Supporting Information

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,[1,2*] Aikaterini Theodori,[1] Anastasios G. Papadiamantis,[2,3] Andreas Tsoumanis,[2,3] Dimitrios Zouraris,[2,3] Maria Antoniou,[2,3] Nikoletta-Maria Koutroumpa,[3] Georgia Melagraki,[4] Iseult Lynch,[5] Antreas Afantitis[1,2,3,6*]

[1] NovaMechanics MIKE, Piraeus 18545, Greece

[2] Entelos Institute, Nicosia 2102, Cyprus

[3] NovaMechanics Ltd, Nicosia 1070, Cyprus

[4] Division of Physical Sciences and Applications, Hellenic Military Academy, Vari 16672, Greece

[5] School of Geography, Earth and Environmental Sciences, University of Birmingham, B15 2TT Birmingham, UK

[6]Department of Pharmacy, Frederick University, Nicosia 1036, Cyprus

Supplementary tables

Table S1*. List of features included in the original Deng et al.[1] "Length" dataset.*

Feature	Description
Carbide	Number of carbon atoms
Metal	Number of metal atoms
MC	Macromolecular compound
Oxide	Number of oxygen atoms
Com1	Component 1 [A_r of Com1] where A_r is the relative atomic mass
Com2	Component 2 [A_r of Com2]
Dim0	Granular
Dim1	One-dimensional
Dim2	Two-dimensional
Hollow	Hollow
sizeTEM	Particle size measured by transmission electron microscopy [nm]
Purity	NP purity [%]
sizeDLS	Particle size measured by dynamic light scattering [nm]
Zeta	Zeta potential [mV]
Species	Plant species (cucumber, bean, wheat, rice, tomato, maize)
EP	NP exposure pathway (seeds, root, leaf)
MT	Measured tissue (root, shoot, plant)
Age	Exposure age [day]
GS	Growth stage of plant (germination, seedling, vegetative)

Cultured	Cultivation method (medium, hydroponic, soil)
Category	Plant carbon fixation metabolic pathway (C_3 or C_4 photosynthetic process)
TC	Total content (dose) [mg]
Duration	The time elapsed from the exposure of the plant to NPs to the measurement [day]
Photoperiod	Hours of plant exposure to light per day [hours/day]
Illumination	Illumination intensity [$\mu mol \cdot m^{-2}s^{-1}$]
Humidity	Humidity [%]
DT	Daytime temperature [°C]
NT	Night temperature [°C]

Table S2. List of corrections to the original Deng et al.[1,2] dataset.

Original dataset rows	Corrections	Referenced study
2-7	Positive labels were normalised based on the control sample.	Cui *et al*. (2014)[3]
13,17	Positive labels were normalised based on the control sample.	Konate *et al*. (2018)[4]
18-29	Rows removed because the NP size was not clearly mentioned.	Gopalakrishnan Nair *et al*. (2015)[5]
30-37	Rows removed because the total NP concentration in mg/L could not be calculated.	Abdel Latef *et al*. (2018)[6]
38-51	Illumination intensity values for medium cultivations were corrected to 500 µmol/m^2s.	Feng *et al*. (2019)[7]
45-51	"Cultured" column correction. The root length was measured after medium cultivation instead of hydroponic cultivation.	Feng *et al*. (2019)[7]
52-58	Positive labels were normalized based on the control sample. Negative labels were re-calculated for consistency.	Feng *et al*. (2019)[7]
67-68, 71-72, 76-77, 80-81	Rows removed because the NP size was not clearly mentioned in the referenced study.	Mahawar *et al*. (2018)[8]
65-66	Rows removed because atomistic descriptors could not be calculated for large Fe_2O_3 NPs.	Mahawar *et al*. (2018)[8]
74-75	Rows removed because atomistic descriptors could not be calculated for large Fe_2O_3 NPs.	Mahawar *et al*. (2018)[8]
82-90	Positive labels were normalized based on the control sample. Negative labels were re-calculated for consistency.	Cui *et al*. (2014)[3]
91-108	Rows removed because the total NP concentration in mg/L could not be calculated.	Wang *et al*. (2019)[9]
109-110	Rows removed because the NP size was not clearly mentioned in the referenced study.	Zuo *et al*. (2017)[10]
111-113	Rows removed because the total NP concentration in mg/L could not be calculated.	Du *et al*. (2015)[11]
114-117	Rows removed because the NP size was not clearly mentioned in the referenced study.	Wang *et al*. (2020)[12]
118-119	Rows removed because the NP shape was not clearly mentioned in the referenced study.	Iftikhar *et al*. (2020)[13]
120-129	Rows removed because the NP shape was not clearly mentioned in the referenced study.	Hussain *et al*. (2018)[14]
130-133	Rows removed because the NP shape was not clearly mentioned in the referenced study.	Khan *et al*. (2019)[15]
134-137	Rows removed because the NP shape was not clearly mentioned in the referenced study.	Wang *et al*. (2020)[16]
138-140	Rows removed because the total NP concentration in mg/L could not be calculated.	Li *et al*. (2020)[17]
141-146	Rows removed because the NP size was not clearly	Adhikari *et al*. (2015)[18]

	mentioned in the referenced study.	
148-149	Positive labels were normalized based on the control sample.	Lian *et al*. (2020)[19]
150-151, 155-156	Rows removed because the total NP concentration in mg/L could not be calculated for foliar NP treatments.	Lian *et al*. (2020)[19]
157-160	Rows removed because the total NP concentration in mg/L could not be calculated.	Toqeer *et al*. (2020)[20]
161-163	Rows removed because the total NP concentration in mg/L could not be calculated.	Yan *et al*. (2020)[21]
165-167	Positive labels were normalized based on the control sample.	Rizwan *et al*. (2019)[22]
175-181	Positive labels were normalized based on the control sample. Negative labels were re-calculated for consistency.	Feng *et al*. (2019)[7]
189-210	Rows removed because atomistic descriptors could not be calculated for large carbon sheets and tubes.	López-Vargas *et al*. (2020)[23]
211-215	Rows removed because the total NP concentration in mg/L could not be calculated.	Zadeh *et al*. (2019)[24]
216-231	Rows removed because the data could not be matched to any referenced study.	No referenced study found
233-234	Positive labels were normalized based on the control sample.	Wang *et al*. (2020)[25]
239-248	Positive labels were normalized based on the control sample. Negative labels were re-calculated for consistency.	Abbas *et al*. (2019)[26]
249-260	sizeTEM and purity of NPs were corrected to 30 nm and 97.5% respectively.	Liu *et al*. (2018)[27]
261-268	Mean illumination intensity, mean humidity and daytime and nighttime temperatures were corrected to 305 µmol/m^2s, 65% and 20^{o}C and 28^{o}C respectively.	Ji *et al*. (2017)[28]
269-272	Rows removed because atomistic descriptors could not be calculated for polymeric NPs.	Zhou *et al*. (2021)[29]
273-277	Rows removed because atomistic descriptors could not be calculated for large polymeric NPs.	Li *et al*. (2020)[30]
278-282	Rows removed because atomistic descriptors could not be calculated for large polymeric NPs.	Li *et al*. (2020)[31]
283-294	Rows removed because atomistic descriptors could not be calculated for large polymeric NPs.	Gong *et al*. (2021)[32]
295-299	Rows removed because atomistic descriptors could not be calculated for large polymeric NPs.	Lian *et al*. (2020)[33]
	7 rows were added to include shoot measurements.	Wang *et al*. (2020)[25]
	5 rows were added to include shoot measurements.	Gopalakrishnan Nair *et al*. (2015)[34]

Plot digitisation error analysis

In order to normalise the (root, shoot, overall plant) length labels relative to the control value, it was required to extract the actual experimental measurement values from the bar plots presented in the referenced papers, whenever these were not clearly referred in the text. Each plot was copied with its original aspect ratio and image resolution preserved: Screenshots of plots were taken directly from PDF files, avoiding any external resizing or compression artifacts and then imported into the WebPlotDigitizer[35] tool. As an initial step, two known, distinct points on the y-axis range were manually selected, and their values were entered into the tool to calibrate the bar plot. Subsequently, a point on the top of each bar was manually identified, from which the tool computed the corresponding numerical value.

During this process, both human systematic error and low image quality were recognised as potential sources of inaccuracy in the extracted numerical length values. To minimise this error, all plots were digitised by the same analyst, and a quality control analysis was conducted. To do so, plots for which the original numerical values were explicitly reported in the referenced articles were selected for digitisation, allowing a direct comparison between the extracted and published values. Although numerical values were not provided for the length parameter, similarly styled plots from the same studies depicting experimental parameters not included in the final dataset were used for this comparison. As these plots originate from the same studies (featuring similar aspect ratios, bar dimensions, and image resolutions) and were subjected to an identical digitisation procedure, including the same axis calibration method, they were considered appropriate proxies for assessing the digitisation process error.

Three examples of differently designed bar plots (different axis scaling, bar width and colour, see Figures S1-S3) were analysed, and their extracted, digitised values were compared to the referenced ones within the text. The relative absolute errors (RAE, Eq. 1) were then calculated as reported in Tables S3-S5. In all cases, the RAE was less than 1% which confirms that the plot digitisation method is suitable for extracting numerical values to be used in model development.

$$\%RAE = \frac{|x_{reported} - x_{digitized}|}{x_{reported}} \cdot 100 \text{ [1]}$$

Example 1. Digitisation of plots from Cui *et al*. (2014)[3]

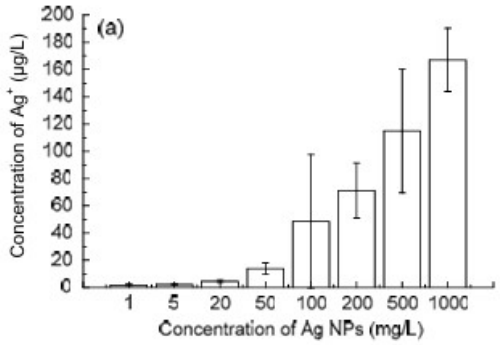

Figure S1. *Free ion concentrations in exposure solutions of Ag NPs after 5-d treatment of Ag NPs at the germination stage of cucumber. Figure 4(a) in the study of Cui et al.*[3]

Table S3. *Reported and Digitiser values comparison for Figure S1 and calculated %RAE.*

Concentration of Ag NPs (mg/L)	Reported value	Digitiser value	%RAE
1	2	2.016	0.806
1000	167	167.742	0.444

Example 2. Digitisation of plots from Abbas *et al*. (2019)[26]

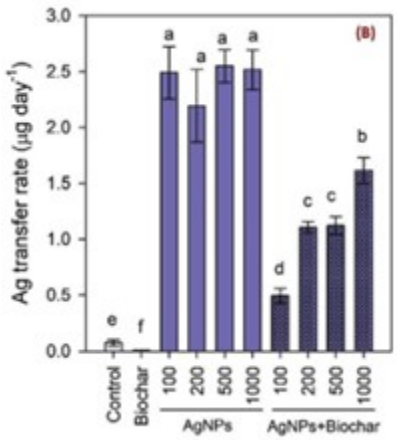

Figure S2. *Ag transfer rate (µg day*$^{-1}$*). Figure 4(B) in the study of Abbas et al.*[26]

Table S4. *Reported and Digitiser values comparison for Figure S2 and calculated %RAE.*

AgNPs+Biochar	Reported value	Digitiser value	%RAE
1000	1.62	1.616	0.225

Example 3. Digitisation of plots from Feng *et al*. (2019)[7]

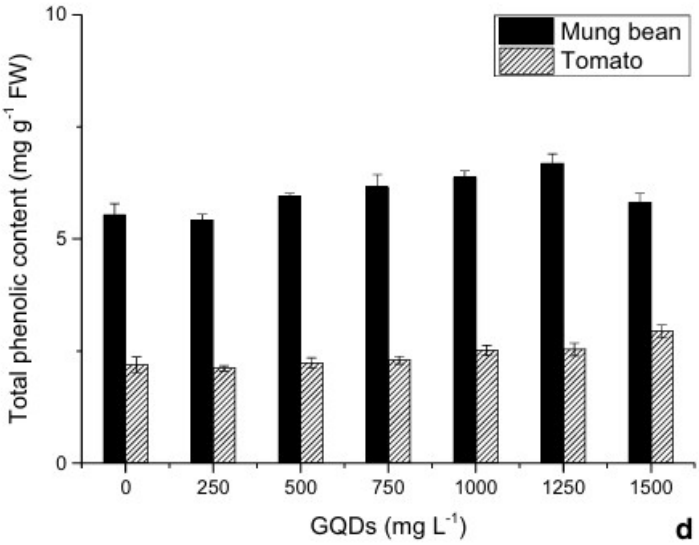

Figure S3. *Effect of GQDs total phenolic contents in mung bean and tomato seedlings after 2 weeks of culture in hydroponic medium containing different concentrations of GQDs. Figure 5d in the study of Feng et al.*[7]

Table S5. *Reported and Digitiser values comparison for Figure S3 and calculated %RAE.*

GQDs (mg L^{-1})	Reported value	Digitiser value	%RAE
1250 Mung Bean	6.69	6.686	0.061
1500 Tomato	2.94	2.948	0.271

Supplementary figures

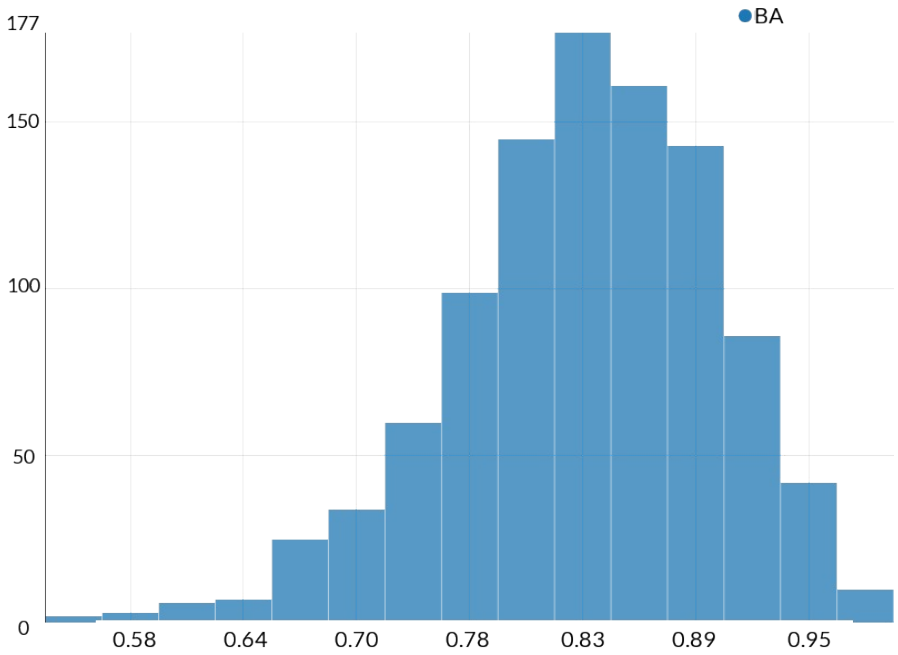

Figure S4: *Histogram of BA values obtained from 1000 bootstrap resamplings of the test set predictions using the XGBoost model.*

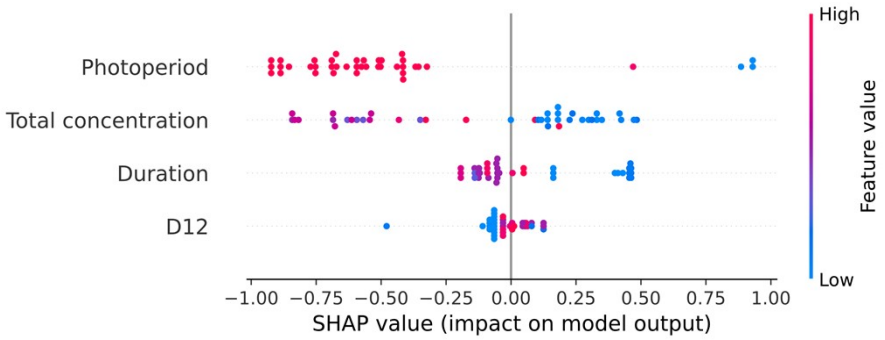

Figure S5: *SHAP analysis summary plot for the XGBoost model applied on the test sample. Only the features with numerical values are included in the analysis.*

References

1. Deng, P. *et al.* Development potential of nanoenabled agriculture projected using machine learning. *Proc. Natl. Acad. Sci. U. S. A.* **120**, (2023).

2. dp1999nku. All-datasets. https://github.com/dp1999nku/All-datasets (2023).

3. Cui, D. *et al.* Phytotoxicity of silver nanoparticles to cucumber (Cucumis sativus) and wheat (Triticum aestivum). *J. Zhejiang Univ. Sci. A* **15**, 662–670 (2014).

4. Konate, A. *et al.* Comparative effects of nano and bulk-Fe3O4 on the growth of cucumber (Cucumis sativus). *Ecotoxicol. Environ. Saf.* **165**, 547–554 (2018).

5. Gopalakrishnan Nair, P. M. & Chung, I. M. Biochemical, anatomical and molecular level changes in cucumber (Cucumis sativus) seedlings exposed to copper oxide nanoparticles. *Biologia (Bratisl).* **70**, 1575–1585 (2015).

6. Abdel Latef, A. A. H., Srivastava, A. K., El-sadek, M. S. A., Kordrostami, M. & Tran, L. P. Titanium Dioxide Nanoparticles Improve Growth and Enhance Tolerance of Broad Bean Plants under Saline Soil Conditions. *L. Degrad. Dev.* **29**, 1065–1073 (2018).

7. Feng, P. *et al.* Graphene quantum dots-induced physiological and biochemical responses in mung bean and tomato seedlings. *Rev. Bras. Bot.* **42**, 29–41 (2019).

8. Mahawar, H. *et al.* Deciphering the mode of interactions of nanoparticles with mung bean (Vigna radiata L.). *Isr. J. Plant Sci.* **65**, 74–82 (2018).

9. Wang, Y. *et al.* Effect of metal oxide nanoparticles on amino acids in wheat grains (Triticum aestivum)in a life cycle study. *J. Environ. Manage.* **241**, 319–327 (2019).

10. Zuo, Z. *et al.* Melatonin improves the photosynthetic carbon assimilation and antioxidant capacity in wheat exposed to nano-zno stress. *Molecules* **22**, (2017).

11. Du, W. *et al.* Physiological and Biochemical Changes Imposed by CeO2 Nanoparticles on Wheat: A Life Cycle Field Study. *Environ. Sci. Technol.* **49**, 11884–11893 (2015).

12. Wang, Z. *et al.* Nano-ZnO priming induces salt tolerance by promoting photosynthetic carbon assimilation in wheat. *Arch. Agron. Soil Sci.* **66**, 1259–1273 (2020).

13. Iftikhar, A. *et al.* Effect of gibberellic acid on growth, biomass, and antioxidant defense system of wheat (Triticum aestivum L.) under cerium oxide nanoparticle stress. *Environ. Sci. Pollut. Res.* **27**, 33809–33820 (2020).

14. Hussain, A. *et al.* Zinc oxide nanoparticles alter the wheat physiological response and reduce the cadmium uptake by plants. *Environ. Pollut.* **242**, 1518–1526 (2018).

15. Khan, Z. S. *et al.* The accumulation of cadmium in wheat (Triticum aestivum) as influenced by zinc oxide nanoparticles and soil moisture conditions. *Environ. Sci. Pollut. Res.* **26**, 19859–19870 (2019).

16. Wang, S. *et al.* Silver nanoparticles with different concentrations and particle sizes affect the functional traits of wheat. *Biol. Plant.* **64**, 1–8 (2020).

17. Li, P. *et al.* Insight into the interaction between Fe-based nanomaterials and maize (Zea mays) plants at metabolic level. *Sci. Total Environ.* **738**, 139795 (2020).

18. Adhikari, T., Kundu, S., Biswas, A. K., Tarafdar, J. C. & Subba Rao, A. Characterization of Zinc Oxide Nano Particles and Their Effect on Growth of Maize (Zea mays L.) Plant. *J. Plant Nutr.* **38**, 1505–1515 (2015).

19. Lian, J. *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* **239**, 124794 (2020).

20. Toqeer, I. *et al.* Synthesis and application of controlled size copper oxide nanoparticles for improving

biochemical and growth parameters of maize seedling. *J. Plant Nutr.* **0**, 2622–2632 (2020).

21. Yan, L., Li, P., Zhao, X., Ji, R. & Zhao, L. Physiological and metabolic responses of maize (Zea mays) plants to Fe3O4 nanoparticles. *Sci. Total Environ.* **718**, 137400 (2020).

22. Rizwan, M. *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.* **248**, 358–367 (2019).

23. López-Vargas, E. R. *et al.* Seed priming with carbon nanomaterials to modify the germination, growth, and antioxidant status of tomato seedlings. *Agronomy* **10**, 1–22 (2020).

24. Zadeh, R. R., Arvin, S. M. J., Jamei, R., Mozaffari, H. & Reza Nejhad, F. Response of tomato plants to interaction effects of magnetic (Fe3O4) nanoparticles and cadmium stress. *J. Plant Interact.* **14**, 474–481 (2019).

25. Wang, W. *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.* **104**, 770–777 (2020).

26. Abbas, Q. *et al.* 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.* **250**, 728–736 (2019).

27. Liu, J., Dhungana, B. & Cobb, G. P. Copper oxide nanoparticles and arsenic interact to alter seedling growth of rice (Oryza sativa japonica). *Chemosphere* **206**, 330–337 (2018).

28. Ji, Y. *et al.* Jointed toxicity of TiO2 NPs and Cd to rice seedlings: NPs alleviated Cd toxicity and Cd promoted NPs uptake. *Plant Physiol. Biochem.* **110**, 82–93 (2017).

29. Zhou, C. Q. *et al.* Response of rice (Oryza sativa L.) roots to nanoplastic treatment at seedling stage. *J. Hazard. Mater.* **401**, 123412 (2021).

30. Li, Z., Li, R., Li, Q., Zhou, J. & Wang, G. Physiological response of cucumber (Cucumis sativus L.) leaves to polystyrene nanoplastics pollution. *Chemosphere* **255**, 127041 (2020).

31. Li, Z., Li, Q., Li, R., Zhou, J. & Wang, G. The distribution and impact of polystyrene nanoplastics on cucumber plants. *Environ. Sci. Pollut. Res.* **28**, 16042–16053 (2021).

32. Gong, W. *et al.* Species-dependent response of food crops to polystyrene nanoplastics and microplastics. *Sci. Total Environ.* **796**, (2021).

33. Lian, J. *et al.* Impact of polystyrene nanoplastics (PSNPs) on seed germination and seedling growth of wheat (Triticum aestivum L.). *J. Hazard. Mater.* **385**, 121620 (2020).

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

35. Rohatgi, A. WebPlotDigitizer. at https://automeris.io.