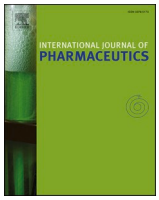
International Journal of Pharmaceutics 667 (2024) 124905

Contents lists available at ScienceDirect

International Journal of Pharmaceutics

journal homepage: www.elsevier.com/locate/ijpharm



Optimising the production of PLGA nanoparticles by combining design of experiment and machine learning

Nidhi Seegobin, Youssef Abdalla, Ge Li, Sudaxshina Murdan, David Shorthouse, Abdul W. Basit *

Department of Pharmaceutics, UCL School of Pharmacy, University College London, 29-39 Brunswick Square, London WC1N 1AX, United Kingdom

ARTICLE INFO

Keywords:
PLGA
Nanotechnology
Nanoprecipitation
Machine learning
Artificial intelligence
Computational modelling
Oral drug delivery

ABSTRACT

Poly(lactic-co-glycolic acid) (PLGA) is a widely used biodegradable polymer in drug delivery and nanoparticle (NP) formulation due to its controlled drug release properties and safety profiles. Among the methods available for NP production, nanoprecipitation is distinguished by its simplicity and scalability. However, it requires careful optimisation to achieve the desired NP characteristics, making the process potentially lengthy and costly. This study aimed to assess and compare the predictive performance of Design of Experiments (DOE) and Machine Learning (ML) models for the optimisation of PLGA nanoparticle size and zeta potential produced by nanoprecipitation. Various ML methods were employed to predict particle size, with Extreme Gradient Boosting (XGBoost) identified as the best performing. The key finding is that integrating ML with DOE provides deeper insights into the dataset than either method alone. While ML outperformed DOE in predictive performance, as evidenced by lower root mean squared error values and higher coefficients of determination, both methods struggled to accurately predict zeta potential, generating models with high errors. However, ML proved more effective in identifying the parameters that most significantly influence NP size, even with a smaller DOE dataset. Combining DOE datasets with ML for parameter importance was particularly advantageous in situations where data is limited, offering superior predictive power and the potential to streamline experimental design and optimisation. These results suggest that the synergistic use of ML and DOE can lead to more robust feature analysis and improved optimisation outcomes, particularly for NP size. This integrated approach can enhance the accuracy of predictions and supports more efficient experimental design, streamlining nanoparticle production processes, especially under resource-constrained conditions.

1. Introduction

Poly(lactic-co-glycolic acid) (PLGA) is a biodegradable and biocompatible polymer that has gained significant attention in the field of drug delivery and nanoparticle (NP) formulation (McCoubrey et al., 2024). Its excellent properties, such as controlled drug release, minimal toxicity, and approval by regulatory agencies (United States Food and Drug Administration and European Medicines Agency), make PLGA an ideal candidate for developing NPs (Operti et al., 2021; Lee et al., 2016). These NPs can be used to deliver a variety of therapeutic agents via a range of administration methods such as oral, ocular, transdermal, intranasal or parenteral routes (Bashir et al., 2021; Ansari and Alshahrani, 2019; Gupta et al., 2010; Baek et al., 2024; Alghareeb et al., 2024; Shah et al., 2020; Seegobin et al., 2024). Polymer-based NPs such as PLGA NPs can be produced via a variety of techniques such as emulsion diffusion, emulsion evaporation or salting out methods, with nanoprecipitation (NPR) standing out for its simplicity and efficiency (Lee et al., 2016; Paliwal et al., 2014; Zielińska et al., 2020; Astete and Sabliov, 2006).

NPR works by dissolving a polymer in a water-miscible organic solvent, followed by the dropwise addition of this solution into an aqueous surfactant solution, resulting in the formation of a colloidal suspension. Upon evaporation of the organic solvent, a stable suspension of nanoparticles is obtained (Lee et al., 2016; Zielińska et al., 2020; Martínez Rivas et al., 2017). NPR offers several advantages, such as simplicity and scalability, making it suitable for large-scale production; mild processing conditions that preserve the integrity of sensitive drugs; and the ability to control particle size and distribution by adjusting process parameters. Despite these advantages, NPR requires careful optimisation of parameters to achieve the desired particle

* Corresponding author.
E-mail address: a.basit@ucl.ac.uk (A.W. Basit).

https://doi.org/10.1016/j.ijpharm.2024.124905
Received 24 September 2024; Received in revised form 28 October 2024; Accepted 1 November 2024
Available online 2 November 2024

characteristics. The optimisation of NPR parameters, such as solvent type, polymer concentration, stirring rate, and temperature, is crucial for obtaining NPs with optimal properties (Bashir et al., 2021; Ansari and Alshahrani, 2019; Martínez Rivas et al., 2017; Shi et al., 2013). However, this process can be lengthy and costly, particularly because PLGA is an expensive material and extensive experimental studies are often required to identify the optimal conditions, leading to increased time and resource consumption (Danhier et al., 2012). Different statistical and modelling methods, such as Design of Experiments (DOE) and Machine Learning (ML), can be used to understand the relationship between different processing parameters and PLGA NP properties which can be used to streamline their production and reduce wastage.

DOE is a systematic and structured approach, primarily used to optimise and control pharmaceutical manufacturing processes. This methodology enables a comprehensive understanding of the relationships between process parameters and product quality attributes, making it invaluable in the development of high-quality pharmaceutical products (Tavares Luiz and Viegas, 2021). DOE's efficiency lies in determining causal relationships between variables in an experimental design. Among the various DOE strategies, screening designs, such as fractional or full factorial designs, identify the most influential factors early in the process. Response surface methodologies such as Central Composite Design (CCD) are then conducted to explore optimal factor levels. This method allows modelling of linear relationships, interactions, and quadratic effects, providing a detailed understanding of the process landscape (Tavares Luiz and Viegas, 2021). DOE has been extensively applied in the optimisation of nanoparticle production, including PLGA-based formulations, due to its ability to identify critical process parameters and their interactions – such as drug amount, polymer amount, and aqueous phase to organic phase ratios – and their effects on particle size and zeta potential (Tavares Luiz and Viegas, 2021; Camacho Vieira et al., 2024; Saka et al., 2020). This identification is crucial for the development of robust and reproducible manufacturing processes, which are essential for scaling up production while maintaining product quality. In nanoparticle synthesis, factors such as polymer concentration or solvent type can significantly influence the particle size and encapsulation efficiency. By systematically varying these parameters within a DOE framework, researchers can optimise these attributes to meet specific therapeutic goals, thus enhancing the efficacy and safety of the final pharmaceutical product. However, it is important to note that DOE has certain limitations, such as difficulties in managing complex data analysis, the constraints imposed by rigid experimental designs, and limitations on scalability (Grangeia et al., 2020).

ML offers an alternative to DOE to analyse data and identify patterns to make predictions, it has drawn a lot of attention as it offers a powerful set of tools which can significantly aid in the development of nanoparticles (Silveira et al., 2024; Zaslavsky et al., 2023). ML algorithms can analyse large datasets to identify relationships that may not be evident through traditional statistical methods. Therefore, ML can be used to model more complex, non-linear relationships, not possible through DOE (Walsh et al., 2022). However, it requires, high-quality labelled datasets to determine these patterns and make accurate predictions (Silveira et al., 2024), additionally, unlike DOEs, ML algorithms do not identify causal relationships, as they are applied on top of existing data and attempt to infer relationships (Silveira et al., 2024). For formulation development, supervised ML algorithms use labelled past experimental data to establish relationships between experimental conditions, such as material composition and processing parameters, and desired properties, such as stability, compatibility, size and yield (Xu et al., 2023; Abdalla et al., 2024). By leveraging these ML models, researchers can potentially optimise NP synthesis by manipulating input features to reach desired outcomes. This has been particularly beneficial for optimising the size of NPs using microfluidic production techniques (Ortiz-Perez et al., 2024; Nathanael et al., 2023; Chen and Lv, 2022), where ML models help adjust flow rates and reagent concentrations to achieve the desired NP size. Similarly, leveraging ML capabilities in NPR can allow researchers to predict and optimise PLGA NP properties based on input parameters, thereby reducing the number of necessary experimental trials and saving both time and resources (Silveira et al., 2024).

This study aims to explore the production of PLGA NPs using both DOE and ML methodologies. By comparing these approaches, this study seeks to determine which method is most suitable for achieving efficient and cost-effective nanoparticle production. The goal is to identify critical process parameters and develop robust predictive models that can streamline the optimisation process and reduce material costs.

2. Materials and methods

2.1. Materials

Resomer® Condensate RG 50:50 MN 2300 (PLGA, acid terminated, 50:50, Mw 2000–2500 g/mol) and Resomer® R 504H, (PLGA, acid terminated Mw 49,000–54,000 g/mol) were purchased from Evonik Industries (Essen, Germany). Acetone, Poloxamer 407 (Kolliphor® P 407) and Tween 80 were purchased from Merck Life Science (Gillingham, UK). Where used, water was of HPLC-grade and obtained via an ELGA HPLC water purification system (ELGA LabWater, High Wycombe, UK).

2.2. Production of nanoparticles by nanoprecipitation

PLGA NPs were prepared in triplicate using the NPR method. PLGA was dissolved in acetone using a range of concentration combinations based on a study DOE JMP® (SAS institute, United Kingdom). An AL-1000 syringe driver (Precision Instruments, Hitchin, UK) was then used to precipitate the polymer-drug organic solutions dropwise at 200 μL/min via a 30 G x 0.5″ needle (Microlance™ 3, Becton Dickinson, New Jersey, USA) in an anti-solvent solution containing either 1 % (w/v) poloxamer 407 or 1.2 % (w/v) Tween 80 in water. The anti-solvent solution was stirred throughout with a magnetic bar rotating at 700 rpm. The final ratio of polymer to anti-solvent solution was 1:2 v/v. Acetone was then allowed to evaporate from the uncovered mixture overnight by stirring at 700 rpm at room temperature (25 °C). The resultant nanosuspension was centrifuged (3-16KL Centrifuge, Sigma Laborzentrifugen, Osterode am Harz, Germany) at 10,500g, for 30 min at 4 °C. After centrifugation the supernatant was discarded, and the NPs were resuspended in 1.5 mL cool (2–8 °C) deionised water, in order to maintain the polymer-based particle in a rigid state, below its glass transition temperature (Sprengholz, 2014).

2.3. Physical characterisation of PLGA nanoparticles

The size and ζ potential of the particles was measured using Dynamic Light Scattering (DLS) with a Malvern ZetaSizer (Malvern Panalytical Ltd., Malvern, UK). Measurements (n = 3) were conducted at 25 °C with an equilibration time of 120 s. The measurement settings were 173° backscatter (NIBS default), with an automatic measurement duration.

2.4. Predictive modelling using DOE

JMP Pro software® (version 17.0, SAS institute, United Kingdom) was used to perform a DOE study. The design was based on a RSM using extreme test values for optimal parameter designs. The selection of the experimental points for each factor was performed by the software and resulted in a table of training experiments, see Tables S1-S4. The investigated process parameters for the nanoprecipitation were assigned as factors with two categorical factors (nature of the PLGA and nature of the anti-solvent) and two continuous factors (PLGA concentration and anti-solvent concentration) were evaluated. The investigated ranges are summarized in Table 1. The selection of experimental points leads to the construction of initial four-factor/two-level factorial dataset consisting of 48 experiments including triplicate experiments. The investigated

Table 1
The investigated ranges of parameters used in the data sets.

Factor		Investigated range
Categorical	PLGA	Resomer condensate or Resomer 504
	Anti-solvent	Aqueous solution of Poloxamer 407 or of Tween 80
Continuous	PLGA concentration	1–60 mg/mL
	Anti-solvent solution concentration	10–50 % (w/v)

responses were particle size and zeta (ζ) potential. The objective was to produce particles controllable in size, expected to be between 100 and 150 nm, and stable with ζ potential ≤ −30 mV for NP uptake and stability (Lu and Gao, 2010; Gupta and Trivedi, 2018). The maximum polymer concentration was fixed at 60 mg/mL based on preliminary experimental data. ANOVA was performed to reveal the effect of parameter estimates on total variance.

2.5. Predictive modelling using machine learning

2.5.1. Machine learning model

To determine which ML model is best able to make predictions based on the datasets, five different supervised ML models were employed. A variety of models were employed, including tree-based ensembles like extreme gradient boosting (XGBoost) and random forest (RF), a memory-based model (k-nearest neighbors, kNN), a kernel-based model (support vector machine regressor, SVM), and a neural network model known as a multilayer perceptron (MLP). Model hyperparameters were optimised using 100 different randomly chosen hyperparameter combinations (random search), evaluated using 5-fold cross validation. The ML models were used for predictions of particle size and ζ potential. All ML models were run on Python (Version 3.10.4) on a Windows desktop (Operating System: Windows 11; Processor: AMD Ryzen Threadripper 7960X 24-core 4.2 GHz; RAM Memory: 128 GB, GPU: RTX 4090 24 GB) using the Scikit-learn (Version 1.1.3) Python package, except for XGBoost (XGBoost Version 1.6.2).

2.5.2. Feature processing

Before being input into the ML models, the type of PLGA and anti-solvent used were one-hot encoded. All other data was normalised to a range of 0–1.

2.5.3. Evaluating model performance

Model performance was evaluated using leave-one-out cross validation (LOOCV) and 5-fold cross-validation (CV). For LOOCV, the data set is split into one observation as the test set and the rest (N-1) observations are considered as the training set. For 5-fold CV data sets are split into 5 equal subsets, or folds, which are approximately the same size. For each fold, the fold is taken as the test data set and the remaining folds as the training data. The models are fitted to the training data and evaluated on the test data, and subsequently, the evaluation score is retained (Abdalla et al., 2023). The overall performance of the model is determined as the average of all the iterations.

2.5.4. Feature importance

XGBoost feature importance was determined by measuring the weights of different parameters used in the trained model. This is determined by identifying the number of times a parameter is used to split a tree, across all trees in the ensemble (Abdalla et al., 2023).

2.6. Data analysis

DOE and ML results included regression analyses, their performance was measured using the coefficient of determination (R^2) and the root mean squared error (RMSE) of predictions. R^2 is a measure of the goodness of fit of the model and is defined as the proportion of the total variance of the experimental points explained by the model. The R^2 was used to measure the reliability and robustness of the model as shown in Eq. (1). We note however, that a high R^2 does not necessarily indicate a good model, only that the model explains the variance in the collected data to a high level.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (1)$$

With y_i as the observed values, $\hat{y}_i$ as the predicted values, $\bar{y}_i$ the mean of the observed values and n as the number of observations.

The RMSE measures the average magnitude of the prediction errors in a model, providing a direct indication of the model's predictive performance by quantifying the square root of the average squared differences between predicted and observed values. The RMSE was used to measure the average magnitude of the prediction errors and assess the predictive performance of the model, as shown in Eq. (2).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (2)$$

With y_i as the observed values, $\hat{y}_i$ as the predicted values and n as the number of observations.

3. Results and discussion

3.1. PLGA nanoparticles

In this study, PLGA NPs were generated in a systematic manner to create DOE models and train ML models based on the experimental training datasets outlined in the Tables S1-S4. Initially, NPs were formulated to assess parameter importance using a four-factor/two-level design dataset (extremes) (see Table S1), followed by those prepared using the two-factor/two level design and centre points data set (CDD) (see Table S2), those generated through a full factorial design (see Table S3), and final an ML training set generated through the random removal of 20 % of the data for external validation (see Table S4) – the factorial and training data sets differ by the addition of 21 data points for the ML training set. As shown in Fig. 1, the data points are well-distributed across different particle sizes and ZP values, providing a comprehensive representation of the variations present in the training data. The particle size obtained from these preparations ranged from 86.9 to 381.4 nm with ζ Potential (ZPs) ranges of –11.0 to −52.2 mV; desirable particles were those under 150 nm in size and with a ZP ≤−30 mV, representing around 10 % of the training datapoints for each dataset. These findings align with the objectives of producing NPs within the target size range for optimal delivery through the enhanced permeation and retention (EPR) effects, leaky gut and permeation through tight junctions to facilitate drug delivery to diseased tissue (Dolai et al., 2021; Lamprecht et al., 2001; Clayburgh et al., 2004; Hartwig et al., 2022). The inclusion of ZP as a critical stability parameter also ensured that the particles exhibited adequate colloidal stability, with values of ±30 mV or greater indicating sufficient stability to prevent particle aggregation (Lu and Gao, 2010; Gupta and Trivedi, 2018). Furthermore, the PDI values for the majority of samples were below 0.3, indicating acceptable monodispersity for pharmaceutical applications (Wu et al., 2011; Musielak et al., 2022; Danaei et al., 2018). Only 6 out of 192 samples (3 %) had PDI values above 0.3, but all were still below 0.4, confirming that they remained within an acceptable range for use. Given these results, further optimisation of PDI was deemed unnecessary, as monodispersity was sufficiently achieved across nearly all samples.

3.2. Optimisation of ML for size prediction

To select the optimal ML model for particle characteristic prediction,

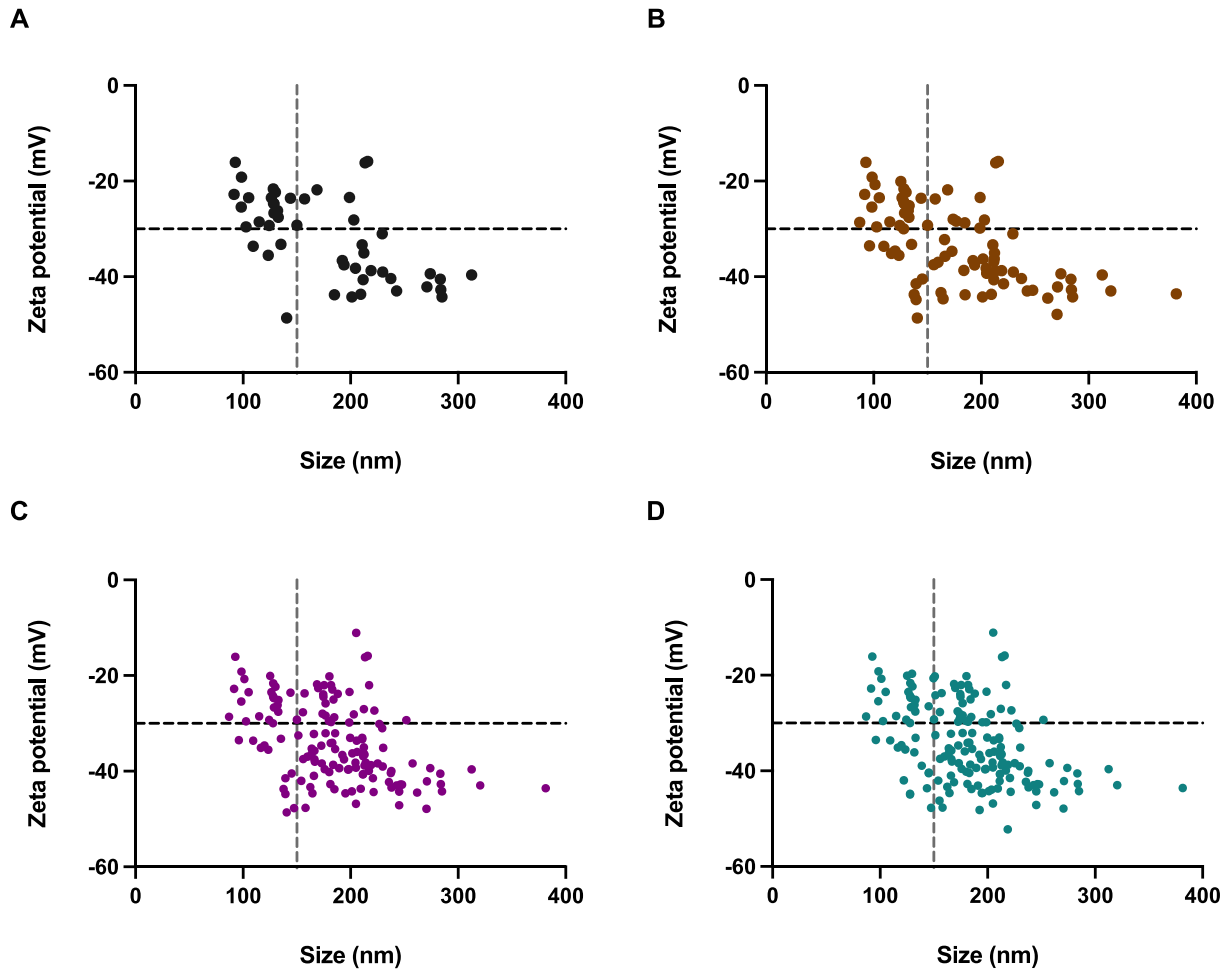


Fig. 1. Scatter plot depicting the size and ζ potential distribution of the training data obtained following the (A) extremes (n = 48), (B) CCD (n = 60), (C) factorial (n = 144) and (D) training data sets (n = 165).

the performance of various ML models was evaluated using both 5-fold CV and LOOCV. LOOCV was trialled as it can reduce the risk of overfitting when used for small datasets and maximise the prediction potential (Abdalla et al., 2023). Five different models were tested: RF, XGBoost, KNN, MLP and SVM. LOOCV, while more computationally demanding, yielded results consistent with those from 5-fold CV. Given the increased computational load without additional performance benefits, further evaluation was carried out using 5-fold cross-validation. MLP yielded consistently negative predictive performance. This was anticipated as neural networks are likely to be overfit with small sample sizes and therefore usually not appropriate for use (Meyer et al., 2002). Therefore, this study focused on training the 4 other models to predict particle size. A random search was carried out to identify the optimal hyperparameters for these models. The

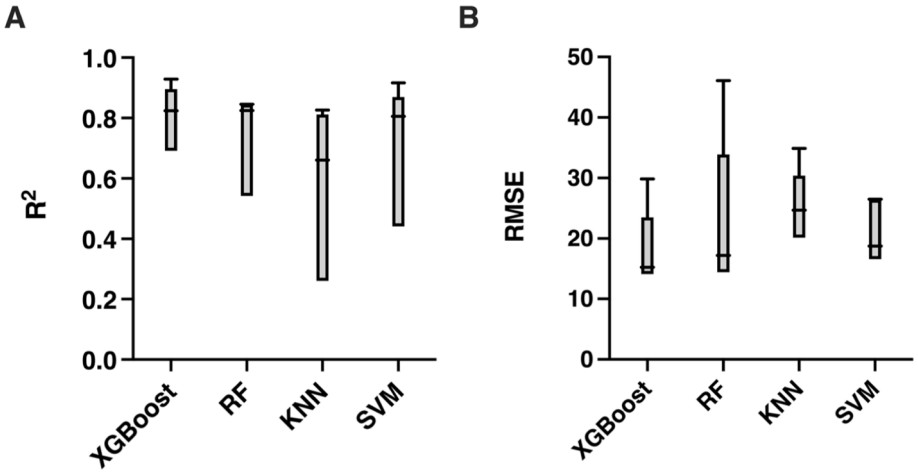


Fig. 2. Box plots illustrating the predictive accuracies of the ML models' cross validation R^2 (A) and RMSE (B) on the training datasets using 5-fold CV.

performance of the tuned models, using 5-fold CV can be seen in Fig. 2, where the R2 indicates how well the model's predictions match the actual data. An R2 value of 1 means the model perfectly predicts the data, while a value of 0 means the model does no better than guessing the mean. Meanwhile, RMSE measures the average magnitude of errors between predicted and actual values, providing an indication of the model's prediction performance. It was observed that RF and XGBoost were found to be the top performing models, with the highest R^2 and lowest RSME values, this is anticipated as tree-based models tend to perform the best for small-to-medium sized tabular datasets (Grinsztajn et al., 2022) and is consistent with the literature for NP optimisation (Ortiz-Perez et al., 2024; Nathanael et al., 2023). XGBoost slightly outperformed RF, therefore, it was selected for further evaluation. These results are in line with the study by Ortiz-Perez et al. (Ortiz-Perez et al., 2024) where XGBoost was used for particle size predictions. To further explore XGBoost's capabilities and determine the optimal number of data points for model training, we iteratively sampled randomised data points in increasing training set sizes and used them to train the XGBoost model, testing each dataset using 5-fold CV (Fig. 3). It was observed that the model started to explain variance in the data when approximately 30 triplicates (R^2 was greater than 0) were taken and plateaued at approximately 40 triplicates.

3.3. The predictions for DOE and ML

3.3.1. Feature and parameter importance for particle size

Utilising the parameters listed in Table 1 and the highest and lowest (extremes) data points shown in Table S1, DOE and ML predictions were conducted to evaluate parameter estimates and feature weights, respectively. The DOE approach not only ranks the importance of test parameters but also employs ANOVA to assess their statistical significance, thereby facilitating the use of RSM predictions focused exclusively on significant parameters. In contrast, XGBoost feature weights provide a ranking of parameter importance, determined by the number of times an individual feature is used to split a tree across all trees in the model, without indicating statistical significance (Chen, 2016). As illustrated in Fig. 4, DOE analysis demonstrated that the concentrations of both the anti-solvent solution ($p < 0.001$) and PLGA concentration ($p < 0.0001$) were significantly influential in determining particle size, whilst the type of anti-solvent solution and the type of PLGA did not exhibit a significant impact. Conversely, ML feature weight analysis identified the type of PLGA and the concentration of PLGA as the most critical features for all datasets, with these factors exhibiting more than twice the importance compared to the type and concentration of the anti-solvent solution. The DOE findings align with previous studies by Hernández-Giottonini et al. (Hernández-Giottonini et al., 2020) and Huang et al. (Huang and Zhang, 2018) which demonstrated that PLGA concentration and the anti-solvent solution concentration significantly influence the size of PLGA nanoparticles produced via nanoprecipitation, highlighting that both models are able to extract meaningful insights from the data. These papers suggest that an increase in the viscosity of the organic phase within the aqueous solvent affects solvent diffusion and evaporation, thereby impacting nanoparticle size. Interestingly, Huang et al. (Huang and Zhang, 2018) also highlighted the importance of the type of organic solvent, temperature, and ionic strength of the aqueous phase in determining particle size. This study suggests that these parameters influence the diffusion coefficient of the solvent in the aqueous media in the presence of PLGA, identifying this coefficient as a key predictor of particle size (Huang and Zhang, 2018). They further suggest that these parameters also affect the diffusion coefficient of the solvent in the aqueous media in the presence of PLGA, and that this coefficient is the main predictor of particle size (Huang and Zhang, 2018).

Herein, interesting discrepancies were found between the DOE and ML models regarding the importance of the parameters. DOE analysis of the parameter estimates (see Fig. 4**A**) indicated that the type of PLGA was not significantly important for further analysis. Conversely, the ML model identified the type of PLGA as the most important feature, assigning it the highest feature weight. When the DOE model was run using the larger DOE factorial dataset, the parameter estimates changed notably, showing statistical significance for the type of PLGA, the concentration of PLGA, and the concentration of the anti-solvent solution (see Fig. 5). These results suggest that the ML model, with the much smaller 'extremes' DOE dataset of 48 data points, could determine the importance of nanoprecipitation parameters more effectively than DOE models. This required a larger dataset of 144 data points for DOE models to do alone.

3.3.2. CCD dataset predictions for particle size

This parameter estimates results were used to design a CCD model, as shown in Table S2. The DOE and ML models developed using the different experimental datasets were evaluated against an external validation set of 11 ($n = 3$) unseen datapoints, as illustrated in Fig. 6. Notably, the RMSE was lower for DOE than for ML using the CCD and factorial datasets, indicating better predictive performance for DOE with a smaller dataset. However, both ML and DOE exhibited negative R^2 values. This negative R^2 indicates that the models performed worse than a simple mean prediction, thus reflecting significant inaccuracies in their models. The negative R^2 values highlight a potential overfitting or underfitting issue within the models, or they may suggest that the dataset is not sufficiently representative to explain any variance in the data. Overall, while the DOE model showed a lower RMSE and thus better performance on this metric, the negative R^2 values for both modelling approaches underscore the need for further refinement and

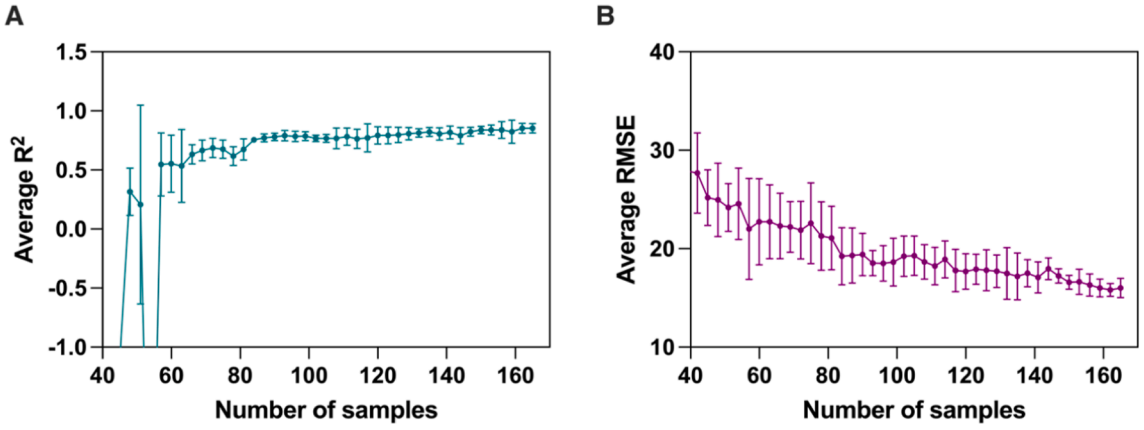


Fig. 3. Evaluation of XGBoost performance A) R^2 and B) RMSE, evaluated using the average of the 5 runs using 5-fold CV as the number of triplicate samples increases.

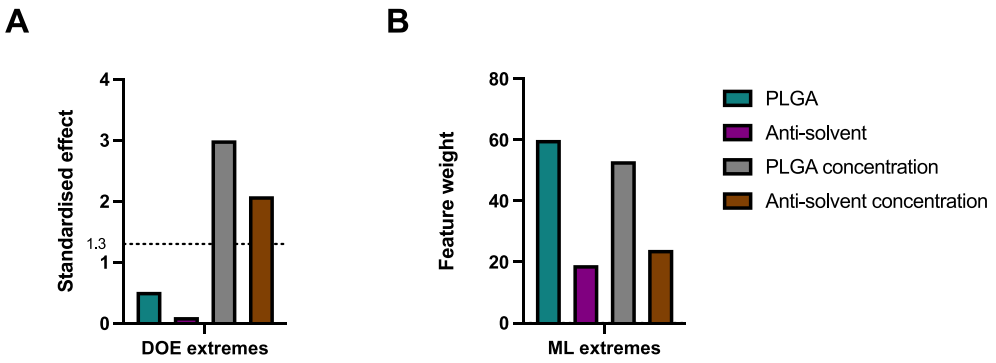


Fig. 4. Bar charts comparing (A) a Pareto chart of the standardised effects obtained from parameter estimates in the DOE analysis, with a significance threshold indicated by the reference line at p = 0.05, and (B) feature importance derived from the ML analysis focused on predicting particle size. The significance threshold in (A) helps identify which parameters have a statistically significant effect, while (B) highlights the relative importance of different features in the machine learning model.

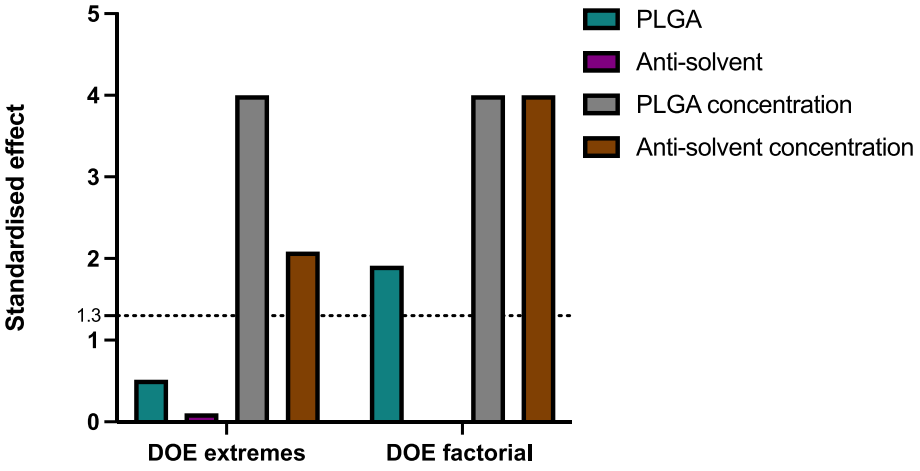


Fig. 5. Pareto chart showing parameter significance for particle size predictions from DOE analysis, comparing a small dataset (48 points) to a larger one (144 points), with a significance threshold indicated by the reference line at p = 0.05.

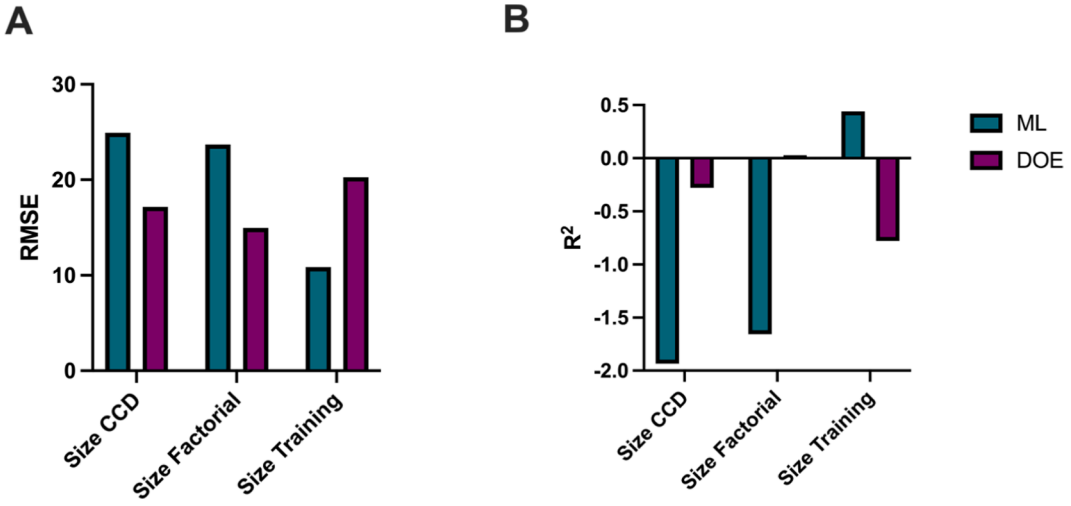


Fig. 6. Bar charts comparing the predictive performance of ML and DOE models for particle size based on RMSE (A) and R^2 (B) values.

potentially larger, more representative datasets to improve predictive performance.

3.3.3. *Full factorial and training datasets predictions for particle size*

To enhance predictive capabilities, further data points were included based on a full factorial analysis (see Table S3) for DOE and a randomized training set for ML (see Table S4) which included the DOE data and 21 extra datapoints. For comparison, each dataset was analysed using both DOE and ML. With these expanded datasets, the RMSE decreased in both the DOE's full factorial design and the ML's training

set, demonstrating improved model performance. The larger datasets yielded positive R2 values for both methods, with the optimum ML model achieving an R2 of 0.441 and an RMSE of 0.035. In contrast, the best DOE model achieved an R2 of 0.027 and an RMSE of 0.048, indicating performance comparable to a model that consistently predicts the mean of the data. Surprisingly, despite the relatively small size of the dataset, often considered insufficient for robust ML modelling, ML outperformed DOE. This suggests its potential in guiding optimisation approaches and experimental design. Despite the typical requirement for large datasets to achieve predictive accuracy, our study demonstrated that ML could extract meaningful insights from a smaller dataset. This suggests that ML can be a valuable tool for guiding optimisation approaches, particularly when large datasets are not available. However, the significant decline in R2 values from the earlier 5-fold CV to the performance test indicates potential overfitting of the ML model, suggesting that the model's predictions are highly specific to this dataset and may not be applicable to different polymer systems. Therefore, to improve the generalisability and robustness of the ML model, additional data is required. These findings indicate that ML can be effectively used to obtain better fitting of feature importance or patterns and trends in data generated using DOE, that may not be immediately apparent. This is likely due to non-linearity in the relationships between factors and output measurements, which can be captured by ML but not DOE.

3.3.4. Zeta potential

Interestingly, both ML and DOE were unable to explain variance for ζ potential as both obtained negative R^2 values (see Figure S1). These negative R^2 values indicate that the models perform worse than a baseline model that predicts the mean of the observed data. This poor performance is likely attributed to the small sample size, or the incorrect selection of features used to train the ML model. Indeed, whilst the literature highlights that the ζ potential of NPs was found to be influenced not only by the PLGA concentration but also by the type of polymer terminated chain and the ionic strength of the solutions (Hernández-Giottonini et al., 2020; Berg et al., 2009). Further studies could investigate incorporating additional data points from external sources or additional features, such as polymer chain termination and ionic strength, to enhance model robustness and improve the predictive accuracy of ζ potential.

4. Conclusion

This study highlights the comparative strengths of DOE ML in parameter analysis and predictive modelling for optimising PLGA NP production using NPR. Initially, DOE and ML provided distinct insights into the factors influencing particle size. DOE analysis identified the concentrations of the anti-solvent solution and PLGA as statistically significant factors, while the ML model highlighted the type and concentration of PLGA as the most important features. As the analysis progressed with a larger dataset, it became evident that ML not only maintained consistent parameter importance but also offered a more stable and insightful analysis than DOE alone, whose findings evolved to align more closely with the ML predictions. When evaluating predictive performance using an external validation dataset, ML outperformed DOE, achieving lower RMSE values and higher R^2 values. However, both methods were inadequate in predicting zeta potential, indicating the need for further refinement in these areas. This study suggests that fitting ML models to DOE-designed factorial datasets can provide deeper insights into the relationships within the data than DOE alone. Rather than viewing these methods as competitors, this combined approach leverages the strengths of both, suggesting that ML could play an increasingly valuable role in experimental design and optimisation, especially in resource-constrained scenarios.

5. Statements & declarations

CRediT authorship contribution statement

Nidhi Seegobin: Writing – review & editing, Writing – original draft, Validation, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Youssef Abdalla:** Writing – review & editing, Writing – original draft, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Ge Li:** Writing – review & editing, Methodology, Investigation. **Sudaxshina Murdan:** Writing – review & editing, Supervision. **David Shorthouse:** Writing – review & editing, Supervision, Conceptualization. **Abdul W. Basit:** Writing – review & editing, Supervision, Conceptualization.

Funding

This project received funding from the EPSRC CDT in Transformative Pharmaceutical Technologies grant 'EP/S023054/1'.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijpharm.2024.124905.

Data availability

Data will be made available on request.

References

Abdalla, Y., Elbadawi, M., Ji, M., Alkahtani, M., Awad, A., Orlu, M., Gaisford, S., Basit, A. W., 2023. Machine learning using multi-modal data predicts the production of selective laser sintered 3D printed drug products. Int. J. Pharm. 633, 122628. https://doi.org/10.1016/j.ijpharm.2023.122628.

Abdalla, Y., Ferianc, M., Awad, A., Kim, J., Elbadawi, M., Basit, A.W., Orlu, M., Rodrigues, M., 2024. Smart laser Sintering: Deep Learning-Powered powder bed fusion 3D printing in precision medicine. Int. J. Pharm. 661, 124440. https://doi.org/10.1016/j.ijpharm.2024.124440.

Alghareeb, S., Asare-Addo, K., Conway, B., Adebisi, A., 2024. PLGA nanoparticles for nasal drug delivery. J. Drug Deliv. Sci. Technol. 95. https://doi.org/10.1016/j.jddst.2024.105564.

Ansari, M.J., Alshahrani, S.M., 2019. Nano-encapsulation and characterization of baricitinib using poly-lactic-glycolic acid co-polymer. Saudi Pharm. J. 27, 491–501. https://doi.org/10.1016/j.jsps.2019.01.012.

Astete, C., Sabliov, C., 2006. Synthesis and characterization of PLGA nanoparticles. J. Biomater. Sci.-Polymer Ed. 17, 247–289.

Baek, S., Hwang, E., Hur, G., Kim, G., An, Y., Park, J., Hong, J., 2024. Intranasal administration enhances size-dependent pulmonary phagocytic uptake of poly (lactic-co-glycolic acid) nanoparticles. Ejnmmi Radiopharmacy Chem. 9. https://doi.org/10.1186/s41181-023-00227-x.

Bashir, S., Aamir, M., Sarfaraz, R.M., Hussain, Z., Sarwer, M.U., Mahmood, A., Akram, M. R., Qaisar, M.N., 2021. Fabrication, characterization and in vitro release kinetics of tofacitinib-encapsulated polymeric nanoparticles: a promising implication in the treatment of rheumatoid arthritis. Int. J.Polym. Mater. Polym. Biomat. 70, 449–458. https://doi.org/10.1080/00914037.2020.1725760.

Berg, J., Romoser, A., Banerjee, N., Zebda, R., Sayes, C., 2009. The relationship between pH and zeta potential of ~ 30 nm metal oxide nanoparticle suspensions relevant to in vitro toxicological evaluations. Nanotoxicology 3, 276–283. https://doi.org/10.3109/17435390903276941.

Camacho Vieira, C., Peltonen, L., Karttunen, A.P., Ribeiro, A.J., 2024. Is it advantageous to use quality by design (QbD) to develop nanoparticle-based dosage forms for parenteral drug administration? Int. J. Pharm. 657, 124163. https://doi.org/10.1016/j.ijpharm.2024.124163.

Chen, X., Lv, H., 2022. Intelligent control of nanoparticle synthesis on microfluidic chips with machine learning. NPG Asia Mater. 14. https://doi.org/10.1038/s41427-022-00416-1.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system, pp. 785-794.

Clayburgh, D.R., Shen, L., Turner, J.R., 2004. A porous defense: the leaky epithelial barrier in intestinal disease. Lab. Invest. 84, 282–291. https://doi.org/10.1038/labinvest.3700050.

Danaei, M., Dehghankhold, M., Ataei, S., Davarani, F., Javanmard, R., Dokhani, A., Khorasani, S., Mozafari, M., 2018. Impact of particle size and polydispersity index on the clinical applications of lipidic nanocarrier systems. Pharmaceutics 10. https://doi.org/10.3390/pharmaceutics10020057.

Danhier, F., Ansorena, E., Silva, J.M., Coco, R., Le Breton, A., Préat, V., 2012. PLGA-based nanoparticles: an overview of biomedical applications. J. Control. Release 161, 505–522. https://doi.org/10.1016/j.jconrel.2012.01.043.

Dolai, J., Mandal, K., Jana, N.R., 2021. Nanoparticle Size Effects in Biomedical Applications. ACS Appl. Nano Mater. 4, 6471–6496. https://doi.org/10.1021/acsanm.1c00987.

Grangeia, H., Silva, C., Simoes, S., Reis, M., 2020. Quality by design in pharmaceutical manufacturing: A systematic review of current status, challenges and future perspectives. Eur. J. Pharm. Biopharm. 147, 19–37. https://doi.org/10.1016/j.ejpb.2019.12.007.

Grinsztajn, L., Oyallon, E., Varoquaux, G., 2022. Why do tree-based models still outperform deep learning on typical tabular data? Adv. Neural Inf. Proces. Syst. 35, 507–520.

Gupta, H., Aqil, M., Khar, R.K., Ali, A., Bhatnagar, A., Mittal, G., 2010. Sparfloxacin-loaded PLGA nanoparticles for sustained ocular drug delivery. Nanomedicine 6, 324–333. https://doi.org/10.1016/j.nano.2009.10.004.

Gupta, V., Trivedi, P., 2018. In vitro and in vivo characterization of pharmaceutical topical nanocarriers containing anticancer drugs for skin cancer treatment. Lipid Nanocarriers Drug Targeting 563–627. https://doi.org/10.1016/b978-0-12-813687-4.00015-3.

Hartwig, O., Loretz, B., Nougarede, A., Jary, D., Sulpice, E., Gidrol, X., Navarro, F., Lehr, C.M., 2022. Leaky gut model of the human intestinal mucosa for testing siRNA-based nanomedicine targeting JAK1. J. Control. Release 345, 646–660. https://doi.org/10.1016/j.jconrel.2022.03.037.

Hernández-Giottonini, K.Y., Rodríguez-Córdova, R.J., Gutiérrez-Valenzuela, C.A., Peñuñuri-Miranda, O., Zavala-Rivera, P., Guerrero-Germán, P., Lucero-Acuña, A., 2020. PLGA nanoparticle preparations by emulsification and nanoprecipitation techniques: effects of formulation parameters. RSC Adv. 10, 4218–4231. https://doi.org/10.1039/c9ra10857b.

Huang, W., Zhang, C., 2018. Tuning the size of poly(lactic-co-glycolic Acid) (PLGA) nanoparticles fabricated by nanoprecipitation. Biotechnol. J. 13. https://doi.org/10.1002/biot.201700203.

Lamprecht, A., Schäfer, U., Lehr, C.M., 2001. Size-dependent bioadhesion of micro- and nanoparticulate carriers to the inflamed colonic mucosa. Pharm. Res. 18, 788–793. https://doi.org/10.1023/a:1011032328064.

Lee, B., Yun, Y., Park, K., 2016. PLA micro- and nano-particles. Adv. Drug Deliv. Rev. 107, 176–191. https://doi.org/10.1016/j.addr.2016.05.020.

Lu, G.W., Gao, P., 2010. Emulsions and Microemulsions for Topical and Transdermal Drug Delivery. *Handbook of Non-Invasive Drug Delivery Systems: Non-Invasive and Minimally-Invasive Drug Delivery Systems for Pharmaceutical and Personal Care Products* **2010**, 59-94, doi:10.1016/b978-0-8155-2025-2.10003-4.

Martínez Rivas, C.J., Tarhini, M., Badri, W., Miladi, K., Greige-Gerges, H., Nazari, Q.A., Galindo Rodríguez, S.A., Román, R., Fessi, H., Elaissari, A., 2017. Nanoprecipitation process: From encapsulation to drug delivery. Int. J. Pharm. 532, 66–81. https://doi.org/10.1016/j.ijpharm.2017.08.064.

McCoubrey, L.E., Ferraro, F., Seegobin, N., Verin, J., Alfassam, H.A., Awad, A., Marzorati, M., Verstrepen, L., Ghyselinck, J., De Munck, J., et al., 2024. Poly(D, l-lactide-co-glycolide) particles are metabolised by the gut microbiome and elevate short chain fatty acids. J. Control. Release. https://doi.org/10.1016/j.jconrel.2024.03.039.

Meyer, D., Balemi, A., Wearing, C., 2002. Neural Networks - Their Use and Abuse for Small Data Sets. In: Abbass, H.A., Newton, C.S., Sarker, R. (Eds.), Heuristic and Optimization for Knowledge Discovery. Hershey, PA, USA, IGI Global, pp. 169–185.

Musielak, E., Feliczak-Guzik, A., Nowak, I., 2022. Optimization of the conditions of solid lipid nanoparticles (SLN) synthesis. Molecules 27. https://doi.org/10.3390/molecules27072202.

Nathanael, K., Cheng, S., Kovalchuk, N., Arcucci, R., Simmons, M., 2023. Optimization of microfluidic synthesis of silver nanoparticles: A generic approach using machine learning. Chem. Eng. Res. Des. 193, 65–74. https://doi.org/10.1016/j.cherd.2023.03.007.

Operti, M.C., Bernhardt, A., Grimm, S., Engel, A., Figdor, C.G., Tagit, O., 2021. PLGA-based nanomedicines manufacturing: Technologies overview and challenges in industrial scale-up. Int. J. Pharm. 605, 120807. https://doi.org/10.1016/j.ijpharm.2021.120807.

Ortiz-Perez, A., van Tilborg, D., van der Meel, R., Grisoni, F., Albertazzi, L., 2024. Machine learning-guided high throughput nanoparticle design. Digital Discov. 3, 1280–1291. https://doi.org/10.1039/d4dd00104d.

Paliwal, R., Babu, R.J., Palakurthi, S., 2014. Nanomedicine scale-up technologies: feasibilities and challenges. AAPS PharmSciTech 15, 1527–1534. https://doi.org/10.1208/s12249-014-0177-9.

Saka, O., Öz, U., Küçüktürkmen, B., Devrim, B., Bozkir, A., 2020. Central composite design for optimization of zoledronic acid loaded PLGA nanoparticles. J. Pharm. Innov. 15, 3–14. https://doi.org/10.1007/s12247-018-9365-6.

Seegobin, N., McCoubrey, L.E., Vignal, C., Waxin, C., Abdalla, Y., Fan, Y., Awad, A., Murdan, S., Basit, A.W., 2024. Dual action tofacitinib-loaded PLGA nanoparticles alleviate colitis in an IBD mouse model. Drug Deliv. Transl. Res. https://doi.org/10.1007/s13346-024-01736-1.

Shah, N., Guzmán, E., Wang, Z., Meenach, S., 2020. *Routes of administration for nanocarriers*; pp. 67-87.

Shi, W., Zhang, Z.J., Yuan, Y., Xing, E.M., Qin, Y., Peng, Z.J., Zhang, Z.P., Yang, K.Y., 2013. Optimization of parameters for preparation of docetaxel-loaded PLGA nanoparticles by nanoprecipitation method. J. Huazhong Univ. Sci. Technol. Med. Sci. 33, 754–758. https://doi.org/10.1007/s11596-013-1192-x.

Silveira, R.F., Lima, A.L., Gross, I.P., Gelfuso, G.M., Gratieri, T., Cunha-Filho, M., 2024. The role of artificial intelligence and data science in nanoparticles development: a review. Nanomedicine (lond) 1–13. https://doi.org/10.1080/17435889.2024.2359355.

Sprengholz, M., 2014. Industrial Ram Extrusion As Innovative Tool For The Development Of Biodegradable Sustained Release Implants. **2014**, 178.

Tavares Luiz, M., Santos Rosa Viegas, J., Palma Abriata, J., Viegas, F., Testa Moura de Carvalho Vicentini, F., Lopes Badra Bentley, M.V., Chorilli, M., Maldonado Marchetti, J., Tapia-Blácido, D.R., 2021. Design of experiments (DoE) to develop and to optimize nanoparticles as drug delivery systems. *Eur. J. Pharm. Biopharm.*, *165*, 127-148, doi:10.1016/j.ejpb.2021.05.011.

Walsh, I., Myint, M., Nguyen-Khuong, T., Ho, Y., Ng, S., Lakshmanan, M., 2022. Harnessing the potential of machine learning for advancing “Quality by Design” in biomanufacturing. MAbs 14. https://doi.org/10.1080/19420862.2021.2013593.

Wu, L., Zhang, J., Watanabe, W., 2011. Physical and chemical stability of drug nanoparticles. Adv. Drug Deliv. Rev. 63, 456–469. https://doi.org/10.1016/j.addr.2011.02.001.

Xu, P., Ji, X., Li, M., Lu, W., 2023. Small data machine learning in materials science. npj Comput. Mater. 9. https://doi.org/10.1038/s41524-023-01000-z.

Zaslavsky, J., Bannigan, P., Allen, C., 2023. Re-envisioning the design of nanomedicines: harnessing automation and artificial intelligence. Expert Opin. Drug Deliv. 20, 241–257. https://doi.org/10.1080/17425247.2023.2167978.

Zielińska, A., Carreiró, F., Oliveira, A.M., Neves, A., Pires, B., Venkatesh, D.N., Durazzo, A., Lucarini, M., Eder, P., Silva, A.M., et al., 2020. Polymeric Nanoparticles: Production, Characterization, Toxicology and Ecotoxicology. *Molecules*, *25*, doi:10.3390/molecules25163731.