.5.[57] Categorical descriptors were one-hot-encoded to be incorporated into the APD limits calculation.

Results

Data curation and cleaning were the first steps towards the development of a robust and interpretable ML model with the aim of accurately predicting the effects of exposure to NPs on the length of the root, shoot or whole plant. Data curation led to the removal of 62% of the initial data points (186 rows) which included missing values or unaccounted for data, or NP compositions for which atomistic descriptors could not be calculated (*e.g.*, too large particles or compositions based on carbon), and the addition of the 53 calculated atomistic descriptors for each of the 113 remaining NPs. While this critical and labour-intensive step was not automated, all downstream tasks, from data transformation and feature engineering to model training, were conducted within the KNIME Analytics Platform, using the Enalos+,[58] the Palladian,[59] and the autoML component[60] to automate preprocessing and modelling. Our tailored autoML pipeline[53] balances classes *via* SMOTE, discards low-variance and intercorrelated features, and ranks variables by information gain prior to model development. This end-to-end automation reduces manual intervention while preserving model interpretability. Finally, the validated predictive model was packaged and deployed on the Isalos Analytics Platform,[61] providing a web-based interface for researchers to upload new NP data and instantly obtain predictions of root, shoot, or whole-plant length effects as long as their NP compositions are within the APD of the model.

To build our predictive workflow the curated dataset was initially split into training (47 NP observations), validation (26 NP observations) and test (40 NP observations) sets following a stratified sampling technique to ensure equal classes representation between sets and with the original dataset. Training data were oversampled using SMOTE (with k = 5), and 23 data points belonging to the minority class ("positive") were generated to correct class imbalance and increase the dataset's size (Table 4).

The oversampled training set was filtered to exclude noisy and intercorrelated descriptors, using a low-variance and a correlation filter, leading to the removal of 16 and 15 descriptors, respectively. To further refine the descriptor set, we calculated information gain scores for the remaining 38 features, retaining only those with non-zero values (19 descriptors). The variable representing total concentration was not subjected to this selection process, as it encodes critical experimental parameters and is considered as a priority driver of plant uptake,[8] and was thus force-included in the final training set (Table 5). To eliminate redundancy that could introduce bias, the filtered dataset was reviewed for duplicate observations. Finally, categorical variables were one-hot-encoded, and numerical descriptors were standardised *via* *z*-score normalisation to ensure comparability across varying scales.

Using the 20 selected features, seven ML models were developed, and their hyperparameters were optimised

Table 5 List of selected variables *via* the information gain method. These variables were used to train the ML models within the autoML workflow

Selected variable	Description	Information gain score
Photoperiod	Hours of plant exposure to light per day in hours per day	0.419
Species	Plant species (cucumber, bean, wheat, rice, tomato, maize)	0.340
Illumination	Illumination intensity in μmol m^{-2} s^{-1}	0.284
NT	Nighttime temperature in °C	0.229
Purity	% nanoparticle material purity	0.225
D24	The average coordination parameter (5Ang) of all atoms	0.198
D12	The average difference of the coordination parameter between core and shell atoms	0.179
D19	The average coordination parameter (4Ang) of all atoms	0.179
D25	The average coordination parameter (5Ang) of the core atoms	0.179
D26	The average coordination parameter (5Ang) of the shell atoms	0.179
D37	Lattice energy of NP in eV	0.179
Duration	NP treatment duration (time elapsed from the exposure of the plant to NPs to the measurement) in days	0.179
DT	Daytime temperature in °C	0.160
D4	The average potential energy of all atoms in eV	0.152
GS	Growth stage of plant (germination, seedling, vegetative)	0.057
EP	NP exposure pathway (seeds, root, leaf)	0.045
Category	Plant carbon fixation metabolic pathway (C_3 or C_4 photosynthetic process)	0.040
MT	Measured tissue (root, shoot, plant)	0.037
Cultured	Cultivation method (medium, hydroponic, soil)	0.008
Total concentration	Total concentration of NP treatment in mg L^{-1}	Forced inclusion

through five-fold cross-validation (Table 6). To select the best-performing among these models, they were applied on the validation set (26 NPs), which was also normalised following the training normalisation function. The accuracy of the optimised models on the validation set is presented in Fig. 2. The XGBoost model yielded the highest ACC and BA values (0.77 and 0.71, respectively) and therefore, was selected as the final model for the prediction of the "Length" class.

Given the limited size of the training data, and to mitigate the risk of overfitting, we applied a permutation-based feature importance to the XGBoost model following its selection within autoML based on the inner validation performance, to further refine the descriptor set. More specifically, to assess the stability and relevance of each of the 20 descriptors of Table 5, permutation importance was calculated within a ten-fold cross-validation scheme using the XGBoost model. In each fold, individual descriptors in the validation subset were randomly shuffled, and the resulting drop in the predictive performance (compared to the baseline/unshuffled validation subset) was recorded. Normalisation 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 were the total concentration, species, MT, cultured, category, duration, photoperiod and D12, and are depicted in Fig. 3. The final XGBoost model was retrained on the reduced descriptor set and subsequently validated.

The XGBoost model was rebuilt with the refined set of descriptors and was validated externally, by applying it to the 40 NPs of the test set, which preserved the original distribution of "positive" and "negative" classes. Test set descriptors were previously normalised with the training set *z*-score function. In Table 7 the confusion matrix of the test set is presented. Table 8 summarises the performance of the XGBoost model on the test set: the model demonstrates strong predictive capabilities and generalizability on unseen data with an ACC of 0.85 and a BA of 0.83.

To evaluate the stability of the model's predictive performance on the test set,[62] a bootstrap resampling strategy was employed. The test set was resampled 1000 times with replacement, and predictions were produced by applying the XGBoost model. For each bootstrap sample, the BA value was calculated, and the mean, median, and standard deviation of these values were used to assess the stability and robustness of the model's performance. The results are summarised in Table 9 and visualised in Fig. S4. The distribution of the BA values is approximately symmetric (shows minimal negative skew), as also indicated by the proximity of the mean (0.830) and median (0.833). The low standard deviation (0.073) further supports the consistency and reliability of the model's predictions.

Further confirmation of the model's robustness comes from the poor predictive BA and MCCs (similar to random estimation accuracy) obtained with the *y*-randomised models as shown in Table 10, demonstrating that the model's learned classification rules are meaningful and not just random noise.

The cross-validation results, summarised in Table 11, prove the model's stability to data inclusion–exclusion.

The APD threshold was calculated equal to 5.071. As all test NPs domain values were below this limit value, the test set predictions are all considered reliable.

Data and model FAIRification and dissemination

Using the FAIR (findable, accessible, interoperable, reusable) principles,[63] their interpretations[64] and the EU's Joint

Table 6 ML methodologies' hyperparameters tuning in the autoML workflow after five-fold cross validation

ML methodologies	Optimised hyperparameters
Gradient boosted trees	nrModels = 50
Naïve Bayes	Threshold = 0.0044
Logistic regression	stepSize = 0.05
Decision tree	minNumberRecordsperNode = 16
Random forest	maxLevels = 4, minNodesize = 20, nrModels = 200
Neural network	Hiddenlayer = 1, nrhiddenneurons = 15
XGBoost trees	Eta = 0.2, Max_depth = 5

Research Centre (JRC) Guidelines for FAIR Data,[65] FAIR guidelines for computational workflows,[66] and FAIR4RS (FAIR for Research Software) principles,[67] FAIRification of the produced dataset and developed model, respectively, has been achieved. The curated dataset has been enriched with metadata, it has been uploaded to the nanoPharos database and has been assigned a unique URI (https://db.nanopharos.eu/Queries/Datasets.zul?datasetID=np31). The dataset is offered in tabular, ready-for-modelling (.CSV, .XLSX) and machine actionable formats (.XML). To support model reproducibility, the training, validation, and test observations as used in this work are explicitly annotated in the dataset. The dataset has been linked to rich scientific, bibliographical, and provenance metadata, which have been published as a machine actionable nanopublication (https://w3id.org/np/RALzBzdG_gTKI1Xn6SubCkA-ZTeK5LyAV_aYjrW-Yyzj8) in nanodash. The data and metadata entries reference the metadata's and data's URIs respectively. Data retrieval is possible using the dataset's ID through the open and freely available nanoPharos API (https://db.nanopharos.eu/swagger-ui/).

To support model FAIRification, the produced model has been FAIRified using the FAIR principles application to computational workflows guidelines.[66] The main steps of model development are documented in this paper and in a straightforward manner through a standardized QMRF report,[68] which is in line with the OECD's Guidelines for the Validation of QSAR Models.[55] The QMRF is available as part of this paper's SI.

Apart from data and model development sharing, which enhances reproducibility, FAIR model dissemination, following the FAIR4RS principles,[67] is equally important to promote and support future advancement of nano-enabled agriculture. For this purpose, the developed XGBoost model was transferred to the Isalos Analytics Platform,[61] a user-friendly software tool for data manipulation and ML model development. This facilitates model deployment and sharing as a web service with a unique identifier, called CeresAI-nano, *via* the Enalos Cloud platform (https://enaloscloud.novamechanics.com/chiasma/agrinano/). The CeresAI-nano can be also accessed remotely using APIs (https://enaloscloud.novamechanics.com/chiasma/swagger-ui/).

In the CeresAI-nano application, users can complete the provided table with the required input descriptors (total

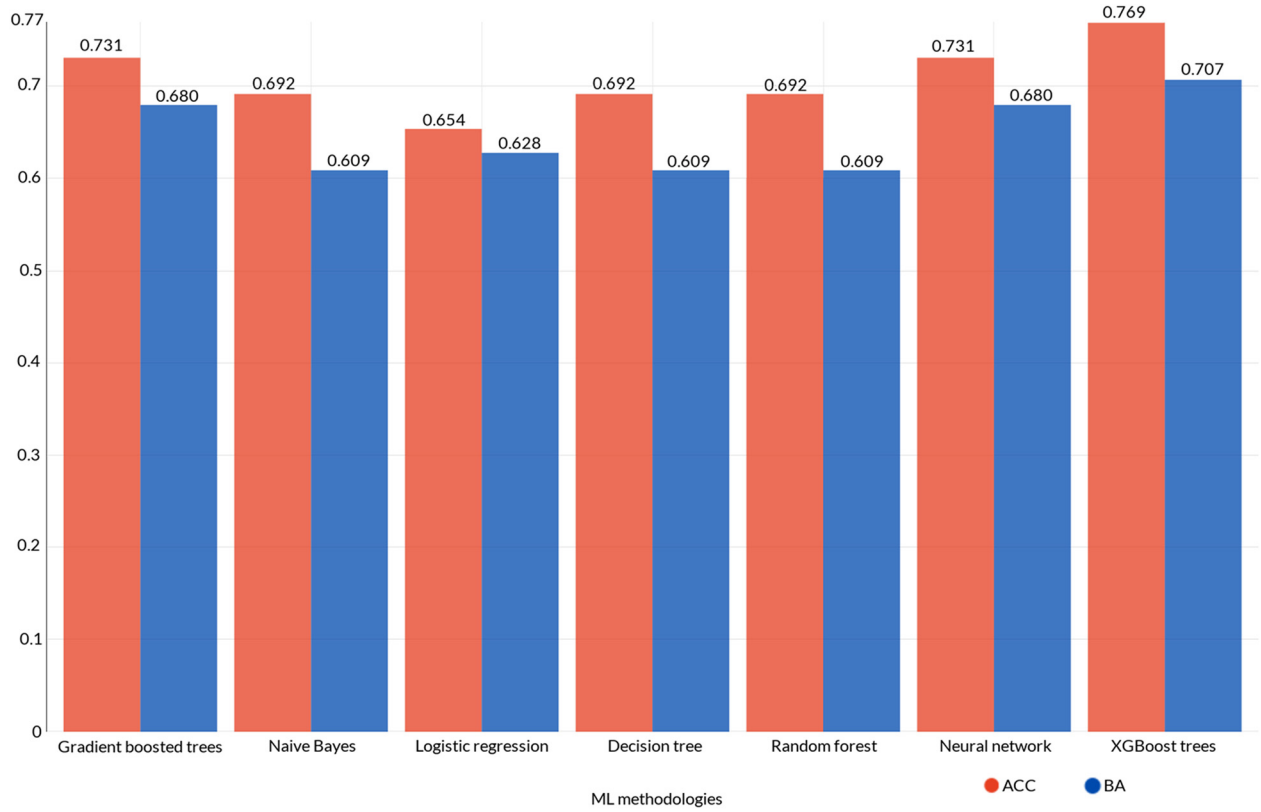

Fig. 2 ACC and BA of the validation set of the ML models within the autoML workflow.

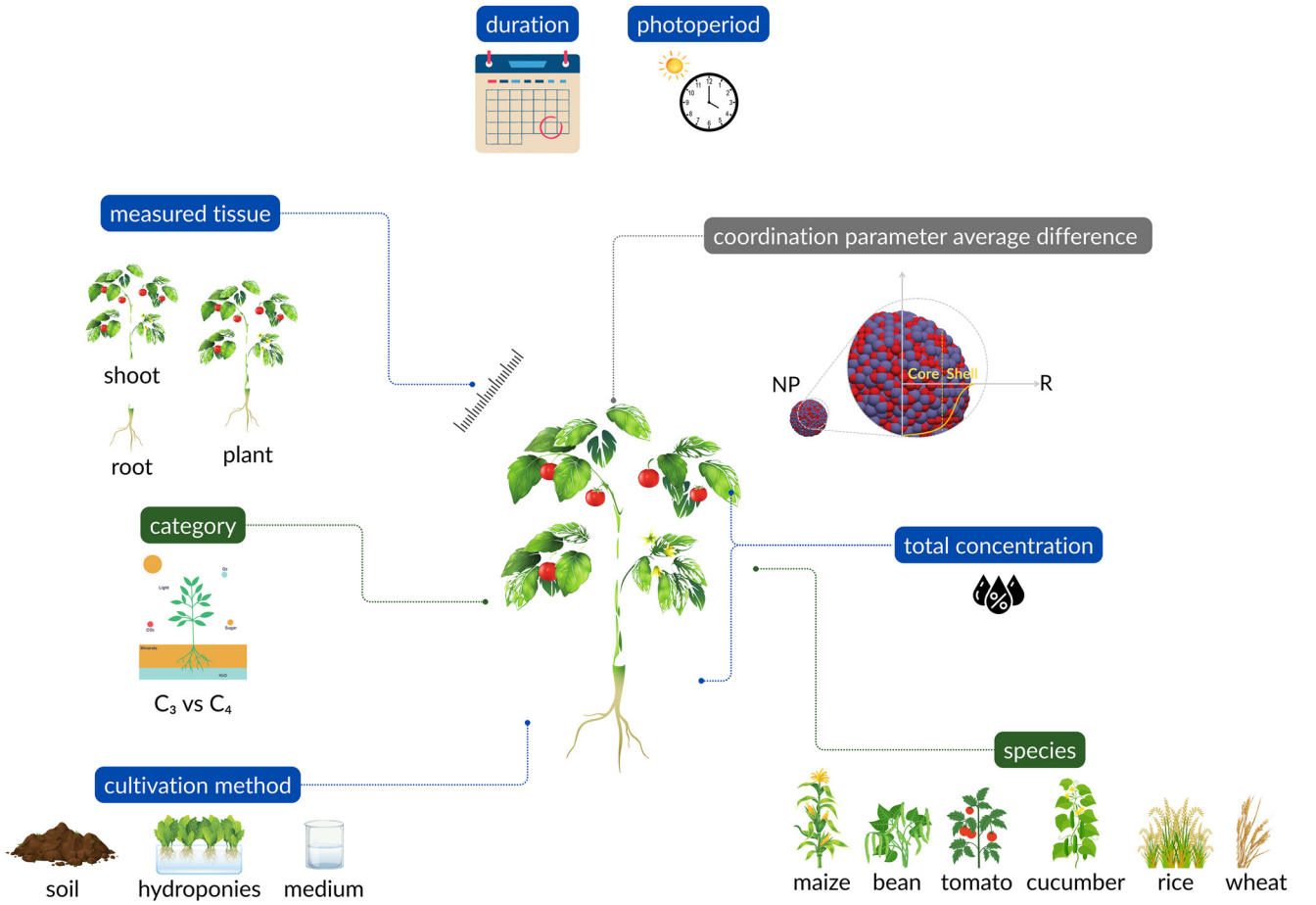

Fig. 3 Overview of the selected descriptors after feature permutation for the development of the final XGBoost model. Blue labels refer to experimental conditions, green labels to plant properties and grey labels to NP related descriptors. Designed with https://Canva.com.

Table 7 Confusion matrix summarising the number of correct predictions and misclassifications from the test set

Actual/predicted class	Negative	Positive
Negative	26	4
Positive	2	8

concentration, species, MT, cultured, category, duration, photoperiod and D12) or upload the input data using the provided template file. The atomistic descriptors can be obtained using the NanoConstruct tool (https://enaloscloud.novamechanics.com/riskgone/nanoconstruct/), which is also freely available. Upon execution, the web application generates predictions for the length class of the input NP–plant treatments, and indicates their reliability based on the APD limits in tabular format. The results are available for download, enabling further analysis. A detailed user-guide is also available to facilitate the use of the CeresAI-nano. An overview of the graphical user interface is presented in Fig. 4.

Table 8 Accuracy statistics of the XGBoost model when applied to the test set which contained data for 40 NPs

Metrics	Values
ACC	0.850
BA	0.833
SEN	0.867
PRE	0.929
SPE	0.800
F1	0.897
MCC	0.630

Discussion

Despite the extensive filtering of the data and the reduction of the initial dataset to less than half of its original size (38% of the original observations remained), the final XGBoost model, through extensive data curation, enrichment with 53 atomistic descriptors, and synthetic data generation, was able to predict the (plant) length effect class with satisfactory accuracy (ACC: 85.0%, BA: 83.3%), as demonstrated by the model's external validation statistics. Although the XGBoost model performed adequately when developed with a small training set, data and meta-data scarcity and heterogenicity remain a central challenge in the nanoinformatics field.[49,69,70] These limitations reduce the range and granularity of the feature space and hence model generalizability. While synthetic data was employed to address class imbalance and expand the input domain, it

Table 9 BA bootstrap validation results on the XGBoost model applied to the test dataset (40 NPs)

BA	Values
Mean	0.830
Median	0.833
Standard deviation	0.073

Table 11 Performance of the XGBoost model in LOO and ten-fold cross-validation

Metrics	LOO	Ten-fold
ACC	0.843	0.886
BA	0.843	0.886
MCC	0.686	0.777

inherently reflects assumptions that may not necessarily be representative of the physicochemical and biological heterogeneity that actually exist in NP–plant interactions. Thus, our effort highlights the critical need for the generation and systematic collection of experimental data and metadata to enable the development of robust predictive models in the nanoinformatics field including the nano-enabled agriculture. The availability of our curated dataset and its rich metadata through the nanoPharos database and nanodash, respectively, contributes to data reusability and overall FAIRness.

The interactions between NPs and plants are characterised by a high level of complexity and ambiguity[6,8] including the complexity and diversity of the environment-plant system[24] and the intricate network of parameters controlling plant responses.[2] 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.[1,7,13,18] Beyond these, additional factors may be considered that will provide a more comprehensive and realistic assessment of NP impacts in agriculture and the environment and will be incorporated in future *in silico* models. Such parameters may include genomic, proteomic, and metabolomic studies to assess long-term effects of NP–plant interactions,[71] soil texture, pH, and organic matter content to unravel NPs fate and biotransformation (*e.g.*, dissolution) due to plant–microbial–soil interactions.[72] The factors influencing plant length, – and plant growth in general, are complex and sometimes not entirely known, and the presence of NPs 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 (see also the following paragraphs).

For instance, it has been shown that identical Ag NPs treatments using different NP shapes result in significantly different root length responses of the *Arabidopsis* seedlings, with decahedral NPs giving maximum length increase compared to the control sample.[73] Moreover, the existence of an applied concentration threshold[74] is reported, beyond which plant growth is adversely affected. This threshold varies depending not only on the NP type, but also the plant species since the same NPs have been known to have opposite effects on different plants. While some NP treatments have the same effect both on the shoot and the root length, it has been reported that NPs can have mixed effects on plant species such as *Coriandrum sativum* L. and *Solanum melongena* L.[75] Another example is the case of Ag NPs that enhance *Zea mays* plant growth but inhibit the root growth of *Lolium multiform*.[76] Plant varieties also modulate the NP–plant interactions. For example, lignin content differences among sweetpotato varieties influence root length development following treatment with CuO NPs.[77] It is evident that the complex biochemical and physiological interactions between NPs and plants, which remain largely unexplored and not yet fully understood, hinder the development of interpretable ML models capable of accurately predicting the quantitative plant length response. Indeed, there are also complex dynamics within plants as to whether to divert energy to below ground growth (root system) *versus* above ground growth (shoots and fruit), that are only beginning to be explored.[78,79] Developing case-by-case models may offer a viable solution when dealing with "conflicting" NP–plant interaction effects, although this approach demands a more targeted data collection and curation approach. Consequently, given the reduced size of the dataset and the limited number of distinct NP–plant observations, it was proposed to simplify this inherently complex problem by categorising the "Length" labels as either "positive" or "negative" and developing a classification ML model. This qualitative approach is not just a forced compromise to facilitate the development of reliable classification models but can provide valuable predictions regarding the effects of NPs on plant growth and support the design of optimal NPs to maximize growth and yield. Furthermore, the proposed classification strategy could act as the initial step of an SSbD workflow, guiding the development of species-specific regression models for refined validation or for identifying hazardous concentration thresholds.

Table 10 BA and MCC statistics of the ten shuffled-endpoint XGBoost models when applied on the test set, all of which are lower than the BA and MCC of the original model (see Table 8)

Randomisation	BA	MCC
1	0.467	−0.061
2	0.467	−0.061
3	0.467	−0.061
4	0.533	0.067
5	0.467	−0.058
6	0.533	0.067
7	0.417	−0.146
8	0.383	−0.204
9	0.550	0.087
10	0.567	0.118

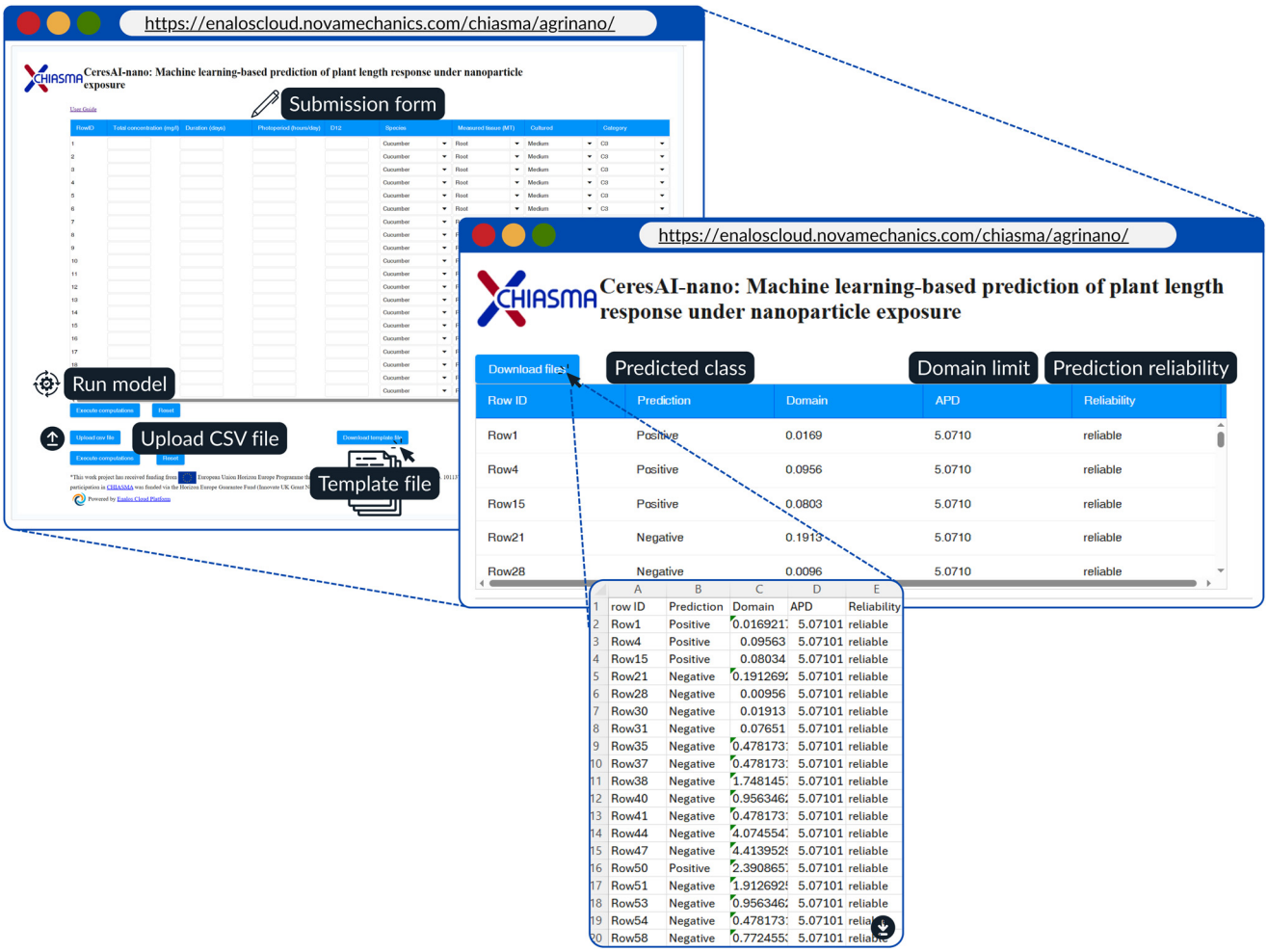

Fig. 4 CeresAI-nano overview. Users can fill in the provided form with NP treatment features or upload a CSV file that follows the template file. Upon submission, the web-service generates a table presenting predictions for each input treatment, along with an indication of their reliability based on the model's APD limits. The predictions can be downloaded for further analysis. Designed with https://Canva.com.

The "positive"/"negative" classification allows the development of simpler and more robust models, but the interpretation of these classes should be performed with caution. Yet, important conclusions can be drawn from the length response prediction for different parts of the plant, since root growth is related to the whole plant growth through an "allometry" relationship.[80] Depending on particle size, dose, composition, and physicochemical characteristics, NPs have been shown to have both beneficial and detrimental impacts on the root length and overall plant growth and development.[7,81,82] The output class ("positive" or "negative") refers to the length of the whole plant, root, or shoot – as specified by the input variable and usually is considered as a beneficial effect.[2] However, a "positive" length response does not necessarily correlate with improved yield-related outcomes. For example, the root length or the root/shoot ratio can be used as an indication of environmental stress. In nutrient-deficient soils or in desert environments, root length usually increases to aid the plant obtain the necessary resources in terms of nutrients[80] and water,[83] respectively. Furthermore, in the case of grain producing crops (such as wheat, rice, maize *etc.*) plant length plays a key role in yield production. Shorter plants (dwarf and semi-dwarf varieties) are in general less prone to lodging than taller ones, allowing in this way higher planting densities. In parallel, these varieties tend to allocate a greater proportion of resources to grain production rather than vegetative growth, resulting in increased yield potential. Nonetheless, excessive dwarfing may lead to decreased yield.[84–86] Therefore, the optimization of plant length through precision agriculture is important to achieve higher yields and consequently support global food demands. In cases where yield related data are scarce, the final XGBoost model can be used as a tool to gain preliminary insights on the plant length effects ("positive" or "negative") under NP exposure, prior to a refined experimental assessment to define the parameters that will optimize plant architecture. Thus, this simple classification of the effects of NPs on plant length can provide valuable insights into the optimal NPs treatment plan, provided that it is interpreted correctly in the context of the specific NP–plant application.

As previously discussed, the type of NPs, their composition, size, and the physicochemical properties derived from their type can either positively or negatively regulate root length.[1,7,13,82,87] In our case, the used atomistic descriptor (D12) encodes information about NP composition, shape, size, and crystallinity into a single value. It should be noted that although the NanoConstruct tool also calculates D12 for ellipsoidal NPs, the predictions derived from the reported model should be interpreted with caution, as the model was trained specifically for spherical NPs. NanoConstruct is also limited to NPs below 60 nm, and cannot calculate these parameters for carbon-based NPs, either graphene and carbon nanotubes or polymeric NPs. While descriptors for nanosheets and nanotubes can be generated with the NanoTube Construct tool,[88] these descriptors are not directly comparable with those from NanoConstruct, which can lead to data gaps. 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 concentrations.[1,13,18,82,87,89] In the final model, the exposure concentration factor is represented by the total concentration variable. Finally, the NP effect varies across plant species and can be both positive and negative.[13,82,87] This is the case of Ag NPs,[89] 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.[82] 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).

To further interpret the influence of the selected numerical features on the final model's decision-making process, a SHAP (SHapley Additive exPlanations) analysis[90] was performed in using Tree SHAP to illustrate how these descriptors influence the prediction probability, specifically for the "positive" class (increased plant length). In Fig. S5 the SHAP summary plot is depicted, reflecting the magnitude and direction of each descriptor's impact on the model's output when compared to the average prediction. Descriptors are ranked by decreasing importance corresponding to their mean absolute SHAP value. Positive SHAP values indicate a contribution towards the "positive" class probability (increased length compared to the control), and negative SHAP values are associated with an increase in the "negative" class probability. Photoperiod contributes most to the classification through a monotonous relationship with higher feature values (red dots) driving the prediction towards the "negative" class. Photoperiod plays a fundamental role in plant studies as it serves as a reliable environmental cue for seasonal timing. It is directly connected to photosynthesis, influencing carbon fixation and biomass accumulation, and regulates a wide range of growth-related, physiological, and stress-response processes, including flowering, branch formation and circadian clock synchronization.[91–94] In our analysis, longer photoperiods expressed in hours per day tend to drive the prediction towards the reduced plant length response class under NP treatment compared to the control. Nonetheless, other growth parameters such as the dry or fresh weight, the leaf area, the number of leaves *etc.* after NP treatment were not included in this model to assess how the plant length is associated with overall biomass accumulation. Again, the synergistic effects of NPs, plant species, and experimental conditions should be further investigated to develop a more focused and optimised framework for precision agriculture. Based on the SHAP summary plot, lower NP total concentration values and exposure times mainly drive the prediction towards the "positive" class (increased plant length), which for the case of concentration is in line with our previous findings. The D12 atomistic descriptor is the lowest-ranked descriptor, with lower values showing a tendency to drive the prediction towards the "negative" class probability.

It can be concluded that highly influential factors for assessing plant length response are included in our model and, as they encode controllable experimental conditions (*e.g.*, laboratory, mesocosms,[95,96] greenhouse,[97] or hydroponic setups,[98] 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, saving time from long-time experimental cycles. In the study by Santos *et al.*,[2] which conducted a meta-analysis of research on the effects of nanomaterials on plants from 2019 to 2022, the median duration of experiments involving plants cultured in soil was 49 days. This time span is relatively short, as it falls below the typical growth cycle of annual crops (90–120 days), meaning that many studies capture only part of the plant's development. In this context, predictive modelling offers a valuable shortcut and can complement experimental work, enabling researchers to generate insights more rapidly and efficiently than through conventional experiments alone.

Conclusions

As the demand for increased agricultural production continues to rise in order to meet the growing needs for food, feed, biofuels, and bio-based materials, the role of nano-enabled agriculture is expanding significantly to enhance crop yield, improve plant resilience, and optimise resource use. The use of nanotechnology as a means to boost agriculture is considered essential to support sustainable agriculture. However, understanding of NP–plant interactions is challenging, and datasets are currently relatively limited, as NPs induce a range of different plant responses including changes in root and shoot length, their ratio, and overall plant length relative to untreated controls. Comparisons of the NP exposed plant to the control, can drive important conclusions regarding the effects of NPs on plants.

ML can be integrated to nano-enabled agriculture applications to support the prediction of plant growth indicators and contribute to the design of safe nano-agrichemicals, as

demonstrated by this work. In brief, based on data available on literature we performed a comprehensive and detailed curation, as quality data are the cornerstone for the development of reliable and interpretable ML models. Data were enriched with calculated atomistic NP descriptors, which were used alongside experimental conditions and plant characteristics as input to an autoML workflow for the prediction of the effects of NP exposure on plant length. To correct class imbalance, synthetic data were generated following the SMOTE methodology and seven ML methodologies were optimised and evaluated in terms of prediction accuracy. The XGBoost methodology was selected as the best performing one, yielding an accuracy of 85% and a balanced accuracy of 83% in external validation. Additional bootstrapping and *y*-randomisation tests demonstrated the stability of the predictions and the model's robustness.

The curated data and model are freely available and documented in a standardized report, aligning with the FAIR data principles, to ensure accessibility and reusability. The final model has been disseminated as a user-friendly web application to facilitate broader use by the interested stakeholders and to support applications in nano-enabled agriculture. The CeresAI-nano tool does not require experimental input, instead, it uses variables such as experimental conditions, NP core, size and purity, and plant details. This enables rapid virtual screening of possible NP–plant treatments to assess their plant length effects, saving time from long-term experiments. With the growing interest in nanomaterial applications in agriculture driving increased experimental assessment, a synergistic integration with ML approaches is essential to advance research efficiency. This integration relies on the systematic collection of data and metadata to support the development of robust and reliable predictive models that can later contribute to the understanding of the NP–plant interaction mechanisms.

Author contributions

Dimitra-Danai Varsou: conceptualization, methodology, software, validation, formal analysis, data curation, writing – original draft, writing – review & editing, visualization; Aikaterini Theodori: software, data curation, writing – original draft, writing – review & editing; Anastasios G. Papadiamantis: writing – original draft, writing – review & editing; Dimitrios Zouraris: data curation; Maria Antoniou: software; Nikoletta-Maria Koutroumpa: writing – review & editing; Andreas Tsoumanis: software; Georgia Melagraki: writing – review & editing; Iseult Lynch: conceptualization, supervision, writing – review & editing; Antreas Afantitis: conceptualization, resources, writing – review & editing, supervision, funding acquisition.

Conflicts of interest

DDV, AT, AGP, AT, DZ, MA, NMK, and AA are affiliated with NovaMechanics, a cheminformatics and materials informatics company.

Abbreviations

ACC	Accuracy
ANN	Artificial neural network
APD	Applicability domain
A_r	Relative atomic mass
AutoML	Automated machine learning
BA	Balanced accuracy
CIF	Crystallographic information file
COD	Crystallography Open Database
DLS	Dynamic light scattering
DT	Daytime temperature
EP	Exposure pathway
FAIR	Findable, accessible, interoperable and reusable
FN	False negatives
FP	False positives
LightGBM	Light gradient boosting machine models
GS	Growth stage of plant
*k*NN	*k*-Nearest neighbours
LOO	Leave-one-out
MCC	Matthews correlation coefficient
ML	Machine learning
MT	Measured tissue
NP	Nanoparticle
NT	Night temperature
OECD	Organisation for Economic Co-operation and Development
QC	Quality control
QMRF	QSAR model report format
QSAR	Quantitate structure–activity relationship
PRE	Precision
RCF	Root concentration factor
RMC	Relative metal/metalloid concentration
SEN	Sensitivity
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic minority over-sampling technique
SPE	Specificity
SSbD	Safe and sustainable-by-design
TC	Total content
TEM	Transmission electron microscopy
thr	threshold (APD)
TF	Transport factor
TN	True negatives
TP	True positives
XGBoost	Extreme gradient boosting

Data availability

The CeresAI-nano web-application is available through the Enalos Cloud Platform: https://enaloscloud.novamechanics.com/chiasma/agrinano/. The model can be also accessed through APIs: https://enaloscloud.novamechanics.com/chiasma/swagger-ui/. Data for this article, including the curated and enriched dataset with the atomistic descriptors are available at the nanoPharos database at https://db.nanopharos.eu/Queries/Datasets.zul?datasetID=np31. Metadata have been published as

a machine actionable nanopublication: https://w3id.org/np/RALzBzdG_gTKI1Xn6SubCkA-ZTeK5LyAV_aYjrW-Yyzj8.

Supplementary information (SI) is available. See DOI: https://doi.org/10.1039/d5en00897b.

Acknowledgements

This work has received funding from the European Union Horizon Europe Programme through the CHIASMA project (Grant Agreement No. 101137613). UoB participation in CHIASMA was funded *via* the Horizon Europe Guarantee Fund (Innovate UK Grant No. 10101594).

References

1 S. Tripathi, K. Tiwari, S. Mahra, J. Victoria, S. Rana and D. K. Tripathi, *et al.*, Nanoparticles and root traits: mineral nutrition, stress tolerance and interaction with rhizosphere microbiota, *Planta*, 2024, **260**(2), 1–19, DOI: 10.1007/s00425-024-04409-y.

2 E. Santos, G. S. Montanha, M. H. F. Gomes, N. M. Duran, C. G. Corrêa and S. L. Z. Romeu, *et al.*, Are nanomaterials leading to more efficient agriculture? Outputs from 2009 to 2022 research metadata analysis, *Environ. Sci.: Nano*, 2022, **9**(10), 3711–3724, Available from: https://xlink.rsc.org/?DOI=D1EN01078F.

3 S. Agrawal, V. Kumar, S. Kumar and S. K. Shahi, Plant development and crop protection using phytonanotechnology: A new window for sustainable agriculture, *Chemosphere*, 2022, **299**, 134465, DOI: 10.1016/j.chemosphere.2022.134465.

4 G. Bhandari, A. Dhasmana, P. Chaudhary, S. Gupta, S. Gangola and A. Gupta, *et al.*, A Perspective Review on Green Nanotechnology in Agro-Ecosystems: Opportunities for Sustainable Agricultural Practices & Environmental Remediation, *Agriculture*, 2023, **13**(3), 668, Available from: https://www.mdpi.com/2077-0472/13/3/668.

5 P. Zhang, Z. Guo, S. Ullah, G. Melagraki, A. Afantitis and I. Lynch, Nanotechnology and artificial intelligence to enable sustainable and precision agriculture, *Nat. Plants*, 2021, **7**(7), 864–876, DOI: 10.1038/s41477-021-00946-6.

6 J. Li, X. Li, M. Kah, L. Yue, B. Cheng and C. Wang, *et al.*, Unlocking the potential of carbon dots in agriculture using data-driven approaches, *Sci. Total Environ.*, 2024, **944**, 173605, DOI: 10.1016/j.scitotenv.2024.173605.

7 V. R. Rajpal, B. Nongthongbam, M. Bhatia, A. Singh, S. N. Raina, T. Minkina, V. D. Rajput, N. Zahra and A. Husen, The nano-paradox: addressing nanotoxicity for sustainable agriculture, circular economy and SDGs, *J. Nanobiotechnol.*, 2025, **23**(1), 314, DOI: 10.1186/s12951-025-03371-5.

8 P. Deng, Y. Gao, L. Mu, X. Hu, F. Yu, Y. Jia, Z. Wang and B. Xing, Development potential of nanoenabled agriculture projected using machine learning, *Proc. Natl. Acad. Sci. U. S. A.*, 2023, **120**(25), e2301885120, DOI: 10.1073/pnas.2301885120.

9 B. Ahmed, A. Rizvi, K. Ali, J. Lee, A. Zaidi, M. S. Khan and J. Musarrat, Nanoparticles in the soil–plant system: a review, *Environ. Chem. Lett.*, 2021, **19**, 1545–1609, DOI: 10.1007/s10311-020-01138-y.

10 F. Huang, L. Chen, Y. Zeng, W. Dai, F. Wu, Q. Hu, Y. Zhou, S. Shi and L. Fang, Unveiling influences of metal-based nanomaterials on wheat growth and physiology: From benefits to detriments, *Chemosphere*, 2024, **364**, 143212, DOI: 10.1016/j.chemosphere.2024.143212.

11 F. Chen, F. Jiang, M. K. Okla, Z. K. Abbas, S. M. Al-Qahtani, N. A. Al-Harbi, M. A. Abdel-Maksoud and L. M. Gómez- Oliván, Nanoparticles synergy: Enhancing wheat (Triticum aestivum L.) cadmium tolerance with iron oxide and selenium, *Sci. Total Environ.*, 2024, **915**, 169869, DOI: 10.1016/j.scitotenv.2024.169869.

12 N. Xu, J. Kang, Y. Ye, Q. Zhang, M. Ke and Y. Wang, *et al.*, Machine learning predicts ecological risks of nanoparticles to soil microbial communities, *Environ. Pollut.*, 2022, **307**, 119528, Available from: https://linkinghub.elsevier.com/retrieve/pii/S0269749122007424.

13 A. Wahab, A. Munir, M. H. Saleem, M. I. AbdulRaheem, H. Aziz and M. F. B. Mfarrej, *et al.*, Interactions of Metal-Based Engineered Nanoparticles with Plants: An Overview of the State of Current Knowledge, Research Progress, and Prospects, *J. Plant Growth Regul.*, 2023, **42**(9), 5396–5416, DOI: 10.1007/s00344-023-10972-7.

14 C. García-Gómez, A. Obrador, D. González, M. Babín and M. D. Fernández, Comparative study of the phytotoxicity of ZnO nanoparticles and Zn accumulation in nine crops grown in a calcareous soil and an acidic soil, *Sci. Total Environ.*, 2018, **644**, 770–780, DOI: 10.1016/j.scitotenv.2018.06.356.

15 N. Al-Amri, H. Tombuloglu, Y. Slimani, S. Akhtar, M. Barghouthi and M. Almessiere, *et al.*, Size effect of iron (III) oxide nanomaterials on the growth, and their uptake and translocation in common wheat (Triticum aestivum L.), *Ecotoxicol. Environ. Saf.*, 2020, **194**, 110377, DOI: 10.1016/j.ecoenv.2020.110377.

16 M. Nekoukhou, S. Fallah, A. Abbasi-Surki, L. R. Pokhrel and A. Rostamnejadi, Improved efficacy of foliar application of zinc oxide nanoparticles on zinc biofortification, primary productivity and secondary metabolite production in dragonhead, *J. Cleaner Prod.*, 2022, **379**(P2), 134803, DOI: 10.1016/j.jclepro.2022.134803.

17 L. Q. Thao, D. T. Kien, N. D. Thien, N. C. Bach, V. Van Hiep and D. G. Khanh, Utilizing AI and silver nanoparticles for the detection and treatment monitoring of canker in pomelo trees, *Sens. Actuators, A*, 2024, **368**, 115127, Available from: https://linkinghub.elsevier.com/retrieve/pii/S0924424724001201.

18 R. Javed, B. Khan, U. Sharafat, M. Bilal, L. Galagedara and L. Abbey, *et al.*, Dynamic interplay of metal and metal oxide nanoparticles with plants: Influencing factors, action mechanisms, and assessment of stimulatory and inhibitory effects, *Ecotoxicol. Environ. Saf.*, 2024, **271**, 115992, DOI: 10.1016/j.ecoenv.2024.115992.

19 R. Rienzie and N. M. Adassooriya, Toxicity of Nanomaterials in Agriculture and Food, in *Nanomaterials: Ecotoxicity, Safety,*

and Public Perception, ed. M. Rai and J. K. Biswas, Springer International Publishing, Cham, 2018, pp. 207–234, Available from: DOI: 10.1007/978-3-030-05144-0.

20 A. M. Santos-Espinoza, D. González-Mendoza, V. M. Ruiz-Valdiviezo, M. C. Luján-Hidalgo, F. Jonapa-Hernández and B. Valdez-Salas, *et al.*, Changes in the physiological and biochemical state of peanut plants (Arachis hypogaea L.) induced by exposure to green metallic nanoparticles, *Int. J. Phytorem.*, 2021, **23**(7), 747–754, DOI: 10.1080/15226514.2020.1856037.

21 D.-D. Varsou and H. Sarimveis, Apellis: An online tool for read-across model development, *Comput. Toxicol.*, 2021, **17**, 100146, DOI: 10.1016/j.comtox.2020.100146.

22 J. Li, L. Yue, F. Chen, X. Cao, B. Cheng and C. Wang, *et al.*, Artificial neural networks to investigate the bioavailability of selenium nanoparticles in soil–crop systems, *Environ. Sci.: Nano*, 2024, **11**(1), 418–430, Available from: https://xlink.rsc.org/?DOI=D3EN00412K.

23 X. Wang, L. Liu, W. Zhang and X. Ma, Prediction of Plant Uptake and Translocation of Engineered Metallic Nanoparticles by Machine Learning, *Environ. Sci. Technol.*, 2021, **55**(11), 7491–7500.

24 L. Rossi, M. Bagheri, W. Zhang, Z. Chen, J. G. Burken and X. Ma, Using artificial neural network to investigate physiological changes and cerium oxide nanoparticles and cadmium uptake by Brassica napus plants, *Environ. Pollut.*, 2019, **2019**(246), 381–389, DOI: 10.1016/j.envpol.2018.12.029.

25 T. Min, T. Lu, S. Zheng, W. Tan, T. Luo and G. Qiu, Efficiency of metal oxides in reducing heavy metal uptake in typical crops: A machine learning-assisted meta-analysis, *J. Cleaner Prod.*, 2025, **491**, 144856, DOI: 10.1016/j.jclepro.2025.144856.

26 H. Yu, S. Tang, S. F. Y. Li and F. Cheng, Averaging Strategy for Interpretable Machine Learning on Small Datasets to Understand Element Uptake after Seed Nanotreatment, *Environ. Sci. Technol.*, 2023, **57**(34), 12760–12770.

27 Organisation for Economic Co-operation and Development (OECD), *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models*, OECD, 2014, pp. 1–154, Available from: https://www.oecd-ilibrary.org/environment/guidance-document-on-the-validation-of-quantitative-structure-activity-relationship-q-sar-models_9789264085442-en.

28 dp1999nku, All-datasets, 2023, [cited 2023 Sep 15], Available from: https://github.com/dp1999nku/All-datasets.

29 P. D. Kolokathis, D. Zouraris, E. Voyiatzis, N. K. Sidiropoulos, A. Tsoumanis and G. Melagraki, *et al.*, NanoConstruct: A web application builder of ellipsoidal nanoparticles for the investigation of their crystal growth, stability, and the calculation of atomistic descriptors, *Comput. Struct. Biotechnol. J.*, 2024, **25**, 81–90.

30 P. D. Kolokathis, E. Voyiatzis, N. K. Sidiropoulos, A. Tsoumanis, G. Melagraki, K. Tamm, I. Lynch and A. Afantitis, ASCOT: A Web Tool for the Digital Construction of Energy Minimized Ag, CuO, TiO_2 Spherical Nanoparticles and Calculation of Their Atomistic Descriptors, *Comput. Struct. Biotechnol. J.*, 2024, **25**, 34–46, DOI: 10.1016/j.jbc.2022.102753.

31 D. Cui, P. Zhang, Y. H. Ma, X. He, Y. Y. Li and Y. C. Zhao, *et al.*, Phytotoxicity of silver nanoparticles to cucumber (Cucumis sativus) and wheat (Triticum aestivum), *J. Zhejiang Univ., Sci., A*, 2014, **15**(8), 662–670.

32 J. Liu, B. Dhungana and G. P. Cobb, Copper oxide nanoparticles and arsenic interact to alter seedling growth of rice (Oryza sativa japonica), *Chemosphere*, 2018, **206**, 330–337, DOI: 10.1016/j.chemosphere.2018.05.021.

33 J. Liu, M. Simms, S. Song, R. S. King and G. P. Cobb, Physiological Effects of Copper Oxide Nanoparticles and Arsenic on the Growth and Life Cycle of Rice (Oryza sativa japonica 'Koshihikari'), *Environ. Sci. Technol.*, 2018, **52**(23), 13728–13737.

34 P. M. Gopalakrishnan Nair and I. M. Chung, Physiological and molecular level studies on the toxicity of silver nanoparticles in germinating seedlings of mung bean (Vigna radiata L.), *Acta Physiol. Plant.*, 2015, **37**(1), 1719, DOI: 10.1007/s11738-014-1719-1.

35 W. Wang, J. Liu, Y. Ren, L. Zhang, Y. Xue and L. Zhang, *et al.*, Phytotoxicity Assessment of Copper Oxide Nanoparticles on the Germination, Early Seedling Growth, and Physiological Responses in Oryza sativa L, *Bull. Environ. Contam. Toxicol.*, 2020, **104**(6), 770–777, DOI: 10.1007/s00128-020-02850-9.

36 Y. Wang, F. Jiang, C. Ma, Y. Rui, D. C. W. Tsang and B. Xing, Effect of metal oxide nanoparticles on amino acids in wheat grains (Triticum aestivum)in a life cycle study, *J. Environ. Manage.*, 2019, **241**, 319–327, DOI: 10.1016/j.jenvman.2019.04.041.

37 A. Rohatgi, *WebPlotDigitizer*, Available from: https://automeris.io.

38 S. Gražulis, A. Merkys and A. Vaitkus, Crystallography Open Database (COD), in *Handbook of Materials Modeling*, Springer International Publishing, Cham, 2020, pp. 1863–1881, Available from: DOI: 10.1007/978-3-319-44677-6_66.

39 E. R. López-Vargas, Y. González-García, M. Pérez-Álvarez, G. Cadenas-Pliego, S. González-Morales and A. Benavides-Mendoza, *et al.*, Seed priming with carbon nanomaterials to modify the germination, growth, and antioxidant status of tomato seedlings, *Agronomy*, 2020, **10**(5), 1–22.

40 E. B. Tadmor, R. S. Elliott, J. P. Sethna, R. E. Miller and C. A. Becker, The potential of atomistic simulations and the knowledgebase of interatomic models, *JOM*, 2011, **63**(7), 17.

41 Q. Abbas, G. Liu, B. Yousaf, M. U. Ali, H. Ullah and R. Ahmed, Effects of biochar on uptake, acquisition and translocation of silver nanoparticles in rice (Oryza sativa L.) in relation to growth, photosynthetic traits and nutrients displacement, *Environ. Pollut.*, 2019, **250**, 728–736, DOI: 10.1016/j.envpol.2019.04.083.

42 P. Feng, B. Geng, Z. Cheng, X. Liao, D. Pan and J. Huang, Graphene quantum dots-induced physiological and biochemical responses in mung bean and tomato seedlings, *Rev. Bras. Bot.*, 2019, **42**(1), 29–41, DOI: 10.1007/s40415-019-00519-0.

43 A. Konate, Y. Wang, X. He, M. Adeel, P. Zhang and Y. Ma, *et al.*, Comparative effects of nano and bulk-Fe3O4 on the

growth of cucumber (Cucumis sativus), *Ecotoxicol. Environ. Saf.*, 2018, **165**, 547–554, DOI: 10.1016/j.ecoenv.2018.09.053.

44 H. Mahawar, R. Prasanna, K. Simranjit, S. Thapa, A. Kanchan and R. Singh, *et al.*, Deciphering the mode of interactions of nanoparticles with mung bean (Vigna radiata L.), *Isr. J. Plant Sci.*, 2018, **65**(1–2), 74–82, DOI: 10.1080/07929978.2017.1288516.

45 J. Lian, L. Zhao, J. Wu, H. Xiong, Y. Bao and A. Zeb, *et al.*, Foliar spray of TiO2 nanoparticles prevails over root application in reducing Cd accumulation and mitigating Cd-induced phytotoxicity in maize (Zea mays L.), *Chemosphere*, 2020, **239**(38), 124794, DOI: 10.1016/j.chemosphere.2019.124794.

46 Y. Ji, Y. Zhou, C. Ma, Y. Feng, Y. Hao and Y. Rui, *et al.*, Jointed toxicity of TiO2 NPs and Cd to rice seedlings: NPs alleviated Cd toxicity and Cd promoted NPs uptake, *Plant Physiol. Biochem.*, 2017, **110**, 82–93.

47 M. Rizwan, S. Ali, M. Zia ur Rehman, M. Adrees, M. Arshad and M. F. Qayyum, *et al.*, Alleviation of cadmium accumulation in maize (Zea mays L.) by foliar spray of zinc oxide nanoparticles and biochar to contaminated soil, *Environ. Pollut.*, 2019, **248**, 358–367, DOI: 10.1016/j.envpol.2019.02.031.

48 D.-D. Varsou, G. Tsiliki, P. Nymark, P. Kohonen, R. Grafström and H. Sarimveis, toxFlow: A Web-Based Application for Read-Across Toxicity Prediction Using Omics and Physicochemical Data, *J. Chem. Inf. Model.*, 2018, **58**(3), 543–549, DOI: 10.1021/acs.jcim.7b00160.

49 D.-D. Varsou, A. Afantitis, G. Melagraki and H. Sarimveis, Read-across predictions of nanoparticle hazard endpoints: a mathematical optimization approach, *Nanoscale Adv.*, 2019, **1**(9), 3485–3498, Available from: http://pubs.rsc.org/en/Content/ArticleLanding/2019/NA/C9NA00242A.

50 P. Xu, X. Ji, M. Li and W. Lu, Small data machine learning in materials science, *npj Comput. Mater.*, 2023, **9**(1), 1–15.

51 J. S. Choi, M. K. Ha, T. X. Trinh, T. H. Yoon and H. G. Byun, Towards a generalized toxicity prediction model for oxide nanomaterials using integrated data from different sources, *Sci. Rep.*, 2018, **8**(1), 1–10, DOI: 10.1038/s41598-018-24483-z.

52 T. X. Trinh, M. K. Ha, J. S. Choi, H. G. Byun and T. H. Yoon, Curation of datasets, assessment of their quality and completeness, and nanoSAR classification model development for metallic nanoparticles, *Environ. Sci.: Nano*, 2018, **5**(8), 1902–1910.

53 D.-D. Varsou, P. D. Kolokathis, M. Antoniou, N. K. Sidiropoulos, A. Tsoumanis and A. G. Papadiamantis, *et al.*, In silico assessment of nanoparticle toxicity powered by the Enalos Cloud Platform: Integrating automated machine learning and synthetic data for enhanced nanosafety evaluation, *Comput. Struct. Biotechnol. J.*, 2024, **25**, 47–60, DOI: 10.1016/j.csbj.2024.03.020.

54 N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, *J. Artif. Intell. Res.*, 2002, **16**(2), 321–357, DOI: 10.1002/eap.2043.

55 OECD, (Q)SAR Assessment Framework: Guidance for the regulatory assessment of (Quantitative) Structure Activity Relationship models, predictions, and results based on multiple predictions Series on Testing and Assessment No. 386, 2023, p. 33, Available from: https://www.oecd.org/chemicalsafety/risk-assessment/qsar-assessment-framework.pdf.

56 A. Gajewicz, How to judge whether QSAR/read-across predictions can be trusted: a novel approach for establishing a model's applicability domain, *Environ. Sci.: Nano*, 2018, **5**(2), 408–421, Available from: http://xlink.rsc.org/?DOI=C7EN00774D.

57 S. Zhang, A. Golbraikh, S. Oloff, H. Kohn and A. Tropsha, A Novel Automated Lazy Learning QSAR (ALL-QSAR) Approach: Method Development, Applications, and Virtual Screening of Chemical Databases Using Validated ALL-QSAR Models, *J. Chem. Inf. Model.*, 2006, **46**(5), 1984–1995, DOI: 10.1021/ci060132x.

58 D.-D. Varsou, S. Nikolakopoulos, A. Tsoumanis, G. Melagraki and A. Afantitis, Enalos+ KNIME Nodes: New Cheminformatics Tools for Drug Discovery, in *Rational Drug Design: Methods and Protocols, Methods in Molecular Biology*, ed. T. Mavromoustakos and T. F. Kellici, Humana Press, New York, NY, 2018, pp. 113–38, Available from: DOI: 10.1007/978-1-4939-8630-9_7.

59 P. Katz, K. Muthmann and D. Urbansky, *Palladian for KNIME*, 2023, Available from: https://nodepit.com/iu/ws.palladian.nodes.feature.feature.group.

60 KNIME AG, *AutoML component*, KNIME, 2023, Available from: https://hub.knime.com/knime/spaces/Examples/00_Components/Automation/AutoML~33fQGaQzuZByy6hE/current-state.

61 D.-D. Varsou, A. Tsoumanis, A. G. Papadiamantis, G. Melagraki and A. Afantitis, *Isalos Predictive Analytics Platform: Cheminformatics, Nanoinformatics, and Data Mining Applications*, Springer International Publishing, 2023, pp. 223–242, Available from: DOI: 10.1007/978-3-031-20730-3_9.

62 N. A. Subramanian and A. Palaniappan, NanoTox: Development of a Parsimonious in Silico Model for Toxicity Assessment of Metal-Oxide Nanoparticles Using Physicochemical Features, *ACS Omega*, 2021, **6**(17), 11729–11739.

63 M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton and A. Baak, *et al.*, The FAIR Guiding Principles for scientific data management and stewardship, *Sci. Data*, 2016, **3**(1), 160018, Available from: https://www.nature.com/articles/sdata201618.

64 A. Jacobsen, R. de Miranda Azevedo, N. Juty, D. Batista, S. Coles and R. Cornet, *et al.*, FAIR Principles: Interpretations and Implementation Considerations, *Data Intell.*, 2020, **2**(1–2), 10–29, DOI: 10.1162/dint_r_00024.

65 H. Lowenthal, T. Austin, L. Bonino Da Silva Santos, C. Chiarelli, A. Cusinato and C. Ferigato, *et al.*, *JRC FAIR Data Guidelines*, European Commission: Joint Research Centre, 2025, Available from: DOI: 10.2760/5646214.

66 S. R. Wilkinson, M. Aloqalaa, K. Belhajjame, M. R. Crusoe, B. de Paula Kinoshita and L. Gadelha, *et al.*, Applying the FAIR Principles to computational workflows, *Sci. Data*,

2025, **12**(1), 328, Available from: http://arxiv.org/abs/2410.03490.

67 M. Barker, N. P. Chue Hong, D. S. Katz, A.-L. Lamprecht, C. Martinez-Ortiz and F. Psomopoulos, *et al.*, Introducing the FAIR Principles for research software, *Sci. Data*, 2022, **9**(1), 622, Available from: https://www.nature.com/articles/s41597-022-01710-x.

68 OECD, Annex I - (Q)SAR model reporting format (QMRF) v.2.1, in *(Q)SAR Assessment Framework: Guidance for the regulatory assessment of (Quantitative) Structure Activity Relationship models, predictions, and results based on multiple predictions Series on Testing and Assessment No 386*, 2023, Available from: https://www.oecd.org/content/dam/oecd/en/topics/policy-sub-issues/assessment-of-chemicals/qsar-assessment-framework-annex-1-qsar-model-reporting-format.docx.

69 G. Mancardi, A. Mikolajczyk, V. K. Annapoorani, A. Bahl, K. Blekos and J. Burk, *et al.*, A computational view on nanomaterial intrinsic and extrinsic features for nanosafety and sustainability, *Mater. Today*, 2023, **67**, 344–370.

70 I. Furxhi, Health and environmental safety of nanomaterials: O Data, Where Art Thou?, *NanoImpact*, 2022, **25**, 100378, DOI: 10.1016/j.impact.2021.100378.

71 S. G. Thabet and A. M. Alqudah, Unraveling the role of nanoparticles in improving plant resilience under environmental stress condition, *Plant Soil*, 2024, **503**(1), 313–330, DOI: 10.1007/s11104-024-06581-2.

72 H. Zhang, T. Zheng, Y. Wang, T. Li and Q. Chi, Multifaceted impacts of nanoparticles on plant nutrient absorption and soil microbial communities, *Front. Plant Sci.*, 2024, **15**, 1497006.

73 Y.-y. Syu, J. H. Hung, J. C. Chen and H.-w. Chuang, Impacts of size and shape of silver nanoparticles on Arabidopsis plant growth and gene expression, *Plant Physiol. Biochem.*, 2014, **83**, 57–64, DOI: 10.1016/j.plaphy.2014.07.010.

74 U. Shafqat, S. Hussain, T. Shahzad, M. Shahid and F. Mahmood, Elucidating the phytotoxicity thresholds of various biosynthesized nanoparticles on physical and biochemical attributes of cotton, *Chem. Biol. Technol. Agric.*, 2023, **10**(1), 1–15, DOI: 10.1186/s40538-023-00402-x.

75 G. Feigl, The impact of copper oxide nanoparticles on plant growth: a comprehensive review, *J. Plant Interact.*, 2023, **18**(1), 2243098, DOI: 10.1080/17429145.2023.2243098.

76 P. Goswami, S. Yadav and J. Mathur, Positive and negative effects of nanoparticles on plants and their applications in agriculture, *Plant Sci. Today*, 2019, **6**(2), 232–242.

77 N. J. Bonilla-Bird, Y. Ye, T. Akter, C. Valdes-Bracamontes, A. J. Darrouzet-Nardi and G. B. Saupe, *et al.*, Effect of copper oxide nanoparticles on two varieties of sweetpotato plants, *Plant Physiol. Biochem.*, 2020, **154**, 277–286.

78 M. N. Umaña, M. Cao, L. Lin, N. G. Swenson and C. Zhang, Trade-offs in above- and below-ground biomass allocation influencing seedling growth in a tropical forest, *J. Ecol.*, 2021, **109**(3), 1184–1193, DOI: 10.1111/1365-2745.13543.

79 E. B. Kopp, N. P. R. Anten, P. A. Niklaus and S. E. Wuest, Belowground competition increases root allocation in agreement with game-theoretical predictions, but only when plants simultaneously compete aboveground, *bioRxiv*, 2025, preprint, DOI: 10.1101/2025.01.29.635491v1.

80 N. K. Fageria and A. Moreira, The Role of Mineral Nutrition on Root Growth of Crop Plants, *Advances in Agronomy*, Elsevier Inc., 1st edn, 2011, vol. 110, pp. 251–331, Available from: DOI: 10.1016/B978-0-12-385531-2.00004-9.

81 S. Wang, B. D. Wu, M. Wei, J. W. Zhou, K. Jiang and C. Y. Wang, Silver nanoparticles with different concentrations and particle sizes affect the functional traits of wheat, *Biol. Plant.*, 2020, **64**, 1–8.

82 M. R. Khan, V. Adam, T. F. Rizvi, B. Zhang, F. Ahamad and I. Jośko, *et al.*, Nanoparticle–Plant Interactions: Two-Way Traffic, *Small*, 2019, **15**(37), 1–20.

83 J. Lynch, Root Architecture and Plant Productivity, *Plant Physiol.*, 1995, **109**(1), 7–13, Available from: https://academic.oup.com/plphys/article/109/1/7-13/6069768.

84 M. J. Paul, A. Watson and C. A. Griffiths, Trehalose 6-phosphate signalling and impact on crop yield, *Biochem. Soc. Trans.*, 2020, **48**, 2127–2137.

85 F. Liu, P. Wang, X. Zhang, X. Li, X. Yan, D. Fu and G. Wu, The genetic and molecular basis of crop height based on a rice model, *Planta*, 2018, **247**, 1–26, DOI: 10.1007/s00425-017-2798-1.

86 S. K. Bhujbal, A. N. Rai and A. J. Saha, Dwarfs standing tall: breeding towards the 'Yellow revolution' through insights into plant height regulation, *Plant Mol. Biol.*, 2025, **115**(2), 1–22, DOI: 10.1007/s11103-025-01565-x%0A.

87 X. Ma, J. Geiser-Lee, Y. Deng and A. Kolmakov, Interactions between engineered nanoparticles (ENPs) and plants: Phytotoxicity, uptake and accumulation, *Sci. Total Environ.*, 2010, **408**(16), 3053–3061, DOI: 10.1016/j.scitotenv.2010.03.031.

88 P. D. Kolokathis, D. Zouraris, N. K. Sidiropoulos, A. Tsoumanis, G. Melagraki and I. Lynch, *et al.*, NanoTube Construct: A web tool for the digital construction of nanotubes of single-layer materials and the calculation of their atomistic descriptors powered by Enalos Cloud Platform, *Comput. Struct. Biotechnol. J.*, 2024, **25**, 230–242, DOI: 10.1016/j.csbj.2024.09.023.

89 L. Wang, J. Sun, L. Lin, Y. Fu, H. Alenius and K. Lindsey, *et al.*, Silver nanoparticles regulate Arabidopsis root growth by concentration-dependent modification of reactive oxygen species accumulation and cell division, *Ecotoxicol. Environ. Saf.*, 2020, **190**, 110072, DOI: 10.1016/j.ecoenv.2019.110072.

90 A. V. Ponce-Bobadilla, V. Schmitt, C. S. Maier, S. Mensing and S. Stodtmann, Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development, *Clin. Transl. Sci.*, 2024, **17**(11), 1–15.

91 J. M. Gendron and D. Staiger, New Horizons in Plant Photoperiodism, *Annu. Rev. Plant Biol.*, 2023, **74**, 481–509.

92 V. M. Roeber, T. Schmülling and A. Cortleven, The Photoperiod: Handling and Causing Stress in Plants, *Front. Plant Sci.*, 2022, **12**, 1–14.

93 M. Osnato, I. Cota, P. Nebhnani, U. Cereijo and S. Pelaz, Photoperiod Control of Plant Growth: Flowering Time Genes Beyond Flowering, *Front. Plant Sci.*, 2022, **12**, 1–20.

94 M. G. Lefsrud and D. A. Kopsell, Biomass production and pigment accumulation in kale grown under different radiation cycles in a controlled environment, *Hortscience*, 2006, **41**(6), 1412–1415.

95 A. Gogos, J. Moll, F. Klingenfuss, M. van der Heijden, F. Irin and M. J. Green, *et al.*, Vertical transport and plant uptake of nanoparticles in a soil mesocosm experiment, *J. Nanobiotechnol.*, 2016, **14**(1), 40, DOI: 10.1186/s12951-016-0191-z.

96 Y. Ge, J. H. Priester, L. C. Van De Werfhorst, S. L. Walker, R. M. Nisbet and Y. J. An, *et al.*, Soybean plants modify metal oxide nanoparticle effects on soil bacterial communities, *Environ. Sci. Technol.*, 2014, **48**(22), 13489–13496.

97 D.-H. Jung, H. S. Kim, C. Jhin, H.-J. Kim and S. H. Park, Time-serial analysis of deep neural network models for prediction of climatic conditions inside a greenhouse, *Comput. Electron. Agric.*, 2020, **173**, 105402, DOI: 10.1016/j.compag.2020.105402.

98 M. Mazaheri-Tirani, S. Dayani and M. I. Mobarakeh, Application of machine learning algorithms for predicting the life-long physiological effects of zinc oxide Micro/Nano particles on Carum copticum, *BMC Plant Biol.*, 2024, **24**(1), 970, DOI: 10.1186/s12870-024-05662-9.

