Machine Learning Approaches to Quantify Nanoscale Variations in the Mechanical Properties of Soft Nanoparticles

Benjamin Baylis and John R. Dutcher*

Department of Physics, University of Guelph, Guelph, ON, Canada N1G 2W1

Supporting Information

This document contains additional technical details, along with eleven supporting data figures.

* correspondence to: dutcher@uoguelph.ca

Approach		
Near the contact point	**Near the maximum applied force**	**Spanning from low to high forces**
Deformation from 0.04 nN to 0.1 nN	Deformation from 1 nN to 1.5 nN	Deformation from 0.04 nN to 1.5 nN
Deformation from 0.04 nN to 0.5 nN	Deformation from 1 nN to 2 nN	Deformation from 0.04 nN to 2 nN
Deformation from 0.04 nN to 1 nN	Deformation from 1.5 nN to 2 nN	Deformation from 0.1 nN to 1.5 nN
Deformation from 0.1 nN to 0.5 nN	Stiffness	Deformation from 0.1 nN to 2 nN
Deformation from 0.1 nN to 1 nN		Deformation from 0.5 nN to 1.5 nN
Deformation from 0.5 nN to 1 nN		Deformation from 0.5 nN to 2 nN
Modulus (10 nm of indentation)		Modulus (entire indentation)

Retraction		
Near the maximum applied force	**Near the surface of the sample**	**Spanning from high to low forces**
Recovery from 1.5 nN to 1 nN	Recovery from 0.1 nN to 0.04 nN	Recovery from 1.5 nN to 0.04 nN
Recovery from 2 nN to 1 nN	Recovery from 0.5 nN to 0.04 nN	Recovery from 2 nN to 0.04 nN
Recovery from 2 nN to 1.5 nN	Recovery from 1 nN to 0.04 nN	Recovery from 1.5 nN to 0.1 nN
	Recovery from 0.5 nN to 0.1 nN	Recovery from 2 nN to 0.1 nN
	Recovery from 1 nN to 0.1 nN	Recovery from 1.5 nN to 0.5 nN
	Recovery from 1 nN to 0.5 nN	Recovery from 2 nN to 0.5 nN
	Max adhesion	
	Height of max adhesion	
	Adhesion energy	
	Height range of adhesion energy	

Table S1: Table of features used in both the particle/substrate classifiers and the inner particle structure classifiers (supervised and unsupervised) and calculated from the force-distance curves. The features are divided into three regimes for the approach and retraction portions of the curves.

Classifying Force-Distance Curves in the Supervised ML Classifier

To classify individual force-distance curves as either corresponding to a PG nanoparticle or the bare substrate in the supervised particle/substrate ML classifier, we used an extreme gradient boosting algorithm available in the library XGBoost (version 2.0.3) implemented in Python [43] (Figure 4). The XGBoost classifier is a widely used classifier known for its computational speed and accuracy [46-48]. It operates on a dataset containing N vectors, each

with k features, building an ensemble of decision trees to classify each vector into one of two target classes. Decision trees are added to the model iteratively and are constructed to minimize a regularized objective function. The regularized objective function is composed of two terms: a loss function (which measures the discrepancy between the predicted and actual target classes) and a regularization term (which helps to avoid over-fitting). This function is minimized using the first and second order gradients. The loss function used in this work was the log-loss (logistic loss) function for binary classification which allows us to determine the probability of force-distance curves belonging to one of two classes [48], where we used a probability of 50 % to differentiate between binary classifications. XGBoost also allows us to quantify the importance of each input feature to the final model by calculating their contributions to the reduction of the regularized objective function expressed as the "gain" [11]. In the present study, we determine the relative importance of each feature in the models by expressing the gain of a feature as a percentage of the total gain from all features.

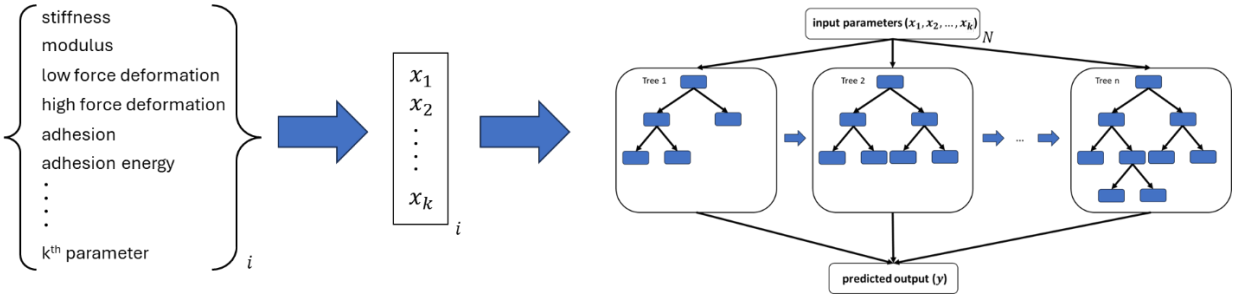

Figure S1: Schematic of the XGBoost algorithm. The $\boldsymbol{k}$ parameters calculated from each $\boldsymbol{i}^{\text{th}}$ force-distance curve are used to construct a feature vector containing the values of the features. The feature vector for $\boldsymbol{N}$ force-distance curves are used as inputs to an XGBoost algorithm to classify each force-distance curve as corresponding to one of two materials. When training the classifier, each $\boldsymbol{i}^{\text{th}}$ feature vector also contains a known binary target class $\boldsymbol{y_i}$.

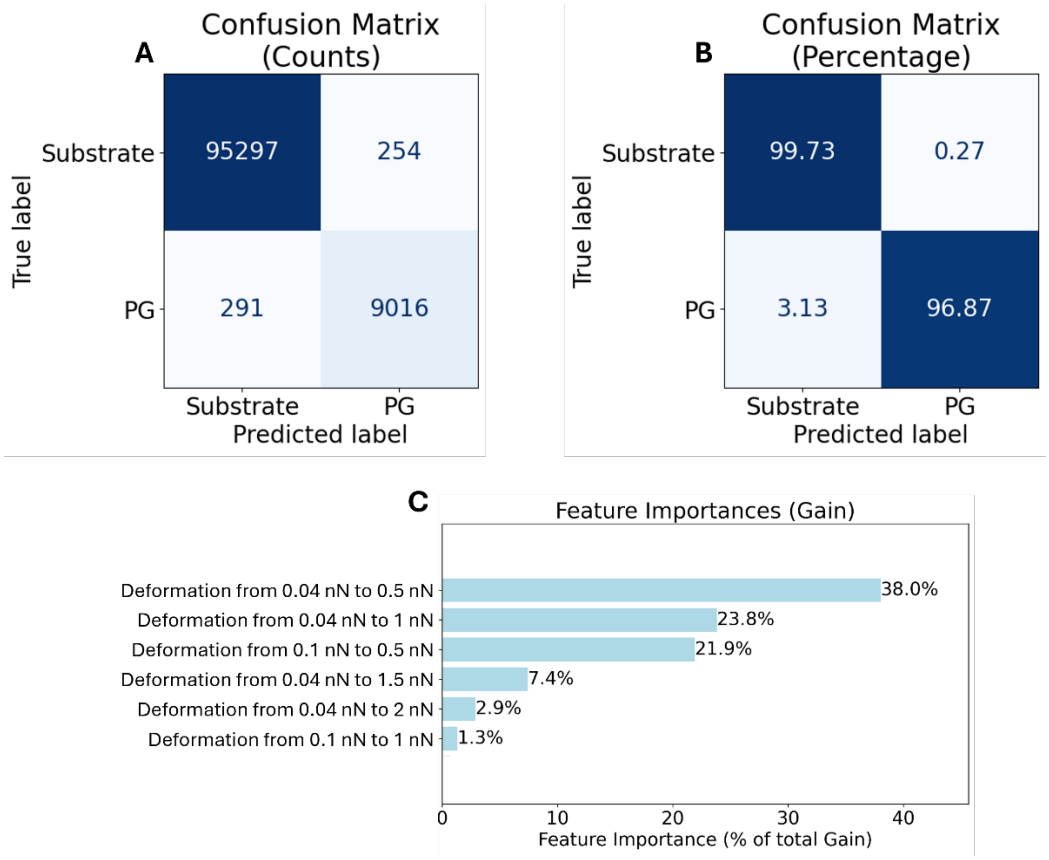

Figure S2: (A) Confusion matrix for the classification of the testing dataset using the supervised particle/substrate ML classifier showing the total counts for each true label. The y-axis of the confusion matrix is the true label, or manual classification, and the x-axis is the predicted label or classification from the output of the supervised particle/substrate ML classifier and the values display the number of pixels/force-distance curves. For example, the bottom row in A indicates pixels that were manually labeled as corresponding to PG in the testing dataset: 291 of these pixels/force-distance curves were classified as corresponding to the substrate, and 9016 of these pixels/force-distance curves were classified as corresponding to PG by the supervised particle/substrate ML classifier. (B) Confusion matrix for the classification of the testing dataset using the supervised particle/substrate ML classifier showing the percentage of pixels/force-distance curves for each true label. For example, the bottom row in A indicates pixels that were manually labeled as corresponding to PG: 3.13 % of these pixels/force-distance curves were classified as corresponding to the substrate, and 96.87 % of these pixels/force-distance curves were classified as corresponding to PG by the supervised particle/substrate ML classifier. (C)

The top six important features identified by the supervised particle/substrate ML classifier as specified by the percentage of the gain (improvement in the classifier accuracy) relative to the total gain across all decision trees displayed for each feature.

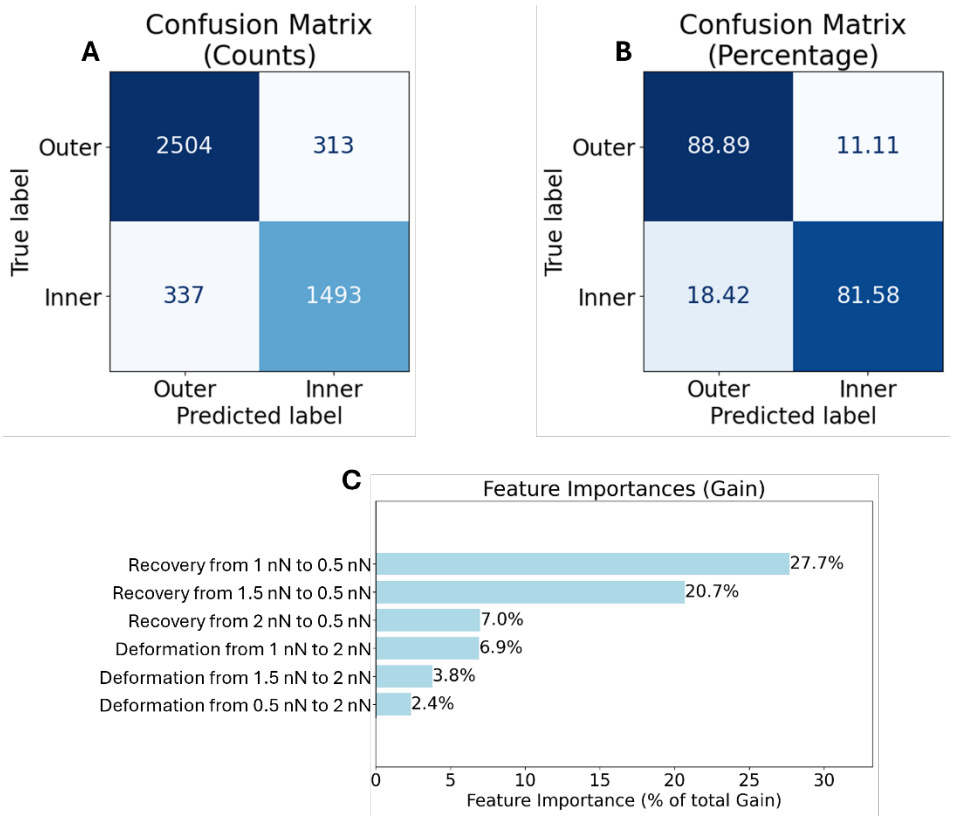

Figure S3: (A) Confusion matrix for the classification of the testing dataset using the supervised inner particle structure ML classifier showing the total counts for each true label. The y-axis of the confusion matrix is the true label, or manual classification, and the x-axis is the predicted label or classification from the output of the supervised inner particle structure ML classifier and the values display the number of pixels/force-distance curves. For example, the bottom row in A indicates pixels that were manually labeled as corresponding to the stiffer inner region in the testing dataset: 337 of these pixels/force-distance curves were classified as corresponding to the softer outer region, and 1493 of these pixels/force-distance curves were classified as corresponding to the stiffer inner region using the supervised inner particle structure ML classifier. (B) Confusion matrix for the classification of the testing dataset using the supervised inner particle structure ML classifier showing the percentage of pixels/force-distance curves for each true label. For example, the bottom row in A indicates pixels that were manually labeled as corresponding to the stiffer inner region: 18.42 % of these pixels/force-distance curves were classified as corresponding to the softer outer region, and 81.58 % of these pixels/force-distance curves were classified as corresponding to the stiffer inner region using the supervised inner

particle structure ML classifier. (C) The top six important features identified by the supervised inner particle structure ML classifier as specified by the percentage of the gain (improvement in the classifier accuracy) relative to the total gain across all decision trees displayed for each feature.

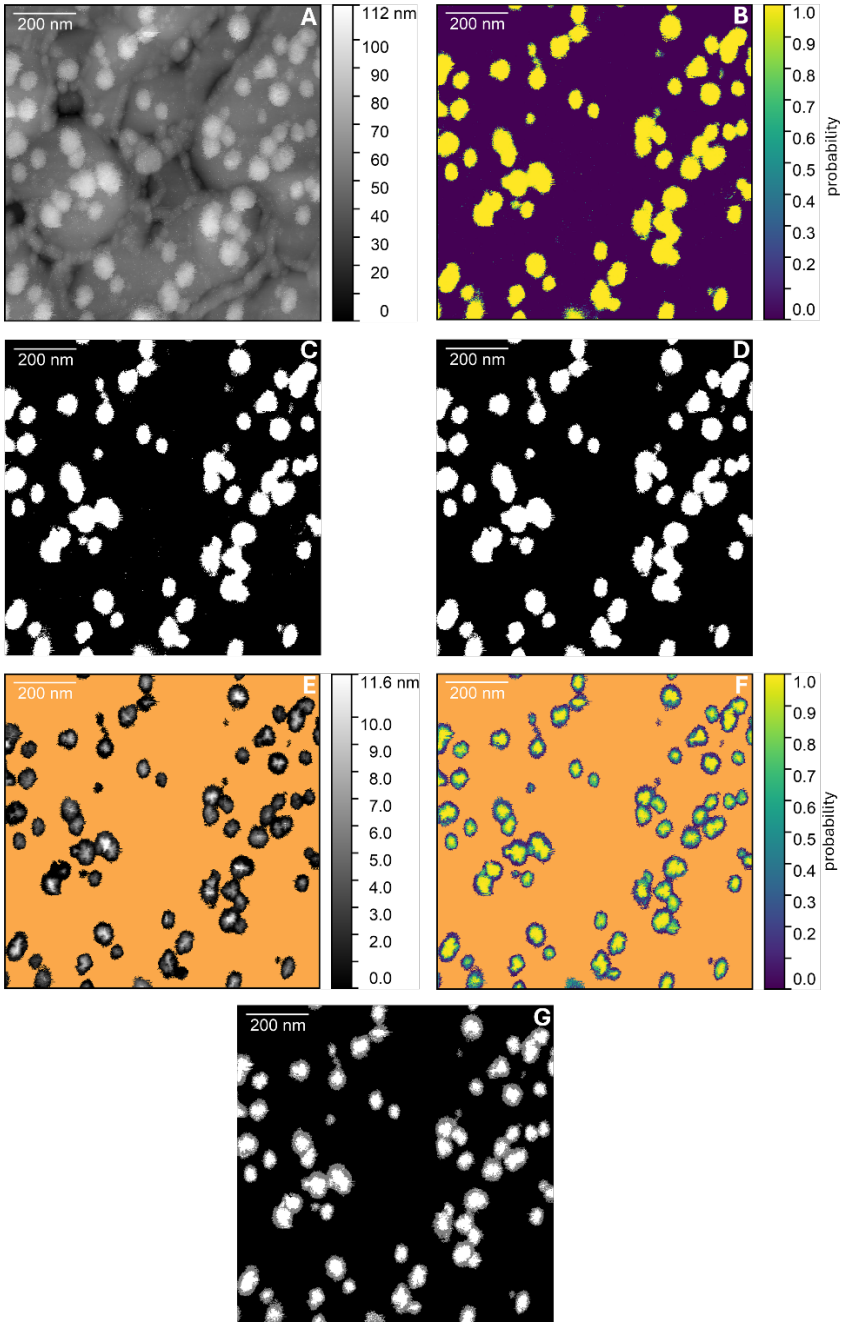

Figure S4: AFM-FS images and supervised ML classifier outputs for a second sample, in addition to those shown in Figures 5 and S5. (A) Contact height image of PG nanoparticles on the 4-MPBA/gold substrate. (B) The probability map output from the supervised particle/substrate ML classifier using the force-distance curves from the measurement in A showing the probability that each pixel corresponds to an interaction of the AFM tip with PG. (C) The classification map output using a threshold probability of 50 % to classify each pixel as

corresponding to either PG (white pixels) or the MPBA/gold substrate (black pixels). (D) The same classification map shown in C with small clusters (< 35 pixels) of white pixels removed to isolate full PG nanoparticles in the classification map. These removed pixels corresponded to only 1.0 % of PG pixels in C. (E) Peak force height image of the same scan in A revealing the stiffer inner structure of the PG nanoparticles when they are compressed onto the surface with an applied force of 2 nN. The particle regions outside the orange mask were determined using the mask in D. (F) The output of the supervised inner particle structure ML classifier for the same region as shown in E, indicating the probability of the stiffer inner region of the PG nanoparticles. (G) The output of the supervised inner particle structure ML classifier for the same region as shown in E and F, indicating the stiffer inner (white) and softer outer (grey) regions of the particles using a probability threshold of 50 % to classify the two regions.

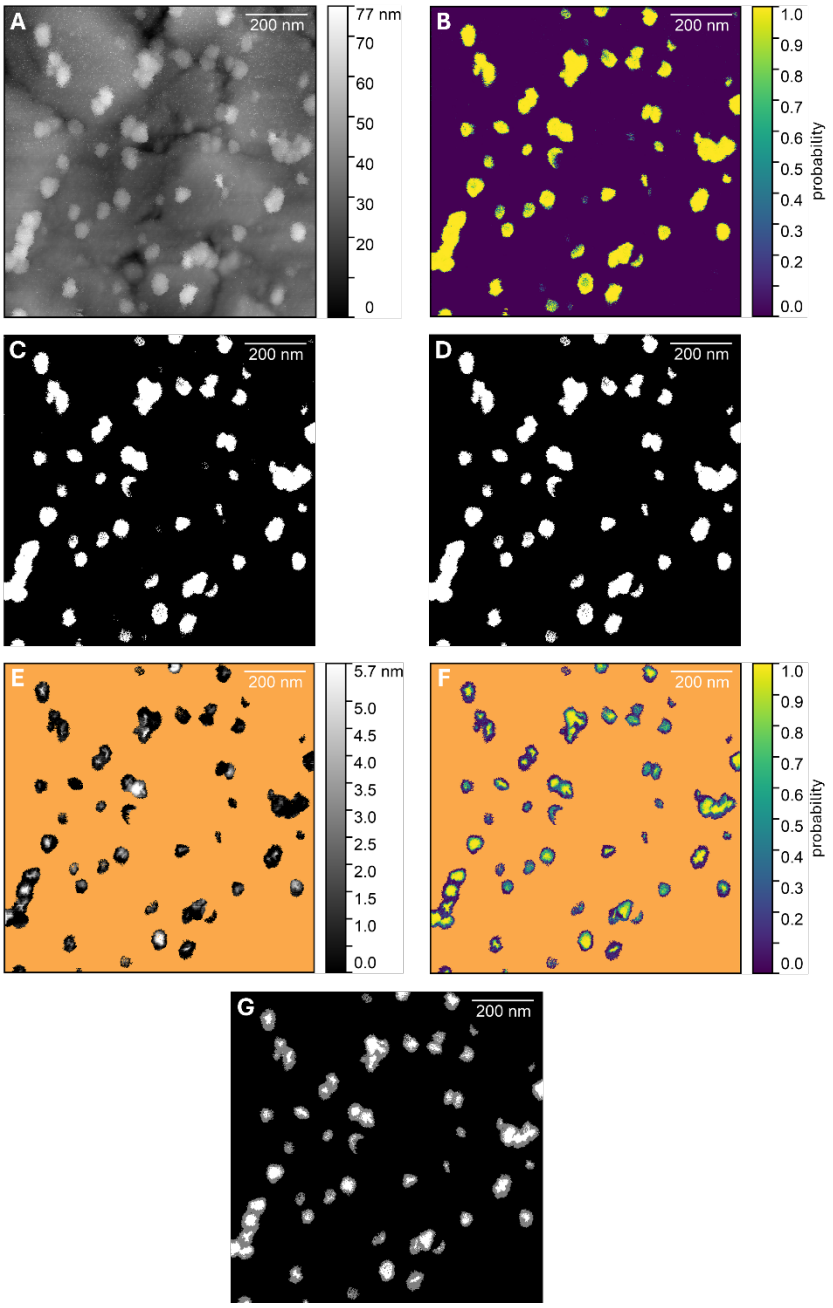

Figure S5: AFM images and supervised ML classifier outputs for a third sample, in addition to those shown in Figures 5 and S4. (A) Contact height image of PG nanoparticles on the 4-MPBA/gold substrate. (B) The probability map output from the supervised particle/substrate ML classifier using the force-distance curves from the measurement in A showing the probability that each pixel corresponds to an interaction of the AFM tip with PG. (C) The classification map output using a threshold probability of 50 % to classify each pixel as corresponding to either PG (white pixels) or the MPBA/gold substrate (black pixels). (D) The same classification map

shown in C with small clusters (< 35 pixels) of white pixels removed to isolate full PG nanoparticles in the classification map. These removed pixels corresponded to only 0.8 % of PG pixels in C. (E) Peak force height image of the same scan in A revealing the stiffer inner structure of the PG nanoparticles when they are compressed onto the surface with an applied force of 2 nN. The particle regions outside the orange mask were determined using the mask in D. (F) The output of the supervised inner particle structure ML classifier for the same region as shown in E, indicating the probability of the stiffer inner region of the PG nanoparticles. (G) The output of the supervised inner particle structure ML classifier for the same region as shown in E and F, indicating the stiffer inner (white) and softer outer (grey) regions of the particles using a probability threshold of 50 % to classify the two regions.

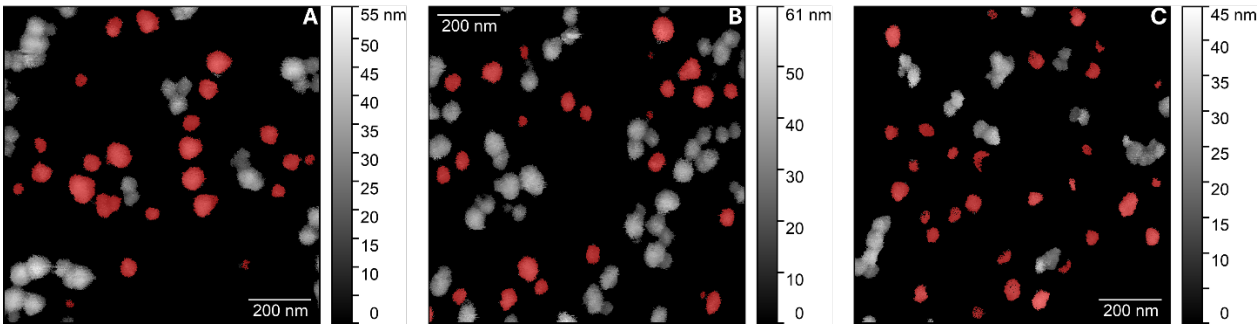

Figure S6: Contact height image of isolated PG nanoparticles for three AFM scans (Figures 5, S4 and S5) determined using the supervised particle/substrate ML classifier on new, unseen data. The red mask indicates individual PG nanoparticles used to determine the average equivalent spherical radius $\bar{r}$.

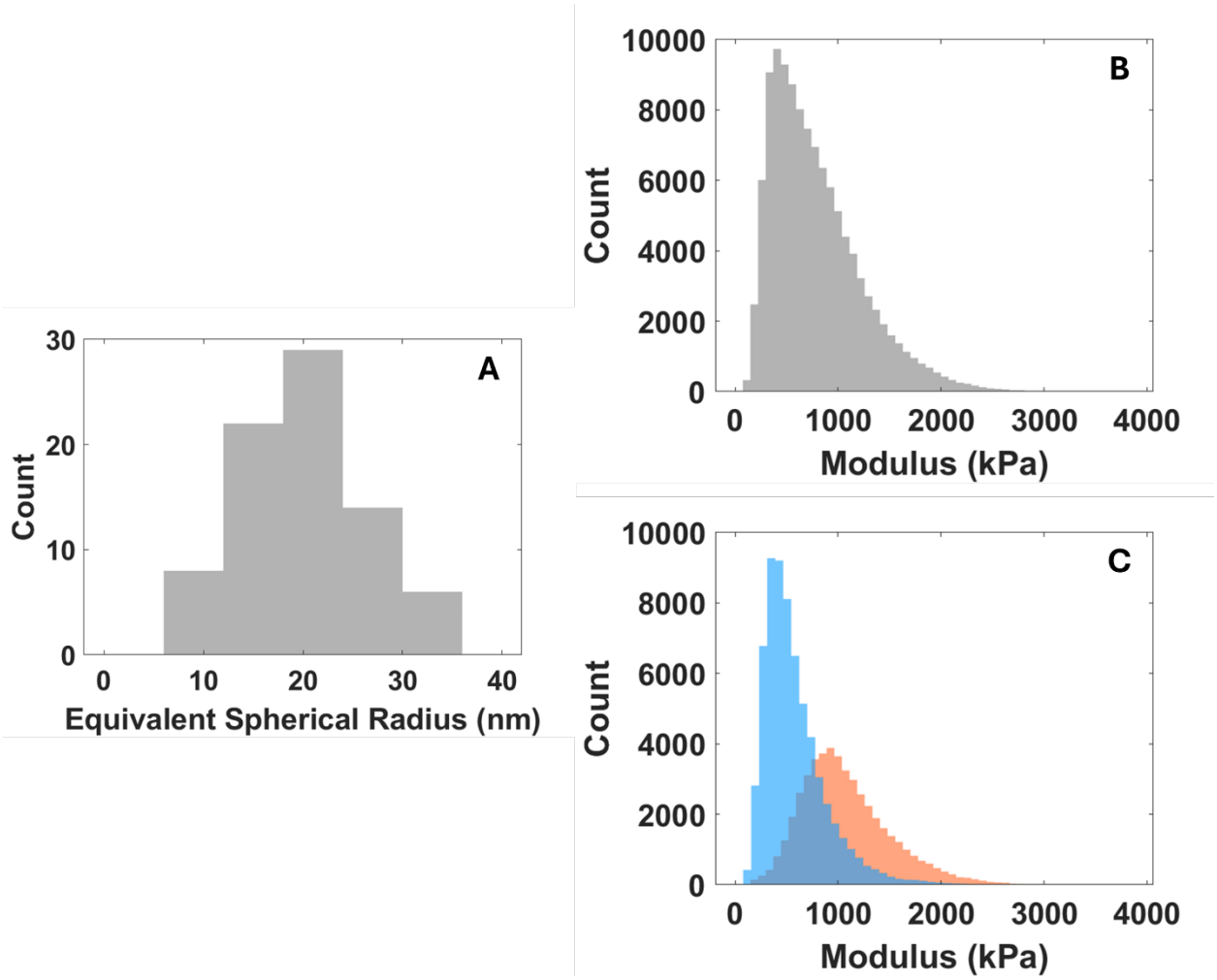

Figure S7: Radius and Young's modulus values for the particles and stiffer inner and softer outer regions determined from the supervised ML classifiers. (A) Histogram of the values of the equivalent spherical radius $\boldsymbol{r}$ calculated for isolated PG nanoparticles measured in Milli-Q water ($\boldsymbol{n}$ = 79) in the new, unseen data shown in Figure S6, corresponding to an average value of $\bar{\boldsymbol{r}}$ = 20.2 nm with a standard error of 0.7 nm. (B) Histogram of the Young's modulus $\boldsymbol{E}$ values measured for pixels ($\boldsymbol{n}$ = 112,897) corresponding to the PG nanoparticles in the new, unseen data shown in Figures 6D, S4D, and S5D. The median $\boldsymbol{E}$ value is 693 kPa with a standard error of the median of 2 kPa. (C) Histogram of the Young's modulus $\boldsymbol{E}$ values measured for the stiffer inner (orange) and softer outer (blue) regions of the PG nanoparticles in the new, unseen data shown in Figures 7D, S4G, and S5G. The median $\boldsymbol{E}$ value is 1006 kPa with a standard error of the median

of 3 kPa for the inner region ($\boldsymbol{n}$ = 47,797) and 503 kPa with a standard error of the median of 2 kPa for the outer region ($\boldsymbol{n}$ = 65,100).

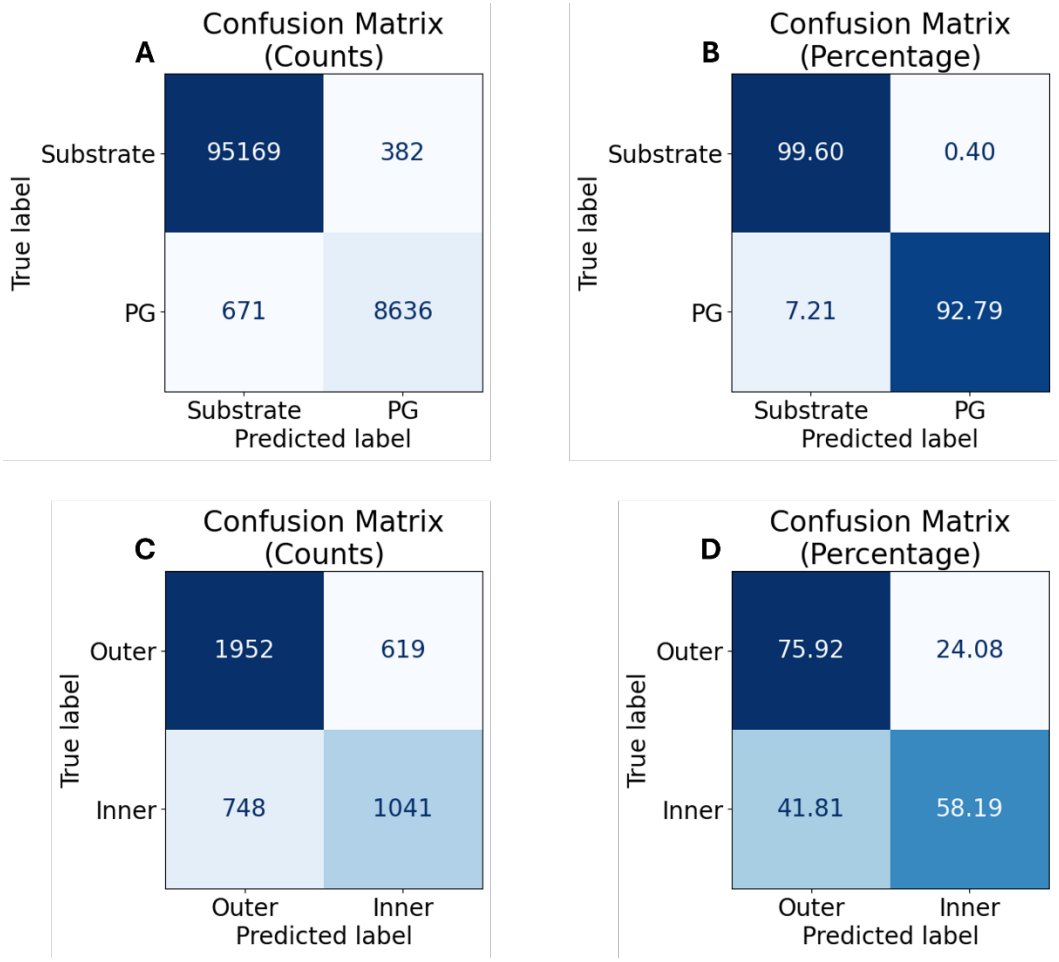

Figure S8: Confusion matrices for the k-means unsupervised ML classifiers. (A) Confusion matrix for the classification of the testing dataset using the unsupervised particle/substrate ML classifier showing the total counts. The y-axis of the confusion matrix is the true label, or manual classification, and the x-axis is the predicted label or classification from the output of the unsupervised particle/substrate ML classifier, and the values display the number of pixels/force-distance curves. For example, the bottom row in A indicates pixels that were manually labeled as corresponding to PG: 671 of these pixels/force-distance curves were classified as corresponding to the substrate, and 8636 of these pixels/force-distance curves were classified as corresponding to PG using the unsupervised particle/substrate ML classifier. (B) Confusion matrix for the

classification of the testing dataset using the unsupervised particle/substrate ML classifier showing the percentage of pixels/force-distance curves for each true label. For example, the bottom row in A indicates the pixels that were manually labeled as corresponding to PG: 7.21 % of these pixels/force-distance curves were classified as corresponding to the substrate, and 92.79 % of these pixels/force-distance curves were classified as corresponding to PG using the unsupervised particle/substrate ML classifier. (C) Confusion matrix for the classification of the testing dataset using the unsupervised inner particle structure ML classifier showing the total counts for each true label. The y-axis of the confusion matrix is the true label, or manual classification, and the x-axis is the predicted label or classification from the output of the unsupervised inner particle structure ML classifier and the values display the number of pixels/force-distance curves. For example, the bottom row in A indicates pixels that were manually labeled as corresponding to the stiffer inner region: 748 of these pixels/force-distance curves were classified as corresponding to the softer outer region, and 1041 of these pixels/force-distance curves were classified as corresponding to the stiffer inner region using the unsupervised inner particle structure ML classifier. (D) Confusion matrix for the classification of the testing dataset using the unsupervised inner particle structure ML classifier showing the percentage of pixels/force-distance curves for each true label. For example, the bottom row in A indicates the pixels that were manually labeled as corresponding to stiffer inner region: 41.81 % of these pixels/force-distance curves were classified as corresponding to the softer outer region and 58.19 % of these pixels/force-distance curves were classified as corresponding to the stiffer inner region using the unsupervised inner particle structure ML classifier.

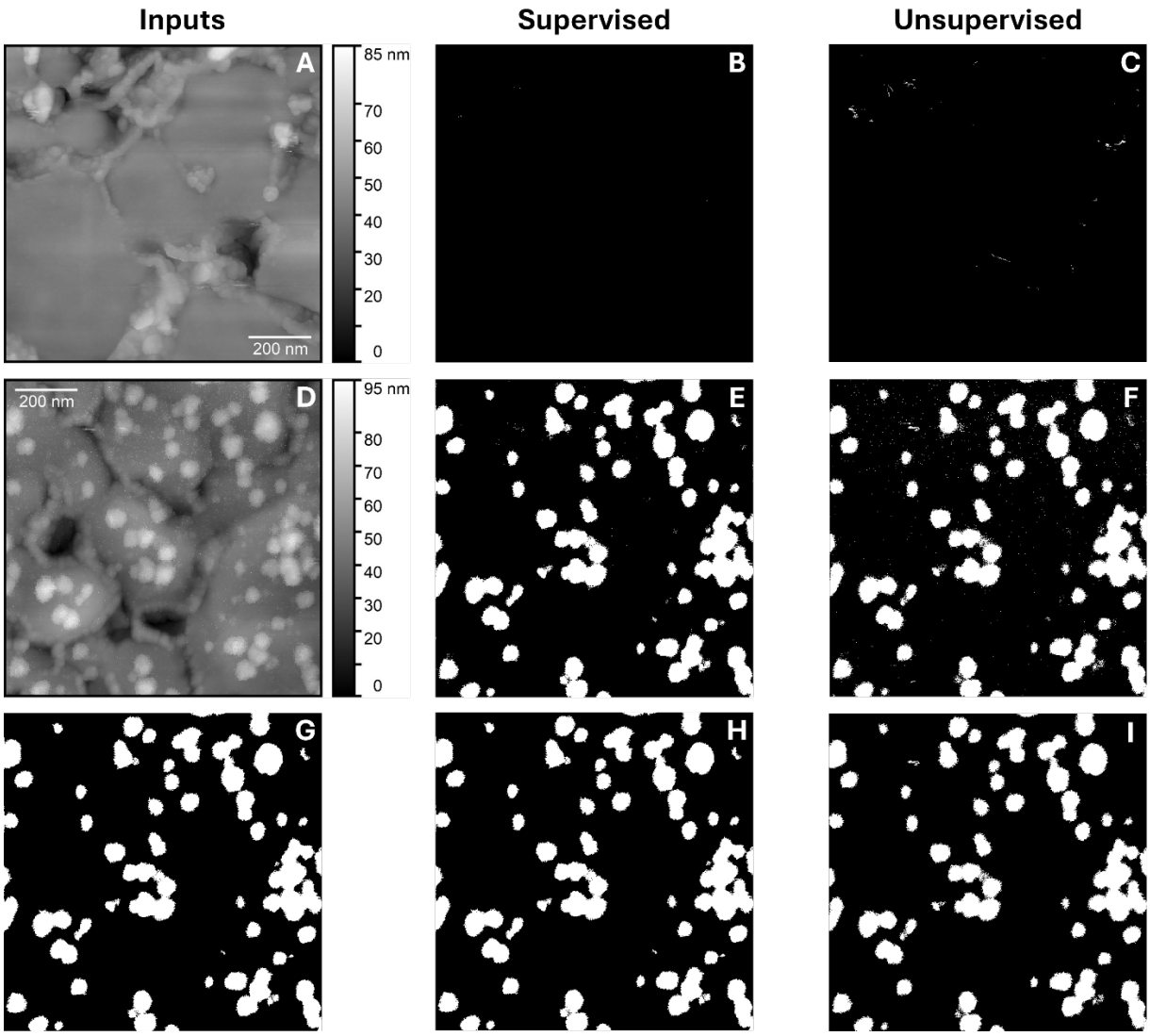

Figure S9: Comparison of the output from the supervised and unsupervised classifiers for the training/testing dataset. (A) AFM height image of the bare hard substrate. (B) The classification map output for the AFM data shown in A for the supervised particle/substrate ML classifier using a threshold probability of 50 % to classify each pixel as corresponding to either PG (white pixels) or the hard substrate (black pixels). (C) The classification map output for the AFM data shown in A for the unsupervised particle/substrate ML classify each pixel as corresponding to either PG (white pixels) or the hard substrate (black pixels). (D) AFM contact height image of PG nanoparticles on a hard substrate (4-MPBA/gold). (E) The classification map output for the AFM data shown in D for the supervised particle/substrate ML classifier using a threshold probability of 50 % to classify each pixel as corresponding to either PG (white pixels) or the hard

substrate (black pixels). (F) The classification map output for the AFM data shown in D for the unsupervised particle/substrate ML classifier where each pixel is classified as corresponding to either PG (white pixels) or the hard substrate (black pixels). (G) Manually created mask of the particles (white) in D with the surrounding hard substrate (black). (H) The same classification map shown in E with the removal of small clusters (< 35 pixels) of white pixels. (I) The same classification map shown in F with the removal of small clusters (< 35 pixels) of white pixels.

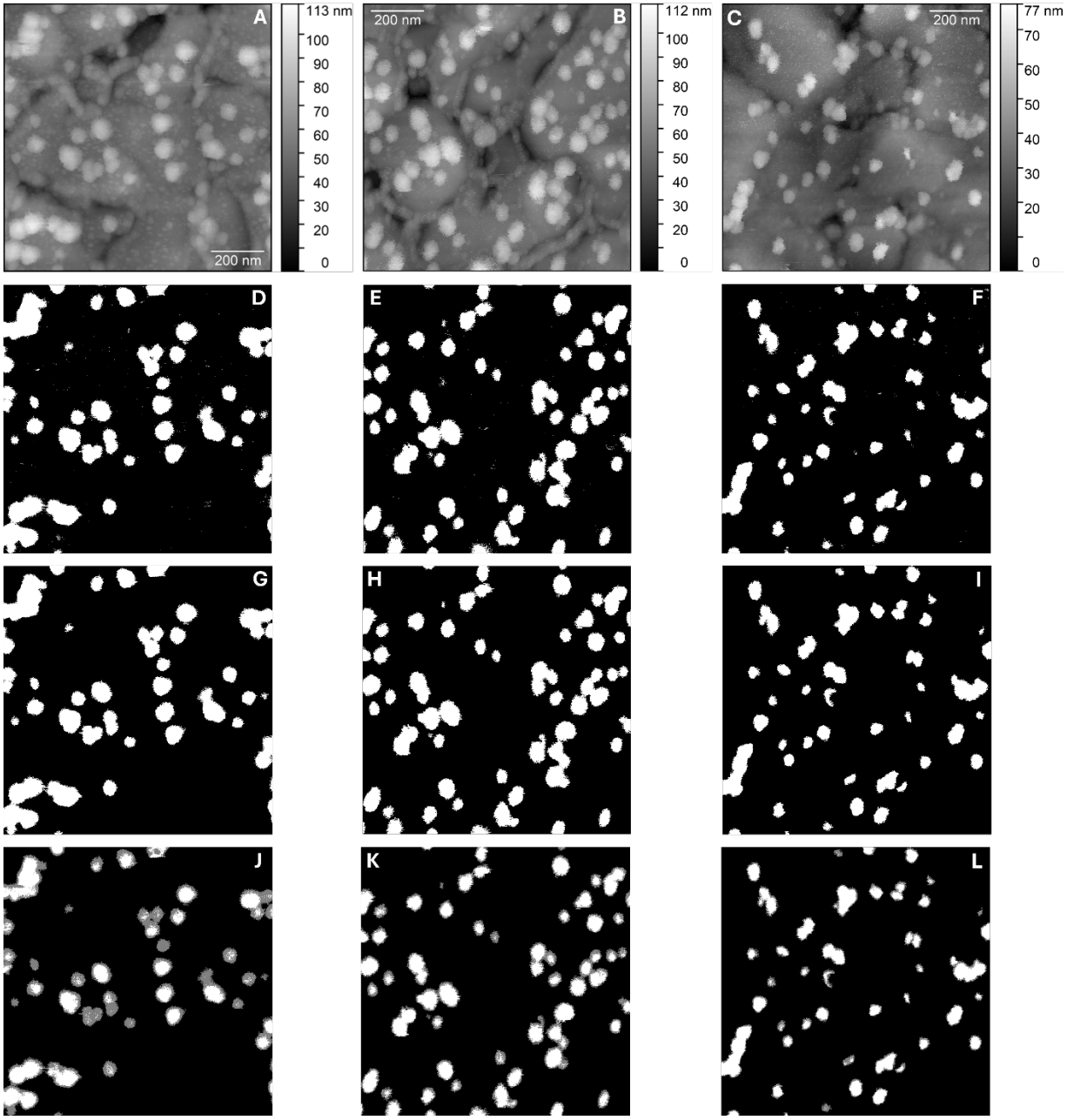

Figure S10: Output results for the unsupervised particle/substrate classifier for the new, unseen data from three AFM-FS measurements (Figures 5, S4 and S5). AFM contact height image of PG nanoparticles on a hard substrate (4-MPBA/gold) are shown for scan 1 (A), scan 2 (B) and scan 3 (C). The classification map outputs for the AFM data in the top row of images from the unsupervised particle/substrate ML classifier are shown for scan 1 (D), scan 2 (E) and scan 3 (F), where each pixel is classified as corresponding to either PG (white pixels) or the hard substrate (black pixels). The same classification maps shown in the second row from the top with the

removal of small clusters (< 35 pixels) of white pixels are shown for scan 1 (G), scan 2 (H) and scan 3 (I). The classification map outputs for the AFM data in the top row of images from the unsupervised internal particle structure ML classifier are shown for scan 1 (J), scan 2 (K) and scan 3 (L), indicating the stiffer inner (white) and softer outer (grey) regions of the particles.

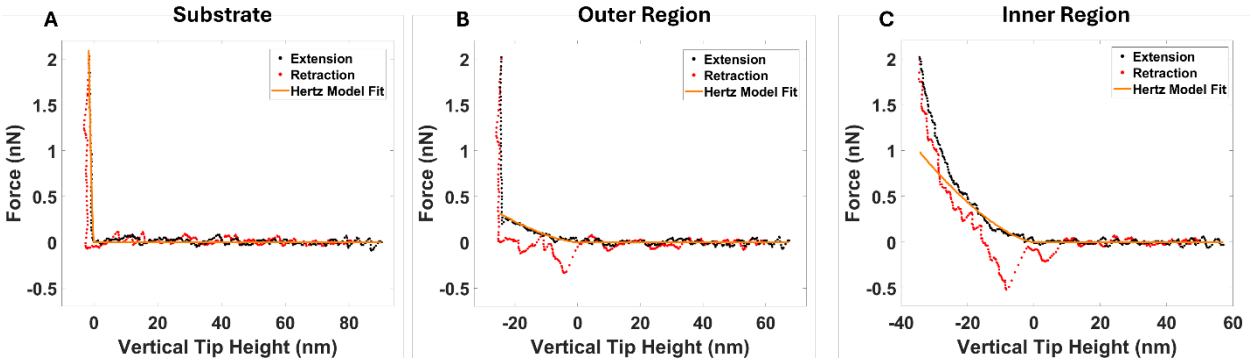

Figure S11: Representative force-distance curves and Hertz model fits for three key regions being classified: the substrate (A), the outer region of the PG nanoparticles (B), and the inner region of the PG nanoparticles (C). The extension (or approach) data is shown in black, the retraction data is shown in red, and the fit to the Hertz model (for 10 nm of the indentation) is shown in orange.