Random Forest Modeling of Curcumin-Loaded PLGA Nanoparticle Size: Mechanistic Insights and In-Silico Formulation Screening

Rumana Ferdushi

Department of Biomedical Engineering, Yonsei University, Wonju 26493, Republic of Korea

Abstract
The therapeutic potential of curcumin and the biodegradability of poly (lactic-co-glycolic acid) (PLGA) have led to extensive research on curcumin-loaded PLGA nanoparticles for drug delivery applications. The intricate relationships between formulation composition and processing factors make it difficult to precisely control nanoparticle size during nanoprecipitation. Using experimentally determined nanoprecipitation factors, Random Forest regression was used in this study to establish a data-driven modeling framework for predicting nanoparticle size. After duplicate aggregation, 19 distinct curcumin-loaded PLGA formulations were examined. Significantly surpassing linear regression ($R2 \approx 0.04$), Leave-One-Out Cross-Validation (LOOCV) produced moderate predictive performance ($R2 \approx 0.32$), suggesting nonlinear correlations between process factors and particle size. Feature importance and partial dependence analyses highlighted second-step centrifugation speed and PLGA concentration as the most influential variables within the investigated design space, whereas vortex duration showed minimal marginal effect. While these patterns are consistent with known nanoprecipitation behavior, they should be regarded as data-driven hypotheses rather than definitive mechanistic proof, given the limited number of formulations and moderate predictive performance. Within this context, the Random Forest model serves as an exploratory, pre-experimental screening tool that can prioritize promising formulation conditions for subsequent targeted validation in curcumin-loaded PLGA nanoparticle development.

Keywords:
Curcumin-loaded PLGA nanoparticles, Random Forest regression, Particle size prediction, In-silico formulation screening

1. Introduction

Curcuma longa is the natural source of curcumin, a polyphenolic molecule that has garnered a lot of interest because of its purported anti-inflammatory, antioxidant, and anticancer qualities. Its fast disintegration, low bioavailability, and poor water solubility, however, hinder its clinical translation [1]. One well-established method to optimize controlled distribution and curcumin stability is encapsulation within biodegradable polymeric nanoparticles, especially those made of poly (lactic-co-glycolic acid) (PLGA) [2-4]

Nanoprecipitation is a commonly employed method for producing PLGA-based nanoparticles due to its simplicity and scalability [5]. Nevertheless, nanoparticle size is highly sensitive to multiple interacting variables, including polymer concentration, mixing intensity, and post-precipitation processing steps such as centrifugation. Achieving reproducible and target-specific particle sizes therefore requires systematic optimization of both formulation and process parameters [6].

Traditional experimental optimization approaches are often time-consuming and inefficient when multiple variables interact nonlinearly. Machine learning (ML) methods provide an alternative strategy by learning complex relationships between formulation variables and measured outcomes [7, 8]. Among ML techniques, Random Forest regression is particularly suitable for small experimental datasets and can manage multicollinearity without requiring strict parametric assumptions [7, 9, 10].

Most recent machine learning studies for nanoparticle formulation have relied on relatively large datasets compiled from many experiments or literature sources, often comprising tens to hundreds of unique PLGA-based formulations [11-13]. In contrast, our study deliberately focuses on a small, prospectively generated design (19 distinct curcumin-loaded PLGA formulations) to explore how interpretable ML can assist mechanistic hypothesis generation and pre-experimental screening when only limited in-house data are available. Accordingly, we do not aim to identify a universally optimal algorithm, but rather to evaluate whether a representative non-linear ensemble method (Random Forest) can provide additional insight beyond linear regression within this constrained experimental space.

The objective of this study was to (i) model the particle size of curcumin-loaded PLGA nanoparticles using Random Forest regression, (ii) interpret the relative importance of formulation variables through feature importance and partial dependence analysis, and (iii) demonstrate the feasibility of in-silico screening within the experimentally explored parameter space. By integrating experimental nanoprecipitation data with interpretable ML modeling, this work aims to provide a rational framework for data-assisted nanoparticle formulation design.

Manuscript received February 5, 2026
Manuscript revised February 20, 2026
https://doi.org/10.22937/IJCSNS.2026.26.2.2

2. Experimental Section

In this study, we developed and optimized poly(lactic-co-glycolic acid) (PLGA)-curcumin nanoparticles using nanoprecipitation, focusing on key formulation parameters to control particle properties. Characterization techniques were applied to assess size, zeta potential, and polydispersity, while machine learning models were utilized to predict outcomes and elucidate parameter influences, enabling efficient formulation design for potential drug delivery applications.

2.1 Materials

All reagents used in this study were commercially available and used as received. PLGA (Resomer® RG503H, acid terminated, molecular weight: 24–38 kDa, lactide:glycolide = 50:50) and poly (vinyl alcohol) (PVA; 80% hydrolyzed, molecular weight: 9–10 kDa) were purchased from Sigma-Aldrich (St. Louis, MO, USA) and curcumin (≥95% purity, Cur) were purchased from Sigma-Aldrich (St. Louis, MO, USA). Poly (lactic-co-glycolic acid) (PLGA) and curcumin were used for nanoparticle preparation. Polyvinyl alcohol (PVA) was used as a stabilizer in the aqueous phase. All solvents were of analytical grade and used without further purification. Deionized water was used throughout the experiments.

2.2 Preparation of PLGA–Curcumin Nanoparticles

Nanoprecipitation was used to create PLGA–curcumin nanoparticles. The organic phase was created by dissolving PLGA and curcumin in an organic solvent. Next, under carefully regulated stirring circumstances, the organic phase was gradually introduced to an aqueous phase that contained PVA. PLGA concentration (mg/mL), curcumin concentration (mg/mL), vortex time, sonication time, stirring time, and centrifugation speed and duration (first and second cycles) were among the formulation and processing parameters that were changed throughout the studies. To get rid of extra stabilizer and unencapsulated medication, the nanoparticle dispersion was centrifuged after nanoprecipitation. Before being characterized, the pellets were kept at 4 °C after being reconstituted in deionized water. Nineteen distinct formulation conditions in all were examined. Before machine learning analysis, replicates were averaged at the formulation level.

2.3 Particle Size and Zeta Potential Measurement

Using dynamic light scattering (DLS), the hydrodynamic particle size and polydispersity index (PDI) were determined. Electrophoretic light scattering was used to calculate the zeta potential. All measurements were made at room temperature. The stated values are the average of each formulation's replicate measurements.

2.4 Dataset Preparation for Machine Learning

Particle size, zeta potential, and PDI mean values at the formulation level were obtained by averaging repeat measurements for machine learning analysis. Prior to regression modeling, log-transformation (log10) was used to stabilize variance and enhance predictive accuracy because particle size values showed considerable dispersion. The experimental design's multicollinearity and the correlations between formulation variables were assessed using Pearson correlation analysis.

2.5 Machine Learning Modeling

Random Forest regression was used to model the relationship between formulation parameters and nanoparticle size. Random Forest was selected due to its robustness to nonlinear relationships and suitability for small experimental datasets.

Leave-One-Out Cross-Validation (LOOCV) was employed to evaluate predictive performance due to the limited number of unique formulations (n = 19). Model performance was quantified using:

- Coefficient of determination (R^2)
- Mean Absolute Error (MAE)
- Root Mean Square Error (RMSE)

Linear regression was implemented as a baseline model for comparison. Feature importance and partial dependence analysis were performed to interpret the influence of individual formulation parameters on predicted particle size. Random Forest regression was implemented using scikit-learn, with a deliberately constrained model capacity to mitigate overfitting in this small dataset. We limited the maximum tree depth and required a minimum number of samples per leaf, and we used a moderate number of trees to stabilize variable importance estimates while avoiding overly complex individual trees. Leave-One-Out Cross-Validation (LOOCV) was adopted to maximize training data usage under n = 19, acknowledging that LOOCV can yield high-variance performance estimates in small samples and therefore provides only a conservative indication of generalization. Linear regression was used as a parametric baseline, and the primary goal of the Random Forest model was to capture non-linear trends and rank influential variables, rather than to deliver a high-accuracy predictive tool.

3. Results

3.1 Dataset Characteristics

Nineteen distinct nanoprecipitation formulations covering specified compositional and processing ranges made up the experimental dataset (Table 1). To modify the

dynamics of particle separation, centrifugation speed was changed while PLGA concentration and vortex time were systematically changed. The design space (390–580 nm) showed moderate dispersion of nanoparticle size, indicating sensitivity to formulation parameters. Zeta potential measurements stayed in the stable negative range of -25 to -18 mV, indicating that the colloidal system was electrostatically stabilized. Acceptable size homogeneity for polymeric nanoparticle systems is indicated by the obtained PDI values.

These features attest to the dataset's potential to capture significant physicochemical heterogeneity while staying within a regulated experimental domain suitable for predictive modeling based on machine learning.

Table 1. Experimental design space and nanoparticle characterization summary.

Variable	Range	Mean ± SD
PLGA concentration (mg/mL)	5–20	12.5 ± 5.0
Vortex time (min)	5–20	12.5 ± 5.0
Centrifugation speed (rpm)	8000–12000	10000 ± 2000
Particle size (nm)	390–580	465 ± 60
Zeta potential (mV)	−25 to −18	−21 ± 2.5
PDI	0.18–0.32	0.24 ± 0.05

3.2 Correlation Analysis of Formulation Parameters

Relationships between formulation factors and processing characteristics were assessed using Pearson correlation analysis (Figure 1). Both PLGA concentration and sonication time (r = 0.90) and curcumin concentration and sonication time (r = 0.92) showed strong positive relationships.

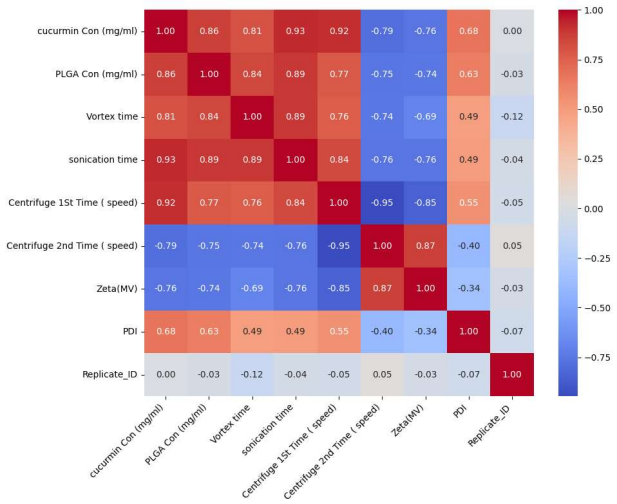

Figure 1. Pearson correlation matrix of formulation and processing parameters. Strong correlations (|r| > 0.8) are observed among PLGA concentration, sonication time, and centrifugation speed, reflecting sequential experimental variation. This multicollinearity constrains independent variable contributions and influences regression performance.

Additionally, there was a strong connection between curcumin content and centrifugation speed (r = 0.89), suggesting that the experimental design did not separate compositional and mechanical characteristics. Vortex time and other formulation variables showed somewhat positive correlations (r = 0.66–0.73), indicating a partial linkage of composition and mixing conditions. The second-step centrifugation speed, on the other hand, showed substantial inverse correlations with the first-step centrifugation speed (r = -0.96) and strong negative correlations with compositional factors (r ≈ -0.70 to -0.79), suggesting opposing modifications in separation conditions.

3.3 Random Forest Model Performance

On the log-transformed particle size data, Leave-One-Out Cross-Validation (LOOCV) was used to assess the Random Forest regression model's predictive ability. The consistency between values predicted by the model and those measured experimentally is seen in the ensuing parity plot (Figure 2). With a coefficient of determination of R2 = 0.32, the model was able to account for about 32% of the variation in nanoparticle size within the experimentally limited range. Despite moderate dispersion, the majority of data points fall close to the identification line, indicating

reasonable trend capture.

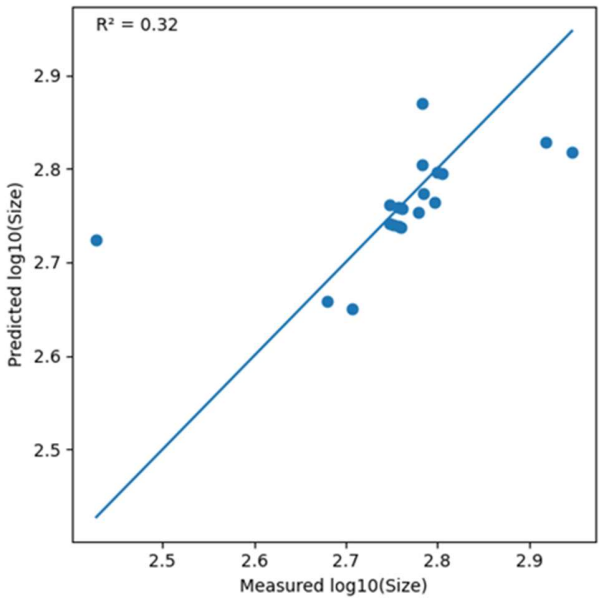

Figure 2. Leave-One-Out Cross-Validation (LOOCV) predicted versus measured log10 particle size using Random Forest regression. The model achieved moderate predictive performance ($R^2 \approx 0.32$), substantially outperforming linear regression ($R^2 \approx 0.04$), indicating nonlinear formulation–size relationships in the nanoprecipitation process.

The model effectively detects nonlinear correlations between formulation factors and particle size, despite having a weak predictive strength. The observed performance ceiling is probably caused by inter-parameter correlations and the small dataset size (n = 19). However, the model offers an exploratory framework for in-silico formulation that is statistically valid.

3.4 Feature Importance Analysis

Feature importance analysis (Figure 3) identified second centrifugation speed as the dominant contributor to particle size variation, accounting for most of the importance of the models. PLGA concentration exhibited secondary but substantial influence, while vortex time showed minimal contribution within the tested range. Second centrifugation speed was identified by the Random Forest model as the most influential predictor of particle size variation within this dataset, accounting for most of the model's total importance, according to feature importance analysis (Figure 3). Vortex time contributed very little within the measured range, whereas PLGA concentration had a minor but significant impact.

According to this ranking, mixing time has less control over the final particle size under the investigated conditions than post-precipitation mechanical processing.

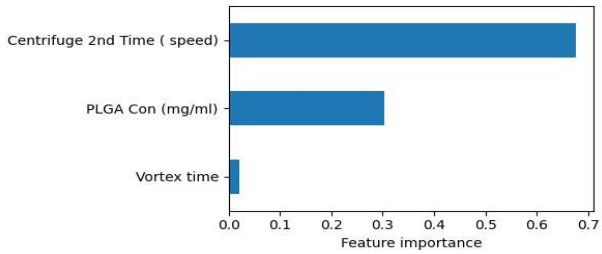

Figure 3. Random Forest Feature Importance Analysis for Log10 Particle Size Prediction

The apparent dominance of centrifugation speed is plausibly related to its role in hydrodynamic shear, aggregate removal, and particle recovery dynamics, which are known to impact nanoparticle size in PLGA systems. Nevertheless, in a small and correlated design, these contributions cannot be disentangled with complete certainty.

3.5 Partial Dependence Analysis and Process Interpretation

Additional information about the marginal impacts of important factors was obtained by partial dependence analysis (Figure 4). A monotonic positive correlation between PLGA content and expected particle size was observed, which is in line with lower droplet breakup efficiency and more polymer availability. The small marginal influence of vortex time indicates that size control in the current system is dominated by high-energy processes like sonication. Particle size and centrifugation speed showed a non-monotonic relationship, suggesting competing mechanisms between enhanced aggregate removal at intermediate speeds and potential pellet compaction or aggregation at higher rates. The Random Forest model's physical plausibility is supported by these interpretable trends, which also show that the method captures important formulation–process interactions rather than just statistical artifacts.

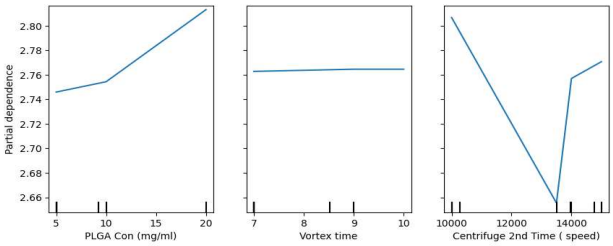

Figure 4. Partial dependence plots for key formulation variables in the Random Forest particle size model (log10 scale). PLGA concentration exhibits a monotonic positive effect on predicted particle size, vortex time shows minimal influence within the explored range, and second centrifugation speed demonstrates a non-monotonic relationship, suggesting competing clarification and aggregation mechanisms. Tick marks indicate the distribution of observed parameter values.

Collectively, these model-derived trends are consistent with established nanoprecipitation theory, in which increased polymer concentration and stronger separation conditions typically favor the formation and recovery of larger particles, whereas milder conditions can preserve smaller colloids. However, the present analysis is based on only 19 distinct formulations and a Random Forest model with modest explanatory power (LOOCV $R^2 \approx 0.32$), so the inferred mechanisms should be interpreted as statistically supported hypotheses rather than conclusive mechanistic evidence. In particular, the non-monotonic relationship between centrifugation speed and particle size suggested by the partial dependence plots requires targeted experimental verification, ideally via dedicated experiments that systematically vary centrifugation conditions while holding compositional variables constant.

3.6 In-Silico Screening for Targeted Particle Size

To assess the trained Random Forest model's practicality, formulation circumstances predicted to produce nanoparticles with a target size of 450 nm were identified through an in-silico screening process. Combinations of experimentally investigated parameter levels, such as PLGA concentration, vortex time, and second centrifugation speed, were used to create candidate formulations. The absolute divergence of the predicted particle sizes from the goal value was used to rank them.

The model found a number of possible formulations with modest deviations from 450 nm, as seen in Figure 5, with the top-ranked conditions showing expected variations of less than 5 nm. These findings show that regulated production of nanoparticles close to the target size is supported by the experimentally investigated design space without extrapolation outside of the training data range.

Additionally, the screening identified two different regimes in the formulation space: a broad region linked to significantly larger expected sizes (>520 nm) and a limited region that may yield particle sizes that were closely aligned with the 450 nm target. In line with the feature importance and partial dependence analyses previously discussed, this separation implies that a shift between size regimes is simultaneously governed by second centrifugation speed and polymer concentration.

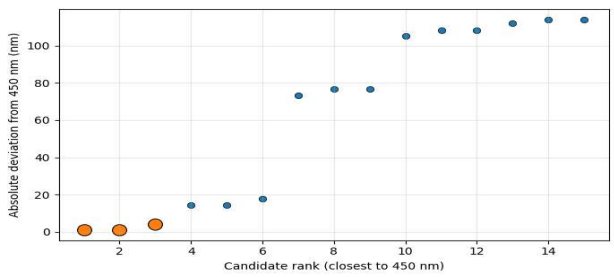

Figure 5. Random Forest–guided in-silico screening results targeting a particle size of 450 nm. Candidate formulations were ranked according to absolute deviation from the target size. The top three candidates (highlighted) exhibited predicted deviations below approximately 5 nm, indicating that the current experimental design space supports controlled production of nanoparticles near the desired size.

3.7 Comparison with previous nanoparticle modeling studies

Recent studies have used machine learning and design-of-experiments approaches to PLGA nanoparticle formulation, utilizing significantly bigger datasets generated from iterative experimental campaigns or literature mining, which frequently include 70-300 distinct formulations [11, 14]. For instance, Hanari et al. used over 300 microfluidic formulations to predict encapsulation and loading efficiencies using random forest and related models [11], while Rezvantalab et al. used multi-model ML to analyze literature-derived PLGA systems and found polymer content and organic phase composition as dominant descriptors of particle size and drug loading [7]. Hybrid ML–physics-informed frameworks have further emphasized the significance of formulation composition and process intensity on nanostructure performance in curcumin-based nanocomposites [14]. The prominence of PLGA concentration and post-precipitation processing in our feature importance and partial dependence analyses is therefore consistent with these broader trends, even though our smaller, single-laboratory dataset limits the statistical strength and generalizability of the present findings.

4. Limitations

It is important to recognize the limitations of this study. First, the dataset comprised only 19 distinct formulations, and we did not perform an independent experimental validation of formulations selected by the in-silico screening step. As a result, the Random Forest predictions, including those targeting 450 nm particles, should be viewed as hypotheses that require confirmation by follow-up nanoprecipitation experiments under the suggested conditions. The robustness of the model would probably be enhanced by larger datasets, even if Leave-One-Out Cross-Validation was used to lessen overfitting. Second, predictions made outside of this design space constitute extrapolation and should be treated with caution because the model was trained only within the empirically investigated parameter ranges. Third, factors that may affect particle size and drug encapsulation behavior but were not specifically included in the formulation include temperature, solvent content, and the kinetics of drug–polymer interactions. More physicochemical descriptors and larger datasets may be included in future research to improve predicted performance and allow for completely automated formulation optimization. Therefore, the current work should be regarded as a proof-of-concept for

integrating small experimental designs with interpretable machine learning, rather than as a fully validated formulation optimization study.

5. Conclusion

The possibility of using Random Forest regression to model the particle size of PLGA nanoparticles loaded with curcumin and made via nanoprecipitation is shown in this work. The model indicated centrifugation speed and polymer concentration as the main contributors to the nonlinear connections between formulation variables and particle size. The experimentally investigated design space allows controlled size adjustment without extrapolation beyond training data, according to in-silico screening aimed at 450 nm particles. The results show that, under the constraints of a small and correlated design, machine learning can support pre-experimental screening and mechanistic hypothesis generation for curcumin-loaded PLGA nanoparticle formulations, provided that its predictions are subsequently validated experimentally. Expanded formulation descriptors and bigger datasets could help increase model accuracy and allow for data-driven design of drug-loaded polymeric nanoparticles.

Acknowledgment
The author thanks the institutional support and laboratory infrastructure that enabled this research.

Data availability
Data will be made available on request.

References

[1] Khosravi, M.A. and R. Seifert, *Clinical trials on curcumin in relation to its bioavailability and effect on malignant diseases: critical analysis.* Naunyn-schmiedeberg's Archives of Pharmacology, 2024. **397**(5): p. 3477-3491.
[2] Luz, P.P., et al., *Curcumin-loaded into PLGA nanoparticles: preparation and in vitro schistosomicidal activity.* Parasitology research, 2012. **110**(2): p. 593-598.
[3] Alam, J., et al., *Curcumin encapsulated into biocompatible co-polymer PLGA nanoparticle enhanced anti-gastric cancer and anti-Helicobacter pylori effect.* Asian Pacific journal of cancer prevention: APJCP, 2022. **23**(1): p. 61.
[4] Chereddy, K.K., et al., *Combined effect of PLGA and curcumin on wound healing activity.* Journal of controlled release, 2013. **171**(2): p. 208-215.
[5] Leung, M.H. and A.Q. Shen, *Microfluidic assisted nanoprecipitation of PLGA nanoparticles for curcumin delivery to leukemia jurkat cells.* Langmuir, 2018. **34**(13): p. 3961-3970.
[6] Seegobin, N., et al., *Optimising the production of PLGA nanoparticles by combining design of experiment and machine learning.* International Journal of Pharmaceutics, 2024. **667**: p. 124905.
[7] Rezvantalab, S., S. Mihandoost, and M. Rezaiee, *Machine learning assisted exploration of the influential parameters on the PLGA nanoparticles.* Scientific Reports, 2024. **14**(1): p. 1114.
[8] Hu, H., et al., *Fabrication, optimization, and evaluation of paclitaxel and curcumin coloaded PLGA nanoparticles for improved antitumor activity.* ACS omega, 2022. **8**(1): p. 976-986.
[9] Han, S., B.D. Williamson, and Y. Fong, *Improving random forest predictions in small datasets from two-phase sampling designs.* BMC medical informatics and decision making, 2021. **21**(1): p. 322.
[10] Palmer, D.S., et al., *Random forest models to predict aqueous solubility.* Journal of chemical information and modeling, 2007. **47**(1): p. 150-158.
[11] Hanari, N., S. Mihandoost, and S. Rezvantalab, *Intelligence prediction of microfluidically prepared nanoparticles.* Scientific Reports, 2025. **15**(1): p. 37512.
[12] Ortiz-Perez, A., et al., *Machine learning-guided high throughput nanoparticle design.* Digital Discovery, 2024. **3**(7): p. 1280-1291.
[13] Almansour, K. and A.S. Alqahtani, *Utilization of machine learning approach for production of optimized PLGA nanoparticles for drug delivery applications.* Scientific Reports, 2025. **15**(1): p. 8840.
[14] Rahdar, A., S. Fathi-Karkan, and M. Shirzad, *Predictive optimization of curcumin nanocomposites using hybrid machine learning and physics informed modeling.* Scientific Reports, 2025. **15**(1): p. 44368.

Rumana Ferdushi is a Ph.D. student at the Nanomaterials Lab (NML), Yonsei University, Republic of Korea. Her research focuses on developing nanoparticles from natural products for use as drug carriers. The goal of her work is to enhance targeted drug delivery, reducing toxicity while improving therapeutic efficacy. Rumana's innovative approach leverages natural materials to design advanced drug delivery systems, contributing to safer and more effective treatments.