Electronic Supplementary Material (ESI) for Nanoscale.

Supplementary Information

Integrating machine learning interpretation methods for investigating nanoparticle uptake during seed priming and its biological effects

Hengjie Yu, Zhilin Zhao, Da Liu and Fang Cheng*

College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310058, P.R. China

* Corresponding author: Fang Cheng

E-mail: fcheng@zju.edu.cn

Tel.: +86 571 88982713

Table of contents

Methods

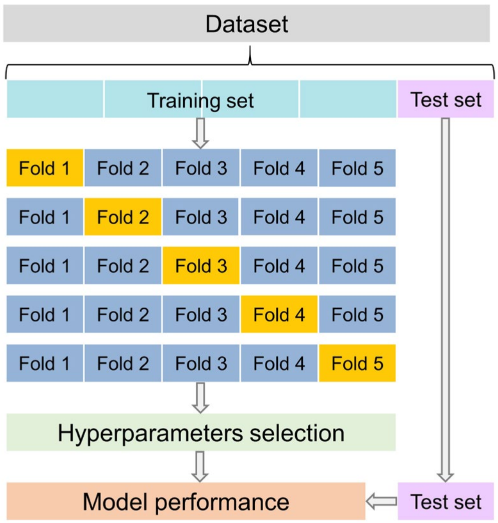


Fig. S1. The workflow of random forest and decision tree modeling for seed Zn content.

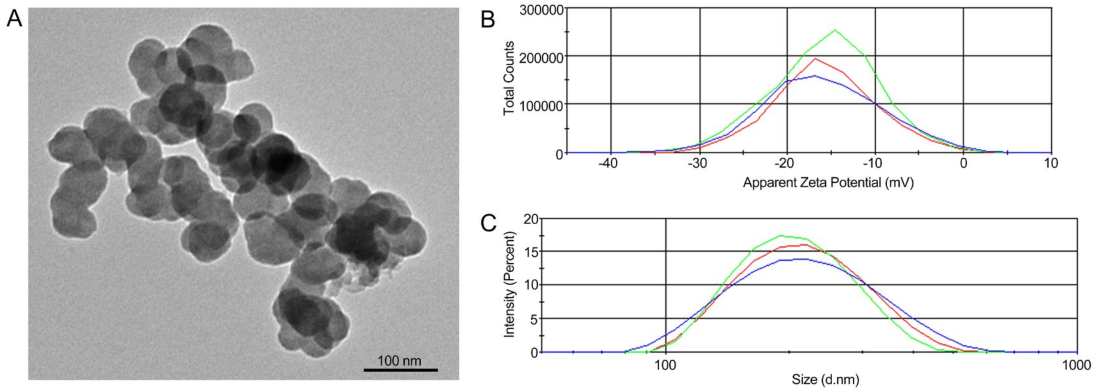


Fig. S2. The characterization of ZnO nanoparticles (50±10 nm, Shanghai Aladdin Reagent Co., Ltd). (A) TEM image. (B) Zeta potential. (C) Hydrodynamic diameter. The green, red, and blue lines in (B) and (C) represent three replicate measurements.

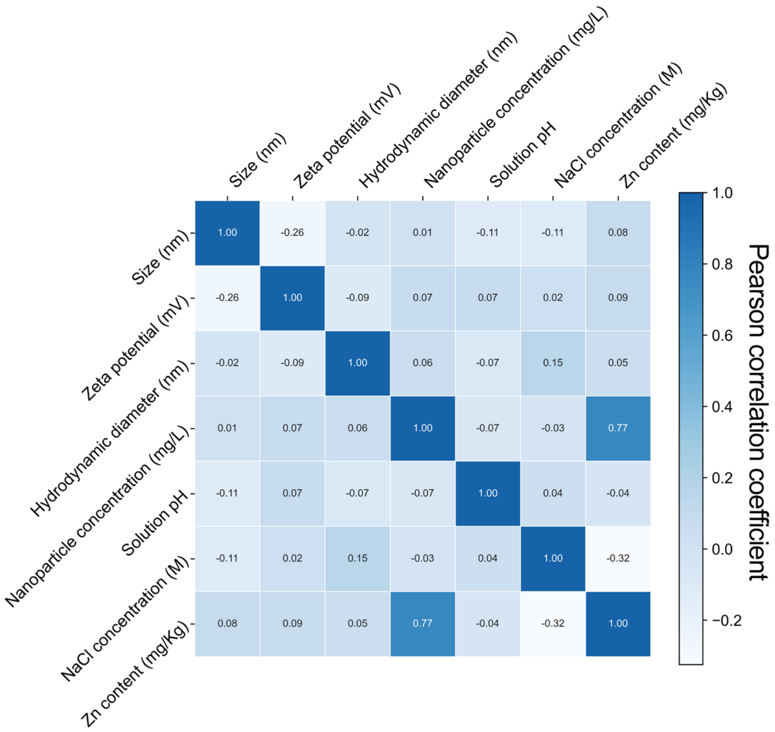


Fig. S3. The heatmap of the Pearson correlation coefficient among numerical factors and seed Zn content in the dataset.

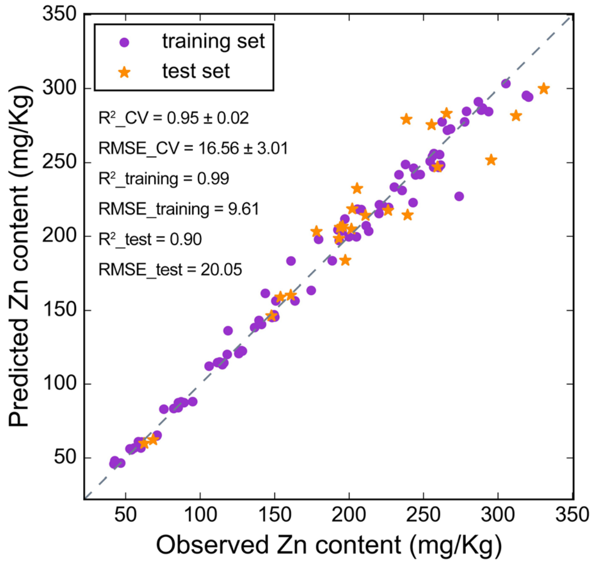


Fig. S4. The performance of the established RF model for seed Zn content. The performance results of 5-fold CV on the training set are presented as mean ± SD (n=5).

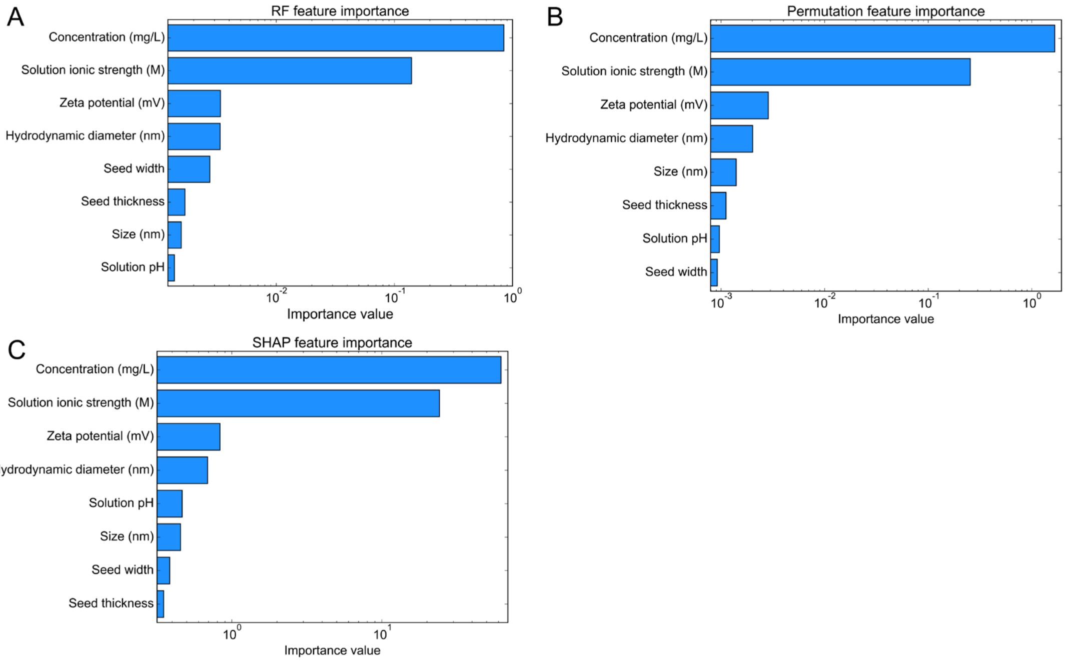


Fig. S5. The absolute importance values of (A) RF feature importance, (B) permutation feature importance, and (C) SHAP feature importance in the established RF model for seed Zn content.

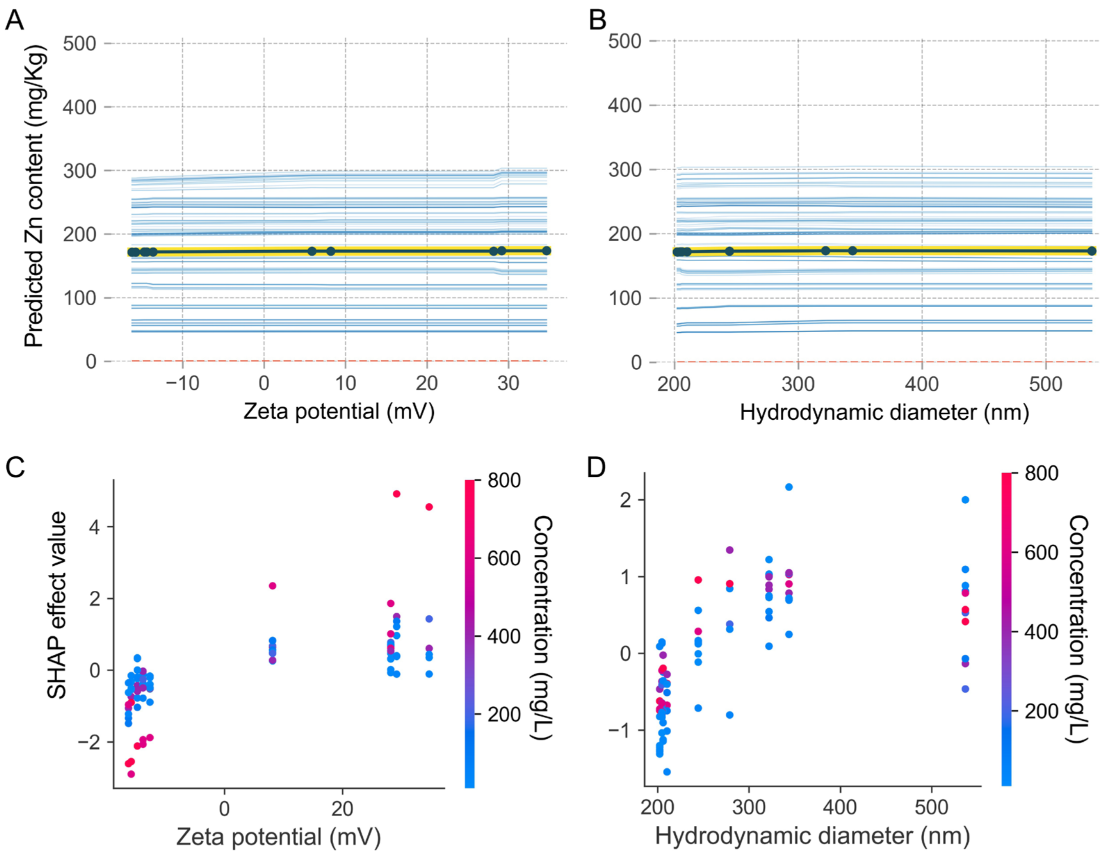
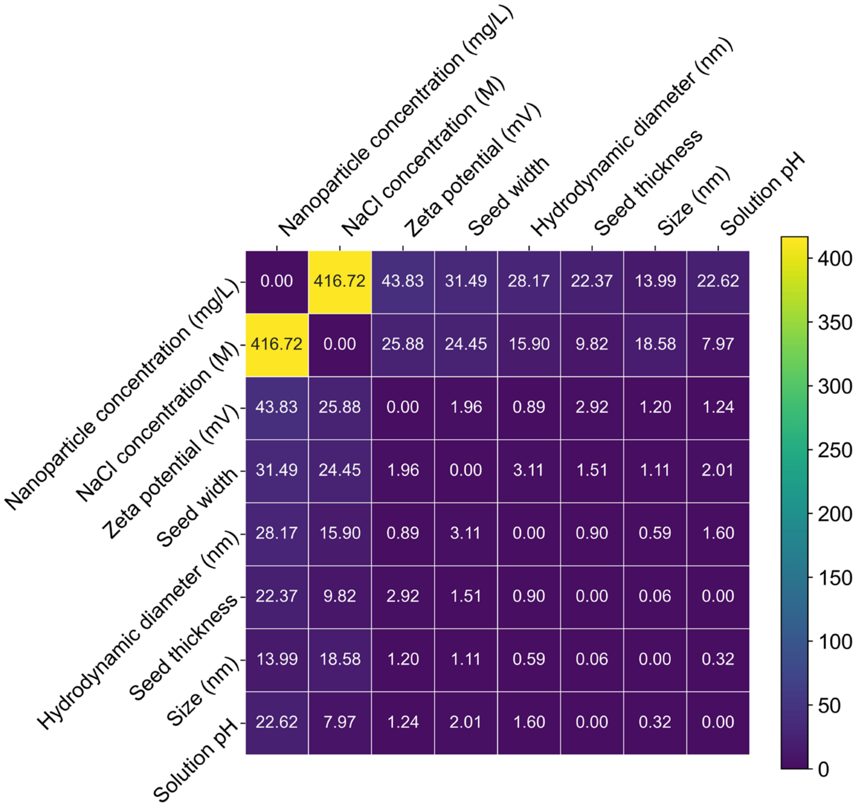


Fig. S6. Feature effects of zeta potential and hydrodynamic diameter in the established RF model for seed Zn content. PDP of feature effects of (A) zeta potential and (B) hydrodynamic diameter. SHAP dependence plots of (C) zeta potential and (D) hydrodynamic diameter, colored by one factor that has the most obvious interactions with it.



	Nanoparticle concentration (mg/L)	NaCl concentration (M)	Zeta potential (mV)	Seed width	Hydrodynamic diameter (nm)	Seed thickness	Size (nm)	Solution pH
Nanoparticle concentration (mg/L)	0.00	416.72	43.83	31.49	28.17	22.37	13.99	22.62
NaCl concentration (M)	416.72	0.00	25.88	24.45	15.90	9.82	18.58	7.97
Zeta potential (mV)	43.83	25.88	0.00	1.96	0.89	2.92	1.20	1.24
Seed width	31.49	24.45	1.96	0.00	3.11	1.51	1.11	2.01
Hydrodynamic diameter (nm)	28.17	15.90	0.89	3.11	0.00	0.90	0.59	1.60
Seed thickness	22.37	9.82	2.92	1.51	0.90	0.00	0.06	0.00
Size (nm)	13.99	18.58	1.20	1.11	0.59	0.06	0.00	0.32
Solution pH	22.62	7.97	1.24	2.01	1.60	0.00	0.32	0.00

Fig. S7. The heatmap of the sum of SHAP interaction values of all instances for every two features in the established model for seed Zn content. The sequence of features is arranged according to the sum of SHAP interaction values with other features.

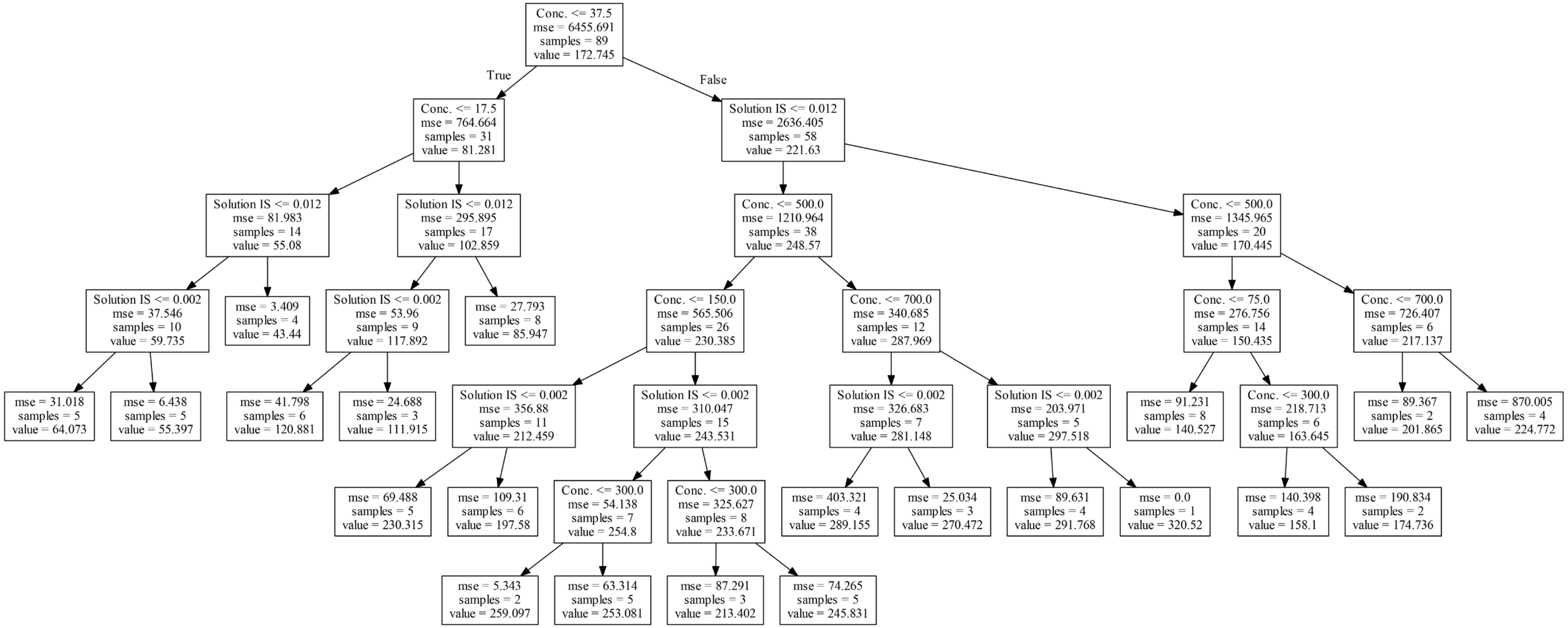
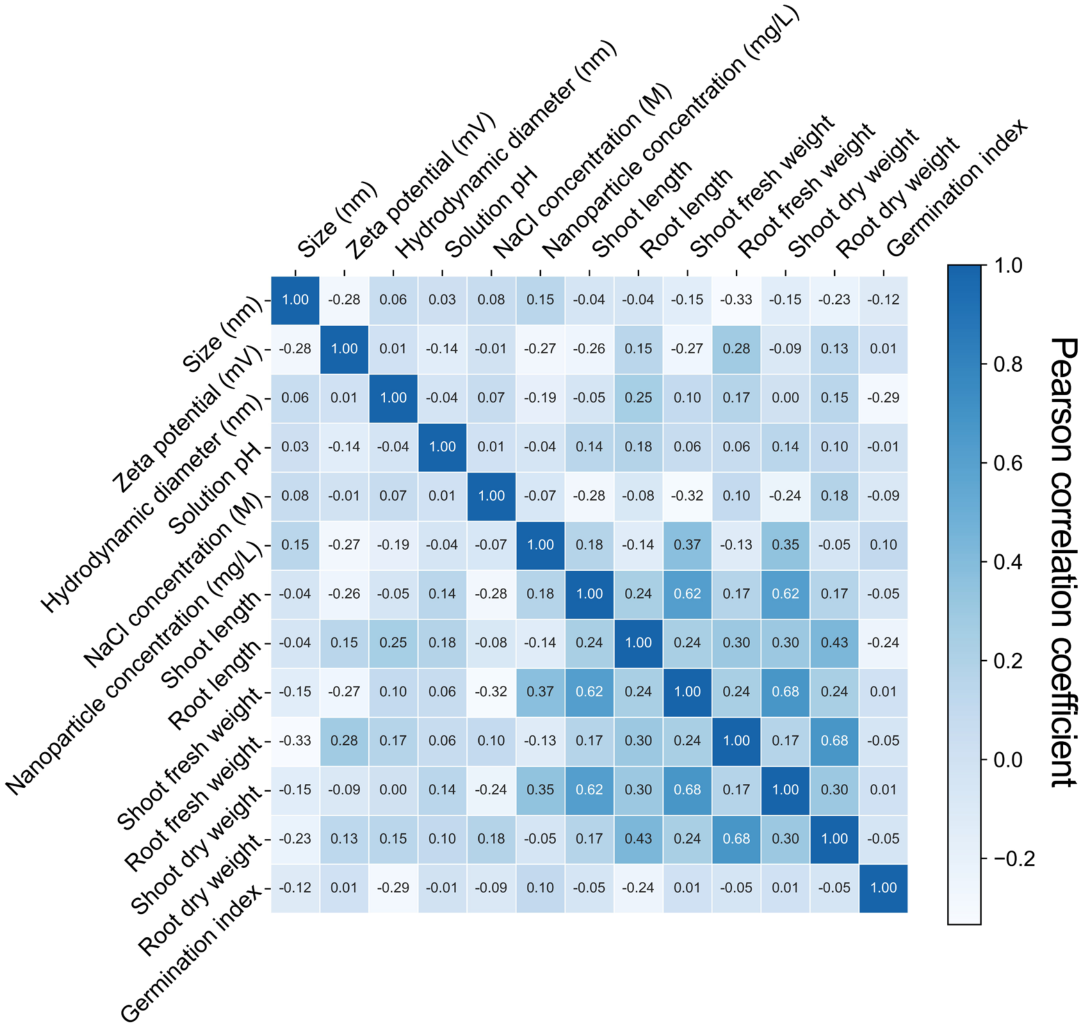


Fig. S8. The visualization of the decision tree model with 21 nodes for seed Zn content. NP, nanoparticle. Conc., concentration. Mse, mean squared error.



	Size (nm)	Zeta potential (mV)	Hydrodynamic diameter (nm)	Solution pH	NaCl concentration (M)	Nanoparticle concentration (mg/L)	Shoot length	Root length	Shoot fresh weight	Root fresh weight	Shoot dry weight	Root dry weight	Germination index
Size (nm)	1.00	-0.28	0.06	0.03	0.08	0.15	-0.04	-0.04	-0.15	-0.33	-0.15	-0.23	-0.12
Zeta potential (mV)	-0.28	1.00	0.01	-0.14	-0.01	-0.27	-0.26	0.15	-0.27	0.28	-0.09	0.13	0.01
Hydrodynamic diameter (nm)	0.06	0.01	1.00	-0.04	0.07	-0.19	-0.05	0.25	0.10	0.17	0.00	0.15	-0.29
Solution pH	0.03	-0.14	-0.04	1.00	0.01	-0.04	0.14	0.18	0.06	0.06	0.14	0.10	-0.01
NaCl concentration (M)	0.08	-0.01	0.07	0.01	1.00	-0.07	-0.28	-0.08	-0.32	0.10	-0.24	0.18	-0.09
Nanoparticle concentration (mg/L)	0.15	-0.27	-0.19	-0.04	-0.07	1.00	0.18	-0.14	0.37	-0.13	0.35	-0.05	0.10
Shoot length	-0.04	-0.26	-0.05	0.14	-0.28	0.18	1.00	0.24	0.62	0.17	0.62	0.17	-0.05
Root length	-0.04	0.15	0.25	0.18	-0.08	-0.14	0.24	1.00	0.24	0.30	0.30	0.43	-0.24
Shoot fresh weight	-0.15	-0.27	0.10	0.06	-0.32	0.37	0.62	0.24	1.00	0.24	0.68	0.24	0.01
Root fresh weight	-0.33	0.28	0.17	0.06	0.10	-0.13	0.17	0.30	0.24	1.00	0.17	0.68	-0.05
Shoot dry weight	-0.15	-0.09	0.00	0.14	-0.24	0.35	0.62	0.30	0.68	0.17	1.00	0.30	0.01
Root dry weight	-0.23	0.13	0.15	0.10	0.18	-0.05	0.17	0.43	0.24	0.68	0.30	1.00	-0.05
Germination index	-0.12	0.01	-0.29	-0.01	-0.09	0.10	-0.05	-0.24	0.01	-0.05	0.01	-0.05	1.00

Fig. S9. The Pearson correlation coefficient among numerical factors and physiological parameters of seed germination.

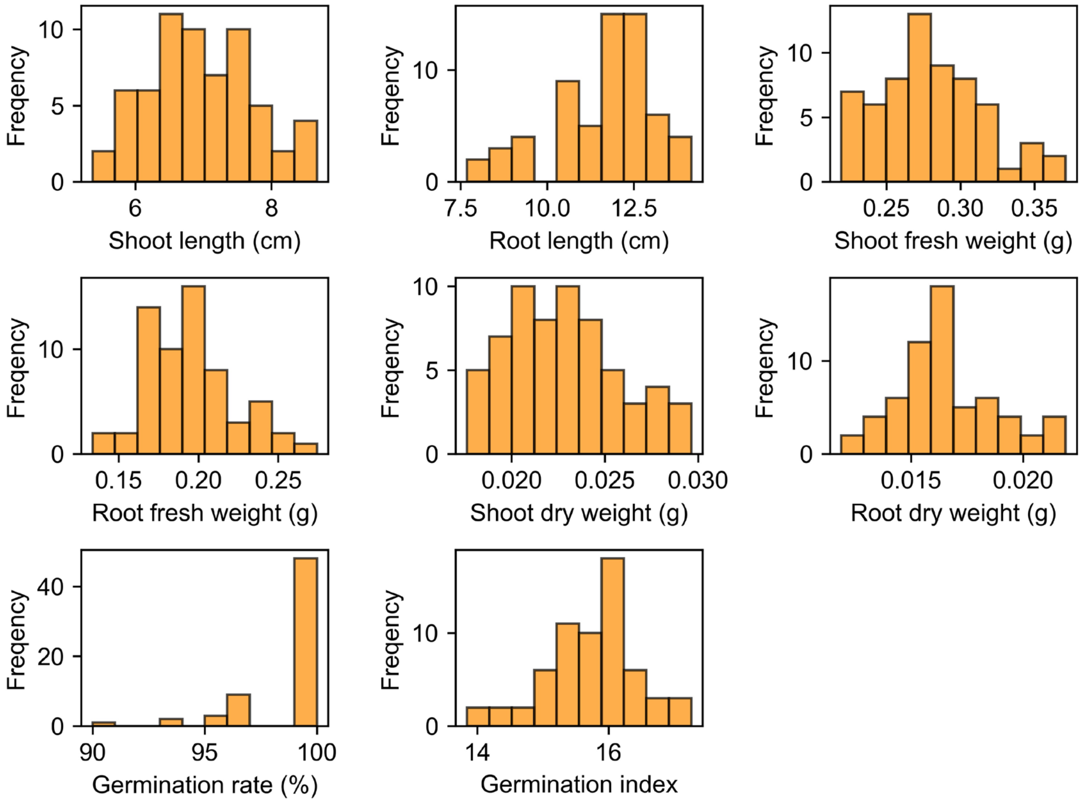


Fig. S10. The data distribution of eight physiological parameters of seed germination.

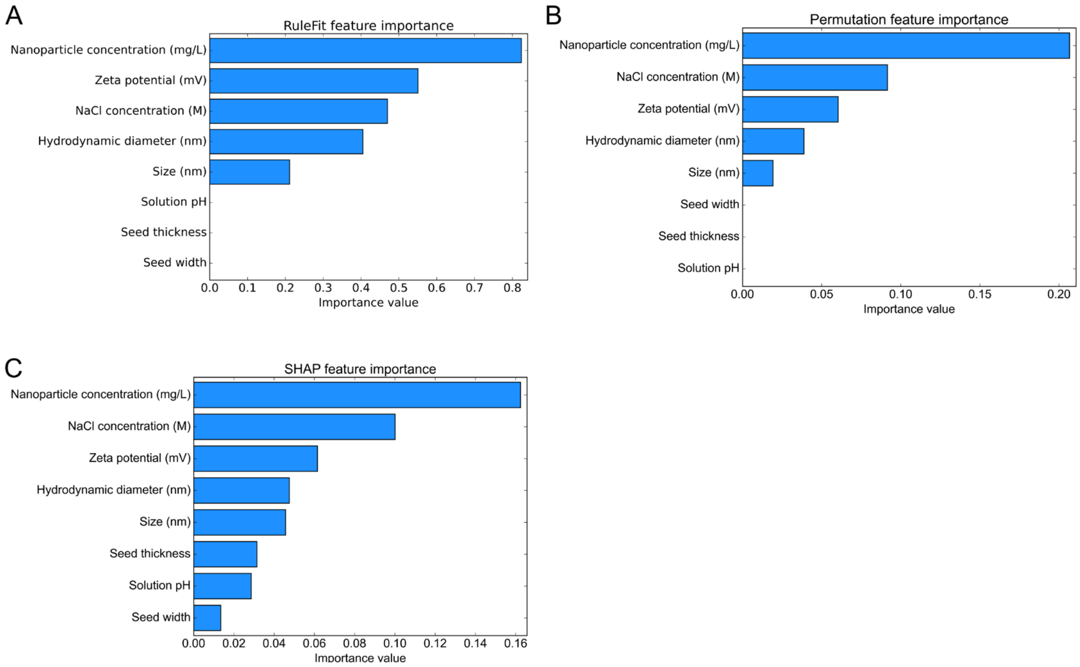


Fig. S11. The absolute importance values of (A) RuleFit feature importance, (B) permutation feature importance for the established RuleFit model, and (C) SHAP feature importance in the established RF model for shoot fresh weight.

Table S1. The determined hyperparameters of the random forest and decision tree models for seed Zn content.

	Random forest	**Decision tree**
max_depth	15	5
min_samples_leaf	2	3
min_samples_split	5	9
max_features	8	7
max_leaf_nodes	19	27
n_jobs	-1	\
n_estimators	100	\

Table S2. The versions of the main software used in this study.

Software/packages	**Version**
Python	3.8.10
scikit-learn	0.24.2
shap	0.39.0
PDPbox	0.2.0
imodels	1.2.5

Table S3. Overview of the datasets used for machine learning.

Feature/Target	**Range/Category**
Size	30, 50, 90 nm
Zeta potential	-16.23 ~ 34.67 mV
Hydrodynamic diameter	202.27 ~ 536.90 nm
Nanoparticle concentration	10, 25, 50, 100, 200, 400, 600, 800 mg/L
Solution pH	5.5, 6.5, 7.5
NaCl concentration	0, 0.005, 0.020 M
Seed thickness	flat, thick
Seed width	narrow, wide
Zn content	42.08 ~ 330.71 mg/Kg
Shoot length	5.37 ~ 8.67 cm
Root length	7.68 ~ 14.16 cm
Shoot fresh weight	0.22 ~ 0.37 g
Root fresh weight	0.13 ~ 0.27 g
Shoot dry weight	0.018 ~ 0.030 g
Root dry weight	0.012 ~ 0.022 g
Germination rate	90 ~ 100 %
Germination index	13.84 ~ 17.26

Methods

Nanosuspension volume for seed nanopriming

The seeds were classified into four categories: wide and thick, wide and flat, narrow and thick, and narrow and flat. The proportion of seed weight to suspension volume was 1:5 g/mL, and each treatment contained 15 seeds. Therefore, the required volumes of nanosuspensions for different categories of seeds were shown in Table S4.

Table S4. The required volumes of nanosuspensions for one centrifuge tube containing 15 seeds. The 100-grain weight of maize seeds is presented as mean ± standard deviation (SD) (n=3).

Seed category	100-grain weight of maize seeds (g)	Required nanosuspension volume for 15 seeds (mL)
wide and thick	48.79±1.21	36.60
wide and flat	44.10±0.02	33.08
narrow and thick	38.80±0.27	29.10
narrow and flat	35.94±0.39	26.96

Composition content of nanosuspensions

Therefore, 110 mL of nanosuspensions was enough for three replicates of seed nanopriming for every factor combination. The 2 M NaCl, 2000 mg/L nanosuspensions, and adjusted deionized (DI) water with pH of 5.5, 6.5, and 7.5 were prepared in advance. To prepare 110 mL of nanosuspensions with NaCl concentration of 0, 0.005, and 0.020 M, the needed volume of 2 M NaCl were 0, 0.275, and 1.1 mL respectively. The required volumes of 2000 mg/L nanosuspensions and adjusted DI water used for the preparation of 110 mL of nanosuspensions are shown in Table S5.

Table S5. The required volumes of 5% nanosuspensions and adjusted DI water used for the preparation of 110 mL of nanosuspensions.

Nanoparticle concentration（mg/L）	Required volume of 2000 mg/L nanosuspensions (mL)	Required volume of adjusted DI water (mL)		
		NaCl concentration = 0 M	NaCl concentration = 0.005 M	NaCl concentration = 0.020 M
10	0.55	109.45	109.175	108.35
25	1.375	108.625	108.35	107.525
50	2.75	107.25	106.975	106.15
100	5.5	104.5	104.225	103.4
200	11	99	98.725	97.9
400	22	88	87.725	86.9
600	33	77	76.725	75.9
800	44	66	65.725	64.9

For every factor combination, the DI water and 2 M NaCl in the required volumes were first added to a 200 mL Erlenmeyer flask, followed by adding 2000 mg/L nanosuspensions that had been dispersed by ultrasonic cell disruptor. Then the 110 mL of the mixture was shaken evenly.

Nitric acid digestion and sample preparation for Zn content measurement

In this study, the 25 mL hydrothermal synthesis autoclave reactor was used for nitric acid digestion. After heat treatment, the autoclave reactor was cooled to room temperature. The mixture in polytetrafluoroethylene (PTFE) liner was transferred into a 20 mL volumetric flask. The PTFE liner was washed 2-3 times with DI water, and the remaining mixture was transferred into the volumetric flask. The volumetric flask was filled to the line marked on the volumetric flask with DI water. The mixture in the volumetric flask was uniformly mixed. Then 1 mL of the diluted mixture was diluted with DI water to a constant volume of 10 mL for Zn measurement.

The used PTFE liner was first cleaned with soapy water and rinsed thoroughly, followed by soaking in 10% nitric acid for 48 h. Then the PTFE liner was washed and rinsed again. After drying and cooling, it could be used for the next treatment.

The measurement of the length of shoot and root by machine vision

Many seedlings were harvested at the same time, and the length and fresh weight would change with time. Therefore, measuring the length of shoot and root by machine vision can save time in the harvest stage and measure the length with the same standard. A scanner was used to capture the image of shoots and roots. The example of the root length measurement is shown in Fig. S12.

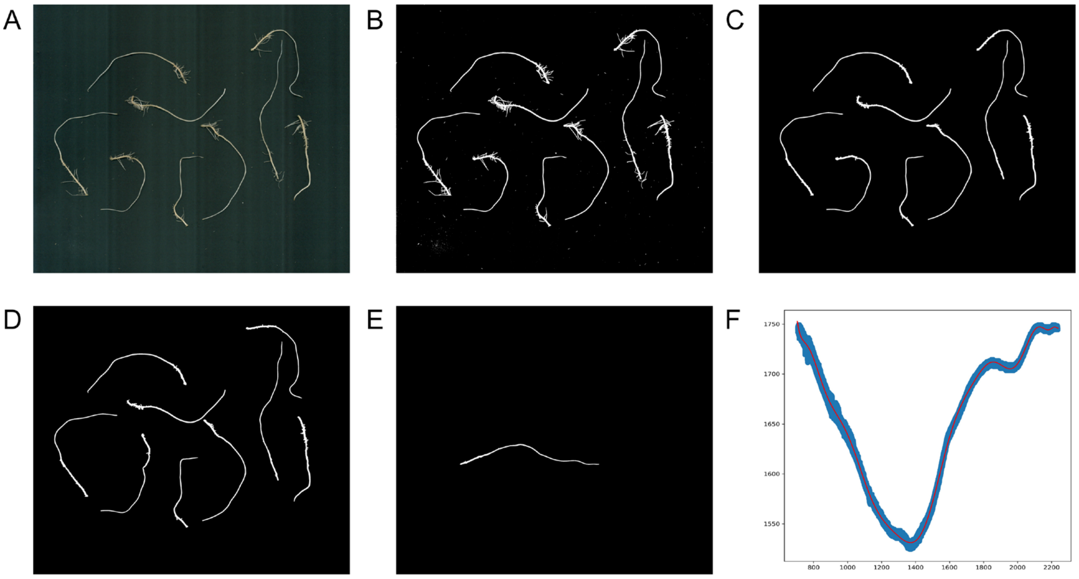


Fig. S12. The example of the root length measurement. (A) The root image obtained by a scanner. (B) The binary mask of the root image. (C) The binary mask without small masks and the mask of root hair. The mask of root hair was manually removed using

PhotoShop. (D) The expansion of root masks with the help of PhotoShop. (E) The rotated root mask for curve fitting. (F) The result of curve fitting. The length of the root is represented by the length of the curve.

The post hoc interpretation of machine learning

Post hoc interpretation was employed to explain the outputs of the established RF model from feature importance, feature effects, and feature interactions based on the training set. Therefore, less overfitting is important for avoiding features that are not predictive of the target on the test set to obtain high importance values. Screening for priority factors that determine nanoparticle uptake during seed priming by ZnO nanoparticles could provide fundamental insight into the mechanisms of nano-seed-solution interactions and avoid overly focusing on those unimportant factors. The visualization of feature effects of important factors can show how these features influence the predicted target. Moreover, the effects of different factors might be non-additive if one factor interacts with other factors. The explanation of interaction effects is therefore needed for interpretable machine learning.

The impurity-based feature importance is provided by the RF model, which is one of the reasons why RF models are so popular. However, the impurity-based importance is biased towards high cardinality features, which may inflate the importance of numerical features. Permutation importance and SHAP importance can mitigate this problem. Permutation importance measures the increase in the model prediction when a single feature value is randomly shuffled.[1] SHAP is a game theoretic method to explain the model output by assigning each feature an importance value for every prediction.[2, 3] SHAP importance is obtained by averaging the absolute Shapley values per feature across the data.

PDP shows the marginal contribution of different features on the model predictions while heterogeneous effects might be hidden due to the average marginal effects.[4] ICE disaggregates this average by highlighting the prediction changes for individual observations.[5] Therefore, PDP and ICE plots can be used in parallel to understand the overall trends of the established model and check problematic individual cases. However, the assumption of feature independence is the same problem for PDP and ICE plots, which leads to some invalid points with unrealistic factor combinations. SHAP solves the problem of feature dependence by explicitly modeling the conditionally expected prediction.[3]

PDP can visualize the combined effects of two factors in one interaction contour plot. However, PDP includes feature interactions and the main effects of these two factors, making it difficult to isolate interaction effects. SHAP provides a method, based on the Shapley interaction index from game theory, to capture interaction effects that can be interpreted as the difference of the SHAP value of one feature when another feature is present or not.

PDP can visualize the combined effects of two factors in one interaction contour plot. However, PDP includes feature interactions and the main effects of these two factors, making it difficult to isolate interaction effects. SHAP provides a method, based on the Shapley interaction index from game theory, to capture interaction effects that

can be interpreted as the difference of the SHAP value of one feature when another feature is present or not.

REFERENCES

1. Fisher, A., Rudin, C. and Dominici, F., *J. Mach. Learn. Res.*, 2019, **20**, 1-81.

2. Lundberg, S.M. and Lee, S.I., *Adv. Neural Inf. Process. Syst.*, 2017, 4766–4775.

3. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. and Lee, S., *Nat. Mach. Intell.*, 2020, **2**, 56-67.

4. Friedman, J.H., *Ann. Stat.*, 2001, **29**, 1189-1232.

5. Goldstein, A., Kapelner, A., Bleich, J. and Pitkin, E., *J. Comput. Graph. Stat.*, 2015, **24**, 44-65.