Nanoscale
Advances

ROYAL SOCIETY
OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Nanoscale Adv.*, 2025, **7**, 5612

Predicting the protein corona on nanoparticles using random forest models with nanoparticle, protein, and experimental features

Nicole Vijgen, Karsten M. Poulsen,† Gustavo Sosa Macias‡ and Christine K. Payne *

Received 30th April 2025
Accepted 23rd July 2025

DOI: 10.1039/d5na00425j

rsc.li/nanoscale-advances

Nanoparticles (NPs) present in any biological environment form a “corona” of proteins on the NP surface. This protein corona, rather than the bare NP, determines the biological response to the protein–NP complex. Experiments, especially proteomics, can provide an inventory of proteins in the corona, but researchers currently lack a method to predict which proteins will interact with NPs. The ability to predict the protein corona would aid the design of NPs by decreasing the time and cost of experiments. We describe the development and use of random forest regression and classification models to predict protein abundance and enrichment, respectively, on the surface of NPs using a dataset of NP, protein, and experimental features. These models were trained using data generated in-house through the synthesis and functionalization of NPs with varied core material, surface ligand, diameter, and zeta potential. NPs were incubated with fetal bovine serum, a common protein source for cultured cells, to form a corona, which was characterized by proteomics. Both models identified protein abundance in the serum used to form the corona as the most significant predictor of corona proteins. NP zeta potential and hydrodynamic diameter emerged as the most important NP factors. The random forest regression model was used to test the ability to predict the protein corona of NPs that were excluded from the training data. We highlight the best and worst predictions. These findings offer a machine learning approach to guide experiments.

Introduction

Humans interact with nanoparticles (NPs) directly, in the form of nanomedicines,[1–5] or indirectly, through industrial and environmental exposures.[6–16] During these interactions, proteins adsorb on the surface of NPs, forming a protein “corona.”[2,17–23] The specific proteins that adsorb on the NP surface determine the biological response to the NPs.[1–6,10,19–37] While previous research,[26,29] including our own,[6,17,32,38] has worked to determine how individual NP features such as diameter and zeta potential influence the formation of the protein corona, the ability to predict the composition of the protein corona based on NP and protein features is lacking.

The ability to predict the protein corona would fill an important gap in the design and use of new nanomaterials. Much previous work,[1–5,10,19–23,25,26,28,29,31,37] including our own,[6,24,27,30,32–36] has shown that the protein corona, rather than the bare NP, determines how cells bind, internalize, and respond to the protein–NP complexes. For example, previous experiments using 105 gold NPs with three different gold cores (15 nm, 30 nm, 60 nm) and 67 different ligands (small molecules, polymers, peptides, surfactants) showed that the protein corona was a predictor of cellular association and pointed towards the importance of hyaluronan-binding proteins in the corona.[31] Predicting the protein corona would provide the first step in predicting the cellular response.

Recent work has brought the tools of machine learning (ML) to the challenge of protein corona prediction. For example, previous work has predicted the protein corona formed on single-walled carbon nanotubes using a random forest classifier (RFC).[39] The RFC model was trained on a dataset of human cerebrospinal fluid and blood plasma proteins, characterized by proteomics and physicochemical features derived from protein sequences in UniProt. The model was then used to predict the adsorption of human cerebrospinal fluid and blood plasma proteins on the same single-walled carbon nanotubes. This was a significant result showing that the RFC model was effective in predicting which proteins would adsorb on single-walled carbon nanotubes. The model also identified key protein features that were associated with a higher binding affinity for single-walled carbon nanotubes. These protein features included a high content of solvent-exposed glycines and a high percentage of non-structure associated amino acids, those amino acids not associated with helices, sheets, or turns. The

Thomas Lord Department of Mechanical Engineering and Materials Science, Duke University, Durham, North Carolina, 27708, USA. E-mail: christine.payne@duke.edu

† Current address: Donaldson Company, Bloomington, Minnesota, USA 55431.

‡ Current address: Duke University School of Medicine, Durham, North Carolina, USA 27710.

Open Access Article. Published on 29 July 2025. Downloaded on 5/10/2026 5:52:52 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 29 July 2025. Downloaded on 5/10/2026 5:52:52 AM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

model was also generalizable to other nanomaterials, such as polystyrene NPs (pNPs). Another RFC model was used to predict the protein corona formed from human serum proteins adsorbed on nanostructures formed from DNA.[40] A separate study using a RFC model examined yeast protein enrichment on silver NPs (10 nm and 100 nm) as a function of protein, NP, and solvent features using a previously published database of yeast protein enrichment on silver NPs.[41] This model illustrated the ability to utilize existing proteomic data to train new ML models, indicating the potential power of data sharing and data scraping. As data sharing becomes increasingly common, researchers have access to large quantities of proteomics data from other research groups, enabling the generation of large datasets that capture multiple protein and NP features for use in model development. This use of external data was again demonstrated using a random forest regression (RFR) model with 652 different NPs (silver, gold, iron oxide, titanium dioxide, silicon dioxide, liposomes, and polystyrene) with coronas formed from human serum, bovine serum, or human plasma.[42] The scraped and analyzed data featured a combination of qualitative factors such as NP type and shape, surface modifications, and dispersion mediums, and quantitative factors such as NP diameter and zeta potential. The RFR model was successful in predicting the composition of the protein corona across NPs and protein sources, suggesting the viability of scraped data to train regression models.

We describe the development and use of two different ML models, RFR and RFC, to predict protein corona composition based on a combination of NP features, protein features, and experimental features. In comparison, previous research using ML to predict the protein corona focused on protein features or a combination of protein features and NP features with proteins as the dominant factor.[39,41] Our models provide a new focus on NP features and experimental features. The RFR model predicts individual protein abundances in the corona as a quantitative, continuous value. The RFC model predicts whether a specific protein is enriched or depleted relative to its abundance in the serum, a categorical value. In addition to building on previous protein corona models,[39–42] random forest models were selected as they provide a connection between model outputs and physical interpretation of the results.

We developed the RFR and RFC models in parallel to compare their performance. To ensure uniform data handling, all of the data used in our models was generated in-house using a semi-automated workflow of corona formation, purification, and characterization using a liquid-handling robot along with a low-cost proteomics protocol to characterize these samples.[43,44] To train and test our ML models, we generated 11 NPs with varying features (core material, surface ligand, diameter, and zeta potential) to probe the relationship between NP properties and the protein corona. Proteomics was used to characterize protein coronas on the 11 NPs following incubation with fetal bovine serum (FBS; 10% and 100%). We selected FBS as the protein source for our protein–NP samples as it is a common nutrient source for cells in culture, well-documented in protein sequence databases, and frequently used in other protein corona studies.[45–48] Protein features (*e.g.* secondary structure, percentage of polar amino acids) were derived from protein sequence data. Experimental features included NP and protein incubation concentrations and separation method. This combination of NP, protein, and experimental features generated an input dataset of 84 total features that was used to train and validate predictions for RFR and RFC models.

We find that both the RFR and RFC models had high performance metrics. The RFR model identified 61 significant features and the RFC model identified 16 significant features for corona prediction. Thirteen features were shared between the two models, with protein abundance in FBS selected as the most significant feature. NP zeta potential and hydrodynamic diameter were the next most important features. We then tested the ability of these models to predict the protein corona of previously unseen NPs using the RFR model. The resulting R^2 values for corona prediction ranged from 0.45–0.88.

These results suggest that both regression and classification models can serve as computational tools to predict protein–NP interactions. The ability to predict a protein corona based on NP features, which are relatively inexpensive to determine compared to full biological experiments, and protein features, which are tabulated in existing databases, would reduce the cost and the time of current experimental methods. Previous work has shown that corona formation can lead to mis-targeting of NPs,[49] masking of targeting ligands,[50] and altered biodistribution of NPs.[51] In comparison to these negative outcomes, NPs can also be designed to select for specific corona proteins with beneficial properties such as targeted drug delivery to specific organs.[3,23,52–54] In the long term, we hope that a detailed NP characterization will allow for the prediction of, for example, the toxicity of new nanomaterials with fewer cell and animal experiments. This would reduce costs and increase throughput in the development and use of new nanomaterials.

Experimental

Synthesis and functionalization of magnetic NPs (mNPs)

Iron oxide magnetic NPs (mNPs) were synthesized using previously published protocols.[55,56] In brief, 40 mL of ethylene glycol (#324558, Sigma-Aldrich, St. Louis, MO), 1.3 g $FeCl_3 \cdot 6H_2O$ (#236489, Sigma-Aldrich), 0.52 grams of trisodium citrate (#S4641, Sigma-Aldrich), and 2.4 grams of sodium acetate (#S2889, Sigma-Aldrich) were mixed in an Erlenmeyer flask (100 mL) with a magnetic stir bar. The addition of deionized (DI) water at this step is used to control the diameter of the NPs. The addition of 4 mL DI water resulted in small NPs (82 nm, Table 1). Without the addition of water, large NPs (182 nm, Table 1) are produced. The Erlenmeyer flask was covered with foil and stirred (1 h). This solution was transferred into a Teflon-lined stainless steel reaction flask (100 mL) and heated to 200 °C for 10 hours. The reaction flask was allowed to cool to room temperature (RT). The resulting NPs were washed three times with ethanol to remove contaminants from previous steps. A magnet was used to remove the NPs from suspension during washes. The washed NPs were suspended in a minimal amount of ethanol (1 mL), transferred to a 1.5 mL centrifuge tube, and dried under a stream of nitrogen overnight.

Open Access Article. Published on 29 July 2025. Downloaded on 5/10/2026 5:52:52 AM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Table 1 Characterization of NP core material, diameter (d_{TEM} and d_h), polydispersity index (PDI), and zeta potential (ZP)

NP	Core	Ligand	d_{TEM} (nm)	d_h (nm)	PDI	ZP (mV)
Citrate-mNP$_S$	mNP	Citrate	82 ± 36	149 ± 3	0.11 ± 0.01	−42 ± 6
Citrate-mNP$_L$	mNP	Citrate	182 ± 48	229 ± 11	0.19 ± 0.02	−49 ± 4
PEI-mNP$_S$	mNP	Polyethyleneimine	82 ± 36	226 ± 62	0.22 ± 0.08	29 ± 4
PEI-mNP$_L$	mNP	Polyethyleneimine	182 ± 48	282 ± 78	0.25 ± 0.09	39 ± 4
PVP-Au-mNP$_S$	Gold-mNP	Polyvinylpyrrolidone	98 ± 60	271 ± 17	0.31 ± 0.04	−12 ± 4
PVP-Au-mNP$_L$	Gold-mNP	Polyvinylpyrrolidone	244 ± 62	316 ± 85	0.22 ± 0.04	−11 ± 3
PEI-Au-mNP$_S$	Gold-mNP	Polyethyleneimine	98 ± 60	229 ± 17	0.19 ± 0.03	12 ± 3
PEI-Au-mNP$_L$	Gold-mNP	Polyethyleneimine	244 ± 53	291 ± 9	0.15 ± 0.04	12 ± 1
PEG-Au-mNP$_L$	Gold-mNP	Polyethylene glycol (5k)	244 ± 53	610 ± 90	0.38 ± 0.04	−3 ± 3
COOH-pNP	Polystyrene	Carboxylate	200 ± 10	221 ± 2	0.02 ± 0.01	−63 ± 9
PEG-pNP	Polystyrene	Polyethylene glycol (2k)	200 ± 23	266 ± 7	0.13 ± 0.06	−7 ± 3

The mNPs were functionalized using previously published protocols.[56] To achieve an adsorbed coating of PEI, dry iron mNPs (5 mg mL^{-1}) were added to polyethyleneimine (PEI; 0.1 mM (aq, #408727, Sigma-Aldrich)). The mixture was shaken (1 h) on a rotary shaker at RT. The mixture was washed three times with water using a magnet to separate the iron mNPs. The PEI coating was verified by a positive zeta potential.

Gold nanoseeds were synthesized following a previously published protocol.[56,57] In brief, 44 mL of DI water, 3 mL of 100 mM sodium hydroxide (aq, #S8045, Sigma-Aldrich), and 1 mL of 50 mM tetrakis-(hydroxymethyl) phosphonium chloride (#404861, Sigma-Aldrich) were mixed in an Erlenmeyer flask (100 mL). After mixing for 5 minutes with a magnetic stir bar, 1.5 mL of 25 mM gold(III) chloride trihydrate (#520918, Sigma-Aldrich) was added. After the addition of the gold salt, the solution turned a deep red, signifying the formation of gold nanoseeds.

The PEI-mNPs were coated with gold nanoseeds by adding the PEI-mNPs (1 mg mL^{-1}) to a solution of gold nanoseeds (50 mL, 10 nM). The gold nanoseeds were grown into a gold shell to get a more complete surface coating. The growth of the gold shell was stabilized by adding NPs functionalized with gold nanoseeds (25 μg mL^{-1}) to polyvinylpyrrolidone (PVP; 9.85 mg mL^{-1}, #PVP40, Sigma-Aldrich). After vortexing, hydroxylamine (75 μg mL^{-1}, #159417, Sigma-Aldrich) and gold(III) chloride trihydrate (75 μg mL^{-1}, #520918, Sigma-Aldrich) were successively added. The color of the solution took on a bluish-purple tint within minutes of adding the gold mixture. The resulting PVP-Au-mNPs were separated using a magnet, washed three times with DI water, and resuspended in DI water. The gold shell growth was confirmed using transmission electron microscopy (TEM), as described in NP characterization.

To obtain NPs with a positive zeta potential and vary the ligand of the NPs, PVP was exchanged for PEI by shaking the PVP-Au-mNPs (1 mg mL^{-1}) in PEI 0.1 mM (aq) for 1 hour. Following ligand exchange, the NPs were removed from suspension using a magnet, washed four times with DI water, and resuspended in DI water. Ligand exchange was confirmed by the zeta potential of the resulting NPs.

To obtain a near-neutral surface charge and vary the ligand of the NPs, PVP was displaced with thiolated polyethylene glycol (PEG) by shaking the PVP-Au-mNPs in a thiol PEG solution (10 mM, A3029-1/M-SH-5000, JenKem Technology, Plano, TX) for 1 hour. Following ligand exchange, the NPs were removed from suspension using a magnet, washed four times with DI water, and resuspended in DI water. Ligand exchange was confirmed by the zeta potential of the resulting NPs.

PEGylation of polystyrene NPs (pNPs)

Commercially available pNPs (200 nm, carboxylate-modified, #C37486, Thermo Fisher Scientific, Waltham, MA) were conjugated with PEG using *N*-(3-dimethylaminopropyl)-*N*′-ethylcarbodiimide (EDC) hydrochloride. pNPs were first washed by diluting 10-fold with DI water and separating *via* centrifuge. To conjugate with PEG, 100 μL of 4 mg mL^{-1} washed pNP were added to 200 μL methoxy-PEG-amine 2k (50 mg mL^{-1}, #A3071, JenKem Technology) in 4-morpholineethanesulfonic acid hemisodium salt (MES; 25 mM, #M0164, Sigma-Aldrich). The NP PEG mixture was vortexed and shaken on a rotary shaker for 5 minutes before adding 40 μL of *N*-(3-dimethylaminopropyl)-*N*′-ethylcarbodiimide hydrochloride (EDC; 45 mg mL^{-1}, #E7750, Sigma-Aldrich). The mixture was rotary shaken for 30 minutes before being washed three times with phosphate-buffered saline (PBS) *via* centrifugation (18 000 rcf, 15 min) to remove excess PEG and EDC. PEGylation was confirmed by zeta potential.

NP characterization

NP diameter was measured with TEM and dynamic light scattering (DLS). TEM was carried out using either a Tecnai G^2 TWIN TEM (FEI, Hillsboro, OR) at the Shared Materials Instrumentation Facility at Duke University or using the Supra 25 FESEM (Zeiss, Oberkochen, DEU) at the UNC Microscopy Services Laboratory. All samples were prepared by drop casting on 400 mesh copper grids (#CF400-Cu, Electron Microscopy Sciences, Hatfield Township, PA) and drying at RT for 12–18 h. NP diameters were measured using ImageJ.[58] Average and standard deviations are reported for all measurements.

Hydrodynamic diameter, polydispersity index, and zeta potential of the NPs (10–100 μg mL^{-1} in PBS diluted 1 : 100 in DI water) were measured using DLS (Zetasizer, Malvern Instruments, Worcestershire, England). Measurements were carried out with three distinct samples. Each measurement was

Open Access Article. Published on 29 July 2025. Downloaded on 5/10/2026 5:52:52 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

performed for 12–30 runs. The average and standard deviation are reported for all measurements. Electrophoretic mobility was converted to zeta potential using the Smoluchowski approximation.

Liquid handling robot

A liquid handling robot (OT-2, Opentrons, Brooklyn, NY) with a magnetic baseplate was used to automate protein corona formation and isolation, as described previously.[43,44] Protocol scripts were written in Python using Opentrons API v2.12. Pipette tips (300 μL, single and multi) and tip racks were purchased from Opentrons to verify compatibility and calibration. The locations of each reagent and sample were designated in the script and appropriately positioned before running the robot. Most experiments used a 96-well plate with three or six replicates, as specified in the text. Two wells were used for background subtraction within a row of eight wells.

Protein corona formation and quantification

A protein corona was formed by incubating NPs (2.4–5 mg mL^{-1}) in 10–100% solutions of FBS (#10437028, Thermo Fisher Scientific) diluted in PBS. The incubations were performed at RT on a microplate shaker for 30 minutes. Coronas formed on the mNPs were generated and purified by the liquid handling robot using magnetic pull-down separation steps. The pNP samples, which are not magnetic, were processed manually. To remove unbound proteins, the NPs were "washed." Each wash step consisted of a magnetic pull-down or centrifugation (18 000 rcf, 15 min), removal of the supernatant, and then resuspension in an equal volume of PBS. When done manually, three wash steps were performed, while when done with the liquid handling robot, six washes were performed. This number of washes has previously been confirmed to remove the excess proteins in the solution.[27,44] The hard corona is defined as the protein that remains bound to the NPs with minimal protein detected in the supernatant, as described previously.[43,44]

Protein concentration was measured with the Pierce 660 nm Protein Assay Reagent (referred to as a 660 nm assay; #2260, Thermo Fisher Scientific) with the addition of Ionic Detergent Compatibility Reagent (#22663, Thermo Fisher Scientific) according to the manufacturer's instructions. The concentration of protein present in the hard corona was determined by removing the proteins from the NPs by incubating with sodium dodecyl sulfate (SDS) buffer (5% w/v, #L3771, Sigma-Aldrich) for 30 minutes at RT. Protein concentration was then determined by measuring absorbance at 660 nm using a plate reader (SpectraMax iD3, Molecular Devices, San Jose, CA). A residual amount of protein is resistant to SDS removal independent of the duration of SDS incubation, as shown previously.[43]

Experimental features

Three experimental features (NP concentration, protein incubation concentration, method of free protein removal) were tracked to investigate how these features would impact the protein corona. NP concentrations ranged from 2.4–5 mg mL^{-1}. Protein concentrations ranged from 10–100% solutions of FBS diluted in PBS (100% FBS corresponds to 40 mg mL^{-1}). Two methods were used to separate NPs from unbound protein: magnetic separation was used for mNPs and centrifugation (18 000 rcf, 15 min) was used for pNPs.

Proteomic analysis

Samples for proteomics were digested using a modified S-Trap mini column (Protifi, Farmingdale, NY) protocol. Proteins were removed from the NP surface by incubating with SDS buffer for 30 minutes. Protein concentration was determined using the 660 nm assay. Samples were pooled to load a minimum of 25 μg of protein on each S-Trap. Two modifications were made to the S-trap protocol: dithiothreitol (DTT; #R0861, Thermo Fisher Scientific) and iodoacetamide (IAM; #I1149, Sigma-Aldrich) were used as the reducer (20 mM) and alkylator (40 mM), respectively. DTT and IAM are commonly used for proteomics and are recommended substitutions. Following the completion of the S-trap protocol, the resulting digested proteins were lyophilized and stored at −20 °C until proteomic analysis.

Proteomic analysis was carried out in the Proteomics and Metabolomics Core Facility, part of the Duke Center for Genomics and Computational Biology, as described previously.[43,44] In brief, digested samples were analyzed using LC-MS/MS with ≤ 25 mg of digested protein injected. MicroFlow LC was performed with an ultra-performance liquid chromatography (UPLC, 1 mm × 100 mm, M-Class, Waters Corporation; 80 μL min^{-1}) column and a 17 minutes total elution time. The column was run with an acetonitrile gradient (5–40%) with 0.1% formic acid. Peptide fragments were analyzed using in-line tandem mass spectrometry (Orbitrap Fusion Lumos, Thermo Fisher).

We analyzed the LC-MS/MS data using MaxQuant (v2.5.2.0, Max Planck Institute, Munich, Germany), an open-source software designed to qualitatively and quantitatively analyze mass spectrometry data.[59,60] The raw LC-MS/MS spectra were searched, using their integrated Andromeda search engine, against the Swiss-Prot Bovine (6046 proteins) canonical protein knowledge base from UniProt, accessed on May 22nd, 2024.[61] A custom contaminants file was used, which contained a relevant subset of the Common Repository of Adventitious Proteins (cRAP) database.[62] For protein and peptide quantification and identification, default MaxQuant parameters were used: a 0.01 false discovery rate, a minimum peptide length of 7 amino acids, a maximum peptide length of 25 amino acids, oxidation, acetyl groups as variable modifications, and carbamidomethyl as a fixed modification. The Intensity method in MaxQuant was used for abundance quantification calculations.

The resulting proteomic data was analyzed and filtered in Perseus (v2.0.11, Max Planck Institute). Proteins were excluded if considered contaminants, quality control standards, or only identified by site. Data normalization was performed in Python (v3.11, Python Software Foundation, Beaverton, OR). A quantitative internal standard was not used for these experiments. To correct for any change in performance or differences in protein loading, each sample was normalized to itself by dividing by the

Open Access Article. Published on 29 July 2025. Downloaded on 5/10/2026 5:52:52 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

mean of the interior 80% of the protein intensities.[63] Each sample was scaled to have the same average. We report these values as percent normalized abundance. Fold change for each protein was calculated by taking the log base 2 of the normalized corona abundance divided by serum abundance. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium *via* the Proteomics Identification Database (PRIDE) partner repository with the dataset identifier PXD053700 and 10.6019/PXD053700.[64] In total, 92 proteins were identified in the sample of FBS used to form the coronas. This is likely due to the overwhelming signal from albumin in the FBS, which limits the detection of lower abundance proteins. In comparison, 369 proteins were identified in the corona samples including the 92 also identified in FBS. Formation of a corona serves as an enrichment step allowing the identification of more unique proteins than in FBS alone.

Data organization and processing for ML

For use in ML, the physical properties of the NPs and proteins are described as features. Our database consists of features grouped into three categories: protein features (Table S1), NP features (Table S2), and experimental features (Table S3). Entry (accession number), sequence, length, and mass of proteins were sourced from the basic canonical protein information for *Bos taurus* (Taxonomy Identifier 9913) from the Swiss-Prot knowledge base from UniProt.[61] The entry feature was used to identify and label the proteins within our samples. NetSurfP3.0 was used to calculate additional protein features. This natural language processing model predicts the protein structure and returns results for each amino acid by feeding in sequence information for each protein.[65,66] The sequence information accessed from UniProt was used with NetSurfP3.0 to predict exposed amino acids, secondary structure, accessible surface area, hydrophobicity, and polarity for each protein sequence. Results for each protein were obtained using a Python script adapted from published code to capture the complete proteome.[39] In comparison to this previous code, we omitted calculations of the proportion of a specific amino acid exposed on the protein surface relative to the total number of that amino acid in the protein, as our model uniquely incorporates NP properties as features, and we aimed to focus on calculations that emphasize the relative composition of the protein surface. By using the percent exposed amino acid of a specific amino acid divided by the total exposed calculation, we specifically investigated how NP features influence the protein corona, as only the exposed amino acids interact with the NPs, providing a more accurate representation of these interactions. The BioPython Protein Analysis library (v.1.8.1) calculated the remaining protein features based on sequence data, such as percent amino acid composition, aromaticity, instability index, flexibility metrics, GRAVY score, and secondary structure.[67] The BioPython Protein Analysis module and NetSurfP predictions cannot account for proteins that have abbreviations for groups of amino acids in their sequence or that contain the amino acid selenocysteine. To account for these instances, the sequence data was cleaned by replacing the unspecified amino acids with the most common amino acid, leucine, and replacing selenocysteine with cysteine. The physicochemical features of the proteins were combined with the proteomic protein abundance data (69 features, Table S1), NP features (12 features, Table S2), and experimental features (3 features, Table S3). NP ligand and core material were One-Hot encoded. These features produced a dataset containing 84 features for training the RFR and RFC models described below. The code is available on GitHub (**https://www.github.com/nvijgen/ProteinCoronaPredict_PayneLab.git**).

Random forest regression (RFR)

A RFR model, implemented using scikit-learn, was used to predict the individual protein abundance values that define the protein corona.[68] The model utilized mean squared error as the scoring criterion. One hundred decision trees and a fixed random seed were used. The fixed random seed kept the same samples divided into data train and test sets splits across runs for reproducibility. The dataset consisted of 369 proteins identified in the coronas across all NP samples. Before use in training, protein abundance values were $\log_2$-transformed to normalize the distribution. Zeros were represented by the smallest nonzero value to prevent errors associated with log transformations.

After data pre-processing, recursive feature elimination with k-fold cross-validation (ten folds) was used to identify the most important predictive features on the training data split (90%) (Table S4). This feature selection process was iterative, with a step size of one, and maintained a minimum of one feature until completion. A custom scoring method was used, which combined mean squared error with a penalty for feature variation and quantity to refine feature selection effectively. This custom approach also integrated evaluation criteria (R^2, mean squared error, Pearson correlation coefficient, and Spearman's rank correlation coefficient) across multiple data folds, providing a comprehensive assessment of model performance on the selected subset of features across ten folds.

The selected features were then used to predict on the test data split (10%), specifically predicting the $\log_2$-transformed abundance values. The performance of the model was evaluated using the custom scoring method which assessed the accuracy of the predictions with respect to the known data values in the test split, hidden during model training.

Random forest classification (RFC)

An RFC model, also implemented using scikit-learn, was employed to predict whether proteins are enriched or depleted in the corona, classified as ‘1’ (enriched) or ‘0’ (depleted), by utilizing the significant features identified during the model feature selection process.[68] Binary classification is necessary for RFC, as it allows the model to categorically distinguish between the two possible states of protein entries, enriched or depleted, based on the identified predictive features. In comparison, the RFR model predicts continuous values. As we expect the continuous values predicted by RFR will be more useful to other

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 29 July 2025. Downloaded on 5/10/2026 5:52:52 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

researchers than the binary value obtained with RFC, the RFC results are provided in the SI.

The RFC model was trained on protein enrichment data, which was calculated based on the $\log_2$-transformed ratio of corona abundance to serum abundance. Handling of protein data was identical to that described for the RFR model. Enrichment values were classified with their respective binary categorization for future RFC classification (*i.e.* thresh parameter set to zero).

Recursive feature elimination with k-fold cross-validation (ten folds) was implemented to identify the most important features for corona prediction (Table S5). Similar to RFR, a step size of one was used to iteratively eliminate features until the optimal number of features was determined. This feature selection process was evaluated with standard classifier model evaluation metrics; area under the receiver operatoring characteristic curve (AUROC; measure of the ability of the model to distinguish between classes, considering both sensitivity and specificity), accuracy, precision (positive predictive value), *F*1 score (harmonic mean of precision and recall), and recall (true positive rate). The optimal features identified were subsequently used to train the classification model and predict on a train-test split of 90/10. Predictions were assessed using the same evaluation metrics used to assess feature selection.

Results and discussion

NP characterization

mNPs were synthesized and functionalized as described in Methods (Table 1). Diameter and functionalization ligand were chosen to provide a range of diameters, functional groups, and zeta potentials for the training data. NPs diameter was measured by both TEM (d_{TEM}) and dynamic light scattering to determine hydrodynamic diameter (d_{h}) and polydispersity index (PDI). Two diameters of magnetic NPs were used for functionalization. The smaller diameter (82 ± 36 nm by TEM) is denoted as $\mathrm{mNP_S}$. The larger diameter (182 ± 48 nm by TEM) is denoted as $\mathrm{mNP_L}$. The mNPs were synthesized with citrate ligands (citrate-mNP) and then functionalized with PEI (PEI-mNP) to provide a positive zeta potential. Functionalization with ligands bound to gold seeds provided additional functional groups and zeta potentials (PVP-Au-mNP, PEI-Au-mNP, PEG-Au-mNP). PEG was of special interest for the resulting ~0 mV zeta potential and relevance to the biomedical community (PEG-Au-$\mathrm{mNP_L}$, PEG-pNP).[69–72] PEG-Au-$\mathrm{mNP_S}$ were generated, but had too little protein present in the corona for proteomics and are not described in the text. pNPs were used to provide an additional core material (COOH-pNP).

Composition of the protein corona

Protein coronas were formed by incubating NPs (2.4–5 mg $\mathrm{mL^{-1}}$) in 10 or 100% (100% FBS equivalent to 40 mg $\mathrm{mL^{-1}}$ of protein) solutions of FBS diluted in PBS for 30 minutes at RT. Samples were then washed three times to remove unbound proteins, as described in Experimental using previously published protocols.[43,44] Proteomic analysis was used to determine

Table 2 Normalized abundance (%) of the top ten most abundant proteins in the protein corona for NPs incubated with FBS (100%, 30 min). Data for NPs incubated with 10% FBS is in the SI (Table S6). The list of proteins is ordered based on protein abundance in FBS alone. The rank order of proteins present in FBS alone is shown in parentheses

Protein	FBS (rank)	Citrate-$\mathrm{mNP_L}$	Citrate-$\mathrm{mNP_S}$	PEI-$\mathrm{mNP_L}$	PEI-$\mathrm{mNP_S}$	PEI-Au-$\mathrm{mNP_L}$	PEI-Au-$\mathrm{mNP_S}$	PVP-Au-$\mathrm{mNP_L}$	PVP-Au-$\mathrm{mNP_S}$	COOH-pNP	PEG-pNP	PEG-Au-$\mathrm{mNP_L}$
Albumin	54.3 (1)	33.7	27.3	42.7	43.5	36.9	23.4	41.4	32.5	19.9	16.8	46.0
Alpha-2-HS-glycoprotein	15.7 (2)	12.6	12.5	13.8	14.1	13.2	8.3	12.1	10.2	7.6	4.0	14.4
Alpha-1-antiproteinase	7.55 (4)	5.5	3.1	6.2	5.9	3.8	2.2	6.0	4.3	3.0	4.1	5.7
Alpha-2-macroglobulin	2.47 (5)	1.8	1.7	3.0	2.5	1.9	2.3	2.8	2.4	0.96	1.1	1.9
Apolipoprotein A-I	1.20 (7)	1.3	0.8	1.7	1.6	0.9	1.0	1.5	1.8	4.9	13.7	1.3
Complement C3	0.78 (9)	3.9	9.1	1.0	1.9	0.5	0.5	1.2	1.1	0.5	0.3	0.8
Hemoglobin subunit alpha	0.51 (14)	1.3	1.5	2.0	1.2	1.0	1.8	3.4	6.8	13.8	40.0	2.9
Inter-alpha-trypsin inhibitor heavy chain H3	0.15 (26)	0.1	0.2	2.8	2.3	7.5	13.2	0.2	0.1	0.1	0.8	0.1
Prothrombin	0.09 (29)	1.8	2.0	2.3	2.3	5.2	7.7	0.8	1.0	6.4	0.7	0.8
Apolipoprotein E	0.02 (45)	3.1	2.1	0.2	1.9	0.15	0.2	5.7	8.6	2.0	4.7	7.2

Open Access Article. Published on 29 July 2025. Downloaded on 5/10/2026 5:52:52 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

the composition of the protein corona, as well as the composition of FBS in the absence of NPs. Single experiments were carried out for each sample. Our previous proteomics studies of triplicate samples confirmed that single experiments are sufficient.[44] To correct for any change in performance or differences in protein loading, each sample was normalized to itself by dividing by the mean of the interior 80% of the protein intensities.[44] Each sample was scaled to have the same average. We report these values as percent normalized abundance (Table 2). The amount of protein in the corona relative to the amount in FBS, in the absence of NPs, was also calculated (Fig. 1), which shows the enrichment of the ten most abundant proteins in the protein corona relative to their abundance in FBS, in fold change, $\log_2$. Proteins are listed in order of their relative abundance (Table 2). Albumin is the most abundant protein in FBS, comprising 54% of the protein in FBS (Table 2). It is present at high abundance (16.8–46%) in all protein coronas and is the dominant protein in the corona of the majority of the NPs (Tables 2 and S6). The exception to this is PEG-pNP, which has a corona dominated by hemoglobin subunit alpha (40%) and COOH-pNP incubated in 10% FBS, which has a corona dominated by alpha-2-HS-glycoprotein (30.6%; Table S6). In comparison to the high abundance of albumin in FBS, albumin is depleted in all protein coronas (Tables 2, S6 and Fig. 1).

While all NPs show a similar interaction with albumin (depletion), the other most abundant proteins show a wider range of protein–NP interactions (Fig. 1). For example, complement C3, which plays a central role in the activation of the complement system,[73,74] exhibits a notably high spread (−2 to +4.5 $\log_2$ fold change) in enrichment and depletion across NPs, in comparison to the more narrow spread for albumin (−2 to −0.25 $\log_2$ fold change). The wide range in enrichment and depletion values for complement C3 suggests that NP features such as diameter, zeta potential, and surface functionalization significantly impact the adsorption of complement C3 onto the NP surface. In contrast, proteins such as albumin (−2 to −0.25 $\log_2$ fold change) and alpha-2-HS-glycoprotein (−2 to +1 $\log_2$ fold change) show narrower $\log_2$ fold changes across different NPs, suggesting less dependence on NP features. Similarly, apolipoprotein A-I and alpha-2-macroglobulin exhibit a narrow spread of $\log_2$ fold changes (−0.5 to +3.5 and −1.5 to +1, respectively), suggesting stable interactions across different NP features. Complement C3 tends to be less enriched with positively charged NPs (*e.g.*, PEI-mNP$_S$, PEI-mNP$_L$), though high variability makes it difficult to establish a clear correlation with zeta potential. Apolipoprotein A-I also shows moderate to high enrichment with negatively charged NPs and consistent or slight depletion with positively charged NPs. While the data does not consistently support the idea that decreasing NP zeta potential reduces protein enrichment, highly negative zeta potentials (*e.g.*, COOH-pNP with −63 mV) may decrease complement C3 enrichment.

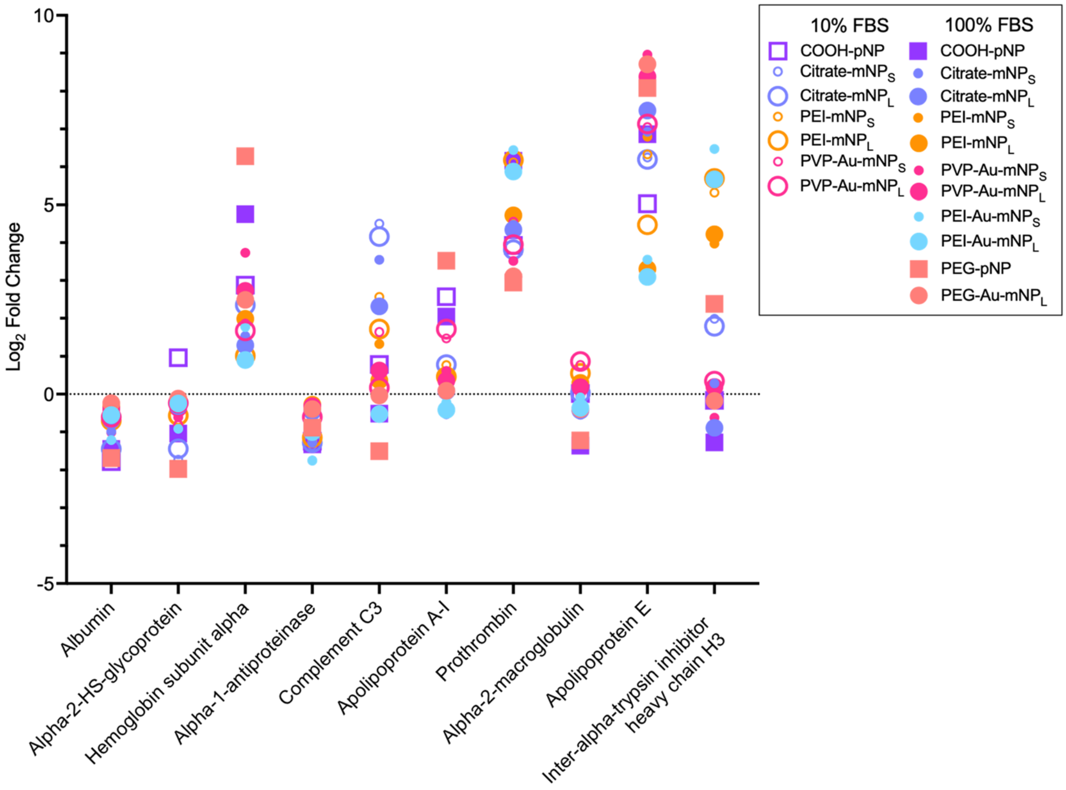

Fig. 1 Enrichment (>0 $\log_2$ fold change) and depletion (<0 $\log_2$ fold change) of the ten most abundant proteins in the protein corona relative to their abundance in FBS. $n = 1$. A value of 0 (dotted line) would reflect a corona protein abundance that matches the abundance of the same protein in the serum used to form the corona. COOH-pNP and PEG-pNPs were centrifuged (squares). All other samples were prepared by magnetic pull-down (circles).

Open Access Article. Published on 29 July 2025. Downloaded on 5/10/2026 5:52:52 AM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

In addition to NP and protein features, previous work has shown that the separation method used in the preparation of the protein corona, magnetic pull-down or centrifugation, is one factor in the composition of the protein corona.[56] We note increased enrichment of hemoglobin subunit-alpha and apolipoprotein A-I on PEG-pNP and COOH-pNP, which underwent centrifugation, in comparison to the mNPs, which are separated using magnetic pull-down (Fig. 1).

Overall, this level of NP, protein, and experimental feature complexity makes it challenging to extract trends. Instead, we use the three NP core materials (mNP, gold-mNP, and polystyrene), two core diameters (82 nm and 182 nm), six surface ligands, and seven effective surface charges as training data for ML.

Protein features

In addition to NP features (core material, surface ligand, diameter, zeta potential), protein corona formation will depend on protein features (Table S1). A protein feature database was built using UniProt,[61] NetSurfP 3.0,[65,66] and BioPython.[67] UniProt was used to extract protein entry accession number, length, mass, and sequence information for the complete bovine Swiss-Prot knowledge base. NetSurfP was used to predict solvent accessibility, secondary structure, and structure disorder on a per amino acid basis. We modified previously published Python code to capture the entire bovine proteome.[39] These protein features were combined with data calculated by the BioPython package, which performs calculations using the protein sequence paired with structural data from the Protein DataBank.

ML model development and evaluation

We constructed a comprehensive dataset of 84 features incorporating NP features, protein features, and experimental features for use with ML models (Tables S1–S3). Two random forest-based supervised learning models, RFR and RFC, were developed and tested for the ability to predict the protein corona.

Random Forest models were selected for their ability to reduce overfitting, due to their ensemble nature, and strengths for this specific application.[75–77] For example, random forest models average results over multiple trees, which leads to higher accuracy than support vector machines.[78] Neural networks require training data on the scale of ∼10 000 or greater data points, making them less useful for this type of data-limited application.[79] Implementation in Python allows for code sharing and use by others. RFR was employed to predict continuous numerical values, specifically the abundance of individual proteins in the corona, which we quantified as peptide intensities. RFC was utilized to predict categorical outcomes, determining whether a protein would be enriched or depleted in the protein corona relative to the serum used to form the corona. Both models were trained to determine the optimal number of input features using recursive feature elimination with k-fold cross-validation across ten folds. Model performance was evaluated using multiple performance metrics.

For RFR, R^2, mean squared error, Pearson correlation coefficient, and Spearman's rank correlation coefficients were used to assess model performance. R^2 provides a measure of how well true dataset values are replicated by the model, on a scale of 0 to 1, with 1 being a perfect replication by the model. An R^2 of 0 would indicate that the predictions are random. Mean squared error measures the difference between the predicted values made by the model and true values in the dataset, with values closer to zero indicating better model performance. The Pearson correlation coefficient measures the correlation between the predicted and observed corona protein abundance values. Spearman's rank correlation coefficient measures the rank-based correlation between predicted and observed corona protein abundance values. Both Pearson correlation and Spearman's correlation coefficients can have values ranging from −1 to 1. A value of +1 indicates perfect positive correlation. For RFC, area under the receiver operating characteristic curve (AUROC), accuracy, precision, *F*1 score, and recall were used as performance metrics. AUROC measures the ability of the RFC model to distinguish between classes, defined as protein enrichment and depletion, considering the true positive rate and false positive rate across all classification thresholds (*i.e.* enriched or depleted). An AUROC value of 0.5 corresponds to a random guess and 1.0 represents perfect classification. Accuracy describes the ratio of correctly predicted proteins to total number of proteins. Precision describes the proportion of true positives to predicted positives. Recall describes the proportion of true positives to actual positives (both true positives and false negatives). Precision and recall are combined in the *F*1 score, which is the harmonic mean of the two values, meaning both precision and recall are equally weighted, which can range from 0 to 1 (perfect precision and recall).

We first used recursive feature elimination with k-fold cross-validation to identify and score the most important NP, protein, and experimental features for the protein corona. Recursive feature elimination with k-fold cross-validation is a feature selection technique that recursively removes less important features while evaluating the performance of the model through cross-validation. This technique identifies the optimal subset of features that balance model accuracy and complexity. The

Table 3 Top ten of the 61 most important NP, protein, and experimental features identified by the RFR model. The percent importance is normalized

Feature	Importance (%)
Protein abundance in FBS	46.7
Zeta potential	4.1
d_{h}	3.2
Protein incubation concentration (10% or 100%)	2.1
d_{TEM}	1.8
% Phenylalanine	1.7
% Non-structure associated amino acids	1.4
% Asparagine	1.3
NP incubation concentration	1.2
% Alanine	1.2

Open Access Article. Published on 29 July 2025. Downloaded on 5/10/2026 5:52:52 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

output is the list of selected features and their associated percentage of importance. Results for the RFR model show the negative mean squared error as a function of the number of features (Fig. S1). The error decreases sharply initially and stabilizes at 61 features, indicating that 61 features are optimal for our model (Table 3). The average and standard deviation of the R^2, mean squared error, Pearson correlation coefficient, and Spearman's rank correlation further validate robustness (Table S7). High scores across these metrics reinforce the utility of the RFR model and demonstrate the efficacy of the selected features (Table S4).

Following feature selection using recursive feature elimination with k-fold cross-validation, the RFR model demonstrated robust predictive performance, as reflected by high evaluation metrics for predictions made on the test data split (Fig. S2 and Table 4). A comprehensive analysis of RFR model and prediction performance is provided in SI (Table S7).

For both models, the abundance of individual proteins in FBS was identified as the most important feature for predicting the protein corona (Table 3 and S5). This is intuitive, as proteins with higher relative abundance are more likely to interact with NPs. However, protein abundance is not the single determinant of protein adsorption on NPs, reflecting the distinction between kinetic and thermodynamic control in protein corona formation. For instance, despite being the most abundant protein in FBS (54.3%), albumin is depleted in the corona of all NPs (Tables 2, S6 and Fig. 1). This is in agreement with previous research showing that initial kinetic adsorption of proteins can be displaced by proteins with greater thermodynamic stability on the NP surface.[22,29,37,80–85] This indicates that while protein abundance is a key factor, NP and protein features also influence protein adsorption. Both models identified zeta potential as the next most significant feature after relative protein abundance values, with importance values of 4.1% and 8.5% for the RFR and RFC models, respectively (Table 3 and S5). The hydrodynamic diameter of the functionalized NPs (d_h) was the third most significant feature, with an importance of 3.2% and 7.6% for RFR and RFC, respectively (Table 3 and S5).

Predicting protein coronas for individual NPs

One goal of this research was to develop models that could be used to predict the protein corona of new NPs based on characteristics of the new NPs and protein features drawn from existing databases. We evaluated the RFR model for the ability to predict proteins coronas on NPs that were excluded from the training data. With our library of 11 NPs (Table 1), we incubated 7 of these NPs with both 10% and 100% FBS for 18 total NP samples. We used 17 NPs, of our 18 total NPs, as training data and then tested the model on the one NP that was excluded from the training data. This training and testing were done iteratively using leave-one-group-out cross-validation to test all 18 NPs. No additional feature selection was performed during this process. The range in R^2 values for the ability to predict the abundance of individual proteins present in the corona was 0.45–0.88 (Table 5). Mean squared error (1.38–6.18), Pearson correlation coefficient (0.73–0.94), and Spearman's rank correlation coefficient (0.63–0.91) were also determined for each NP (Table S10). The RFC model was not used to predict the corona of individual NPs as it provides only relative enrichment and depletion of individual proteins rather than abundance. The RFR model showed the best predictive ability for citrate-mNP$_S$ ($R^2 = 0.88$) (Table 5 and Fig. 2A). PEG-pNP ($R^2 = 0.45$) were the worst in terms of predictive ability (Table 5 and Fig. 2B). To investigate the dependence of these predictions on the specific training data used, we tested two scenarios: (1) RFR models trained solely on citrate-mNP samples to predict the corona formed on a citrate-mNP$_S$ (100% FBS) and (2) RFR models trained solely on non-citrate-mNP samples to predict the corona formed on a citrate-mNP$_S$ (100% FBS). R^2 values remained consistent with values obtained from use of the full data set (Table S11).

To determine if higher abundance corona proteins, such as albumin (Table 2), were more likely to be predicted correctly, we measured the correlation of the top ten most abundant corona proteins with the accuracy of prediction. We found no correlation between protein corona abundance and predictive ability (Fig. S6).

Table 4 Evaluation metrics and scores assessing predictive ability of the RFR model for 61 features. The performance of the RFC model was also optimized through recursive feature elimination with k-fold cross-validation and described in SI (Fig. S3–S5 and Tables S8 and S9)

Evaluation metric	Score
R^2	0.81
Mean squared error	2.02
Pearson	0.90
Spearman	0.87

Table 5 Prediction of protein coronas of individual NPs using the RFR model. NPs ranked from best to worst predictions in terms of R^2

NP	FBS incubation (%)	R^2
Citrate-mNP$_S$	100	0.88
Citrate-mNP$_S$	10	0.88
Citrate-mNP$_L$	10	0.87
PVP-Au-mNP$_L$	100	0.86
PVP-Au-mNP$_L$	10	0.85
PEI-mNP$_S$	10	0.85
PVP-Au-mNP$_S$	10	0.84
PEI-mNP$_S$	100	0.84
Citrate-mNP$_L$	100	0.83
PVP-Au-mNP$_S$	100	0.79
PEI-mNP$_L$	100	0.78
PEI-mNP$_L$	10	0.78
COOH-pNP	10	0.77
PEG-Au-mNP$_L$	100	0.76
COOH-pNP	100	0.71
PEI-Au-mNP$_L$	100	0.69
PEI-Au-mNP$_S$	100	0.63
PEG-pNP	100	0.45

Open Access Article. Published on 29 July 2025. Downloaded on 5/10/2026 5:52:52 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

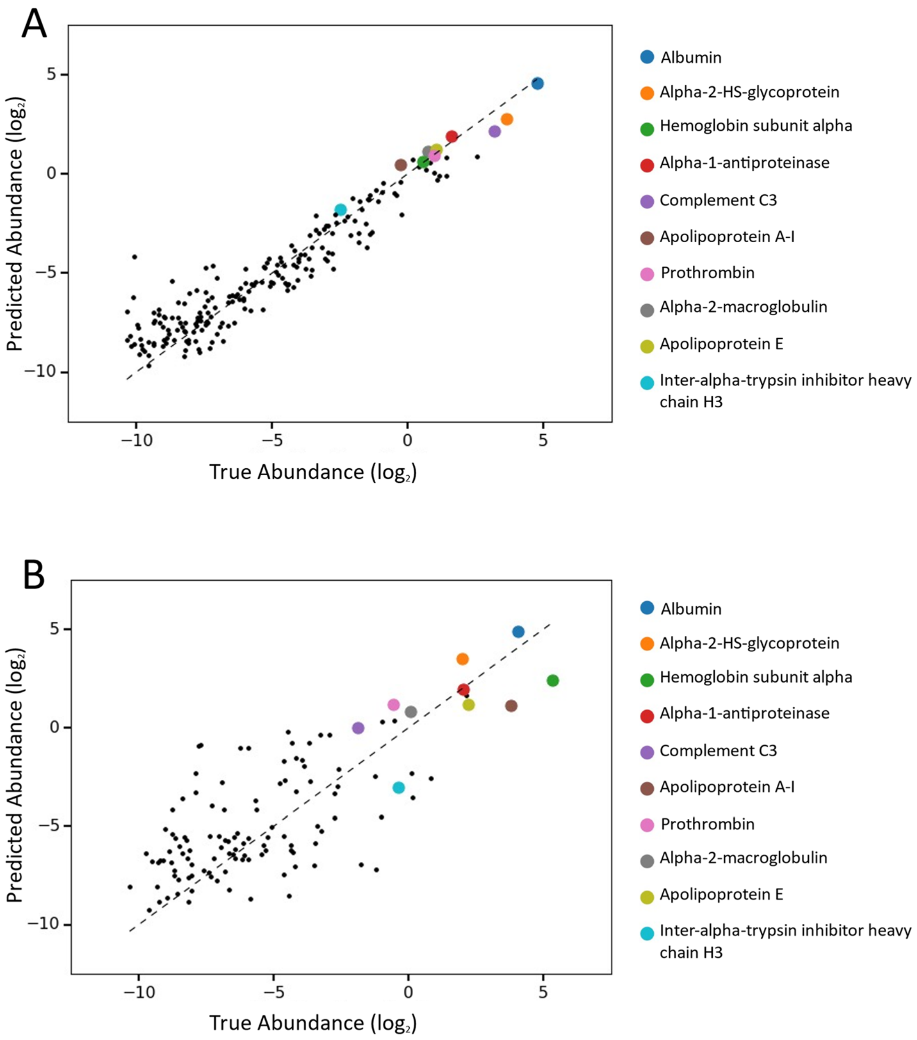

Fig. 2 Comparison of predicted and true (observed) protein abundance determined by the RFR model. (A) Citrate-mNP$_S$ had the best predictive ability. (B) PEG-pNP had the worst predictive ability. The top ten most abundant proteins present in each corona are shown in color.

Conclusions

The goal of this research was to develop regression and classification ML models capable of identifying the relevant NP, protein, and experimental features necessary to predict the protein corona and then using ML to predict the protein corona of individual NPs. We generated NPs with a range of features (core material, surface ligand, diameter and zeta potential) (Table 1). Protein features (*e.g.* secondary structure, percentage of polar amino acids) were derived from proteomics data (Table 2, S6 and Fig. 1), UniProt, NetSurfP3.0, and BioPython (Table S1). We selected FBS as the protein source for our protein–NP samples as it is a common nutrient source for cells in culture, well-documented in protein sequence databases, and frequently used in other protein corona studies.[45–48] Other protein sources, such as mouse or human plasma or serum, could be incorporated into our code, which is shared on GitHub (**https://www.github.com/nvijgen/ProteinCoronaPredict_PayneLab.git**), either experimentally or computationally, through the use of NetSurfP and BioPython to translate FBS protein features into protein features from another protein source. NP and protein features were combined with experimental features (NP concentration, protein incubation concentration, method of free protein removal) to build the dataset (Tables S1–S3) for the RFR and RFC models.

Both RFR and RFC models showed excellent performance metrics (Tables 4, S7–S9 and Fig. S1–S5). Our models identified the top feature in predicting the protein corona composition to be the protein abundance in FBS (46.7% and 27.7% for RFR and RFC, respectively) (Table 3 and S5). The next most important feature for both models was zeta potential (4.1% and 8.5% for RFR and RFC, respectively). The hydrodynamic diameter of the functionalized NP was the third most significant feature, with an importance of 3.2% for RFR and 7.6% for RFC. To evaluate RFR model predictions for new NPs, we implemented leave-one-group-out cross-validation, where each NP was iteratively excluded from training and then used as the test set. This approach allowed us to assess the predictive ability for each NP individually, identifying which were the best and worst performing (Fig. 2 and Tables 5 and S10). The model had the

Open Access Article. Published on 29 July 2025. Downloaded on 5/10/2026 5:52:52 AM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

highest accuracy in predicting protein abundances for citrate-mNP$_S$ (Fig. 2 and Tables 5 and S10) and the lowest accuracy in predicting protein abundances for PEG-pNP (Fig. 2 and Tables 5 and S10). Future work will explore the specific protein–NP interactions that underlie these predictions.

Our work builds on previous ML models with the common goal of protein corona prediction.[39–42] In comparison to the previous model developed for single-walled carbon nanotubes,[39] our models incorporate NP and experimental features, in addition to protein features, into the training set. Our models also incorporated the abundance of individual proteins in the FBS used to form the corona, which was ultimately the most important feature identified by both models (Tables 3, S4 and S5). In comparison to the model developed with silver NPs and yeast proteins,[41] FBS provides a more relevant protein source, especially for NPs used in applications with cultured cells. Additionally, our RFC model achieved an AUROC of 0.99 and *F*1 score of 0.93, whereas the RFC run on the yeast dataset achieved an AUROC of 0.83 and *F*1 score of 0.81. We also included core composition as an NP feature, a feature suggested in their work to be important for future consideration. While our models were constructed from data obtained from 18 protein–NP combinations, previous work has used a dataset with >600 NPs.[42] This work focused on NP and experimental features, lacking protein features. The inclusion of protein features in our dataset led to a total feature count of 84. In comparison, this previous work had a feature count of only 21. In comparison to this previous work, our RFR model enables a quantitative prediction of specific protein abundance in the corona. In addition, our leave-one-group-out cross-validation tests the predictive ability of the RFR model on NPs that were not present in the training data, a key goal in the use of corona prediction for the design of novel nanomaterials.

One limitation of this current predictive ability is that a totally novel NP with features well-outside of our existing data set of NPs (mNP, gold-mNP, polystyrene) and ligands (citrate, polyethyleneimine, polyvinylpyrrolidone, polyethylene glycol, carboxylate) may need additional experimental data for predictions with high accuracy.

Previous work has shown that the protein corona determines the cellular and physiological response to NPs.[1–6,10,19–37] The ability to predict the protein corona could provide a first step in predicting the cellular response. For example, previous work using 105 different gold NPs showed that the protein corona determined the interaction of NPs with cells, but this was determined experimentally in a tour de force experimental study.[31] We hope the RFR and RFC models described above will provide a computational tool for pre-screening NP candidates for use in biological applications prior to experiments. For example, in the longer term, we could envision the use of a ML-predicted protein corona to reduce the need for cell and animal experiments to determine NP toxicity.

Author contributions

K. M. P., G. S. M., and C. K. P. conceived the research. K. M. P. and G. S. M conducted the experiments. K. M. P., G. S. M., and N. V. analyzed the data. N. V, K. M. P., and G. S. M developed the machine learning models. N. V., K. M. P. and C. K. P. wrote the manuscript. All authors reviewed the manuscript.

Conflicts of interest

There are no conflicts to declare.

Data availability

Mass spectrometry proteomic data is available from the ProteomeXchange Consortium *via* the Proteomics Identification Database (PRIDE) partner repository with the dataset identifier PXD053700 and 10.6019/PXD053700 at **https://www.ebi.ac.uk/pride/**. The code associated with this publication can be found on GitHub at **https://www.github.com/nvijgen/ProteinCoronaPredict_PayneLab.git**.

Supplementary information including additional data, as referenced in the text, and a comparison of the RFR and RFC models is available. See DOI: **https://doi.org/10.1039/d5na00425j**.

Acknowledgements

This research was funded by NSF (CBET-1901579, CBET-2405017) and the Duke University MEDx program. We thank the Duke University School of Medicine for the use of the Proteomics and Metabolomics Core Facility, which performed LC/MS molecular weight determination of our samples. This research was also performed in part at the Duke University Shared Materials Instrumentation Facility (SMIF), a member of the North Carolina Research Triangle Nanotechnology Network (RTNN), which is supported by the National Science Foundation (ECCS-2025064) as part of the National Nanotechnology Coordinated Infrastructure (NNCI). We thank J. Haws, Director of OIT Data and Analytics Practice at Duke University for the helpful advice on ML and data science and N. Hoang for helpful advice on NP synthesis.

References

1 O. Bondarenko, M. Mortimer, A. Kahru, N. Feliu, I. Javed, A. Kakinen, S. Lin, T. Xia, Y. Song, T. P. Davis, I. Lynch, W. J. Parak, D. T. Leong, P. C. Ke, C. Chen and Y. Zhao, *Nano Today*, 2021, **39**, 101184.
2 D. Docter, S. Strieth, D. Westmeier, O. Hayden, M. Gao, S. K. Knauer and R. H. Stauber, *Nanomedicine*, 2015, **10**, 503–519.
3 K. Hamad-Schifferli, *Nanomedicine*, 2015, **10**, 1663–1674.
4 S. Soares, J. Sousa, A. Pais and C. Vitorino, *Front. Chem.*
5 P. Foroozandeh and A. A. Aziz, *Nanoscale Res. Lett.*, 2015, **10**, 221.
6 S. Runa, M. Hussey and C. K. Payne, *J. Phys. Chem. B*, 2018, **122**, 1009–1016.
7 A. Mohajerani, L. Burnett, J. V. Smith, H. Kurmus, J. Milas, A. Arulrajah, S. Horpibulsuk and A. A. Kadir, *Materials*, 2019, **12**, 3052.

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 29 July 2025. Downloaded on 5/10/2026 5:52:52 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

8 J. P. Giraldo and S. Kruss, *Nat. Nanotechnol.*, 2023, **18**, 107–108.

9 G. M. Newkirk, P. de Allende, R. E. Jinkerson and J. P. Giraldo, *Front. Plant Sci.*, 2021, **12**, 691295.

10 L. Xu, M. Xu, R. Wang, Y. Yin, I. Lynch and S. Liu, *Small*, 2020, **16**, 2003691.

11 T. Hofmann, G. V. Lowry, S. Ghoshal, N. Tufenkji, D. Brambilla, J. R. Dutcher, L. M. Gilbertson, J. P. Giraldo, J. M. Kinsella, M. P. Landry, W. Lovell, R. Naccache, M. Paret, J. A. Pedersen, J. M. Unrine, J. C. White and K. J. Wilkinson, *Nat. Food*, 2020, **1**, 416–425.

12 M. J. Hanus and A. T. Harris, *Prog. Mater. Sci.*, 2013, **58**, 1056–1102.

13 J. Lee, S. Mahendra and P. J. J. Alvarez, *ACS Nano*, 2010, **4**, 3580–3590.

14 K. E. Wheeler, A. J. Chetwynd, K. M. Fahy, B. S. Hong, J. A. Tochihuitl, L. A. Foster and I. Lynch, *Nat. Nanotechnol.*, 2021, **16**, 617–629.

15 Z. Zahra, Z. Habib, S. Chung and M. A. Badshah, *Nanomaterials*, 2020, **10**, 1469.

16 M. M. Nabi, J. Wang, M. Erfani, E. Goharian and M. Baalousha, *Environ. Sci.: Nano*, 2023, **10**, 718–731.

17 C. C. Fleischer and C. K. Payne, *Acc. Chem. Res.*, 2014, **47**, 2651–2659.

18 C. K. Payne, *J. Chem. Phys.*, 2019, **151**, 130901.

19 C. D. Walkey and W. C. W. Chan, *Chem. Soc. Rev.*, 2012, **41**, 2780–2799.

20 M. P. Monopoli, C. Åberg, A. Salvati and K. A. Dawson, *Nat. Nanotechnol.*, 2012, **7**, 779–786.

21 P. C. Ke, S. Lin, W. J. Parak, T. P. Davis and F. Caruso, *ACS Nano*, 2017, **11**, 11773–11776.

22 P. del Pino, B. Pelaz, Q. Zhang, P. Maffre, G. Ulrich Nienhaus and W. J. Parak, *Mater. Horiz.*, 2014, **1**, 301–313.

23 C. Rodriguez-Quijada, M. Sánchez-Purrà, H. de Puig and K. Hamad-Schifferli, *J. Phys. Chem. B*, 2018, **122**, 2827–2840.

24 R. D. Bartone, L. J. Tisch, J. Dominguez, C. K. Payne and J. C. Bonner, *ACS Nano*, 2024, **18**, 26215–26232.

25 M. Mahmoudi, M. P. Landry, A. Moore and R. Coreas, *Nat. Rev. Mater.*, 2023, **8**, 422–438.

26 K. Nienhaus and G. U. Nienhaus, *Small*, 2023, **19**, 2301663.

27 K. M. Poulsen, M. C. Albright, N. J. Niemuth, R. M. Tighe and C. K. Payne, *Environ. Sci.: Nano*, 2023, **10**, 2427–2436.

28 L. Kobos and J. Shannahan, *Wiley Interdiscip. Rev.:Nanomed. Nanobiotechnol.*, 2020, **12**, e1608.

29 K. Nienhaus and G. U. Nienhaus, *Curr. Opin. Biomed. Eng.*, 2019, **10**, 11–22.

30 D. T. Jayaram, S. M. Pustulka, R. G. Mannino, W. A. Lam and C. K. Payne, *Biophys. J.*, 2018, **115**, 209–216.

31 C. D. Walkey, J. B. Olsen, F. Song, R. Liu, H. Guo, D. W. H. Olsen, Y. Cohen, A. Emili and W. C. W. Chan, *ACS Nano*, 2014, **8**, 2439–2455.

32 C. C. Fleischer and C. K. Payne, *J. Phys. Chem. B*, 2014, **118**, 14017–14026.

33 A. Hill and C. K. Payne, *RSC Adv.*, 2014, **4**, 31735–31744.

34 C. C. Fleischer, U. Kumar and C. K. Payne, *Biomater. Sci.*, 2013, **1**, 975–982.

35 G. W. Doorley and C. K. Payne, *Chem. Commun.*, 2012, **48**, 2961–2963.

36 G. W. Doorley and C. K. Payne, *Chem. Commun.*, 2010, **47**, 466–468.

37 S. Lindman, I. Lynch, E. Thulin, H. Nilsson, K. A. Dawson and S. Linse, *Nano Lett.*, 2007, **7**, 914–920.

38 C. C. Fleischer and C. K. Payne, *J. Phys. Chem. B*, 2012, **116**, 8901–8907.

39 N. Ouassil, R. L. Pinals, J. T. Del Bonis-O'Donnell, J. W. Wang and M. P. Landry, *Sci. Adv.*, 2022, **8**, eabm0898.

40 J. Huzar, R. Coreas, M. P. Landry and G. Tikhomirov, *ACS Nano*, 2025, **19**, 4333–4345.

41 M. R. Findlay, D. N. Freitas, M. Mobed-Miremadi and K. E. Wheeler, *Environ. Sci.: Nano*, 2018, **5**, 64–71.

42 Z. Ban, P. Yuan, F. Yu, T. Peng, Q. Zhou and X. Hu, *Proc. Natl. Acad. Sci. U. S. A.*, 2020, **117**, 10492–10499.

43 K. M. Poulsen, T. Pho, J. A. Champion and C. K. Payne, *Anal. Bioanal. Chem.*, 2020, **412**, 6543–6551.

44 K. M. Poulsen and C. K. Payne, *Anal. Bioanal. Chem.*, 2022, **414**, 7265–7275.

45 W. Ngo, J. L. Y. Wu, Z. P. Lin, Y. Zhang, B. Bussin, A. Granda Farias, A. M. Syed, K. Chan, A. Habsid, J. Moffat and W. C. W. Chan, *Nat. Chem. Biol.*, 2022, **18**, 1023–1031.

46 D. Glancy, Y. Zhang, J. L. Y. Wu, B. Ouyang, S. Ohta and W. C. W. Chan, *J. Controlled Release*, 2019, **304**, 102–110.

47 H. T. R. Wiogo, M. Lim, V. Bulmus, J. Yun and R. Amal, *Langmuir*, 2011, **27**, 843–850.

48 R. Tedja, M. Lim, R. Amal and C. Marquis, *ACS Nano*, 2012, **6**, 4083–4093.

49 Q. Dai, Y. Yan, J. Guo, M. Björnmalm, J. Cui, H. Sun and F. Caruso, *ACS Macro Lett.*, 2015, **4**, 1259–1263.

50 A. Salvati, A. S. Pitek, M. P. Monopoli, K. Prapainop, F. B. Bombelli, D. R. Hristov, P. M. Kelly, C. Åberg, E. Mahon and K. A. Dawson, *Nat. Nanotechnol.*, 2013, **8**, 137–143.

51 R. Cai and C. Chen, *Adv. Mater.*, 2019, **31**, 1805740.

52 A. Cifuentes-Rius, H. de Puig, J. C. Y. Kah, S. Borros and K. Hamad-Schifferli, *ACS Nano*, 2013, **7**, 10066–10074.

53 J. C. Y. Kah, J. Chen, A. Zubieta and K. Hamad-Schifferli, *ACS Nano*, 2012, **6**, 6730–6740.

54 W. Kim, N. K. Ly, Y. He, Y. Li, Z. Yuan and Y. Yeo, *Adv. Drug Delivery Rev.*, 2023, **192**, 114635.

55 H. Deng, X. Li, Q. Peng, X. Wang, J. Chen and Y. Li, *Angew. Chem., Int. Ed.*, 2005, **44**, 2782–2785.

56 K. N. L. Hoang, K. E. Wheeler and C. J. Murphy, *Anal. Chem.*, 2022, **94**, 4737–4746.

57 D. G. Duff, A. Baiker and P. P. Edwards, *Langmuir*, 1993, **9**, 2301–2309.

58 C. A. Schneider, W. S. Rasband and K. W. Eliceiri, *Nat. Methods*, 2012, **9**, 671–675.

59 S. Tyanova, T. Temu and J. Cox, *Nat. Protoc.*, 2016, **11**, 2301–2319.

60 J. Cox and M. Mann, *Nat. Biotechnol.*, 2008, **26**, 1367–1372.

61 U. P. Consortium, *Nucleic Acids Res.*, 2023, **51**, D523–D531.

62 cRAP protein sequences: The Global Proteome Machine, https://www.thegpm.org/crap/, accessed 5 June 2024.

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 29 July 2025. Downloaded on 5/10/2026 5:52:52 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

63 J. P. Kastan, E. Y. Dobrikova, J. D. Bryant and M. Gromeier, *Sci. Adv.*, 2020, **6**, eaba0745.

64 Y. Perez-Riverol, A. Csordas, J. Bai, M. Bernal-Llinares, S. Hewapathirana, D. J. Kundu, A. Inuganti, J. Griss, G. Mayer, M. Eisenacher, E. Pérez, J. Uszkoreit, J. Pfeuffer, T. Sachsenberg, S. Yilmaz, S. Tiwary, J. Cox, E. Audain, M. Walzer, A. F. Jarnuczak, T. Ternent, A. Brazma and J. A. Vizcaíno, *Nucleic Acids Res.*, 2019, **47**, D442–D450.

65 M. H. Høie, E. N. Kiehl, B. Petersen, M. Nielsen, O. Winther, H. Nielsen, J. Hallgren and P. Marcatili, *Nucleic Acids Res.*, 2022, **50**, W510–W515.

66 M. S. Klausen, M. C. Jespersen, H. Nielsen, K. K. Jensen, V. I. Jurtz, C. K. Sønderby, M. O. A. Sommer, O. Winther, M. Nielsen, B. Petersen and P. Marcatili, *Proteins*, 2019, **87**, 520–527.

67 P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski and M. J. L. de Hoon, *Bioinformatics*, 2009, **25**, 1422–1423.

68 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

69 K. Rahme, L. Chen, R. G. Hobbs, M. A. Morris, C. O'Driscoll and J. D. Holmes, *RSC Adv.*, 2013, **3**, 6085–6094.

70 U. Wattendorf and H. P. Merkle, *J. Pharm. Sci.*, 2008, **97**, 4655–4669.

71 L. Shi, J. Zhang, M. Zhao, S. Tang, X. Cheng, W. Zhang, W. Li, X. Liu, H. Peng and Q. Wang, *Nanoscale*, 2021, **13**, 10748–10764.

72 Y. R. Perera, J. X. Xu, D. L. Amarasekara, A. C. Hughes, I. Abbood and N. C. Fitzkee, *Molecules*, 2021, **26**, 5788.

73 J. V. Sarma and P. A. Ward, *Cell Tissue Res.*, 2011, **343**, 227–235.

74 M. Pekna and M. Pekny, *Cells*, 2021, **10**, 1812.

75 T. Hastie, R. Tibshirani and J. Friedman, in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, ed. T. Hastie, R. Tibshirani and J. Friedman, Springer, New York, NY, 2009, pp. 587–604.

76 L. Breiman, *Mach. Learn.*, 2001, **45**, 5–32.

77 M. Kuhn and K. Johnson, in *Applied Predictive Modeling*, ed. M. Kuhn and K. Johnson, Springer, New York, NY, 2013, pp. 173–220.

78 M. Pal, *Int. J. Remote Sens.*, 2005, **26**, 217–222.

79 D. J. Benkendorf and C. P. Hawkins, *Ecol. Inform.*, 2020, **60**, 101137.

80 D. Baimanov, J. Wang, J. Zhang, K. Liu, Y. Cong, X. Shi, X. Zhang, Y. Li, X. Li, R. Qiao, Y. Zhao, Y. Zhou, L. Wang and C. Chen, *Nat. Commun.*, 2022, **13**, 5389.

81 J. Hühn, C. Fedeli, Q. Zhang, A. Masood, P. del Pino, N. M. Khashab, E. Papini and W. J. Parak, *Int. J. Biochem. Cell Biol.*, 2016, **75**, 148–161.

82 S. Milani, F. Baldelli Bombelli, A. S. Pitek, K. A. Dawson and J. Rädler, *ACS Nano*, 2012, **6**, 2532–2541.

83 W. Liu, J. Rose, S. Plantevin, M. Auffan, J.-Y. Bottero and C. Vidaud, *Nanoscale*, 2013, **5**, 1658–1668.

84 S. Kihara, N. J. van der Heijden, C. K. Seal, J. P. Mata, A. E. Whitten, I. Köper and D. J. McGillivray, *Bioconjugate Chem.*, 2019, **30**, 1067–1076.

85 O. Vilanova, J. J. Mittag, P. M. Kelly, S. Milani, K. A. Dawson, J. O. Rädler and G. Franzese, *ACS Nano*, 2016, **10**, 10842–10850.

© 2025 The Author(s). Published by the Royal Society of Chemistry