Supplementary Information

Single-particle multi-element fingerprinting (spMEF) using inductively coupled plasma time of flight mass spectrometry (ICP-TOFMS) to identify engineered nanoparticles against the elevated natural background in soils

Antonia Praetorius[a,b], Alexander Gundlach-Graham[c], Eli Goldberg[d], Willi Fabienke[a], Jana Navratilova[a], Andreas Gondikas[a,e], Ralf Kaegi[f], Detlef Günther[c], Thilo Hofmann[a,b*], Frank von der Kammer[a*]

[a]University of Vienna, Department of Environmental Geosciences and Environmental Science Research Network, Althanstr. 14, UZA II, 1090 Vienna, Austria

[b]University of Vienna, Research Platform Nano-Norms-Nature, Vienna, Austria

[c]ETH Zurich, Laboratory of Inorganic Chemistry, Vladimir-Prelog-Weg 1, 8093 Zurich, Switzerland

[d]ETH Zurich, Institute for Chemical and Bioengineering, Vladimir-Prelog-Weg 1, 8093 Zurich, Switzerland

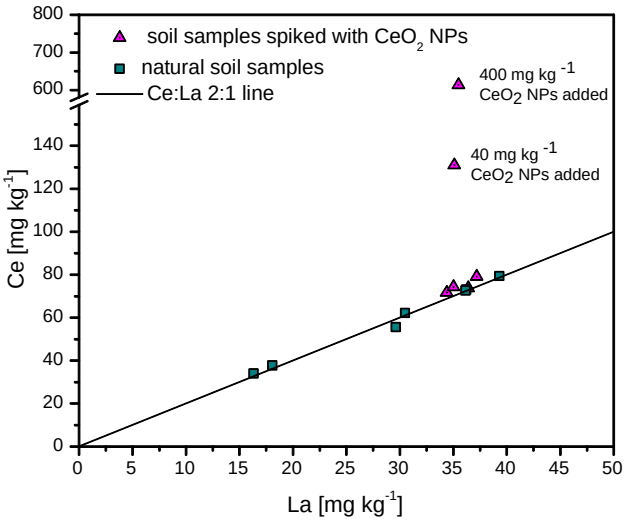
[e]University of Gothenburg, Department of Marine Sciences, Kristineberg 566, 45178 Fiskebäckskil, Sweden

[f]Eawag, Swiss Federal Institute of Aquatic Science and Technology, Überlandstr. 133, 8609 Dübendorf, Switzerland

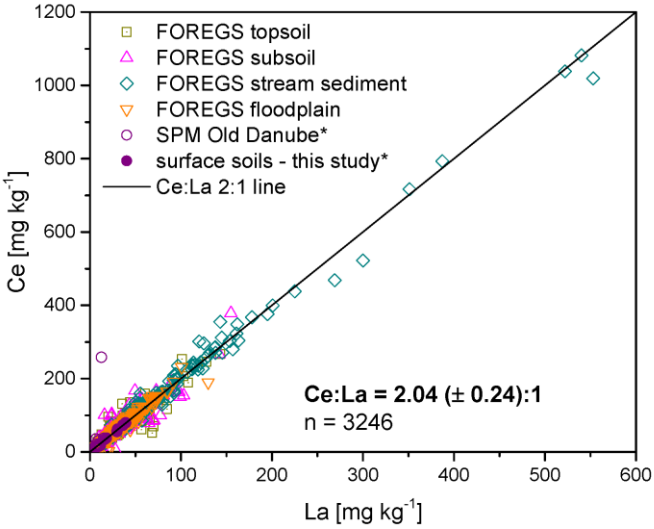
*email: thilo.hofmann@univie.ac.at, frank.kammer@univie.ac.at

1. Bulk ICP-MS measurements

Total Ce and La content in bulk soil samples and colloid extracts was determined by ICP-MS (7700 series ICP-MS, Agilent Technologies, CA, USA). Bulk soil samples were subjected to microwave digestion (Multiwave 3000, PerkinElmer, MA, USA) before ICP-MS measurement. Samples were diluted with Milli-Q water prior to ICP-MS analysis.



Supplementary Figure 1: Cerium (Ce) versus Lanthanum (La) concentrations (bulk) in different natural soil samples collected in Vienna (Austria) (green squares) and in the soils spiked with CeO_2 NPs (pink triangles).



Supplementary Figure 2: Cerium (Ce) versus Lanthanum (La) concentrations (bulk) in European soils, sediments and suspended particulate matter (SPM). Data obtained from the Geochemical Atlas of Europe, Forum of the European Geological Surveys (FOREGS) (Geochemical Atlas of Europe; copyright © 2005 the Association of the Geological Surveys of The European Union (EuroGeoSurveys))[1] as well as from own measurements (marked with *), from a study of the Old Danube River in Vienna (Austria)[2] and the soils analyzed in the present study, collected within Vienna (Austria).

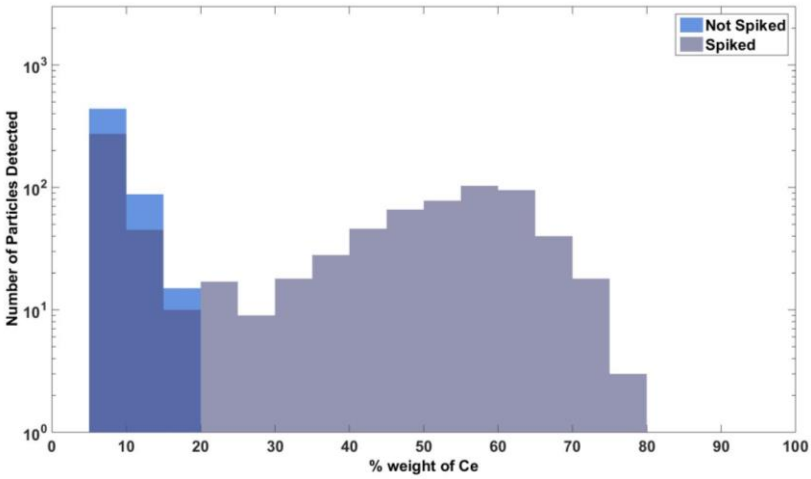
2. Automated scanning electron microscopy (SEM)

Electron microscopy is one of the most commonly used techniques to measure the size and the composition of nanomaterials. It offers very high spatial resolution and the ability to determine elemental composition of individual nanoparticles using energy dispersive x-ray spectroscopy (EDX). However, the presence of a multitude of organic compounds, microorganisms, nano- and micro-particles in complex matrices (e.g. soil samples), and the possible excitation of atoms from neighboring particles, make the detection of ENPs very challenging and results remain in general on a qualitative level. A major drawback of the method is the high time consumption due to manual measurements which results in a rather low number of particles that can be investigated. A significant improvement of this technique is achieved through the use of automated particle detection methods for identifying particles based on contrast differences (e.g. density contrast) and using EDX to determine elemental composition of individual particles. Using automated EM, several thousands of individual particles can be detected and classified (based on their elemental composition) within a few hours.

In this work, automated scanning electron microscopy (SEM) analysis was employed to investigate spiked and pristine soil colloidal extracts. Following the automated particle analyses, particles were grouped based on their cerium content (Supplementary Figure 3). A population of particles with high cerium content (> 20 wt % Ce) was observed in the spiked sample only. The particle size limit of detection (LOD) under the applied conditions was around 70 nm (diameter). Assuming particles consisting purely of CeO_2, this size translates into a mass-per-particle LOD of 1.1 fg Ce, which falls about 1 order of magnitude above the mass-per-particle LODs obtained for the spICP-MS (Supplementary Table 3) and spICP-TOFMS (Supplementary Table 6). For both the spiked and non-spiked samples, a comparable amount of 'low cerium containing' particles (< 20 wt% Ce) was detected; however, manual inspection of the individual spectra of these 'low cerium containing' particles revealed no Ce signal. The calculated Ce contents, thus, represent an analytical artifact and indicate the detection limits of the automated SEM method under the applied operational conditions. Nevertheless, our data suggest that the automated algorithm for identifying particles in combination with EDX analysis is a promising method for samples where a multitude of different types of particles are present.

However, SEM requires a sample preparation procedure that is more time-consuming, complex and susceptible to errors than the simple dilution steps that are required for ICP-TOFMS. First, automated SEM analysis requires the sample to be diluted in order to avoid

particles depositing on top of each other when loaded on the sample holder; at the same time the dilution should not be too high in order to avoid having too few particles on the sample holder. This trial and error process is dependent upon the properties of the sample (e.g. composition, particle size distribution, etc.). In addition, a high content of organic material in the sample would cause contamination during the analysis and thus interfere with the identification of individual particles with the SEM. This effect was observed during our analysis and an oxidation pre-treatment step using hydrogen peroxide was deemed necessary in order to reduce the organic content. This additional pre-treatment step increases the risk of particle loss and alteration. Finally, the measurement time for one particle in the SEM is in the order of 10 s and thus almost 4 orders of magnitudes larger than the corresponding spICP-TOFMS measurement time (0.3 ms). Furthermore, assuming a ratio of ENPs:NNPs of 1:100 – 1:1000, it would take 1'000 – 10'000 s to detect one ENP, which makes the quantification of ENP in complex matrices using automated EM currently unrealistic.

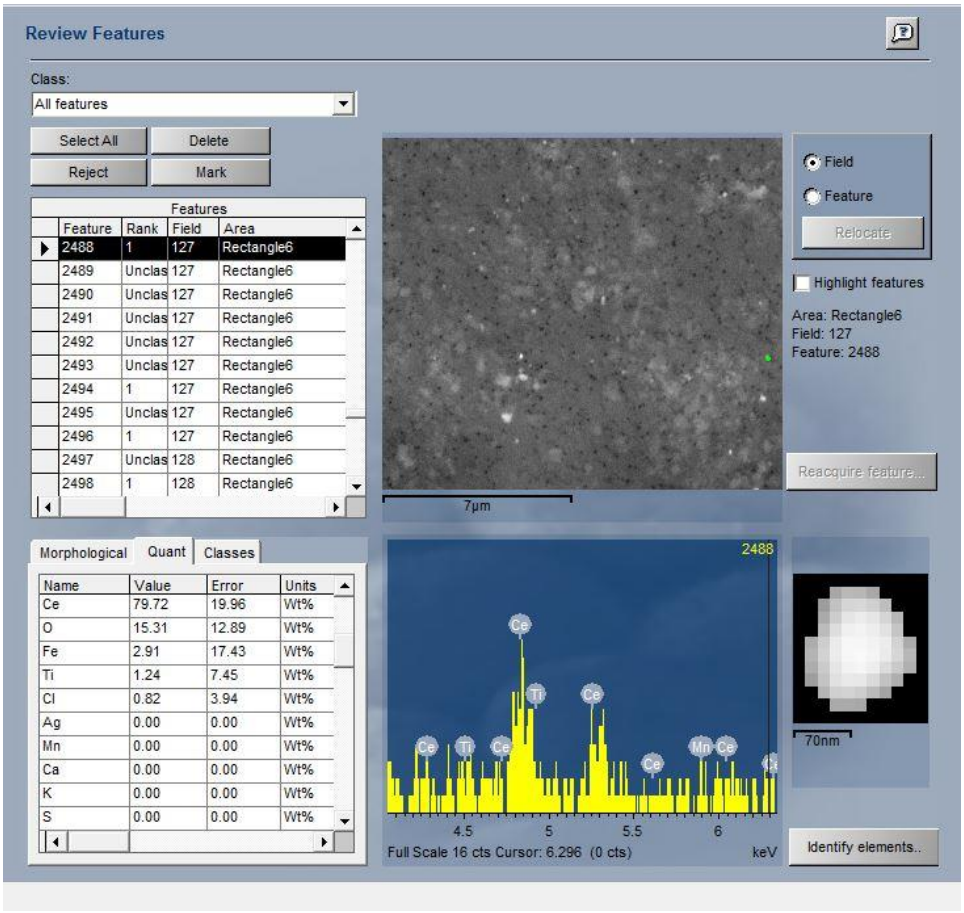


Supplementary Figure 3: Number of Ce-containing particles and the percentage of Ce content in each particle, based on an automated SEM imaging with EDX analysis. The colloid extract without added Ce (Not Spiked) and with the addition of CeO_2 ENPs at a concentration of 400 mg kg^{-1} (sample E) in the soil are shown. Samples were pre-treated with hydrogen peroxide to remove organic material, which was interfering with the SEM analysis**.**

Method

To remove organic material, which may mask the presence of small nanoparticles and cause contamination during SEM analyses, 13.5 mL of the colloidal extracts were mixed with 15 mL of hyperoxide solution (Suprapore Hydrogen Peroxide 30 %, Millipore) in a 100 mL glass volumetric flask and were reacted for 36 hours. After that time it was assumed that no hydroxyl radicals were left. The volumetric flask was filled up to 100 mL with MilliQ water and shaken. An aliquot of the resulting suspension was further diluted with MilliQ to a final volume of 100 mL and mass concentration of 10 mg L^{-1}. An aliquot of 5 mL from the

resulting suspension was filtered through polycarbonate membranes (Nucleopore) with 50 nm pore size and were left to dry in polypropyelene containers for at least 48 hours, prior to analysis. The filters were coated with a thin layer (~ 10 nm) of Carbon to make them electrically conductive. The contrast (atomic weight contrast resulting from backscattered electrons) between particles and the background was used to identify particles on a pre-selected area and EDX analysis was performed on every particle. The elemental composition and the particle dimensions were tabulated along with low resolution images (Supplementary Figure 4).



Supplementary Figure 4: Automated particle analysis conducted in the SEM. Particles are first identified on a selected area (top image). EDX analysis is subsequently conducted on each particle (bottom images).

3. Concentration range of CeO_2 ENPs spiked to soil samples

Supplementary Table 1: Total added CeO_2 NP concentration for all five samples, the original pristine sample (P) (used for training the GBC model) and control sample (O), another pristine soil from the same sample area. Note that in all cases the CeO_2 NP suspensions were stabilized with Suwanee River Natural Organic Matter (SRNOM) before spiking resulting in an added SRNOM concentration of 2 mg kg_{soil}^{-1} TOC. The control sample is a soil sample collected at the same site as the soil employed for the spiking experiments.

Sample	**CeO_2 ENP added [mg kg_{soil}^{-1}]**
P (pristine)	0
O (control)	0
A	4×10^{-2}
B	4×10^{-1}
C	4×10^{0}
D	4×10^{1}
E	4×10^{2}

4. Single-particle ICP-MS analysis of CeO_2 ENP spiked and pristine colloid extracts

Single-element spICP-MS measurements were performed on an Agilent 7700 ICP-MS instrument (Agilent Technologies, CA, USA), with ^{139}La and ^{140}Ce monitored consecutively. Samples were measured at a 10,000-fold dilution and a dwell-time of 5 ms and spectra were recorded for 60 s. Instrumental settings of the spICP-MS measurements are summarized in Supplementary Table 2. Calibration with dissolved Ce and La standards was performed prior to sample analysis. The nebulization efficiency was measured in triplicates using a well-characterized gold 60 nm dispersion (BBI, UK) following the particle frequency method described by Pace et al.[3] The sample flow rate was measured with a TruFlo flow meter (GlassExpansion, Australia).

Supplementary Table 2: Instrument setting for conventional single-particle ICP-MS measurements

Instrument	**Agilent 7700 ICP-MS**
RF power	1550 W
Plasma gas flow rate	15 L min^{-1}
Carrier gas flow rate	0.8 mL min^{-1}
Dilution gas flow rate	0.4 mL min^{-1}
Nebulizer	Micromist
Isotopes monitored	^{139}La or ^{140}Ce
Integration time	5 ms
Nebulization efficiency	5 %
Sample flow rate	344 µl min^{-1}
Acquisition time	60 s

The spICP-MS data processing was done using Matlab R2012b. A threshold for particle detection was set at 5 times the standard deviation of the dissolved Ce or La signal, σ_{diss}, obtained with an iterative algorithm to separate the background from particle events.[3-5] The resulting mass-per-particle limits of detection (LOD) for Ce and La are listed in Supplementary Tables 3 and 4).

Supplementary Table 3: Mass-per-particle LODs obtained after applying a threshold for Ce-particles at 5σ of the dissolved background distribution for spICP-MS measurements of pristine soil sample (P) used for training the GBC model, control sample (O) and the spiked samples (A to E).

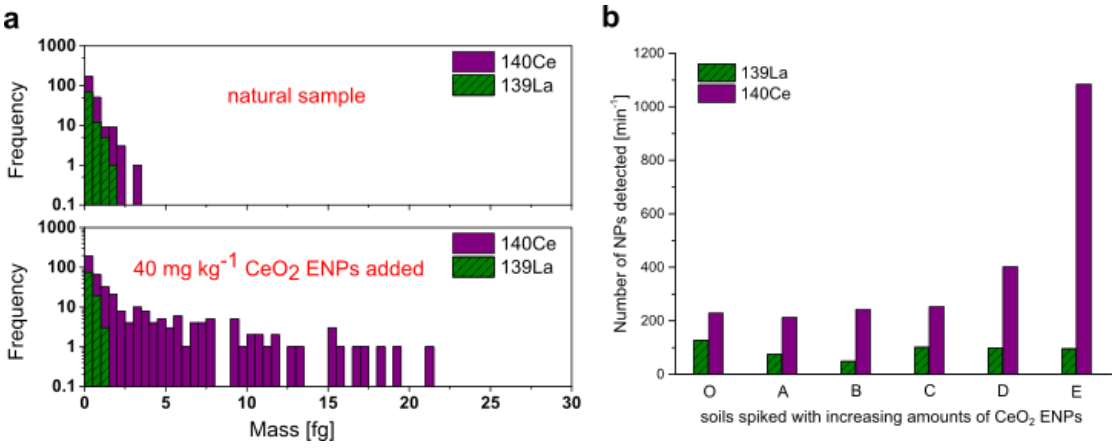
Sample	CeO_2 ENPs added [mg kg^{soil-1}]	Ce $5\sigma_{diss}$ LOD [counts]	Ce $5\sigma_{diss}$ LOD [fg]	Ce NPs detected [min^{-1}]
P	0	8.79	0.14	242
O	0	10.8	0.17	226
A	4×10^{-2}	8.90	0.14	210
B	4×10^{-1}	8.91	0.14	240
C	4×10^{0}	8.16	0.13	250
D	4×10^{1}	9.01	0.14	397
E	4×10^{2}	35.4	0.57	1071

Supplementary Table 4: Mass-per-particle LODs obtained after applying a threshold for La-particles at 5σ of the dissolved background distribution for spICP-MS measurements of pristine soil sample (P) used for training the GBC model, control sample (O) and the spiked samples (A to E).

Sample	CeO_2 ENPs added [mg kg^{soil-1}]	La $5\sigma_{diss}$ LOD [counts]	La $5\sigma_{diss}$ LOD [fg]	La NPs detected [min^{-1}]
P	0	5.66	0.21	86
O	0	6.26	0.22	125
A	4×10^{-2}	6.28	0.22	74
B	4×10^{-1}	11.9	0.34	48
C	4×10^{0}	5.68	0.21	100
D	4×10^{1}	6.25	0.22	96
E	4×10^{2}	6.30	0.22	94

Separate measurements on masses ^{139}La and ^{140}Ce using conventional spICP-MS confirmed the presence of Ce and La containing (nano-) particles in all samples (Supplementary Figure 5a, b). The natural background is composed of detectable Ce- and La-containing particles between 0.15 and 3.1 fg Ce and 0.21 and 1.5 fg La per particle. (Note that the lack of information on the total composition, shape or density of the individual natural particles makes it impossible to derive a corresponding particle size from the Ce or La mass signal.[4]) On average, this corresponds well to the observed bulk Ce:La ratio of 2:1 found in the uncontaminated soils (Supplementary Figures 1 and 2) and suggests that this ratio is largely maintained even on the single particle level. In the spiked samples, the presence of CeO_2 ENPs is indicated by the detection of higher mass Ce-containing particles with particles

containing up to 21 fg Ce (Supplementary Figure 5a), which corresponds to CeO_2 ENPs of about 180 nm (assuming a spherical particle shape). With increasing amounts of added CeO_2 ENPs, the number of detected Ce-containing particles increases, while the number of La-particles remains constant (Supplementary Figure 5b) A significant increase is, however, only observed at CeO_2 ENP concentrations of 40 mg kg^{-1}, corresponding to about half of the natural background of Ce in the investigated soils (72.6 mg kg_{soil}^{-1}). Even in worst-case scenarios, expected concentrations of released CeO_2 ENPs are at least one order of magnitude below this level.[6] Most importantly, a distinction between individual Ce-containing NPs was not possible based on the results from the conventional spICP-MS. To identify individual Ce-containing NPs as either natural or engineered we need to measure multiple elements simultaneously on each particle.



Supplementary Figure 5: Conventional single-particle ICP-MS analysis. **a**, Histograms of Ce- and La-containing single-particle signals from the pristine (top) and a CeO_2 ENP spiked soil sample (bottom). Added CeO_2 ENPs result in the presence of Ce-containing particles of higher mass compared to the pristine sample. **b**, Overview of number of detected Ce- and La-containing particles in all samples. Samples A to E are spiked a 10-fold increase of added CeO_2 ENP concentration per sample spanning from 0.04 mg kg_{soil}^{-1} (A) to 400 mg kg_{soil}^{-1} (E) and control sample O has no added CeO_2 ENPs (Table S1). A significant increase in Ce to La ratio is only observed in samples with more than 40 mg kg^{-1} added CeO_2 ENPs (samples D and E).

5. Single-particle ICP-TOFMS Measurements

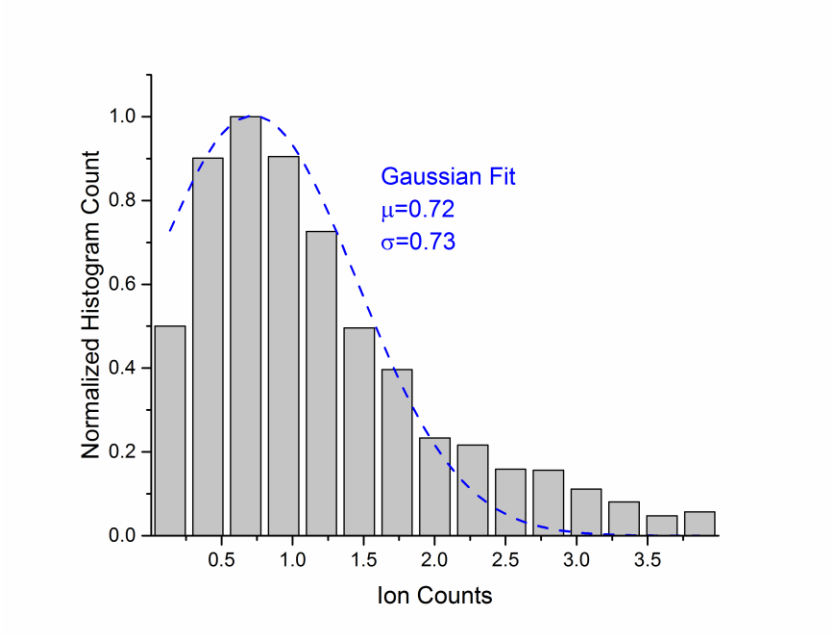
Single-particle ICP-TOFMS measurements were made on a prototype ICP-TOFMS that features the front end (ICP, extraction optics) of an ELAN 6000 ICPMS instrument and a TOFWERK® reflectron-TOFMS mass analyzer. A quadrupole notch filter is used between ion extraction and TOF mass analysis to guide the ion beam and selectively attenuate major background ions (e.g. H_2O^+, $^{40}Ar^+$, etc.). Operating conditions for spICP-TOFMS are provided in Supplementary Table 5. To calibrate single particle masses, standard working curves for solutions of ^{139}La and ^{140}Ce in 1% HNO_3 were generated. To determine the mass of water uptake to the plasma, a blank solution was weighed before and after aspiration with the pneumatic nebulizer; in addition, the waste drained from the cyclonic nebulizer was weighed. The mass the difference between aspirated and drained solution per unit time was taken to be the uptake rate into the plasma. For single-particle ICP-TOFMS experiments, the TOFMS was run at a spectral generation rate of 3333 Hz (300 µs per spectrum) and 120 s of data was collected for each sample. For each TOFMS spectrum, the signal of each m/z channel was integrated to give the peak value for each isotope.

Supplementary Table 5: Instrument setting for conventional single-particle ICP-TOFMS measurements

Instrument	**Prototype ICP-TOFMS**
RF power	1400 W
Outer ICP gas flow (Ar)	16 L min^{-1}
Intermediate ICP gas flow (Ar)	1.1 L min^{-1}
Nebulizer gas flow (Ar)	1.2 L min^{-1}
Sample flow rate	750 µl min^{-1}
Liquid uptake rate to ICP	34 µL min
Mass range	7 – 254 amu
Isotopes attenuated with notch filter	$^{14}N^+$, $^{16}O^+$, H_2O^+, $^{40}Ar^+$
Spectral acquisition time resolution	300 µs
Acquisition time	120 s

Data analysis of the spICP-TOFMS peak values was accomplished with LabVIEW®. Briefly, a peak detector algorithm was used to find all ^{140}Ce single-particle events with a peak value greater than 2 counts and all found events were integrated across 900 µs (i.e. 3 mass-spectral acquisition data points) to ensure complete integration of single-particle events split between adjacent acquisition windows (the typical duration of a single-NP is ~300 µs). The threshold value of 2 counts used for single-particle event finding is low enough that no single-NP events were missed and, indeed, some false events due to fluctuation of the background were

found. Background signals (not associated with single-particle events) were also integrated across three data points (i.e. across 900-µs). A histogram of the found ^{140}Ce single-particle-event signals combined with integrated background noise signals was created. The background portion of this histogram was fit to a Gaussian to determine the standard deviation of the dissolved background signal, σ_{diss} (Supplementary Figure 6). The entire histogram was then thresholded at a count value of $5\sigma_{diss}$ to separate detectable Ce-NPs from background signals. The threshold for Ce-NP detection depended on the amount of dissolved Ce in each sample, and ranged from 4.3–6.9 ion counts (0.10–0.17 fg Ce) (Supplementary Table 6). For machine learning studies, the signals for the isotopes of interest were integrated across the same time windows as each identified Ce-NP signal. For manual ENP vs. NNP classification, the same $5\sigma_{diss}$ thresholding process was used to identify ^{139}La signals above the mass-per-particle LOD.

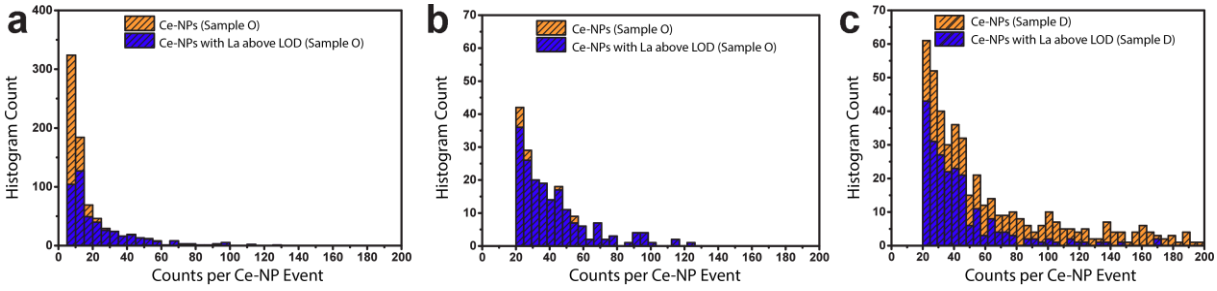


Supplementary Figure 6: Histogram of ^{140}Ce-background in the pristine soil sample (P) fitted to a Gaussian to determine σ_{diss}.

Supplementary Table 6: Mass-per-particle LODs obtained after applying a threshold for Ce-particles at 5σ of the dissolved background distribution for spICP-TOFMS measurements pristine soil sample (P) used for training the GBC model, control sample (O) and the spiked samples (A to E).

Sample	CeO_2 ENPs added [mg $kg^{soil\text{-}1}$]	Ce $5\sigma_{diss}$ LOD [counts]	Ce $5\sigma_{diss}$ LOD [fg]	Ce NPs detected [min^{-1}]
P	0	4.36	0.11	544
O	0	5.68	0.14	438
A	4×10^{-2}	6.16	0.15	336
B	4×10^{-1}	5.93	0.15	328
C	4×10^{0}	4.28	0.10	492
D	4×10^{1}	6.94	0.17	589
E	4×10^{2}	6.00	0.15	1912

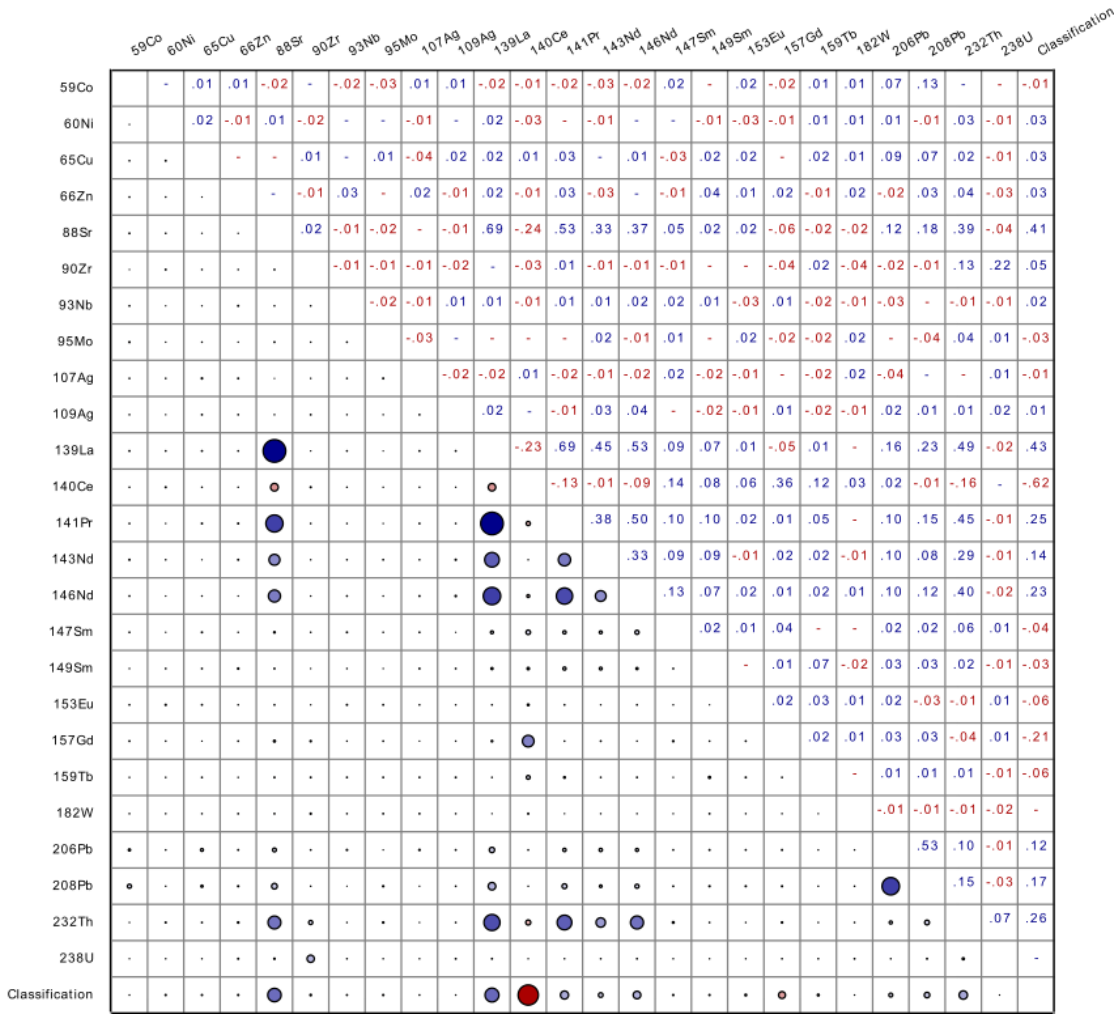
As shown in Supplementary Figure 7, La is not detected in many of the low intensity natural Ce-containing nanoparticles events. This poor correlation of La and Ce at low Ce intensities is a result of La signal near or below the mass-per-particle LOD because La is about half as abundant as Ce in the NNPs. For manual Ce-NNP vs. Ce ENP classification, only Ce single-particle events with more than 20 counts were considered; above this signal threshold, the Ce and La signals are well correlated and are no longer significantly affected by ion-counting statistics and the combined LODs of Ce and La. As shown in Supplementary Figure 7, good correlation between the presence of Ce and La in Ce-NNPs is apparent in the histograms. As CeO_2 ENPs are spiked into the soil extract containing Ce-NNPs, more Ce-NP events are detected without an accompanying ^{139}La signal, which is a characteristic of pure CeO_2 ENPs.



Supplementary Figure 7: a, In the pristine soil sample, Ce-NPs are detectable above ~6 counts. However, La signals are often not above the detection limit for low-intensity Ce-containing NP signals. As the single-particle Ce signal increases, the proportion of Ce-NPs that contain detectable levels of La increases to unity. The fact that La detectability increases with the intensity of Ce indicates that ion counting statistics are limiting the detection of La for all Ce-containing NPs. **b**, If a threshold of 20 counts for Ce is used almost all Ce-NPs have detectable levels of La. **c**, For the soil sample spiked with 40 mg kg^{-1} of CeO_2 ENPs, almost the same number—and a similar distribution shape—of Ce+La NPs (blue bars) are detected compared to the control with no CeO_2 ENPs (**b**). The rest of the Ce-NP signals (visible orange bars) can therefore be classified as CeO_2 ENPs.

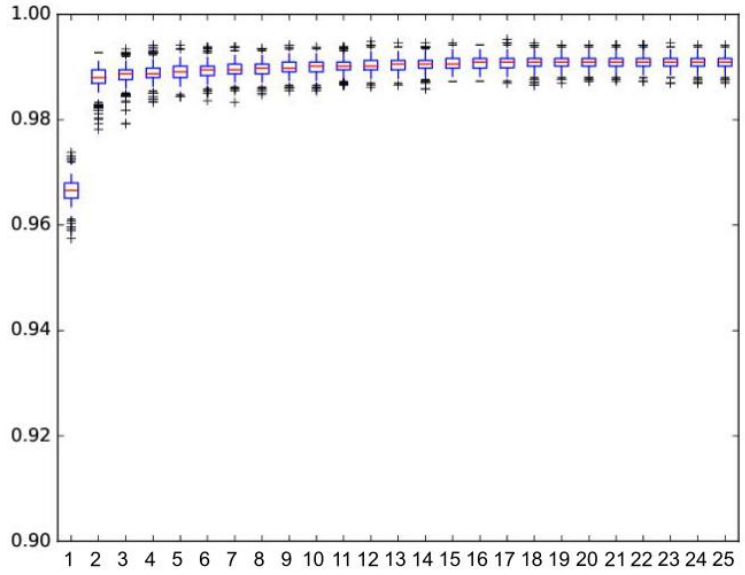
6. Machine Learning - Detailed description

A supervised gradient decision tree boosting classification (GBC) machine learning model was developed to classify individual particles as ENP or NNP. GBC is an ensemble machine learning algorithm in which decision trees are sequentially improved through a forward stage-wise process.[7] Data employed to train the model consisted of several thousand multi-element spICP-TOFMS signals from reference suspensions containing NNPs (natural soil colloid extract) or CeO_2 ENPs only. To improve the general model performance and prevent over fitting, which occurs when unimportant features are employed to train the model, several feature selection steps were applied. First, 25 isotopes were selected as potential relevant features for predictions (see list of element in Supplementary Figure 8) and the spICP-TOFMS traces were extracted for these elements for all Ce-particle signals. Recursive feature elimination with cross validation (RFECV) was employed to identify the elements critical to prediction (Figure 3, main text).[8] During RFECV, the minimum number of training features required to maximize model performance is identified.[8, 9] Training features that contain little or no predictive information are 'trimmed' from the training database. Once the optimal feature set composition is identified for a given model run, a feature mask was applied to remove non-optimal features from both the training and holdout sets. Following RFECV, a grid search hyper parameter optimization with cross validation (GSHPOCV) was performed to identify the best performing combination of model parameters (i.e. the parameter combination that resulted in the highest R^2 cross-validation score).[9] Pedregosa et al.[9] and Goldberg et al.[8] provide critical information about RFECV and GSHPOCV. The generic classification performance (i.e. the ability to classify data not included within the training set) was evaluated using an iterative holdout set partition and validation scheme, and yielded a weighted F1 score of 98%. [8] Aggregated performance results from 1000 model runs are included in Supplementary Figure 9.

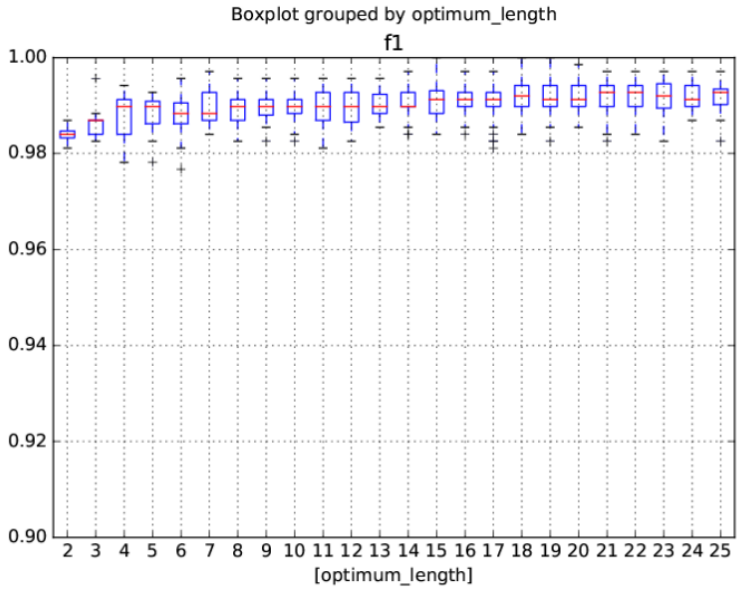


Supplementary Figure 8: Correlation of individual isotopes to each other and towards the classification. Blue "-" represent correlations below 0.01 and red "-" represent correlations smaller than -0.01.

Once the optimum feature set and optimum parameters have been identified the model is re-trained. This final trained version of the GBC model is then used to classify data outside the training set, namely the spICP-TOFMS signals of the mixed NNP and ENP samples (colloid extracts of natural soil spiked with different concentrations of CeO_2 ENPs).



Supplementary Figure 9: Aggregate model performance (F1-Score) to predict the NP classification of the cross validation test fold as a function of the number of features employed to train. Each box plot represents aggregated model scores for five cross validation iterations for all 1000 model runs. This results in 5000 data points for each feature set size. From visual inspection, only marginal improvements in prediction performance are gained past 8-11 features.



Supplementary Figure 10: Aggregate performance of the model in predicting the ENP classification for the holdout set as a function of the number of RFECV-selected features employed to train the model. The best median performance to predict the classification was obtained when all features were employed to train (median = 0.992732543). However, the optimal feature set size using the optimality criteria considering the feature set size, the median performance, the variance, and the frequency of observation of that feature set, was 17 features (median = 0.99128025). The feature set size was never below 2, indicating that Ce and La were always required to optimize prediction performance.

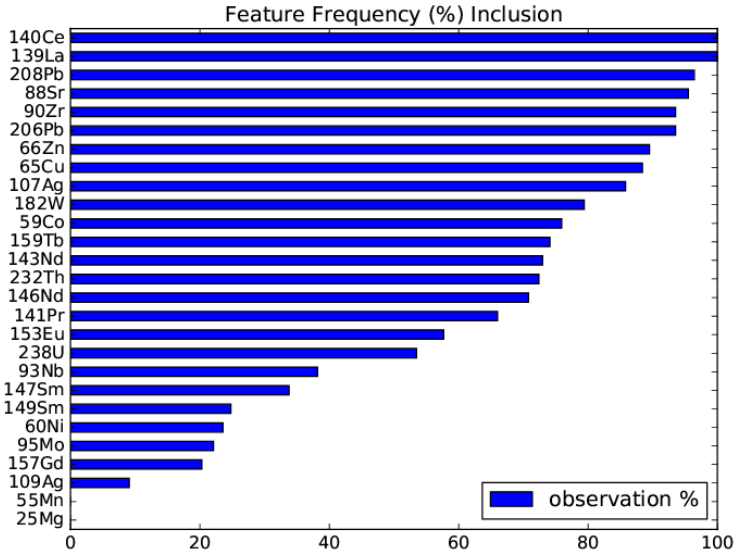
Supplementary Table 7: Aggregate holdout performance and optimality score by feature set size for 1000 model runs.

Group	Quartile 1	Median	Quartile 3	Inner quartile range	Number observed in group	Optimality score
2	0.983287165	0.984014027	0.98474081	0.001453645	4	1398.203435
3	0.984014162	0.98691913	0.986920678	0.002906516	9	1045.844173
4	0.984013352	0.989827194	0.991279733	0.007266382	9	312.8274962
5	0.986192911	0.989826936	0.990916187	0.004723275	14	598.9016168
6	0.986194519	0.988373666	0.990553206	0.004358687	15	580.3141667
7	0.986919573	0.988373666	0.992733096	0.005813523	15	372.934563
8	0.986921341	0.98982513	0.99128025	0.004358908	29	840.1789777
9	0.988008594	0.989826678	0.991279955	0.003271361	24	823.5331836
10	0.988371552	0.989827194	0.991280618	0.002909066	31	1076.586024
11	0.986921341	0.989826678	0.992733342	0.005812	29	458.2690513
12	0.986556364	0.989827108	0.99273285	0.006176486	40	545.2277167
13	0.988374157	0.989827538	0.992369792	0.003995635	34	661.2874148
14	0.98982513	0.989827194	0.99273285	0.00290772	47	1166.42759
15	0.988372978	0.991279733	0.993096794	0.004723816	72	1025.06643
16	0.989825517	0.991279992	0.992733342	0.002907825	72	1561.161817
17	0.98982599	0.99128025	0.992733342	0.002907352	85	1734.906222
18	0.98982599	0.992006427	0.994186489	0.004360499	88	1130.210926
19	0.98982599	0.99128025	0.994186489	0.004360499	80	974.1004702
20	0.989827366	0.99128025	0.994186268	0.004358902	75	867.8761493
21	0.989827194	0.992732235	0.994186489	0.004359295	65	715.2299725
22	0.989826678	0.992732235	0.994186047	0.004359368	69	724.7208106
23	0.989463622	0.992006427	0.994549702	0.00508608	48	413.633496
24	0.989823192	0.991279992	0.994186489	0.004363297	38	366.0675894
25	0.99018997	0.992732543	0.99346021	0.00327024	8	98.56847405

The following optimality criterion was employed to determine the optimal number of features.

$$\boldsymbol{optimality}_i = \frac{\mathbf{1}}{\boldsymbol{i}} * \frac{\mathbf{1}}{\boldsymbol{median\ performance}_i} * \frac{\mathbf{1}}{\boldsymbol{IQR}_i} * \#\ \boldsymbol{observed}_i \qquad \textbf{Supplementary Equation 1}$$

Where $optimality_i$ is the optimality score for feature set size i, $median\ performance_i$ is the median classification performance for feature set size i, IQR_i is the inner quartile range for the classification performance for feature set size i, and $\#\ observed_i$ is the number of times that a feature set of size i was selected by RFECV as the optimum feature set size.



Supplementary Figure 11: Frequency of features included for the prediction of ENP versus NNP through the RFECV (in percent of number of model runs). 55Mn and 25Mg were determined to be exclusively background elements and removed from the assessment prior to training. Observation frequency perfectly correlates with the feature set composition for each size (i.e., a feature set size of 5 chosen by RFECV always contains the 140 Ce, 139 La, 208 Pb, 88 Sr, and 90 Zr). Other REEs (e.g. Nd, Pr and Th), which are strongly inter-correlated in the NNPs, but absent in the ENPs, were infrequently identified as critical during RFECV. Due to their strong correlation with La (Supplementary Figure 8) the presence of any of these four isotopes in a particle signal does not provide new unique information for the classification.

The machine learning analysis was performed using python 2.7.10 and the scikit-learn package 0.17.0. [9] In scikit-learn, the GBC method employs regression decision trees for both classification and regression tasks, as it successively trains regression trees to maximally correlate with the negative gradient for a given loss function at each stage in a steepest descent procedure.[9] Through this process, the model iteratively improves its ability to classify observations modeled poorly by the previous steps. Other machine learning methods were investigated (random forest, adaptively boosted decision trees, and decision trees) before selecting GBC on the basis of its consistently higher classification performance.

References

1. R. Salminen, M. J. Batista, M. Bidovec, A. Demetriades, B. De Vivo, W. De Vos, M. Duris, A. Gilucis, V. Gregorauskiene, J. Halamic, P. Heitzmann, A. Lima, G. Jordan, G. Klaver, P. Klein, J. Lis, J. Locutura, K. Marsina, A. Mazreku, P. J. O'Connor, S. Å. Olsson, R.-T. Ottesen, V. Petersell, J. A. Plant, S. Reeder, I. Salpeteur, H. Sandström, U. Siewers, A. Steenfelt and T. Tarvainen, *Geochemical Atlas of Europe. Part 1: Background Information, Methodology and Maps*, Geological Survey of Finland, 2005.
2. A. P. Gondikas, F. v. d. Kammer, R. B. Reed, S. Wagner, J. F. Ranville and T. Hofmann, *Environ. Sci. Technol.*, 2014, **48**, 5415-5422.
3. H. E. Pace, N. J. Rogers, C. Jarolimek, V. A. Coleman, C. P. Higgins and J. F. Ranville, *Anal. Chem.*, 2011, **83**, 9361-9369.
4. J. Navratilova, A. Praetorius, A. Gondikas, W. Fabienke, F. von der Kammer and T. Hofmann, *Int. J. Env. Res. Public Health*, 2015, **12**, 15756–15768.
5. J. Tuoriniemi, G. Cornelis and M. Hassellöv, *Anal. Chem.*, 2012, **84**, 3965-3972.
6. A. A. Keller and A. Lazareva, *Environ. Sci. Technol. Lett.*, 2014, **1**, 65-70.
7. T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*, Springer, New York, USA, 2 edn., 2009.
8. E. Goldberg, M. Scheringer, T. D. Bucheli and K. Hungerbühler, *Environ. Sci.: Nano*, 2015, **2**, 352-360.
9. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. Dubourg, *J. Mach. Learn. Res.*, 2011, **12**, 2825-2830.