REVIEW ARTICLE ASIAN JOURNAL OF MEDICAL SCIENCES

A decade of machine learning in protein corona research: Innovations and challenges

Berkant Yığın[1], Nazmi Özer[2]

[1]M Pharm Student, [2]Professor, Department of Biochemistry, Faculty of Pharmacy, Girne American University, Girne, Mersin, Turkey

Submission: 10-12-2025 Revision: 02-02-2025 Publication: 01-03-2025

ABSTRACT

Nanomaterials, with their diverse biomedical applications spanning drug delivery to molecular imaging, undergo the adsorption of a protein corona (PC) layer upon exposure to biological environments. This dynamic layer, shaped by intricate interactions, significantly influences immune recognition, biodistribution, and nanoparticle toxicity. Traditional proteomic methods, such as liquid chromatography-tandem mass spectrometry, are effective but limited by low throughput, high costs, and the requirement for specialized expertise. The transition from unintentional PC analysis during polymer evaluations to a deliberate investigation of its role in drug targeting underscores the need for more efficient analytical approaches. The integration of machine learning (ML) into PC research has emerged as a promising solution. This computational methodology, which learns from datasets of characterized protein layers on specific nanoparticles, offers a more streamlined and resource-efficient alternative to traditional methods. Recent studies highlight ML's ability to predict PC dynamics and biological effects, achieving notable accuracy in forecasting organ accumulation patterns. However, challenges remain, including the need for larger and more diverse datasets, significant computational demands, and the necessity for interdisciplinary collaboration between biologists, chemists, and data scientists. In addition, the development of standardized experimental protocols is crucial to ensure reproducibility and comparability across studies. Ethical considerations, such as potential job displacement in traditional fields, such as chemistry, also warrant careful attention as ML continues to evolve in this domain. In summary, while ML shows immense potential to revolutionize PC research, further refinement of methodologies and enhanced collaboration across disciplines are essential to fully realize its application in clinical nanomedicine.

Key words: Nanomaterials; Drug delivery; Molecular imaging; Nanotoxicology; Machine learning

Access this article online

Website: https://ajmsjournal.info/index.php/AJMS/index

DOI: 10.71152/ajms.v16i3.4339

E-ISSN: 2091-0576

P-ISSN: 2467-9100

INTRODUCTION

The versatility of nanomaterials and the role of the protein corona (PC)

The remarkable versatility of nanomaterials has driven their development for a wide array of biomedical applications, including drug delivery, precision medicine, vaccines, molecular imaging, and bio-detection.[1] When nanoparticles are introduced into a biological system, a biomolecular corona forms on their surface. This corona is a dynamic layer resulting from the interaction of proteins and other biomolecules with the nanoparticle surface. Its formation is influenced by multiple factors, such as the concentration of proteins and nanoparticles, the affinity and structure of proteins, and the physicochemical properties of the nanoparticles. This dynamic layer, in turn, profoundly impacts immune recognition, biodistribution, receptor interactions, and the local and systemic toxicity of the nanoparticles.[2-4]

The role of proteins in forming the surface layer of the corona is particularly critical. Proteins dominate the

Address for Correspondence:
Nazmi Özer, Professor, Department of Biochemistry, Faculty of Pharmacy, Girne American University, 98428 Girne, Mersin 10, Turkey.
Mobile: +90-5338418381. **E-mail:** nazmiozer@gau.edu.tr

adsorption process, effectively preventing the binding of other biomolecules to some extent.[2] Consequently, most studies focus on the PC. Importantly, the PC is not a static entity. Some proteins bind reversibly, forming what is termed the "soft" corona, while others form stronger, more stable complexes, constituting the "hard" corona.[5,6]

For the successful clinical translation of nanoparticle-based drug delivery systems, it is essential to understand both the initial formation of the corona and the dynamic changes it undergoes as the nanoparticle traverses the biological environment. However, traditional proteomic methods used to identify the PC, such as liquid chromatography-tandem mass spectrometry (LC-MS/MS), are limited by low throughput, high costs, and the need for highly skilled personnel.[5,7,8]

Objective of this work

This review aims to provide a concise overview of advancements in the use of ML for PC analysis over the past decade. It examines key findings, highlights the challenges faced, and explores the potential of ML to revolutionize our understanding and application of PC dynamics. What are the critical insights and obstacles in leveraging ML for PC analysis? This work seeks to address these questions while shedding light on future directions in this rapidly evolving field.

RESULTS

Machine learning (ML) as a solution

ML offers a promising approach to overcoming the limitations of traditional proteomic methods. ML is a computational technique that does not rely on explicit programming to achieve specific outcomes. Instead, it learns from existing data to generalize patterns and predict outcomes for previously unseen data. In the context of PC analysis, ML models can be trained on datasets comprising previously characterized PC s on specific nanoparticles. This enables the development of predictive models that emulate the decision-making and information-gathering processes observed in biological systems.[9]

10 years of ML in PC research

In 2014, Walkey et al., successfully demonstrated that the PC fingerprint (PCF) provides a more accurate prediction of cell association than the physicochemical properties of nanomaterials.[10] They synthesized gold (AuNPs) and silver nanoparticles (AgNPs) and observed that the core material of the nanoparticles significantly affects the composition of the PC. Notably, only 36.9% of the serum PC formed around silver nanoparticles overlapped with that of gold nanoparticles modified with the same surface ligand. The study highlighted that the specific identities and abundances of proteins within the corona, rather than the total amount of adsorbed protein, play a critical role in determining nanoparticle-cell interactions.

Validation results revealed that a model trained on gold nanoparticles failed to accurately predict the cell association of silver nanoparticles, indicating that the relationship between the PC and cell association is highly dependent on the nanoparticle core material. Although a combined model incorporating both gold and silver nanoparticles showed a slight improvement in prediction accuracy, the overall performance remained low. Conversely, a model trained exclusively on silver nanoparticles accurately predicted their cell association, leading the authors to conclude that separate models may be required for different nanoparticle classes based on their PCF.

Furthermore, while the study identified 785 serum proteins within the PC, only 129 high-abundance proteins were included in defining the protein fingerprint, with 656 low-abundance proteins excluded. These excluded proteins, although not utilized in the model, could still influence cell association or other biological interactions.

Liu et al., further advanced this work by developing a model to identify specific serum proteins in the PC that are most relevant for predicting nanoparticle (NP) cell association.[11] To enhance prediction accuracy and minimize overfitting, the model incorporated only the most significant parameters. The dataset included 84 gold NPs (AuNPs) from Walkey et al., library of 105 NPs, excluding 21 NPs due to negligible protein adsorption. In addition, 129 PCFs were omitted due to insufficient data.

Among the descriptors analyzed, apolipoprotein B (APOB) emerged as the most frequently selected and had the greatest impact on the correlation with AuNP cell association. The study also addressed limitations in validation methodologies, noting that leave-one-out cross-validation (Q^2LOO), which had been used in earlier research, can yield overly optimistic estimates of prediction accuracy, especially for small datasets.

Papa et al., conducted a comparative study to evaluate multiple ML models and identify the most effective approach for PC research.[18] External validation was performed by dividing the original dataset into training and prediction sets, ensuring that the models' predictive performance was assessed within a representative domain for gold NPs (Au-NPs). Key innovations in this study included the direct use of spectral counts of high-abundance serum proteins as experimental descriptors

and a comparative analysis of linear versus non-linear techniques. This methodology provided valuable insights into the strengths and limitations of various ML-based approaches, offering alternatives to previously established models.

The study identified six critical descriptors for modeling cell association: the hydrodynamic diameter of Au-NPs after serum exposure, which correlated positively with increased cell association, and the spectral counts of five proteins-Alpha 1 Antitrypsin (A1AT), CO4B (Complement C4B), KNG1 (Kininogen-1), VNTC (Vitronectin), and glial fibrillary acidic protein (GFAP). Among these, A1AT, VNTC, and CO4B were found to enhance cell association, while KNG1 and GFAP served as inhibitors. Importantly, the roles of A1AT (promoter) and KNG1 (inhibitor) were consistent with prior studies, reinforcing their significance.

Projection pursuit regression, enhanced adaptive regression through hinges, and random forest (RF) emerged as the top-performing ML techniques in this analysis. Conversely, previously developed models by Walkey et al., and Liu et al., demonstrated inferior performance. This comprehensive evaluation emphasized the critical importance of selecting ML models suited to the specific characteristics of datasets and research objectives in PC investigations.

Helma et al., explored the optimal combination of descriptors and ML algorithms for predicting NP associations, contributing to the field of nanomaterial informatics.[19] Utilizing the eNanoMapper online database, they determined that combining PC descriptors with a weighted RF algorithm produced the best results, as measured by root mean square error (RMSE) and R^2 metrics. To ensure reproducibility, the study provided a publicly available, self-contained Docker image that included all necessary software and data.

However, the study acknowledged certain limitations. Since cross-validation folds were generated randomly, replication of the validations might not yield identical results. In addition, comparing the findings with other published models was challenging due to differences in datasets, validation protocols, and performance metrics. For instance, some studies applied global models to subsets of the PC dataset and performed feature selection on the entire dataset before validation, complicating direct comparisons.

Advancements in PC prediction and their biological implications

To date, research has largely focused on the relationship between the PCF and nanomaterial-cell associations, but the analysis of these fingerprints has relied on low-throughput methods. Findlay et al., made a significant advance by developing a RF model capable of predicting which proteins would adsorb onto a NP surface, thereby enabling the prediction of the PCF itself.[12]

The model utilized a balanced dataset that incorporated logarithmic enrichment factors and various protein properties. However, compared to protein features, engineered nanomaterial (ENM) and solvent properties were underrepresented in the training features. This imbalance arose because variations in ENM and solvent properties are challenging to study, as investigating new ENMs or reaction conditions requires fresh protein-ENM reactions and proteomics analyses. In addition, key factors potentially influencing PC formation – such as protein-protein interactions, ENM surface coating exchanges, and unmeasured variables, such as ENM shape – were excluded from the dataset. ENMs with hydrophobic coatings were also omitted due to solubility issues, limiting the model's capacity to evaluate hydrophobicity.

The study observed a strong emphasis on protein biophysical characteristics over ENM and solvent characteristics, consistent with earlier findings by Walkey et al., Within the protein feature set, factors such as the percentage of hydrophilic and aromatic amino acids and cysteine content were highly influential, with cysteine content alone contributing nearly 25%. However, the limited representation of ENM and solvent properties in the dataset restricted the ability to draw robust conclusions about their relative importance. Expanding the database to include a broader range of ENM and solvent features, including hydrophobic ENMs, would improve the model's applicability and robustness, enabling a more comprehensive understanding of PC formation drivers.

Lazarovits et al., were the first to employ PCF in an *in vivo* study, utilizing a neural network to predict the organ accumulation and clearance of gold NPs (AuNPs) in rats.[13] To ensure AuNP stability in blood and extended circulation times for effective isolation and analysis, the NP surfaces were saturated with polyethylene glycol. Their experiments identified five-time points as optimal for achieving the highest prediction accuracy in the neural network model.

The study found that protein patterns, rather than individual proteins, dictated the clearance of NPs from circulation, highlighting the role of the body in altering NP surface chemistry. AuNPs with extended circulation times avoided clearance due to the absence of specific protein combinations required for uptake by the liver and spleen. Notably, these protein combinations could not be artificially replicated, underscoring the value of the model.

ML elucidated this clearance mechanism while leveraging the body as a bioreactor to develop NP surface chemistries that evade liver and spleen uptake.

In addition, the researchers noted that NPs of a specific size represent a Gaussian distribution, with the reported size reflecting the mean. NPs with closely related mean sizes may share overlapping protein profiles, leading to similar biodistributions and limiting statistical differentiation. Improving training and prediction accuracy would require NP synthesis with sub-nanometer standard deviations – a challenge under current bulk synthesis methods.

Duan et al., introduced a RF model incorporating a novel descriptor, fluorescence change (FC), for predicting the PCF.[14] This approach was tested on a diverse range of ENMs, including metals, metal oxides, nanocellulose, and 2D materials, demonstrating its broad applicability.

A high-throughput fluorescamine labeling method, which eliminated the need for washing steps, was employed to generate the FC descriptor. Traditional physicochemical properties such as hydrodynamic diameter and zeta potential were used as benchmarks for comparison, with FC outperforming these descriptors. The physicochemical properties of proteins identified in the corona, including molecular weight, isoelectric point, GRAVY score, and amino acid composition, were calculated using ProtParam to supplement the analysis.

A significant finding, consistent with Walkey et al., was the material-dependent performance of the model.[10] Prediction accuracy improved when the training set included ENMs with properties similar to those in the test set. For instance, using cellulose and spherical ENMs in the training set reduced model accuracy, suggesting that the model's applicability domain may be constrained by ENM shape.

Ban et al., compiled the largest database to date for PC research through an extensive literature review.[15] Using this dataset, they developed a RF model to predict the composition of the PC layer and its impact on cellular recognition. However, the model's performance was inconsistent; while it successfully predicted the presence of certain proteins on NP surfaces, accuracy varied across different cases. These inconsistencies aligned with previous reports by Duan et al.,[14] and Walkey et al.,[10] which highlighted the challenges of creating generalized models capable of handling diverse ENM types.

The variability in physicochemical properties and interactions across different nanomaterials likely contributed to Ban et al., model inconsistencies. This underscores the significant challenge in developing universal predictive models for PC composition and its biological implications. Future efforts should focus on expanding datasets to capture the diversity of ENM properties and interactions, thereby enhancing model generalizability and reliability.

Innovative approaches to predicting PC formation using ML

Yan et al., proposed a groundbreaking approach to nanomaterial annotation, inspired by facial recognition technology, to improve the predictive modeling of protein adsorption on NPs.[16] This method transformed three-dimensional nanostructure data into image formats suitable for convolutional neural network (CNN) analysis. Using a custom program named "ViNAS," they processed data from 147 NPs (36 for protein adsorption and 77 for cellular uptake) into standardized images through Visual Molecular Dynamics software. Written in a tool command language, the program streamlined the conversion process, ensuring high efficiency.

The study utilized a CNN model trained on nanostructure images that integrated features such as core material, size, surface ligand chemistry, and density, bypassing traditional descriptors, such as zeta potential and ligand properties. The model achieved strong performance, with R^2 values exceeding 0.68 for both 5-fold cross-validation and external validation, outperforming conventional physicochemical predictors. Validation methods, including cross-validation, external testing, and Y-scrambling, confirmed the model's robustness.

The reliance on image-based nanostructure representations underscores the potential of this approach to circumvent traditional experimental measurements, opening a promising avenue for innovative applications in nanotechnology.

Quassil et al., explored factors influencing protein binding to single-walled carbon nanotubes (SWCNTs) and developed a RF model to predict the end-state PC composition.[17] Their findings revealed that proteins with high solvent-exposed glycine residues or amino acids lacking secondary structures (e.g., alpha helices or beta sheets) were more likely to bind to the curved SWCNT surface, suggesting that structural flexibility enhances binding. Conversely, proteins rich in leucine or amino acids associated with planar beta-sheet domains exhibited weaker binding, indicating that structural rigidity impedes interactions. The GRAVY score (grand average hydropathy) also emerged as a key determinant, with lower scores correlating with a reduced likelihood of PC formation (Figure 1).

The study focused on predicting the final PC composition rather than the dynamic stages of corona formation. Although the model accurately predicted the end-state binding, it struggled to capture intermediate protein adsorption dynamics, resulting in a poor correlation

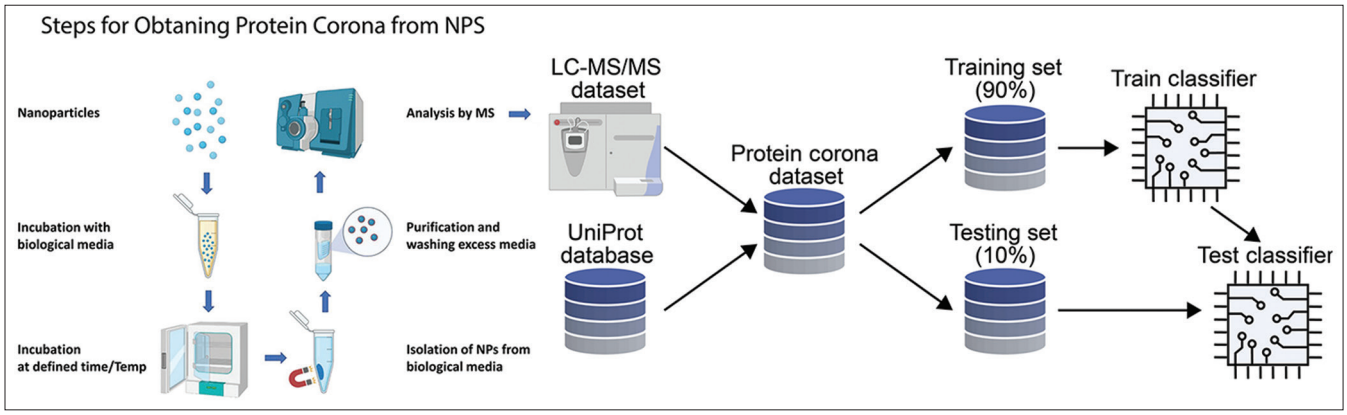

Figure 1: Schematic representation of machine learning development based on protein corona analysis using liquid chromatography-tandem mass spectrometry. (Reproduced with permission)[17,23]

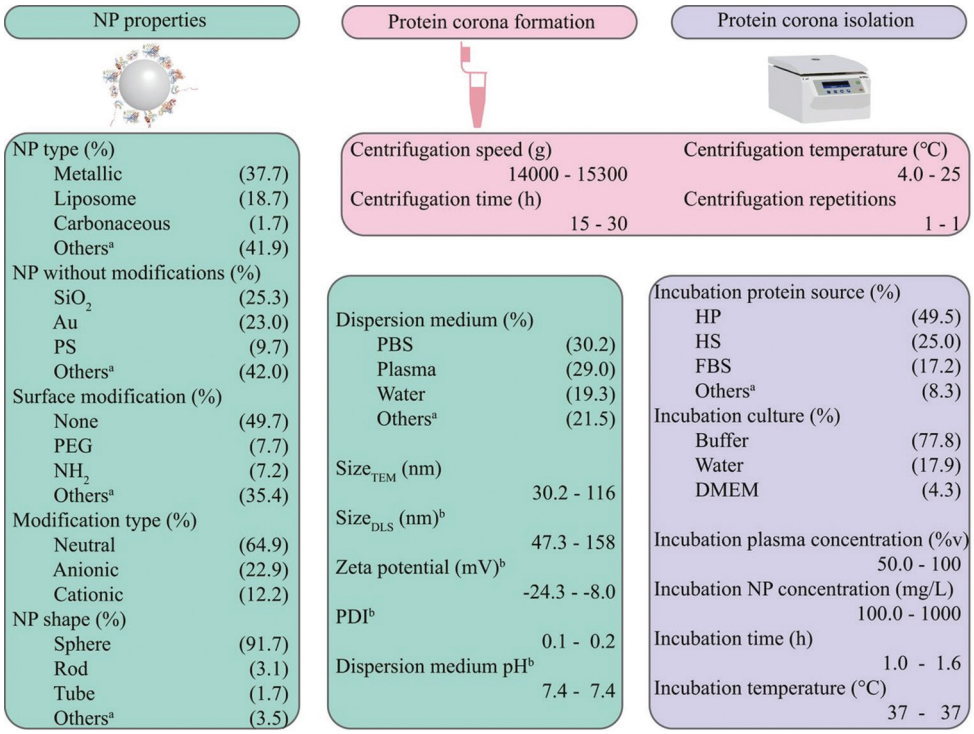

Figure 2: Descriptors by Ban et al., with permission.[15] Bio interactions

between predicted and experimentally observed temporal protein-SWCNT interactions. Despite these limitations, the RF classifier provided valuable insights into the physicochemical factors influencing protein-nanotube interactions, offering a promising tool for predicting corona formation across diverse NP systems (Figure 2).

In 2024, Liao et al., addressed imbalanced data distributions in predicting PC composition, emphasizing the importance of data quality in enhancing model performance.[20] While prior studies focused on predictive models, Liao et al., employed resampling techniques to balance dense and sparse data distributions, significantly improving prediction accuracy.

Three resampling strategies were implemented: Random Oversampling, Synthetic Minority Oversampling Technique for Regression, and the Weighted Relevance-based Combination Strategy. These methods improved dataset distribution, increasing R^2 by 0.06 and reducing RMSE by 0.11. However, the model performed less effectively for proteins with low relative protein abundance (RPA), a limitation attributed to competitive adsorption effects. These findings underscore the critical role of resampling in addressing data imbalance and improving the predictive accuracy of PC models.

Finally, Fu et al., conducted a comprehensive study to predict the RPA of multiple proteins within the PC using

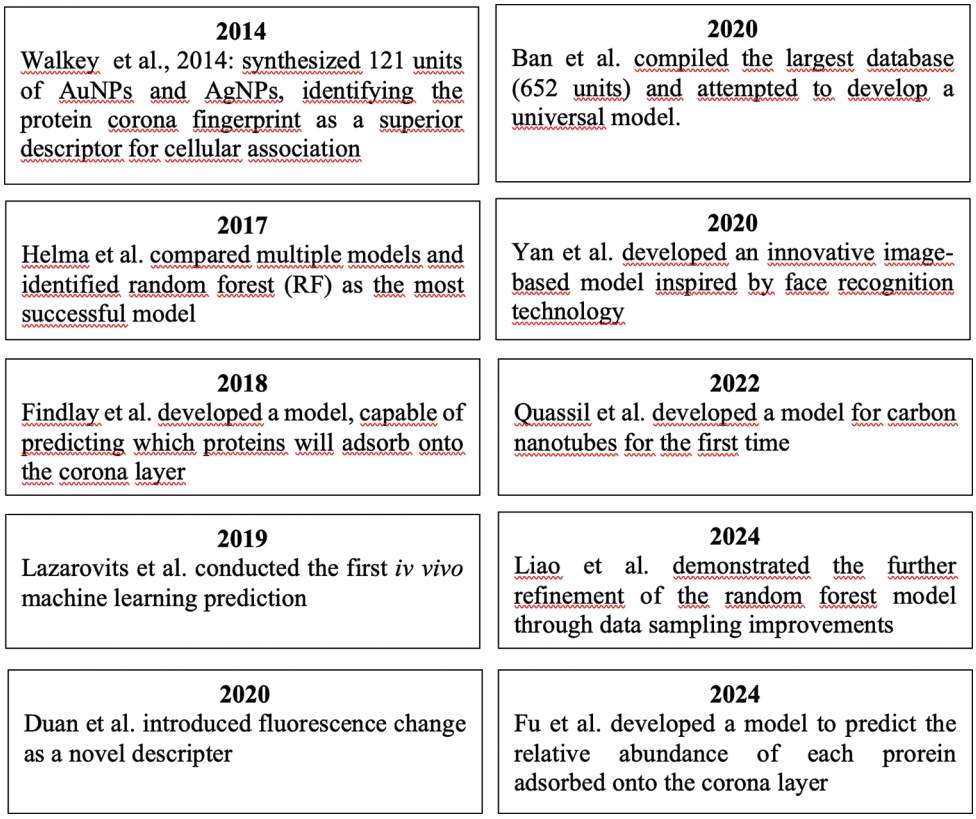

Figure 3: Key findings from 10 years of machine learning in protein corona research

ML models (Figure 3).[21] This research uniquely combined classification and regression tasks to evaluate algorithmic performance and identify key features influencing RPA through interpretable analysis. It represents a pioneering effort to provide interpretable predictions for the RPA of multiple proteins in the PC.

The Extremely Randomized Trees (ERT) algorithm excelled in binary classification tasks, accurately predicting whether a protein adsorbs onto a NP. For regression tasks, RF achieved the highest R^2 performance, while ERT minimized RMSE, demonstrating complementary strengths across different tasks. Feature importance analysis identified the NP core material as the most significant determinant of protein adsorption. In addition, centrifugation speed emerged as a critical factor in determining the relative abundance of adsorbed proteins. These results provide valuable insights into the physicochemical parameters driving PC formation and highlight the potential of ML to unravel complex nano

DICUSSION

The importance of predicting PC formation in nanotechnology

The ability to accurately predict PC formation is critical for understanding cellular responses, assessing local and environmental toxicity, and optimizing the functionality of nanotechnologies. Although ML has significantly enhanced predictive efficiency in this domain, further research and expanded datasets are essential to address existing challenges.

Current challenges in ML for PC research

1. Material-specific limitations in predictive models
   Developing a universal predictive model for diverse NPs remains a significant challenge due to the material-specific nature of PC formation. Walkey et al., demonstrated that the core material of NPs (e.g., gold vs. silver) strongly influences PC composition, with minimal overlap observed between similar NPs.[10] Their findings revealed that models trained on one NP type often fail to predict cell associations for other types, underscoring the dependency on NP material.
   Similarly, the model developed by Ban et al., faced inconsistent performance, despite leveraging a diverse dataset comprising various ENMs.17 This variability reflects the differences in physicochemical properties and interactions among ENMs, complicating the creation of generalized models. Duan et al., further highlighted the issue, showing reduced model accuracy when training and test datasets involved NPs with differing shapes or properties.[14] These findings collectively emphasize the necessity of

tailored or hybrid modeling approaches to address the heterogeneity of NP types (Table 1).

2. Variability in performance metrics
   The use of inconsistent performance metrics in PC ML studies hinders direct comparisons across research. Liu et al., critiqued the reliance on leave-one-out cross-validation, noting its tendency to produce overly optimistic accuracy estimates, particularly for small datasets.[11] This metric evaluates model performance using subsets of data that may not be sufficiently independent from the training set, increasing the risk of overfitting (Table 2).
3. Dependence on secondary data sources
   Many researchers rely on data from published studies rather than conducting independent NP synthesis or purchasing materials from established suppliers. This approach, driven by the complexity and expense of generating new datasets, is exemplified by Ban et al., who compiled the largest PC database to date by aggregating literature data.[15] Similarly, Yan et al., converted data from 147 NPs into image-based formats for CNN analysis.[16]
   While leveraging existing datasets reduces the time and resources required, it also introduces challenges such as dataset inconsistencies, reporting biases, and a lack of standardization. These issues can compromise model reliability and generalizability, underscoring the need for transparent reporting and harmonized protocols to enable robust model development and meaningful comparisons.
4. Variability in key descriptors
   Different studies have identified distinct descriptors as critical for predicting NP-PC interactions, reflecting variations in experimental setups and modeling approaches. For instance, Liu et al.,[11] identified APOB as the most significant descriptor for correlating gold NP (AuNP) cell associations, whereas Papa et al.,[18] highlighted hydrodynamic diameter after serum exposure and A1AT as key predictors. These discrepancies reflect the diversity in datasets, NP types, and feature selection methods, complicating efforts to establish universally important descriptors.
5. In *vitro* versus *in vivo* limitations
   Many studies rely on *in vitro* methods to predict PC interactions, which may not accurately represent *in vivo* conditions.[10-12,14-21] This discrepancy highlights the need for models that better account for the complexities of biological environments.
6. Lack of standardization
   The absence of standardized protocols for experimental procedures and data collection in PC research poses a significant barrier to developing robust ML models.

Table 1: Overview of datasets created by authors, including detailed information on the descriptors used for model development[10,12-17]

Refs	Materials	Dataset	Descriptors (Detailed)	Inc. Envir./Cell Lines	Type of Data
Walkey et al. [10]	• Surface modified gold NPs with 15, 30, 60 nm cores • Surface modified silver NPs with core size of approximately 40 nm • 67 different anionic, cationic or neutral ligands attached to both type NPs	• Self-made 121 units	• All adsorbed proteins (Protein corona fingerprint) • Relative abundance of individual proteins • Hyaluronan-binding proteins • Physicochemical properties: • Size: NP core size and hydrodynamic diameter • Surface charge: Zeta potential • Aggregation state • Localized surface plasmon resonance index • Total adsorbed protein density	• Human serum • A549 human lung epithelial carcinoma cells	• Experimental
Findlay et al. [14]	• 10 and 100 nm sizes of Ag NPs • Two different surface coatings were used: anionic citrate and cationic branched polyethylenimine (BPEI)- coated NPs	• 6 units from Eigenheer et al.	• Protein characteristics: • Isoelectric point (pI), Protein weight (MW), Protein abundance, % Positive charged AAs, % Negative charged AAs, % Hydrophilic AAs, % Aromatic AAs, % Cysteine AA • Physicochemical properties • Size, Surface charge • Solvent characteristics • Cysteine concentration, NaCl concentration	• Isolated soluble protein from 1 L BY4 yeast cells	• Literature review

(Contd..)

Table 1: (*Continued*)

Refs	Materials	Dataset	Descriptors (Detailed)	Inc. Envir./Cell Lines	Type of Data
Lazarovits et al. [15]	• Pegylated gold NPs, 8, 15, 35, 50, and 80 nm • A consistent surface density of 5 PEG/nm • The study also used two "unknown" NPs, termed UK1 and UK2, to test the neural network's prediction accuracy	• 5 units	• Label-Free Quantitative (LFQ) intensities of surface-adsorbed proteins: The primary inputs fort he neural network are the LFQ intensities of proteinsthat adsorb to the NP surface • Time points: Protein data is collected at multiple time points (1, 2, 4, 8, and 24 hours) pos-injection to capture the dynamic changes in protein adsorption on the NP surface over time • NP size: Different sizes	• Rat	• Experimental
Duan et al. [16]	• 15 Metallic NPs CuO, TiO2, ZnO, V_2O_5, Citrate-capped (AuNPs, and AgNPs 15 nm), CeO_2 (10 and 30 nm), SiO_2 (1 and 10% Ag-doped SiO_2 NPs: 8, 11, and 1 5 nm), Al_2O_3 (30 nm), Fe2O3 (10 nm), AgNPs (20 nm), and WO_3 NPs • 3 Cellulose-based ENMs Cellulose nanofibrils (CNF-50 and CNF-80) with single-fibril diameters of 50 and 80 nm, respectively. Cellulose nanocrytals (CNC-250) with a length and diameter of 250 nm • 4 two dimentional (2D) ENMs were used Graphene	• 22 units	• ENM descriptors: Fuorescence change (FC) is the primary novel descriptor used in this study. FC change observed when a protein interacts with an ENM. FC is measured by labeling a protein with fluorescamine, a non-fluorescent dye that becomes fluorescent when it reacts with primary amines on the protein surface • Physicochemical properties of ENMs: The study also used conventional descriptors such as size and surface charge. However, these descriptors were found to be less effective than FCs in predicting protein corona formation. • Other physicochemical properties, such as hydrodynamic diameter and zeta-potential, were measured and reported for the ENMs, but they were not used as descriptors	• Human serum	• Experimental
Ban et al. [17]	• 40 types of ENMs: Metallic NPs: Au, Ag, Fe_3O_4, TiO_2 • Nonmetallic NPs: SiO_2, Si • Carbonaceous NPs: Multi- and single-valled carbon nanotubes • Liposomal NPs: Cholesterol-phosphatidylcholine and thiolated amino-PEG • Other ENMS: Calcium phosphate, Polystyrene, and Zeolite • 50 types of surface modifications	652 units	• The 8 qualitative factors: NP type, size, core material, surface modification, modification type, dispersion medium, incubation plasma source and incubation culture • 13 quatitative factors: Size (measured by transmission electorn microscopy), size (measured by dynamic light scattering), pH of dispersion medium, zeta potential, polydispersity index, incubation plasma concetration, incubation NP concentration, incubation time, incubation temperature, centrifugation speed, centrifugation time and temperature	• Human plasma • Human serum • Fetal bovine serum • Rat plasma • Rat serum • Murine macrophage cell line RAW264.7 • Human leukemic cell line THP-1 • Dendritic cell line DC2.4	• Literature review
Yan et al. [18]	• 123 gold NPs • 12 platinium NPs • 12 Palladium NPs • 91 unique surface ligands attached	• 36 units for protein adsorption • 77 units for cellular uptake	• Core material • Size of NP core • Chemical structure of surface ligand • Surface ligand density	• Unspecified type of cream • Human alveo-based epithelial (A549) cells	• Literature review
Quassil et al. [19]	• (GT) 15-functionalized SWCNTs	• 2 units	• Physicochemical properties of protein: • Amino acid composition • Seconday structure features, which are predicted from the AA sequence using BioPython. These include the % of AAs associated with: Non-secondary structure, b-sheet, a-helix • GRAVY score • Solvent accessibility, which estimates the exposed protein surface area using NetSurfP 2.0	• Human blood plasma • Cerebrospinal fluid of the brain	

Table 2: Summary of machine learning applications in PC Research. This table provides an overview of the developed models, their specific objectives, key findings, and the validation methods employed to assess their robustness[10-21]

References	Key objective	Material	Data source	Descripters used	ML algorithm	Performance	Key Findings	Robustness validation
Walkey et al. [10]	The development of a quantitative model predicting the binding efficiency of gold nanoparticles to cells based on their protein corona layer	Surface modified by gold and silver nanoparticles	Self-made 121 units	· Adsorbed proteins (Protein corona fingerprint) · Nanopartical physicochemical properties · Total adsorbes protein density	· Partial least square regression	· Q^2LOO (0.86)	· This model showed that the protein corona encodes more biologically relevant information than physical properties for predicting cellular associations. Up to 50% more model accuracy compared t to model made with physical properties	· Leave One-Out validation (Q^2LOO) · External validation with silver nanoparticles
Liu et al. [11]	Identify specific corona serum proteins and/or NP physicochemical most relevant to predicting NP-cell association, and develop a model based on these key parameters to minimize overfitting	Surface modified gold and silver nanoparticles	84 units from Walkey data set	· Zeta potential · 5 adsorbed proteins: APOB, AIAT, IGLL5, HRG, APOE (Identified by the model itself)	· Non-lineer epsilon support vector regression (e-SVR)	· R^2 (0.89)	· APOB was the most frequently selected descripter and demonsrated the greatest influence on the correlation with AuN cell association	· Y-randomization · Bootstrap resampling
Papa et al. [12]	Compare multiple modeling techniques to determine the most effective approach for predicting bioactivity	Surface modified with gold silver nanoparticles	84 units from Walkey data set	· Zeta potential · 5 adsorbed proteins (VINC, KNG1, AIAT, CO4B, GFAB)	Top three performers · Projection persuit regression (PPR) · EARTH · RF	· PPR model: Q^2LOO (0.81), R^2 (0.79)	· PPR, EARTH, and RF were the best models · AIAT was the most important descriptor, · Introduction of data splitting (4-fold cross validation)	· Leave One-Out cross validation · Y-randomization · 4-fold cross validation
Helma et al. [13]	Identify which descripter + machine learning (ML) algorithm combination results in best prediction nanoparticle association	Surface modified with gold silver nanoparticles	8121 units from Walkey data set	· Adsorbed proteins · Nanoparticle size · Absorbance spectrophotometry · Zeta potential · Total adsorbed protein density	· RF	· R^2 (0.62) · RMSE (1.62)	· Protein corona descriptors combined with a weighted random forest algorithm yielded the best ML algorithm + descriptor combination	· 10-fold cross validation
Findlay et al. [14]	Predict which proteins will adsorb onto protein corona layer (predict the protein corona fingerprint itself)	Surface modified silver nanoparticles	6 units Eigenheer et al.	· Protein characteristics (UniProt) · Nanoparticle physicochemical properties · Solvent characteristics	· RF	· Precision (0.76) · Recall (0.86) · F1 (0.81) · Accuracy (0.75) · AUROC (0.81)	· Protein isoelectric point and %charged amino acids were found to be the most important features detemining which proteins will adsorb onto corona	Y-randomization

(Contd...)

Table 2: (*Continued*)

References	Key objective	Material	Data source	Descripters used	ML algorithm	Performance	Key Findings	Robustness validation
Lazarovits et al. [15]	Predict nanoparticle clearence and organ accumulation	PEGylated gold nanoparticles	Self-made 5 units	· Nanoparticle size · Time points · Adsorbed proteins	· Neural network	· Accuracy (half-life: 0.77) · Accuracy (accumulation: 0.93)	· It was found that clearence is dictate by combinations of proteins that from patients on the NP surface not by individual proteins. These combinations indicate to the body whether or not that NP should be cleared	· K-fold cross validation · Early stopping rule · Double-blind external validation
Duan et al. [16]	Introduce the use of fluoresce change (FC) from fluorescamine labeling as a novel descriptor for predicting protein corona fngerprint	22 different metal and metal oxides nanomaterials, nanocellulose, and 2D ENMs	Self-made 24 units	· Fluorescence change (FC) · Protein characteristics (ProParam)	· RF	· F1 (0.84) · Precision (0.85) · Recall (0.85)	· First study to include a wide range of nanomaterials in the model. The FCs were shown to be more effective at predicting the the protein corona composition when compared to the traditional descriptors of size and surface charge	· Jackknife resampling
Ban et al. [17]	Establish a comprehensive dataset for machine learnine training.Create and test a model with this dataset	40 different types of nanomaterials including metal and metal oxides, lipid-based, polymer-based, carbon nanotubes	652 units from 56 different papers	· Centrifugation parameters · NP parameters · Incubation medium	· RF	· R^2 mostly > 0.6	· The model was not consistent in its prediction capability. While it could predict some proteins on surface with good performance, it varied heavily · Model with widest data set up to date · Most important features: NP without modification and surface modification	· 10-fold cross-validation
Yan et al. [18]	Predict protein adsorption and cellular uptake	12 gold, 12 platinium, and 12 palladium nanoparticles	36 units for protein adsorption, 77 for cellular uptake	· Nanoparticle images generated by inhouse coded tool	· Neural network	· Protein adsorption: R^2 (0.91), RSME (7.14) · Cellular uptake : R^2 (0.76), RMSE (3.04)	· Drawing inspiration from facial recognition technology, a novel nanostructure annotation method was developed to automatically convert nanostructures into images, eliminating the need for complex nanodescriptor calculations	· Y-randomization · 5-fold cross validation

(Contd...)

Table 2: (*Continued*)

References	Key objective	Material	Data source	Descripters used	ML algorithm	Performance	Key Findings	Robustness validation
Quassil et al. [19]	Prediction protein adsorption	(GT) 5-funtionalized SWCNTs and (GT) 6-functionalized SWCNTs	2 units synthesized by Pinals et al.	· Amino acid composition · Protein secondary structure · Protein hydrophobicity · Amino acid solvent accessibility	· RF	· Accuracy (0.751) · AUC (0.727) · Precision (0.656) · Recall (0.647)	· It was found that proteins with a high content of solvent-exposed glycine rsidues or AAs not associated with secondary structures (e.g. alpha helices or beta sheets) were more likely to bind to the curved SWCNT surface, suggesting that flexibility enhances binding	· Stratified shuffle split · Oversampling · Cross-biofluid validation
Liao et al. [20]	Improve the accuracy of predicting the protein corona fingerprint by improvement of the distribution of the dataset by resampling methods	40 different types of nanomaterials including metal and metal oxides, lipid-based, polymer-based. Carbon nanotubes	652 units from Ban et al.	· 21 factors from Ban et al. descriptors	· RF	· R^2 (0.62) · RMSE (1.o1)	· Resampling demonstrated an improvement in model performance, with the R^2 increasing by (0.06) and the RMSE decreasing by (0.11) · The four most important features identified were; o Incubation plasma concentration (importance: 0.27) o Polydispersity index (PDI) (Importance: 0.26) o Surface modification (Importance: 0.23) o NPs without modification (Importance: 0.22)	· 10-fold cross-validation
Fu et al. [21]	Predict the relative protein abundance (RPA)	40 different types of nanomaterials including metal and metal oxides, lipid-based, polymer-based. Carbon nanotubes	652 units from Ban et al.	· 21 factors from Ban et al. descriptors	· RF · Extremely randomized trees (ERT)	· ERT: AUROC: (0.969), Recall (0.877), Precision (0.915), F1 (0.893), Accuracy (0.919) · RF: RMSE (0.760), R^2 (=.534)	· NP core material was found to be the most important factor for determining protein adsorbtion · Centrifugation speed was the most critical factor for identifying relative abundancy of adsorbed proteins	· 10-fold cross-validation · Oversampling

Standardization is essential for ensuring the reliability and reproducibility of results. For example, Ban et al.,[15] encountered inconsistent outcomes despite adopting rigorous literature review measures, likely due to variations in underlying experimental methodologies.

7. Ethical considerations
   Ethical concerns regarding the potential of ML to displace scientists' roles in research warrant careful consideration.[22] While automation can accelerate discoveries, it is crucial to balance technological advancements with the preservation of scientific expertise.

CONCLUSION

The integration of ML into PC research represents a transformative approach, enabling efficient, accurate, and scalable analysis of NP interactions with biological systems. Traditional methods such as LC-MS/MS, while precise, are limited by their high resource demands, reliance on expert operators, and low throughput. ML addresses these challenges by leveraging computational power to analyze and predict PC dynamics, opening pathways to tailor nanomaterial designs for improved drug delivery, diagnostics, and toxicity mitigation.

Recent advancements, including supervised learning models and RF Classifiers, have showcased promising results in correlating PC compositions with biodistribution patterns, immune responses, and cellular interactions. However, achieving clinical translation requires addressing several challenges. These include the scarcity of large, high-quality datasets, the complexity of identifying relevant features amidst noisy data, and the variability inherent in biological systems. The need for standardized experimental protocols and ethical considerations further underline the multidisciplinary nature of the task at hand.

Despite these hurdles, ML's potential to streamline PC research holds great promise. By reducing dependency on labor-intensive methodologies, it paves the way for rapid, *in silico* predictions of NP behaviors in diverse biological contexts, thereby accelerating the development of safer and more effective nanomedicines.

FUTURE DIRECTIONS

1. Standardization and protocol development
   Establish universally accepted protocols for data collection, NP preparation, and PC isolation to improve reproducibility and model reliability.
2. Data expansion and sharing
   Foster international collaboration to create larger, more diverse datasets. Open-access repositories could enhance data availability and drive innovation.
3. Hybrid approaches
   Combine traditional proteomic methods with ML to validate predictions and improve model accuracy, particularly in unexplored scenarios.
4. Feature optimization
   Develop methods to identify key features of protein-NP interactions, such as surface chemistry, protein structure, and environmental conditions, to reduce noise and enhance model interpretability.
5. Computational resource management
   Explore cloud computing and distributed systems to overcome computational limitations when handling large datasets.
6. Interdisciplinary collaboration
   Strengthen ties between biologists, chemists, and data scientists to integrate domain-specific insights into machine-learning frameworks effectively.
7. Clinical translation frameworks
   Develop pipelines to bridge laboratory findings with clinical applications, ensuring that machine-learning predictions align with real-world biological complexities.
8. Ethical considerations and workforce evolution
   Address the ethical implications of replacing traditional roles with ML by fostering reskilling programs and highlighting new opportunities in data-driven nanomedicine.

By addressing these directions, the field can make significant strides toward leveraging ML for practical and impactful advancements in PC research, paving the way for breakthroughs in nanomedicine.

REFERENCES

1. Pelaz B, Alexiou C, Alvarez-Puebla RA, Alves F, Andrews AM, Ashraf S, et al. Diverse applications of nanomedicine. ACS Nano. 2017;11(3):2313-2381.
   https://doi.org/10.1021/acsnano.6b06040
2. Dawson KA and Yan Y. Current understanding of biological identity at the nanoscale and future prospects. Nat Nanotechnol. 2021;16(3):229-242.
   https://doi.org/10.1038/s41565-021-00860-0
3. Sing AV, Maharjan RS, Kanase A, Siewert K, Rosenkranz D, Singh R, et al. Machine-learning-based approach to decode the influence of nanomaterial properties on their interaction with cells. ACS Appl Mater Interfaces. 2021;13(1):1943-1955.
   https://doi.org/10.1021/acsami.0c18470
4. Mahmoudi M, Landry MP, Moore A and Corea R. The protein corona from nanomedicine to environmental science. Nat Rev Mater. 2023;8(7):422-438.
   https://doi.org/10.1038/s41578-023-00552-2
5. Kokkinopoulou M, Simon J, Landfester K, Mailänder V and Lieberwirth I. Visualization of the protein corona: Towards a

biomolecular understanding of nanoparticle-cell-interactions. Nanoscale. 2017;9(25):8858-8870.
https://doi.org/10.1039/C7NR02977B

6. Ashkarran AA, Dararatana N, Crespy D, Caracciolo G and Mahmoudi M. Mapping the heterogeneity of protein corona by *ex vivo* magnetic levitation. Nanoscale. 2020;12(4):2374-2383.
https://doi.org/10.1039/C9NR10367H
7. Kelly PM, Åberg C, Polo E, O'Connell A, Cookman J, Fallon J, et al. Mapping protein binding sites on the biomolecular corona of nanoparticles. Nat Nanotechnol. 2015;10(5):472-479.
https://doi.org/10.1038/nnano.2015.47
8. Xu M, Soliman MG, Sun X, Pelaz B, Feliu N, Parak WJ, et al. How entanglement of different physicochemical properties complicates the prediction of *in vitro* and *in vivo* interactions of gold nanoparticles. ACS Nano. 2018;12(10):10104-10113.
https://doi.org/10.1021/acsnano.8b04906
9. Naqa IE and Murphy MJ. What is machine learning? In: El Naqa I, Li R and Murphy MJ, editors. Machine Learning in Radiation Oncology: Theory and Applications. Berlin: Springer; 2015. p. 3-11.
https://doi.org/10.1007/978-3-319-18305-3_1
10. Walkey CD, Olsen JB, Song F, Liu R, Guo H, Olsen DW, et al. Protein corona fingerprinting predicts the cellular interaction of gold and silver nanoparticles. ACS Nano. 2014;8(3):2439-2455.
https://doi.org/10.1021/nn406018q
11. Liu R, Jiang W, Walkey CD, Chan WC and Cohen Y. Prediction of nanoparticles-cell association based on corona proteins and physicochemical properties. Nanoscale. 2015;7(21):9664-9675.
https://doi.org/10.1039/C5NR01537E
12. Findlay MR, Freitas DN, Mobed-Miremadi M and Wheeler KE. Machine learning provides predictive analysis into silver nanoparticle protein corona formation from physicochemical properties. Environ Sci Nano. 2018;5(1):64-71.
https://doi.org/10.1039/C7EN00466D
13. Lazarovits J, Sindhwani S, Tavares AJ, Zhang Y, Song F, Audet J, et al. Supervised learning and mass spectrometry predicts the *in vivo* fate of nanomaterials. ACS Nano. 2019;13(7):8023-8034.
https://doi.org/10.1021/acsnano.9b02774
14. Duan Y, Coreas R, Liu Y, Bitounis D, Zhang Z, Parviz D, et al. Prediction of protein corona on nanomaterials by machine learning using novel descriptors. NanoImpact. 2020;17:100207.
https://doi.org/10.1016/j.impact.2020.100207
15. Ban Z, Yuan P, Yu F, Peng T, Zhou Q and Hu X. Machine learning predicts the functional composition of the protein corona and the cellular recognition of nanoparticles. Proc Natl Acad Sci U S A. 2020;117(19):10492-10499.
https://doi.org/10.1073/pnas.1919755117
16. Yan X, Zhang J, Russo DP, Zhu H and Yan B. Prediction of nano-bio interactions through convolutional neural network analysis of nanostructure images. ACS Sustain Chem Eng. 2020;8(51):19096-19104.
https://doi.org/10.1021/acssuschemeng.0c07453
17. Ouassil N, Pinals RL, Del Bonis-O'Donnell JT, Wang JW and Landry MP. Supervised learning model predicts protein adsorption to carbon nanotubes. Sci Adv. 2022;8(1):eabm0898.
https://doi.org/10.1126/sciadv.abm0898
18. Papa E, Doucet JP, Sangion A and Doucet-Panaye A. Investigation of the influence of protein corona composition on gold nanoparticle bioactivity using machine learning approaches. SAR QSAR Environ Res. 2016;27(7):521-538.
https://doi.org/10.1080/1062936X.2016.1197310
19. Helma C, Rautenberg M and Gebele D. Nano-Lazar: Read across predictions for nanoparticle toxicities with calculated and measured properties. Front Pharmacol. 2017;8:377.
https://doi.org/10.3389/fphar.2017.00377
20. Liao R, Zhuang Y, Li X, Chen K, Wang X, Feng C, et al. Unveiling protein corona composition: Predicting with resampling embedding and machine learning. Regen Biomater. 2024;11:rbad082.
https://doi.org/10.1093/rb/rbad082
21. Fu X, Yang C, Su Y, Liu C, Qiu H, Yu Y, et al. Machine learning enables comprehensive prediction of the relative protein abundance of multiple proteins on the protein corona. Research (Wash D C). 2024;7:0487.
https://doi.org/10.34133/research.0487
22. Maryasin B, Marquetand P and Maulide N. Machine learning for organic synthesis: Are robots replacing chemists? Angew Chem Int Ed Engl. 2018;57(24):6978-6980.
https://doi.org/10.1002/anie.201803562
23. Hajipour MJ, Safavi-Sohi R, Sharifi S, Mahmoud N, Ashkarran AA, Voke E, et al. An overview of nanoparticle protein corona literature. Small. 2023;19(36):2301838.
https://doi.org/10.1002/smll.202301838

Authors' Contribution:
BY- Collected the articles and drafted the manuscript; **NÖ-** Reviewed, revised, and finalized the manuscript

Work attributed to:
Girne American University, Girne, Mersin, Turkey.

Orcid ID:
Nazmi Özer - https://orcid.org/0000-0002-2630-741X
Berkant Yığın - https://orcid.org/0009-0003-5308-3268

Source of Support: Nil, **Conflicts of Interest:** None declared.