www.nature.com/scientificdata

scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A formulation dataset of poly(lactide-co-glycolide) nanoparticles for small molecule delivery

Anita Goren[1], Zeqing Bao[2], Juan Pablo Martinez Lozano[1] & Christine Allen[1,2,3] ✉

Poly(lactide-co-glycolide) (PLGA) nanoparticles are promising drug delivery systems, widely recognized for their ability to overcome various limitations associated with conventional formulations. However, designing and optimizing such formulations is a complex and non-trivial process that heavily relies on a lengthy, iterative approach, often involving trial and error. To address the limitations of traditional approaches, formulation scientists are increasingly incorporating artificial intelligence, particularly machine learning, to rationalize and accelerate the process. Despite decades of intensive research into PLGA nanoparticles, a notable shortage remains in the availability of comprehensive open-source datasets essential for driving this accelerated development process forward. Here, we present a literature-curated dataset of 433 PLGA nanoparticle formulations encompassing 65 small molecules. The dataset aims to bridge existing data gaps and provide a comprehensive resource for research on nanoparticle formulations.

Background & Summary

Over the past few decades, considerable research has been dedicated to the development of advanced drug delivery systems to enhance the safety and efficacy of medications[1]. Among these, polymeric nanoparticles (PNPs), have demonstrated potential to address the various challenges of some therapeutics, including limited stability and solubility[2], poor membrane transport[3–5], and insufficient targeting[6]. PNPs are nano-sized drug carriers made from a polymeric matrix, capable of encapsulating both hydrophobic and hydrophilic molecules. These carriers can be engineered to exhibit specific properties, including particle size, payload capacity, and drug release kinetics[7,8]. By tailoring these properties, PNPs can be designed to optimize the delivery of drugs for a wide range of applications[9].

Despite decades of research, the development of PNPs remains a complex and non-trivial process. Typically, this process involves the selection of appropriate formulation parameters, including materials (e.g., polymers and surfactants), preparation method (e.g., nanoprecipitation or emulsion-based methods), and processing parameters (e.g., solvents, and initial composition ratios). These factors can vary widely and significantly impact the performance of the formulation. The goal of formulation optimization is to explore this extensive design space and identify formulation candidates that achieve the desired performance[10].

Traditionally, optimizing PNPs relied on an iterative process, which was time-consuming, resource-intensive, and costly. To overcome these challenges, recent advances in artificial intelligence (AI), particularly machine learning (ML), have been proposed as more efficient alternatives to streamline and rationalize the optimization process[11]. In these studies, ML typically serves as a data-driven approach by leveraging existing data to build predictive computational models that can aid in decision-making[12–14]. By integrating ML into formulation development, researchers can accelerate the process while also allowing for a wider exploration of the design spaces that were previously inaccessible[15,16].

A major challenge in applying ML to drug formulation design is the limited availability of comprehensive and high-quality data. While more datasets are becoming available, a significant gap persists in open-source

[1]Leslie Dan Faculty of Pharmacy, University of Toronto, Toronto, ON, M5S 3M2, Canada. [2]Acceleration Consortium, Toronto, ON, M5S 3H6, Canada. [3]Department of Chemical Engineering and Applied Chemistry, University of Toronto, Toronto, ON, M5S 3E5, Canada. ✉e-mail: cj.allen@utoronto.ca

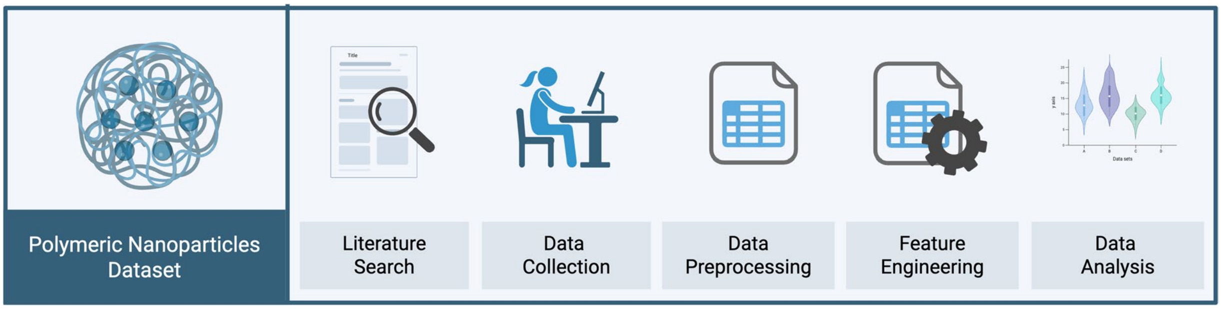

Fig. 1 A summary of the workflow used to generate a dataset of small molecule loaded PLGA nanoparticles. Literature review was conducted to identify studies with relevant data, followed by data collection. The collected data was processed and feature-engineered to generate a comprehensive and structured dataset. Data analysis was then conducted to examine the distribution of features and identify correlations between features.

Feature	Units	Description
polymer_MW	Da	Molecular weight of the PLGA polymer
LA/GA		Ratio of lactide to glycolide in the PLGA polymer.
mol_MW	kDa	Molecular weight of the small molecule
mol_logP		LogP of the small molecule
mol_TPSA	Å²	Topological polar surface area of the small molecule
mol_melting_point	°C	Melting point of the small molecule
mol_Hacceptors		Count of the number of hydrogen acceptors on the small molecule
mol_Hdonors		Count of the number of hydrogen donors on the small molecule
mol_heteroatoms		Count of the number of heteroatoms on the small molecule
drug/polymer		Initial weight ratio of small molecule to polymer
surfactant_concentration	%w/v	Concentration of the surfactant in the aqueous phase
surfactant_HLB		Hydrophilic–lipophilic balance of the surfactant in the aqueous phase
aqueous/organic		Initial volume ratio of the aqueous to organic phase
pH		pH of the aqueous phase
solvent_polarity_index		Polarity index of the solvent used as the organic phase
particle_size	nm	Diameter of the particles
EE	%w/w	Percentage of the small molecule encapsulated by weight relative to the total weight of the PNPs
LC	%w/w	Percentage of the small molecule encapsulated by weight relative to the total weight of the loaded PNPs

Table 1. Description of features in the dataset, including each feature's name, units, and definition.

datasets that the broader research community can easily access and utilize[17–19]. For instance, in the context of PNP formulations, even for well-established and extensively studied polymers such as poly(lactic-co-glycolic) acid (PLGA), no such open-access dataset currently exists[20]. To advance data-driven approaches in formulation development, we present a curated dataset of PLGA nanoparticles loaded with small molecules, sourced from published literature (Fig. 1). The dataset includes 433 formulations, consisting of 65 small molecules, primarily drugs and drug-like compounds. For each formulation, the dataset includes 18 associated features (Table 1) that describe properties of the small molecule, excipients and overall formulation characteristics. Additionally, three key performance metrics are provided: particle size, encapsulation efficiency (EE), and loading capacity (LC). These features were selected to capture a broad range of variables that are critical both to the formulation process and to the *in vitro* or *in vivo* performance of the resulting PNP systems. For example, the physiochemical properties of the small molecule and the processing parameters such as choice of solvent for dissolving both PLGA and the small molecule are included, as these factors can influence formulation properties and performance.

To gain a more in-depth understanding of the dataset, data analysis was conducted to evaluate feature distributions and correlations. The feature distributions in the dataset (Fig. 2) reflect a relatively narrow scope of exploration in the published literature. For instance, the lipophilicity (logP) of the small molecules tend to fall within a limited range, indicating constrained chemical diversity. Correlation analysis (Fig. 3) revealed several expected relationships between certain features, such as a strong positive correlation between the drug to polymer ratio and LC. Additionally, a moderate correlation was observed between the PLGA LA/GA ratio and particle size, however, further investigation is needed to determine if this relationship holds significance.

The dataset presented is intended to serve as a comprehensive resource for researchers. It aims to provide insight into the design of PLGA nanoparticles and address the existing gap in available datasets. Furthermore, it is intended to be a readily accessible tool that supports data-driven approaches and accelerates PNP development.

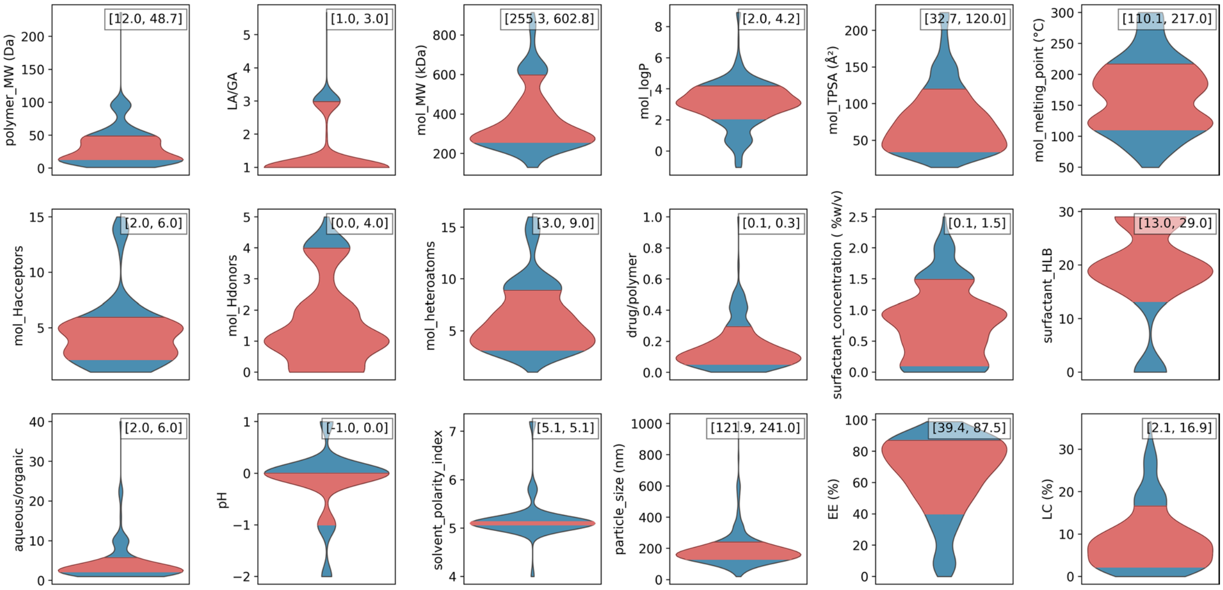

Fig. 2 Violin plots depicting the distribution of the formulation parameters. The central 70% (15th to 85th percentile) is shown in red, while the remaining 30% is shown in blue.

Methods

Literature review and data collection.

A dataset of formulation compositions for small molecule loaded PLGA based nanoparticles was curated from published literature. A search was conducted in May 2024 using Web of Science with the keywords: "PLGA", "drug delivery", "nanoparticles" or "nanospheres", and "nanoprecipitation" or "interfacial deposition" or "solvent displacement" or "solvent injection". This search yielded 812 articles, which were then manually screened for relevance and completeness of the required data. Articles were only included if they met the following criteria: (1) nanoparticles were prepared using the nanoprecipitation method, wherein a small molecule and PLGA polymer are dissolved in an organic solvent and then dispersed into an aqueous phase, with or without a surfactant, (2) the formulations were designed for small molecules, excluding biologics, and (3) no active targeting mechanisms were employed. In addition, only articles that reported or enabled calculation of the features listed in Table 1 were included. This manual screening resulted in 59 articles for data collection. Data collected from these articles included features describing the polymer, small molecule, formulation parameters, and the performance of the formulation.

Data preprocessing and feature engineering.

After data collection, preprocessing and feature engineering were performed to clean the dataset, transform data, and incorporate additional features related to the properties of the solvents, excipients, and small molecules. In cases where either EE or LC was not reported, it was calculated using the following equations.

$$EE\% = \frac{LC \times \textit{mass of the polymer used}\,(mg)}{\textit{mass of the small molecule used}\,(mg)} \times 100\%$$

$$LC\% = \frac{EE \times \textit{mass of the small molecule used}\,(mg)}{\textit{mass of the polymer used}\,(mg)} \times 100\%$$

The pH values of the aqueous phase were categorized into discrete ranges: values below 4 were assigned a label of −1, values between 4 and 6 were labeled as 0, values between 6 and 8 were also assigned a label of 0, and values above 8 were labeled as 1. In cases where only the inherent viscosity of the PLGA polymer was reported, the molecular weight was estimated using the Mark-Houwink equation[21]. Subsequently, additional descriptors for the small molecules were calculated using the RDKit toolkit based on their Simplified Molecular Input Line Entry System (SMILES).

Data analysis.

Data analysis was performed using a custom codebase previously developed to analyze microparticle datasets, focusing on distribution and correlation analyses[19]. The distribution analysis (Fig. 2) is represented as violin plots to represent the central 70% of the data for each feature. Additionally, a correlation matrix (Fig. 3) was generated to display the pairwise Pearson correlations between all formulation parameters.

Data Records

The final dataset, which includes all relevant features, is provided alongside the initial dataset curated from the literature with appropriate references. Additional datasets containing data for small molecules, excipients, and solvents are also provided. The data is openly accessible on Mendeley Data (https://data.mendeley.com/datasets/sbjf5csrdm/1)[22]. Table 2 presents an overview and detailed description of all the datasets provided.

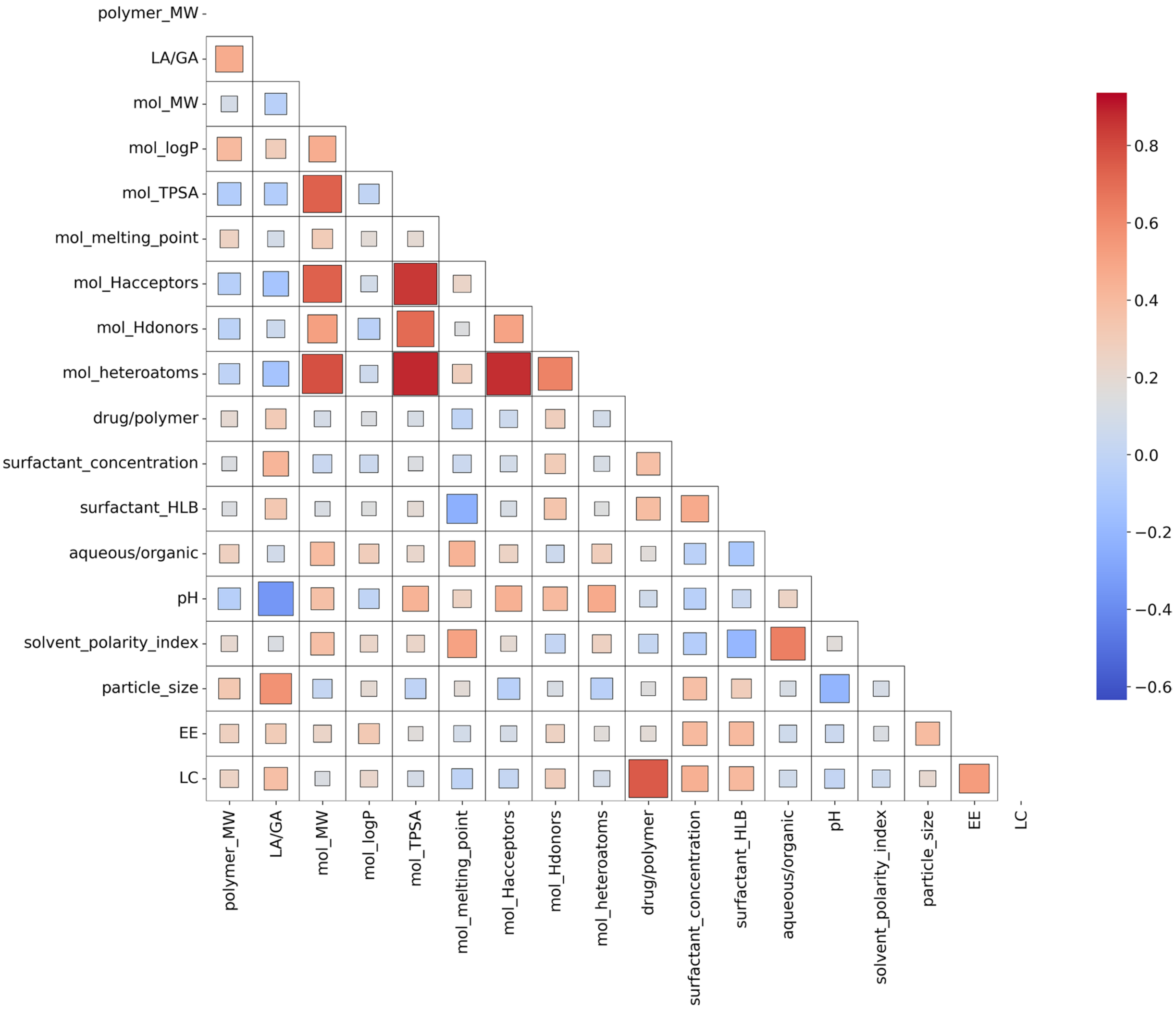

Fig. 3 Correlation matrix showing pairwise Pearson correlations between all the formulation parameters. The colour intensity represents the magnitude of the correlation, where red indicates positive correlations and blue indicates negative correlations.

File name	Description
NP_dataset.csv	Final and complete dataset with all features
NP_dataset_formulations.csv	Initial dataset of formulation compositions sourced from literature
NP_dataset_small_molecules.csv	Dataset of small molecules with their chemical structures
NP_dataset_surfactants.csv	Dataset of surfactants with hydrophilic-lipophilic balance (HLB)
NP_dataset_solvents.csv	Dataset of solvents with polarity index

Table 2. The datasets provided with a description of the file's content.

Technical Validation

Data collection from literature was conducted independently by two individuals. The resulting datasets were then cross validated to identify and resolve any discrepancies. This approach was taken to ensure consistency with reported data and to minimize bias in the manual selection process.

Code availability

The code used in this work was originally developed in a previously published study and is publicly available on Mendeley Data (https://data.mendeley.com/datasets/zzvtdrcy76/2)[19].

Received: 31 March 2025; Accepted: 2 July 2025;
Published online: 10 July 2025

References

1. Bae, Y. H. & Park, K. Advanced drug delivery 2020 and beyond: Perspectives on the future. *Advanced Drug Delivery Reviews* **158**, 4–16 (2020).
2. Ravichandran, R. Nanoparticles in Drug Delivery: Potential Green Nanobiomedicine Applications. *International Journal of Green Nanotechnology: Biomedicine* **1**, B108–B130 (2009).
3. Blanco, E., Shen, H. & Ferrari, M. Principles of nanoparticle design for overcoming biological barriers to drug delivery. *Nat Biotechnol* **33**, 941–951 (2015).
4. Ensign, L. M., Cone, R. & Hanes, J. Oral Drug Delivery with Polymeric Nanoparticles: The Gastrointestinal Mucus Barriers. *Advanced Drug Delivery Reviews* **64**, 557 (2011).
5. Patel, T., Zhou, J., Piepmeier, J. M. & Saltzman, W. M. Polymeric Nanoparticles for Drug Delivery to the Central Nervous System. *Advanced Drug Delivery Reviews* **64**, 701 (2011).
6. Manzari, M. T. *et al.* Targeted drug delivery strategies for precision medicines. *Nat Rev Mater* **6**, 351–370 (2021).
7. Mitchell, M. J. *et al.* Engineering precision nanoparticles for drug delivery. *Nat Rev Drug Discov* **20**, 101–124 (2021).
8. Wang, S. *et al.* Nanoparticle-based medicines in clinical cancer therapy. *Nano Today* **45**, 101512 (2022).
9. Liu, Y., Tan, J., Thomas, A., Ou-Yang, D. & Muzykantov, V. R. The shape of things to come: importance of design in nanotechnology for drug delivery. *Ther Deliv* **3**, 181–194 (2012).
10. Crucho, C. I. C. & Barros, M. T. Polymeric nanoparticles: A study on the preparation variables and characterization methods. *Materials Science and Engineering: C* **80**, 771–784 (2017).
11. Bannigan, P. *et al.* Machine learning directed drug formulation development. *Advanced Drug Delivery Reviews* **175**, 113806 (2021).
12. Noorain, L., Nguyen, V., Kim, H.-W. & Nguyen, L. T. B. A Machine Learning Approach for PLGA Nanoparticles in Antiviral Drug Delivery. *Pharmaceutics* **15**, 495 (2023).
13. Rezvantalab, S., Mihandoost, S. & Rezaiee, M. Machine learning assisted exploration of the influential parameters on the PLGA nanoparticles. *Sci Rep* **14**, 1114 (2024).
14. Seegobin, N. *et al.* Optimising the production of PLGA nanoparticles by combining design of experiment and machine learning. *International Journal of Pharmaceutics* **667**, 124905 (2024).
15. Tao, H. *et al.* Nanoparticle synthesis assisted by machine learning. *Nat Rev Mater* **6**, 701–716 (2021).
16. Li, J. *et al.* AI Applications through the Whole Life Cycle of Material Discovery. *Matter* **3**, 393–432 (2020).
17. Zaslavsky, J. & Allen, C. A dataset of formulation compositions for self-emulsifying drug delivery systems. *Sci Data* **10**, 914 (2023).
18. Bannigan, P. *et al.* Machine learning models to accelerate the design of polymeric long-acting injectables. *Nat Commun* **14**, 35 (2023).
19. Bao, Z., Kim, J., Kwok, C., Le Devedec, F. & Allen, C. A dataset on formulation parameters and characteristics of drug-loaded PLGA microparticles. *Sci Data* **12**, 364 (2025).
20. Jones, D. E., Ghandehari, H. & Facelli, J. C. A review of the applications of data mining and machine learning for the prediction of biomedical properties of nanoparticles. *Comput Methods Programs Biomed* **132**, 93–103 (2016).
21. Wagner, H. L. The Mark–Houwink–Sakurada Equation for the Viscosity of Linear Polyethylene. *Journal of Physical and Chemical Reference Data* **14**, 611–617 (1985).
22. Goren, A., Bao, Z., Martinez Lozano, J. P. & Allen, C. A formulation dataset of poly(lactide-co-glycolide) nanoparticles for small molecule delivery. *Mendeley Data* https://doi.org/10.17632/sbjf5csrdm.1 (2025).

Acknowledgements

This research was supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant (RGPIN-2022-04910) awarded to CA, along with funding from the Acceleration Consortium at the University of Toronto, which is supported by the Canada First Research Excellence Fund (CFREF-2022-00042).

Author contributions

A.G. performed the literature review, data collection, data analysis, and wrote the first manuscript draft. Z.B. performed data analysis and edited the manuscript. J.P.M.L. performed literature review and data collection. C.A. supervised the work, edited and reviewed the manuscript, and secured funding for the project.

Competing interests

C.A. is a cofounder and CEO of Intrepid Labs Inc.

Additional information

Correspondence and requests for materials should be addressed to C.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.