Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

COMMUNICATION

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Digital Discovery*, 2024, 3, 1280

Received 12th April 2024
Accepted 2nd June 2024

DOI: 10.1039/d4dd00104d

rsc.li/digitaldiscovery

Machine learning-guided high throughput nanoparticle design†

Ana Ortiz-Perez, ‡[a] Derek van Tilborg, ‡[ab] Roy van der Meel, [a] Francesca Grisoni *[ab] and Lorenzo Albertazzi *[a]

Designing nanoparticles with desired properties is a challenging endeavor, due to the large combinatorial space and complex structure–function relationships. High throughput methodologies and machine learning approaches are attractive and emergent strategies to accelerate nanoparticle composition design. To date, how to combine nanoparticle formulation, screening, and computational decision-making into a single effective workflow is underexplored. In this study, we showcase the integration of three key technologies, namely microfluidic-based formulation, high content imaging, and active machine learning. As a case study, we apply our approach for designing PLGA-PEG nanoparticles with high uptake in human breast cancer cells. Starting from a small set of nanoparticles for model training, our approach led to an increase in uptake from ~5-fold to ~15-fold in only two machine learning guided iterations, taking one week each. To the best of our knowledge, this is the first time that these three technologies have been successfully integrated to optimize a biological response through nanoparticle composition. Our results underscore the potential of the proposed platform for rapid and unbiased nanoparticle optimization.

Introduction

Nanomedicines are relevant for a variety of biomedical applications,[1] from diagnosis[2] and disease prevention[3] to novel therapeutic approaches.[4] Nanomedicine platforms with a wide range of physicochemical properties can be engineered using a variety of materials,[5,6] by tuning nanoparticle composition and formulation variables.[7,8] These properties, in turn, influence nanoparticle fate and their ability to cross biological barriers.[5,9,10] This versatility opens opportunities to build tailored carriers for a specific application and patient populations[5] but also poses a great challenge towards the design of optimal materials. The resulting enormous combinatorial design space – realistically consisting of thousands of formulations for a single nanoparticle type – makes formulation exploration a daunting task. Thus, we need efficient ways to navigate this vast space, in a time- and cost-effective manner. Novel tools for high-throughput formulation and screening, as well as data-driven computational methods for nanoparticle design hold a great promise to revolutionize the current landscape of material discovery.

However, integrating these tools into a single robust, rapid, and effective workflow is still an open question. In this study, we combined three key technologies: microfluidic formulation, high content imaging, and active machine learning into an iterative workflow to accelerate nanoparticle design (Fig. 1).

Microfluidics offers a versatile platform for rapid and reproducible production of highly monodispersed nanoparticles[11,12] compared to standard bulk formulation. Control over formulation parameters, such as the solvent mixing rate, is achieved by handling small volumes of liquids in highly controlled environments. The solvent mixing rate drives the formulation of several self-assembling nanoparticles including amphiphilic lipids and polymers[13] and controls physical properties like size.

In parallel, the spread of fluorescence-based microscopy together with the rapid development of bio-image analysis tools[14] and automation has enabled the high throughput screening of nanocarriers using high content imaging (HCI). HCI combines automated fluorescence imaging and analysis, providing quantitative multiparametric data from images.[15,16]

HCI-based assays can then be used to understand the impact of the nanoparticle on the cell, including uptake,[17,18] endosomal escape,[19] or cytotoxicity,[20] assisting the rational design of nanoparticles.

Finally, machine learning can be used to guide nanoparticle development[21,22] with the aim of reducing the number of nanoparticle formulations needed to optimize a response. Since

[a] *Institute for Complex Molecular Systems (ICMS), Department of Biomedical Engineering, Eindhoven University of Technology, PO Box 513, 5600 MB Eindhoven, The Netherlands. E-mail: l.albertazzi@tue.nl*

[b] *Centre for Living Technologies, Alliance TU/e, WUR, UU, UMC Utrecht, Princetonlaan 6, 3584 CB, Utrecht, The Netherlands. E-mail: f.grisoni@tue.nl*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d4dd00104d

‡ These authors contributed equally to this work.

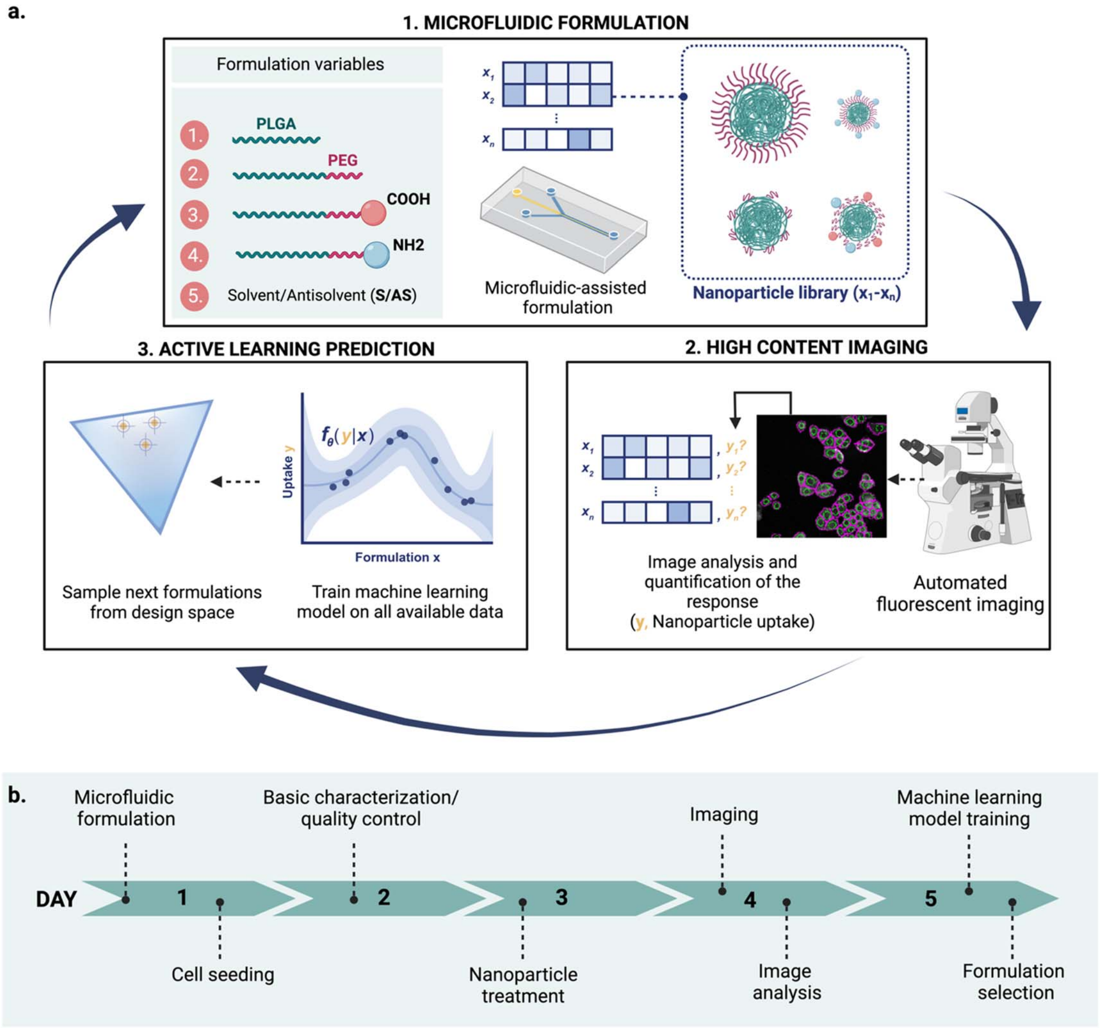
a.
1. MICROFLUIDIC FORMULATION
Formulation variables
1. PLGA
2. PEG
3. COOH
4. NH2
5. Solvent/Antisolvent (S/AS)
x1
x2
xn
Microfluidic-assisted formulation
Nanoparticle library (x1-xn)
2. HIGH CONTENT IMAGING
x1
x2
xn
y1?
y2?
yn?
Image analysis and quantification of the response
(y, Nanoparticle uptake)
Automated fluorescent imaging
3. ACTIVE LEARNING PREDICTION
fθ(y|x)
Uptake y
Formulation x
Sample next formulations from design space
Train machine learning model on all available data
b.
DAY
1
2
3
4
5
Microfluidic formulation
Basic characterization/ quality control
Imaging
Machine learning model training
Cell seeding
Nanoparticle treatment
Image analysis
Formulation selection

Fig. 1 Conceptual overview of the proposed iterative nanoparticle design pipeline. (a) The three key integrated technologies: (1) nanoparticles are formulated using microfluidics-assisted nanoprecipitation by controlling different formulation variables x_i, (2) the formulations are screened with high content imaging (HCI) to determine their properties y_i (*e.g.*, their uptake in MDA-MB-468 cells, as in this proof of concept), and (3) a machine learning model learns the relationship between nanoparticle formulations (*x*) and their corresponding property (*y*), and is used to guide the next cycle. (b) Overview of the experimental cycle: from microfluidic formulation to formulation selection for the following cycle in five days.

the number of available data is often highly limited, specific machine learning strategies like active machine learning are particularly suited for this task.[23–27] By operating in an iterative fashion, active machine learning uses model predictions to decide which samples should be screened and added to the training data to update the model in the next cycle.[28,29] This allows models to reach a desired response faster by screening fewer samples. Furthermore, the iterative nature of active learning makes it fitting for integration with automated design platforms where nanoparticles designs are optimized sequentially.

Although these techniques have been widely explored on their own, combining their advantages can potentially accelerate nanoparticle design. Here, we demonstrate an integrated and semi-automated iterative workflow for rapid nanoparticle design (Fig. 1a), combining the strengths of (1) microfluidic-assisted nanoparticle formulation, (2) HCI, and (3) active machine learning. We apply this iterative approach to find poly(lactic-*co*-glycolic acid)-polyethylene glycol (PLGA-PEG) compositions that yield a high uptake in MDA-MB-468 human breast cancer cells. Owing to its modular character, the approach can be adapted to explore other nanoparticle formulations and responses of interest beyond uptake.

Results & discussion

As a case study, we focused on PLGA-PEG nanoparticles. PLGA-PEG is an amphiphilic block copolymer that self-assembles into nanospheres *via* nanoprecipitation. This type of formulation can be easily adapted to the microfluidic format, which offers several advantages over traditional formulation, such as size tunability by controlling the fluidic parameters.[8,30,31] In addition, PLGA-PEG has excellent biocompatibility and high tunability. The base polymer can be manufactured with different properties, such as molecular weight or functional end-groups. For creating a library of non-targeted PLGA-PEG based nanoparticles, we chose to vary four different building blocks (PLGA, PLGA-PEG, PLGA-PEG-COOH, PLGA-PEG-NH_2) and one process variable (the flow rate ratio between the solvent and antisolvent). By varying building components that directly influence physicochemical properties (size, PEGylation, and charge), we aim to maximize their uptake in a model of human breast cancer.

Platform for nanoparticle design

Our proposed workflow is constituted of three components (microfluidics formulation, HCI and machine learning), each of which contributes to the 'experimental cycle' represented in Fig. 1a. Each cycle can be performed in a week (Fig. 1b), allowing for rapid design iterations. The three components of our platform are the following.

Microfluidics device. Microfluidic systems have been reported for the rapid and controllable formulation of several self-assembling nanoparticles.[12,13] Here, we chose a hydrodynamic flow focusing (HFF) microfluidic device with a Y-junction geometry to manufacture PLGA-PEG nanoparticles. HFF has shown remarkable control over the size of the carriers, by tuning the flow rate ratio between the solvent (S) and anti-solvent (AS).[30,31] Fluidic control is achieved using two syringe pumps connected to the middle and side inlets of the Y-junction, for the solvent (S) and anti-solvent (AS) streams, respectively (Fig. 2a). Different nanoparticle compositions can be prepared automatically using a syringe pump with a rotary valve connected to the sample reservoirs, containing the

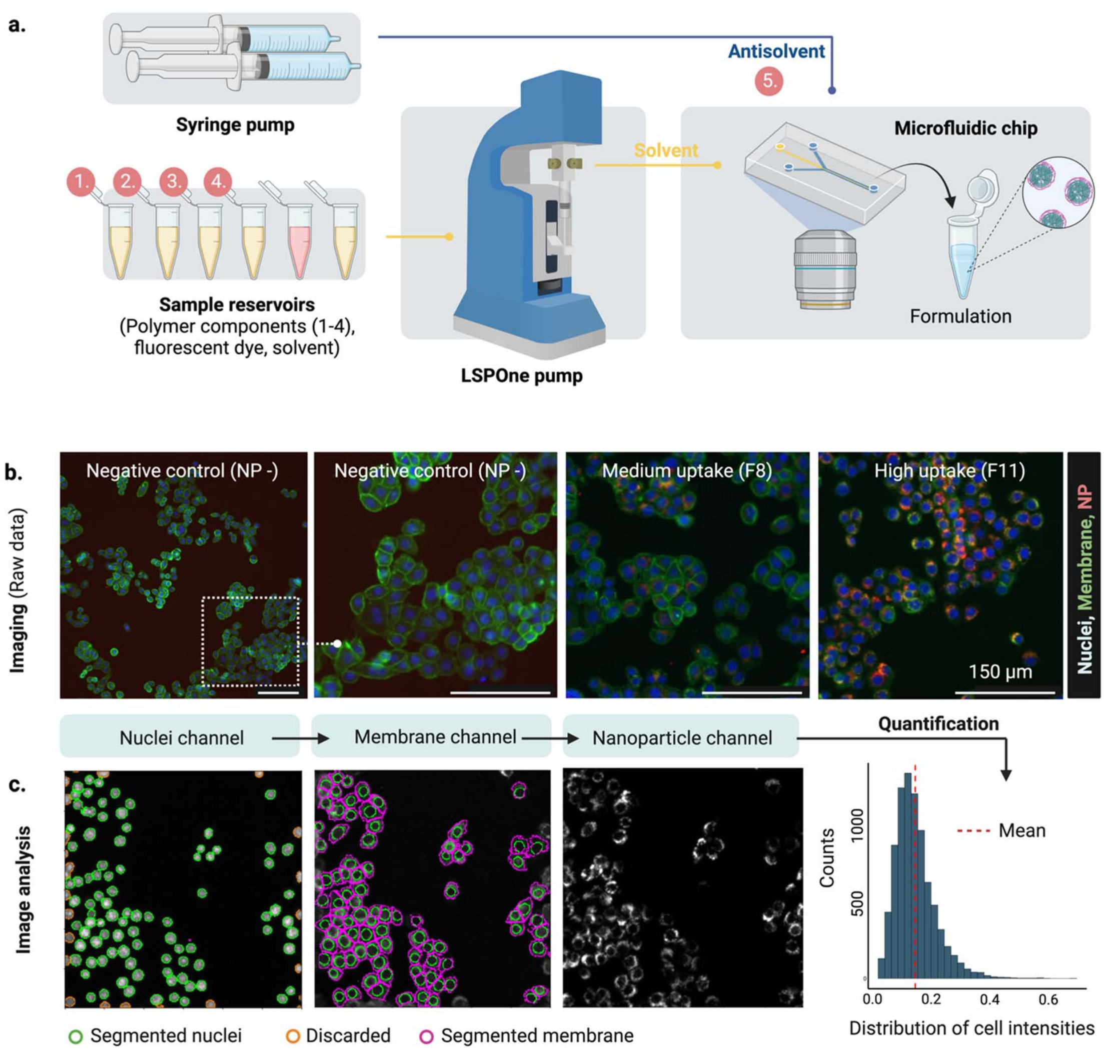

Fig. 2 Microfluidic set-up and high content screening. (a) Formulation of PLGA-PEG nanoparticles varying four different components (1–4, PLGA polymers) and one process variable (5, Solvent/Antisolvent S/AS flow rate ratio). Polymer mixtures and their injection into the middle channel of the hydrodynamic flow focusing (HFF) device is achieved with the LSPOne pump and different levels of S/AS are accomplished by changing the antisolvent (water) flow rate with a syringe pump. (b) For imaging of nanoparticle uptake in MDA-468, the raw data is composed of three channels (nuclei, membrane, nanoparticle (NP)), with each field of view of 804 × 804 px, 1.123 μm px^{-1}. Examples qualitatively illustrating three levels of uptake (negative, medium, high). Scale bars 150 μm. (c) Image analysis by segmentation of the nuclei, followed by membrane segmentation and quantification of mean intensity on the nanoparticle channel per cell per area. Distribution of cell intensities shows a gamma distribution.

different polymer blocks. The polymer mixture is then injected into the middle channel of the microfluidic chip at a constant flow rate, and the anti-solvent rate is adjusted to meet the desired S/AS flow rate ratio (FRR). During formulation, the nanoparticles are labeled *in situ* by encapsulation of a fluorescent dye on their hydrophobic core. This setup enables the automated formulation of nanoparticles with controllable size and composition. The current port configuration (10-port valve, see ESI†) enables mixing up to 6 building blocks or components. Each formulation takes less than 20 min, for 1 mg of material (with variable concentration depending on S/AS).

High content imaging (HCI). High-content imaging is used to acquire and process widefield fluorescence images in an automated way in 96-well plates (Fig. 2b). After acquisition, a three-step bio-image analysis pipeline is used, based on CellProfiler,[32] consisting of (1) nuclear segmentation, (2) membrane segmentation, and (3) intensity quantification (Fig. 2c). In this assay, we measure the nanoparticle intensity per cell per area, which fits an expected gamma distribution, in accordance with theoretical and experimental reports.[17] As highlighted earlier, many fluorescent-based assays can be adapted to this format, expanding the assay capabilities to interrogate cell state,[33] cytotoxicity,[20] or nanoparticle fate.[19] For our proof-of-concept, we chose to measure uptake as a response of interest.

Active machine learning. Active learning is based on the principle that a machine learning model can achieve better performance with less data if it is allowed to choose the data from which it can learn in the next cycles.[34] The two common strategies for selecting the next samples to screen are known as exploration and exploitation.[28,29] In exploration, new regions from the design space are investigated. Here, the samples that are expected to be the most informative to learn from are selected, with the aim of getting a better model. We assume that screening samples with high prediction uncertainty will add the most information to the model. Exploitation, on the other hand, aims to identify nanoparticles with desired experimental properties. This is often done by selecting nanoparticles from the areas in the design space that can be predicted with high certainty. Based on preliminary experiments (see ESI†) and our modelling requirements, we chose to use a Bayesian neural network[35] to predict nanoparticle response in our workflow. A Bayesian neural network is a probabilistic model that, instead of predicting a single value, outputs a distribution of predictions for any input. Compared to feasible alternatives like Gaussian processes or an ensemble of point estimate models (*e.g.*, random forest), Bayesian neural networks enable robust predictions in a low-data setting due to its high expressivity, innate ability to estimate prediction uncertainty, and resilience to dataset shifts[36] and overfitting.[37]

At each cycle, the three technologies work complementary as follows: (a) microfluidics technology is used for nanoparticle production, (b) the obtained nanoparticles are analyzed using HCI for property determination, and (c) the experimental results are used to train the machine learning model, which is then used to suggest what to formulate next. The optimal learning strategy (exploration *vs.* exploitation) over cycles is not predetermined and can be adjusted upon learned insights. Choosing between exploration and exploitation is case-dependent and it is ultimately decided by the scientist.

Designing PLGA-PEG nanoparticles for high uptake

The proposed design platform was used to perform three cycles. Per each cycle, nanoparticles were produced in the microfluidics platform with the chosen formulation. Hydrophobic fluorescent dyes (DiD) were incorporated into the polymer mixture to allow for estimation of cell accumulation. Cell uptake was determined *via* HCI and expressed as fold-increase accumulation (compared to the uptake control, a 100% PLGA-PEG nanoparticle formulated in bulk).

The measured response per nanoparticle was used to train the Bayesian neural network for uptake prediction. The trained model was then used to select the next cycle formulations from a virtual library of 100 000 nanoparticles spanning the entire design space homogenously. Formulations were considered only if their predicted polydispersity index (PDI) was lower than a predetermined threshold (PDI < 0.2, predicted with a different machine learning model, see Materials and methods). Filtering *via* PDI is a form of quality control ensuring the produced nanoparticles are colloidally stable and suitable for biomedical applications such as drug delivery. As a learning strategy, we started by exploring the uncertain areas of the design space (exploration), after which we aimed to find high response nanoparticles (exploitation). As a result, the study was executed in three cycles, as described below (Fig. 3).

Cycle 0 (dataset generation). Machine learning needs training data to start with. When the dataset is limited by size, it has been shown that machine learning algorithms can benefit from starting with a diverse dataset.[38] We selected 29 formulations (cycle 0) using a Design of Experiments (DoE) methodology. A mixture-process variable design allowed us to pick formulations that were distributed homogeneously within the design space (Fig. 3a), yielding a starting set with diverse compositions spread over the whole design space.[39] These formulations were produced and characterized for their cell uptake. The nanoparticle uptake ranged from 0.40 to 4.77-fold with respect to the uptake standard, with an average uptake of 2.03 ± 1.28-fold (Fig. 3d and e, dark blue color). Polymer composition is visually represented in Fig. 3f, and corresponding characterization of physicochemical properties are available in the ESI.† Experimentally determined uptake values, together with the nanoparticle's corresponding formulation variables, were used to train the first neural network model (see Materials and methods).

Cycle 1 (exploration). Predictions from the neural network model trained with cycle 0 data were used to guide the next design cycle. Here, we primarily aimed at exploring the regions of the design space where the model is most uncertain about to increase overall model performance. Therefore, we used the model's prediction uncertainty to drive formulation selection. We selected ten formulations that were: (a) as diverse from each other as possible (determined *via* clustering, see Materials and methods), and (b) with a high prediction uncertainty and

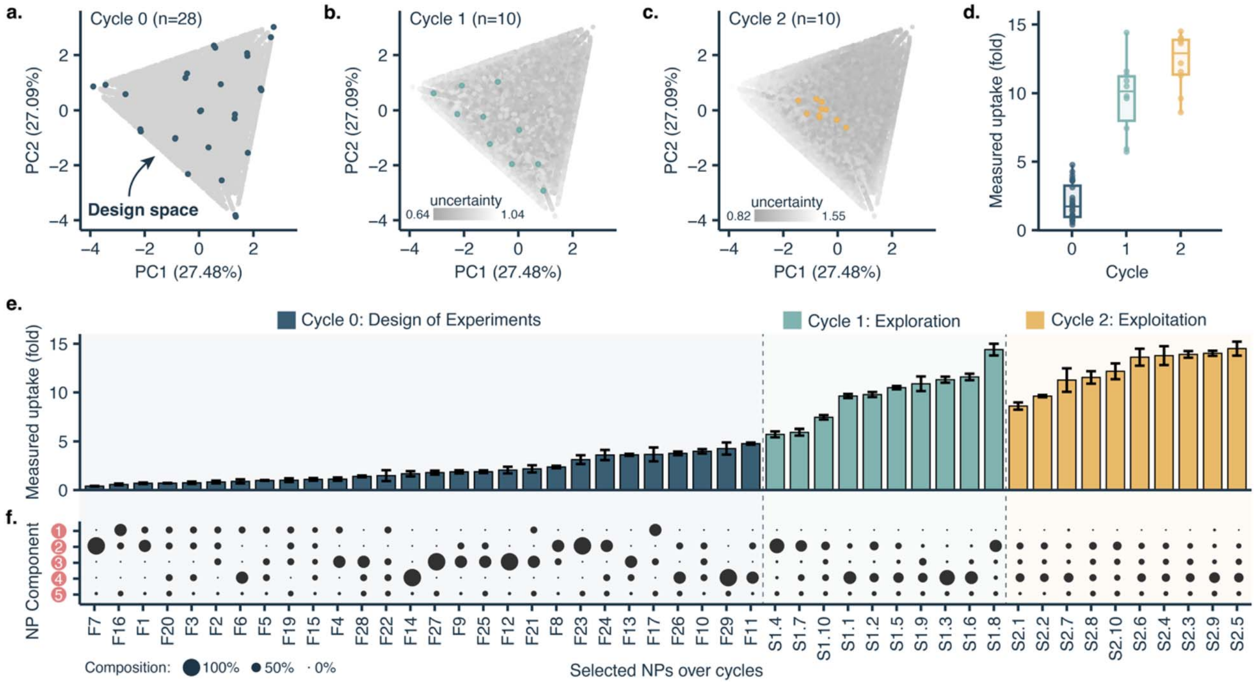

Fig. 3 Optimizing PLGA-PEG nanoparticle uptake in MDA 468 cells with machine learning guided formulation. (a) Principal component analysis (PCA) of the nanoparticle (NP) design space, projecting the range of all five formulation variables into two dimensions. Each point represents a nanoparticle formulation, with grey representing all formulations of the *in silico* screening library ($n = 100\,000$), and blue representing formulated nanoparticles in cycle 0 (DoE, $n = 28$). (b) PCA illustrating the selection of nanoparticle formulations for cycle 1 (exploration, $n = 10$). (c) PCA illustrating the selection of nanoparticle formulations for cycle 2 (exploitation, $n = 10$). (d) Boxplots of the measured uptake of formulated nanoparticles over cycles. (e) Measured uptake over screening cycles. Error bars represent standard deviation. Nanoparticles are sorted by uptake for illustrative purposes. (f) Composition of the formulated nanoparticles. Circle size represents the percentage of each nanoparticle formulation component. Components used are: 1; pure PLGA, 2; PLGA-PEG, 3; PLGA-PEG-COOH, 4; PLGA-PEG-NH2, and 5; solvent/antisolvent ratio. nanoprecipitation).

moderately high uptake (Fig. 3b). The resulting nanoparticle designs were experimentally assessed and had an uptake ranging from 5.72 to 14.40-fold, with an average of 9.72 ± 2.70-fold: a considerable leap in uptake compared to cycle 0 (where the best nanoparticle resulted in 4.77-fold uptake, Fig. 3d and e). This newly obtained data was combined with the data from cycle 0 and used to re-train the model, to guide cycle 2 formulations.

Cycle 2 (exploitation). Having explored the most uncertain areas of the design space, we aimed at obtaining high uptake nanoparticles in MDA-MB-468 cells by selecting formulations in an exploitative manner. Instead of acquiring more knowledge about the uncertain areas in the design space, we selected ten nanoparticle formulations with a high predicted uptake and a low uncertainty (Fig. 3c) for formulation and HCI screening. We enforce some degree of diversity among the selected formulations (see Materials and methods). These nanoparticles were found to have an uptake between 8.60 and 14.50-fold, with an average of 12.30 ± 2.02-fold (Fig. 3d and e). This cycle yielded a remarkable improvement in the average uptake over all ten nanoparticles, and to a slightly higher maximum uptake.

With only three full cycles we were able to move from a mean uptake of 2.03 ± 1.28-fold in the initial set to 12.30 ± 2.02-fold in the last cycle. The maximal uptake improved from 4.77-fold in the first cycle to 14.50-fold in the last cycle.

Model interpretation

To fully leverage what the machine learning model learned from the data, we applied it to interrogate nanoparticle composition–function relationships. We retrained the model with all the data generated from all cycles and used it to select five nanoparticles with low predicted uptake and five with high predicted uptake from the virtual library for further formulation and screening (Fig. 4a). Although the model was better attuned to high-uptake formulations, it was able to identify both high-uptake (10.54 ± 0.66-fold) and low-uptake nanoparticles (2.74 ± 0.99-fold), with statistically significant differences ($p < 0.001$, two-tailed *t*-test, Fig. 4b). This shows that the model has learnt relevant formulation–uptake relationships.

Low- and high-uptake nanoparticles showed statistically significant differences ($p < 0.001$, two-tailed *t*-test) in the content of three polymers (Fig. 4c): (1) PLGA (low-uptake: 53 ± 9%, high-uptake: 5 ± 5%), (2) PLGA-PEG-COOH (low-uptake: 3 ± 5%, high-uptake: 24.4 ± 4.8%), and (3) PLGA-PEG-NH_2 (low-uptake: 10 ± 6%, high-uptake: 25 ± 5%). This is also reflected in the predictions over the whole design space (see ESI†). Furthermore, low-uptake nanoparticles were found to be more monodisperse (PDI = 0.059 ± 0.013) and bigger (size = 154.5 ± 21.4 nm) than high-uptake nanoparticles (PDI = 0.122 ± 0.015, size = 114.0 ± 5.2 nm). Polydispersity or nanoparticle heterogeneity was traditionally seen as an undesired property. However, this

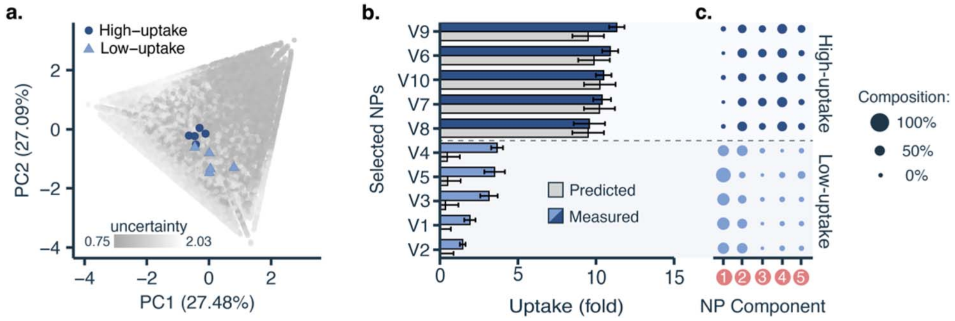

Fig. 4 PLGA-PEG nanoparticle uptake in MDA 468 cells for nanoparticles with low predicted uptake and high predicted uptake. (a) Principal component analysis (PCA) representing all nanoparticle (NP) formulations of the *in silico* screening library (grey, $n = 100\,000$), the selected low-uptake formulations (light blue triangles, $n = 5$), and the selected high-uptake formulations (dark blue circles, $n = 5$). (b) Comparison between predicted (grey) and measured uptake for nanoparticles with low- and high predicted uptake. Error bars represent standard deviation. (c) Composition of the formulated nanoparticles. Circle size represents the percentage of each nanoparticle formulation component. Components used are: 1; pure PLGA, 2; PLGA-PEG, 3; PLGA-PEG-COOH, 4; PLGA-PEG-NH_2, and 5; solvent/antisolvent ratio.

intrinsic heterogeneity can be considered a structural parameter contributing to nanoparticle fate and biological function.[40]

Conclusions and outlook

In this work, we demonstrate a nanoparticle design platform combining three complementary technologies, namely microfluidics-assisted formulation, high content imaging, and machine learning. These three technologies have been tuned to work synergistically within an active learning framework, where the results of each experimental cycle are used to inform the next.

As a proof-of-concept, we applied our approach for designing PLGA-PEG nanoparticles with high uptake in MDA-MB-468 human breast cancer cells. With only two experimental cycles of 5 days each, we were able to triple the measured uptake from ~5-fold to ~15-fold. The resulting model was able to generate low uptake and high uptake nanoparticles based on their composition. Such a model could be used for exploring the relationships between the nanoparticle components and their function as an 'hypothesis generator'. These results demonstrate the potential of this approach to efficiently navigate complex design spaces of multicomponent nanoparticles.

Owing to its modularity, this approach can be further expanded to tackle virtually any nanoparticle formulation. In the future, we will apply this approach for designing nanoparticles with relevant translational properties, such as selective cytotoxicity in cancer cells, or the capability to deliver functional cargo to target cells. Moreover, the approach is generalizable to a range of nanomaterials and can be expanded to models with different biological complexities, *e.g.*, cell lines, patient-derived organoids, or organs-on-a-chip. Our approach demonstrates the potential of closed-loop platforms for rapid and iterative nanoparticle optimization driven by machine learning.

Materials and methods

Nanoparticle formulation

Chemicals. All Poly Lactic-*co*-Glycol Acid (PLGA)-Polyethylene Glycol (-PEG) based polymers were purchased from Akina Inc. division PolySciTech (West Lafayette, USA): PLGA (#AP082), PLGA-PEG (#AK102), PLGA-PEG-COOH (#AI078), PLGA-PEG-NH_2 (#CAI189). The encapsulated dye was a DiD solution from Invitrogen Vybrant™ Multicolor Cell-Labelling kit (Cat no. V22889), purchased from Fisher Scientific (Landsmeer, Netherlands). High-grade acetonitrile (>99%) was used as organic solvent.

Microfluidic chip manufacturing. The microfluidic chips were manufactured in polydimethylsiloxane (PDMS) from a SU-8 patterned Silicon wafer master mold. The design of the corresponding microstructures (a Y-junction, with 200 μm of channel width and 60 μm height) and their manufacturing process are described in detail in our previous work, by Mares *et al.*[31] Each chip (or PDMS replica) was prepared from the master mold by standard soft lithography. First, a PDMS base polymer and elastomer from a two-component kit (Sylgard 184, Dow Corning) were thoroughly mixed in 10 : 1 wt : wt ratio. The mixture was degassed in a desiccator, poured over the master mold, degassed once more and baked overnight at 60 °C. After elastomer curation, the PMDS chips were peeled off from the master mold, the inlets and outlets were punched with a 1.2 mm biopsy puncher and stored in a dust-free environment. On the same day of formulation, to keep surface hydrophilicity, the PDMS replica was freshly bonded to a clean 25 × 75 mm glass slide using oxygen plasma (at 20 W for 30 s), achieved with an Emitech K1050X Plasma Asher from Quorum (East Sussex, UK).

Microfluidic-assisted nanoparticle formulation. Nanoparticles were formulated by microfluidic-assisted nanoprecipitation, in which an acetonitrile stream (solvent, S) containing all polymer components is hydrodynamically focused by an aqueous phase (anti-solvent, AS) in a Y-junction. The AS phase was ultra-pure water pumped into the lateral inlets of the chip by a Fusion 200 Two-channel Chemyx Syringe Pump (Stafford, USA), while the acetonitrile was pumped inside the central channel by a LSPOnePump with a 10-port valve and a 250 μl syringe from Advanced MicroFluidics SA (Ecublens, Switzerland). This last pump was also used to make mixtures of polymer components prior injection of the organic phase into the device. This was achieved by using the following port configuration: 1 waste, 2 output, 3 mixing, 4 buffer, and the remaining ports (5–9) for dye and polymer components reservoirs. Schematics and port configuration shown in ESI Fig. 5.† All solvents used were filtered with a Whatman's polyvinylidene fluoride (PVDF) 0.2 μm membrane filter. Filtered acetonitrile was used to make the polymer stocks (reservoirs) at a concentration of 15 mg ml^{-1} and to dilute the commercial DiD dye from 1 mM to 500 μM. The polymer component mix was injected at a total polymer concentration of 10 mg ml^{-1}, with an S flow rate of 15 μl min^{-1}, and a variable AS flow rate, depending on the desired S/AS ratio (S/AS values ranging from 0.1 to 0.25). DiD was added into the polymer mix at a concentration of 50 μM, to label the particles fluorescently, by *in situ* encapsulation of the hydrophobic dye into the core of the nanoparticles during the process of nanoprecipitation. For each formulation, 0.5 mg of material was collected (for example, for 0.1 S/AS: 0.5 ml of 1 mg ml^{-1} nanoparticle). The nanoparticle solutions were diluted to a concentration of 1 mg ml^{-1} and the nanoparticle solutions were left on the shaker at room temperature overnight to allow evaporation of acetonitrile. All tubing (REF: BL-1815-04 & BL-PTFE-1602-20), fluidic connections (REF: CIL_XP-245X) and PDMS couplers (REF: PN-STN-20G-20) were purchased from Darwin Microfluidics (Paris, France). The Chemyx pump was actuated manually, using the touch screen, while the LSPOne pump was actuated using a custom-made MATLAB script.

Bulk formulation of uptake standard. Nanoparticle uptake standard was formulated by bulk nanoprecipitation. A polymer mixture of 10 mg ml^{-1} containing PLGA-PEG and 50 μM of DiD in pure acetonitrile was added dropwise to ultra-pure water, in a ratio of 1 : 10, at room temperature under stirring (700 rpm). The resulting nanoparticle solution (1 mg ml^{-1}) was left under stirring (400 rpm) on a shaker overnight at room temperature, protected from light, to let the acetonitrile evaporate.

Nanoparticle characterization

Bulk physicochemical characterization. Polydispersity index (PDI) and hydrodynamic diameter were determined by Dynamic Light Scattering (DLS) using a Zetasizer Nano-ZS (Malvern Panalytical), with a 633 nm laser and 173° Backscatter detector. Bulk fluorescence spectrum (Ex. 605 nm, Em. 646–700 nm, 5 nm step) was recorded for each nanoparticle batch using a BioTek Synergy H1 microplate reader (Agilent), in a black-well 96-well plate. Nanoparticle solutions were diluted in ultra-pure water (1 : 10) before measurement. Nanoparticle uptake standard was always included on the plate. Fluorescence coefficients to correct for differences in fluorescence intensity were calculated for each batch in comparison to the uptake standard.

Cell culturing and nanoparticle *in vitro* screening. Breast cancer epithelial cells MDA-MB-468 (HTB-132) were obtained from American Tissue Culture Collection (ATCC) and cultured under standard conditions (37 °C, 5% CO_2) in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% FBS (Fetal Bovine Serum) and 1% penicillin-streptavidin. Standard culture reagents (DMEM, FBS, pen-strep, DPBS 1x, EDTA–trypsin), nuclear (Hoechst 33342, Cat no. 62249) and membrane (Alexa Fluor™ 488 -Wheat germ agglutinin (WGA) conjugate, Cat no. W11261) stains, 16% methanol-free paraformaldehyde (PFA) (Cat no. 043368.9M) and human serum (MP Biomedicals™ Serum, Type AB, Cat no. 11425055) were purchased from Fisher Scientific.

Cells were seeded at a density of 25 000 cells per cm^2 in an "ibiTreat" μ-Plate 96 well back (Cat no. 89626) from IBIDI (Gräfelfing, Germany), cultured for 38–48 h before nanoparticle treatment. Half an hour before starting the treatment, nanoparticle stock solutions were pre-incubated with human serum (1 : 1, v : v) for 30 min at 37 °C. Cells were washed three times with serum-free phenol-free DMEM media, and each nanoparticle condition (pre-incubated with serum) was added to each well at a working concentration of 50 μg ml^{-1}. The resulting "incubation media" contained 5% human serum. After 23.5 h, the cells were counterstained with Hoechst and Alexa Fluor™ 488 WGA at 37 °C. After 24 h, cells were washed with serum-free media 3 times, fixed with PFA 2% (diluted in DPBS 1x), for 10 min, at room temperature. After fixation, cells were washed three times with DPBS 1x and stored at 4 °C protected from light until imaging.

High content imaging (HCI). Widefield fluorescence imaging was performed using a Nikon Eclipse Ti2 microscope, equipped with an automated focus system, an automated piezo stage, a 25 mm primΣ 95B sMOS camera from Teledyne photometrics (Arizona, USA) and a Spectra X light engine from Lumencor (Oregon, USA). The microscope was operated using Nikon Instrument Software (NIS) elements (v. 5.21.03). Pipeline for automated imaging of well-plates were set using Nikon's High-content dedicated macro (JOBS). For each condition, 10 Field of view (FOV), 16 bit images of 804 × 804 px (1.123 μm px^{-1}) in three channels (nuclei, membrane, nanoparticle), were recorded, using a 20x objective. The optical configuration was as follows: (1) for nuclei, laser excitation at 387 nm, DAPI filter cube (Ex: 379–450, Em: 414–480), 5% laser power, 10 ms; (2) for membrane, laser excitation at 470 nm, with a FITC filter cube (Ex: 461–488 – Em: 503–548), 20% laser power, 75 ms; (3) for nanoparticle, laser excitation at 628 nm, with Cy5 filter cube (Ex: 509–645, Em: 659–736), 40% laser power, 200 ms. To account for variability between days (including possible small variations on laser intensity), the same particle (uptake standard) was always included in the plate. Measurements were taken at room temperature.

HCI post-processing. Microscopy images (16 bit greyscale, 3 channels, .tiff) were batch processed using CellProfiler[32] (CP), version 4.2.1. The CP pipeline included segmentation of the cells and quantification of fluorescence signal from the nanoparticle channel. For this, nuclei segmentation ('IdentifyPrimaryObjects' module), followed by membrane segmentation ('IdentifySecondaryObjects' module) and cell identification ('IdentifyTertiaryObjects' module) was performed. Following segmentation, the 'MeasureObjectIntensity' module was used to compute the mean fluorescent signal per cell per unit of area; and the 'MeasureImageQuality' module was also used for quality control checks (see ESI†). The data was exported in .csv files, per single object (nuclei, cytoplasm, cell) and per image. MATLAB was used to calculate means and standard deviations of the features of interest.

Machine learning and computation

Design of experiments (DoE). The starting dataset (initial formulation runs, cycle 0) of PLGA-PEG nanoparticles was proposed using design of experiments (DoE), with Statgraphics Centurion 19. An augmented simplex lattice mixture design was created with one response variable (Uptake in MDA-468 cells), one process variable (S/AS with two levels), and 4 components (PLGA, PLGA-PEG, PLGA-PEG-COOH, PLGA-PEG-NH_2). In mixture designs, all components need to sum up to one (100% of the mixture) and can take up values from zero to one, unless stated otherwise. Process variables are discrete values. A linear model was selected for the process variable and a special cubic model for the mixtures, resulting in a design with 28 coefficients. Using the backward selection exchange algorithm implemented in the software, the number of runs was then set to 31. As a rule of thumb, the minimum number of runs to fit a model in DoE is the number of coefficients +3. Two runs were manually removed from the resulting dataset (100% PLGA, at 0.1 and 0.25 S/AS) since pure PLGA nanoparticles cannot form without the addition of any surfactant, and one run (F15) was manually added. The resulting list of initial formulations, including their physicochemical characterization is available in ESI Table 2.†

Nanoparticle uptake prediction. A Bayesian neural network was used, denoted as $p_\theta(y|x)$, to predict nanoparticle uptake (y) from nanoparticle formulation (x). The model parameters θ were initiated as probability distributions with Gaussian priors. To approximate the posterior distribution, stochastic variational inference (SVI) was used. Following the standard SVI approach,[41] a simpler guide model $q_\theta(x)$ was deployed that uses a multivariate normal distribution to approximate the true posterior distribution of p_θ. The model consisted of three hidden layers with ReLU activation functions. The model was trained with the ADAM optimizer,[42] to maximize the evidence lower bound (ELBO). The ELBO is calculated as follows (eqn (1)):

$$\mathrm{ELBO} = \mathbb{E}_{q_\theta(z|x)}[\log p_\theta(y|x,z)] - \mathrm{KL}(q_\theta(x)\|p_\theta) \tag{1}$$

where z represents the latent variables that aim to capture the data's structure, and KL is the Kullback-Leibler divergence[43] between the guide distribution and the true posterior distribution. To estimate the prediction uncertainty, the predictive distribution was constructed by taking 500 Monte Carlo samples for each data point. Nanoparticles with a PDI greater than 0.2 were excluded for training the uptake prediction models. Additionally, Gaussian processes, and ensembles of random forest or XGBoost models were used in preliminary uptake predictions. Ensembles of $n = 10$ models with different random seeds were used, where the standard deviation of the predictions served as a measure of prediction uncertainty.

Size and PDI prediction. Nanoparticle polydispersity index (PDI) and size (hydrodynamic diameter) predictions were performed with an Extreme Gradient Boosting[44] (XGBoost) model, which is based on decision trees. These models were trained on all available data for each design cycle. PDI and size values were log-transformed.

Model evaluation. Due to the low data setting, all available data was used to train the models in each cycle. Five-fold cross-validation[38] using 80/20 train/test splits was applied to ensure model robustness and stability. The model performance with cross-validation was computed *via* the root mean squared error (RMSE), calculated as follows (eqn (2)):

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

where $\hat{y}_i$ is the predicted value for the i-th nanoparticle (when it is not used to train the model) and y_i is its 'true' experimentally measured value (uptake, PDI, or size). To reduce the stochastic influence of random splitting, five-fold cross-validation is performed a total of three times with different random splits. Finally, the estimated model performance was calculated by taking the average RMSE over all splits.

Model training and optimization. Model hyper-parameters were optimized with cross-validation at each screening cycle, by choosing those leading to the lowest RMSE (eqn (2)). For Bayesian neural network models, we optimized the learning rate (values: 1×10^{-3}, 1×10^{-4}, 1×10^{-5}) and the number of neurons per hidden layer (16, 32, 64) using a grid search. For XGBoost models, Bayesian optimization was used to choose among 500 sets of hyperparameters. The specific sets of hyperparameters to try were selected by maximizing the expected improvement of a Gaussian Process estimator. The following hyperparameters were optimized: learning rate/eta = [0–1], maximal tree depth = [2–20], minimal child weight = [1–20], minimal loss split γ = [0–20], number of trees = [50–500], subsample ratio = [0.1–1], column subsample ratio by tree = [0.1–1], L2 regularization λ = [0–10], L1 regularization α = [1–10]. For random forest models (see ESI†), the same hyperparameter optimization was used using the following hyperparameters: number of trees = [50–500], maximal tree depth = [10–50], and minimal samples per split = [2–10]. Gaussian process models (see ESI†) were optimized using a grid search. The following hyperparameters were optimized: regularization parameter α (values: 10^{-3}, 10^{-2}, 10^{-1}, 1) and the kernel (radial basis function, length scale = 1 within the range [10^{-2}–10^{2}]; Matern kernel, length scale = 1 within the range [10^{-2}–10^{2}], smoothness = 1.5; rational quadratic kernel, length scale = 1 within the range [10^{-2}–10^{2}], α = 1; exponential sine squared kernel, length scale = 1 within the range [10^{-2}–10^{2}], periodicity = 3 within the range [10^{-2}–10]; dot product kernel, σ = 1 within the range [10^{-2}–10]). All kernels used a constant kernel multiplier bounded within the range [10^{-3}–10^{3}].

Experimental error calculation. The experimental error of the nanoparticle formulation was computed by considering the balance error (±0.01 mg), the volume-dependent systematic and random pipetting error (see ESI Table 7†), and a pump maximum dispensing error (1%). These values were taken from their corresponding manuals. The errors were considered as additive and independent and were calculated as the quadrature of the individual errors. The estimated errors were: 1.25% (PLGA), 1.21% (PP-L), 1.24% (PP-COOH), and 1.24% (PP-NH_2). Possible errors or variations in the flow rates of the solvent or antisolvent were not considered for this calculation, neither were errors introduced by dead volumes or by the carryover volume of the valve in the LSPOne pump.

Data augmentation based on experimental error. To artificially inflate the training data and simultaneously incorporate a notion of measurement error into the model, all training data was augmented 5 times (1*x* the original data + 4*x* augmented data), resulting in a slight increase in performance in preliminary experiments. Data augmentation was applied throughout cross-validation and model fitting. For each nanoparticle, PLGA, PLGA-PEG, PLGA-PEG-COOH, and PLGA-PEG-NH_2 values were multiplied with random samples from a normal distribution parameterized by their corresponding experimental error, as determined above. Similarly, nanoparticle uptake, PDI, and size were augmented using the standard deviation of the respective measurements.

Virtual screening library. A virtual library of nanoparticle formulations was generated to span the design space, by sampling PLGA, PLGA-PEG, PLGA-PEG-COOH, and PLGA-PEG-NH_2 composition ratios from a Dirichlet distribution (all variables range from zero to one, adding up to one). We considered sampled variables with values lower than 6% as 0%, taking the carryover error of the pump into account. These discarded values were added to another non-zero variable at random to enforce that all ratios still add up to one. Nanoparticle formulations that overlapped in experimental error or had a PLGA ratio higher than 0.7 were discarded. Finally, solvent/antisolvent ratios of [0.10, 0.15, 0.20, and 0.25] were sampled from a uniform distribution for each virtual nanoparticle. For practical reasons, a total of 100 000 virtual formulations were sampled from of the 1.85×10^8 theoretically possible formulations (see ESI†).

Formulation selection. At each screening cycle, uptake, PDI, and size were predicted for every formulation in the virtual screening library. To select cycle 1 formulations (exploration phase), we enforced diversity *via* clustering.[45] For a batch of $k = 10$ formulations, the subset of the 10% most uncertain predictions was selected. On this subset, K-means clustering with Euclidean distance was performed for *k* clusters. The closest formulation to each cluster centroid was then selected as the formulations to produce. In cycle 2 (exploitation phase), for a batch of $k = 10$ formulations, the formulations from the virtual screening library with the top 10% highest predicted uptake were selected. From this subset, the *k* most certain samples were selected. The same strategy was employed for both high-uptake and low-uptake predictions for formulation-function elucidation. However, for the low-uptake predictions, the 10% lowest predicted uptake particles were selected instead.

Humans-in-the-loop. Nanoparticle production, cell imaging, image analysis, training of machine learning models, and next formulation selection were all automated. However, several handling procedures were done manually. For instance, collecting nanoparticles from the microfluidic device was done by hand, as well as bulk physicochemical analysis (dynamic light scattering) for quality control, and cell seeding and treatment. Human intervention between all automated steps is still required in our setup (*e.g.*, physically moving samples or deciding on the number of active learning cycles) but could in principle be fully automated.

Software and code. All code was implemented in Python (v. 3.9.15). The Bayesian neural network model was implemented using the Python packages PyTorch (v. 1.12.1)[46] and Pyro (v. 1.8.4).[47] XGBoost models were implemented using sklearn (v. 1.2.1)[48] and xgboost (v. 1.7.3)[44] Python libraries. Graphs and figures were made in R (v. 4.3.0)[49] using ggplot2 (v. 3.4.2),[50] Adobe illustrator, and Biorender.com (academic license). Comparisons between means were performed using a standard two-tailed *t*-test in R and the resulting *p*-values are reported in text.

Data availability

All code, trained models, and results are available on GitHub: https://github.com/molML/Nano_Particles_Active_Learning. All raw screening data are available on: https://zenodo.org/records/8289605.

Author contributions

Conceptualization: AOP, DvT, LA, FG; methodology: AOP, DvT, LA, FG; experiments (wet-lab): AOP; experiments (computational): DvT; formal analysis and investigation: AOP, DvT, with contributions from all authors; writing – original draft: AOP, DvT; writing – review & editing: all authors.

Conflicts of interest

The authors declare no conflict of interest.

Acknowledgements

We would like to thank Emiel Visser for providing the MATLAB script to automate the LSPOne pump and for an Eppendorf holder CAD design. We would also like to thank Stijn Haenen for 3D printing the Eppendorf holders. AOP and LA are supported by NWO through a Vidi Grant (192.028). RvdM is supported by a Vidi grant (19861) from the Dutch Research Council (NWO). FG acknowledges the support of the Centre for Living Technologies.

References

1 B. Pelaz, C. Alexiou, R. A. Alvarez-Puebla, F. Alves, A. M. Andrews, S. Ashraf, L. P. Balogh, L. Ballerini, A. Bestetti, C. Brendel, S. Bosi, M. Carril, W. C. W. Chan, C. Chen, X. Chen, X. Chen, Z. Cheng, D. Cui, J. Du, C. Dullin, A. Escudero, N. Feliu, M. Gao, M. George, Y. Gogotsi, A. Grünweller, Z. Gu, N. J. Halas, N. Hampp, R. K. Hartmann, M. C. Hersam, P. Hunziker, J. Jian, X. Jiang, P. Jungebluth, P. Kadhiresan, K. Kataoka, A. Khademhosseini, J. Kopeček, N. A. Kotov, H. F. Krug, D. S. Lee, C.-M. Lehr, K. W. Leong, X.-J. Liang, M. Ling Lim, L. M. Liz-Marzán, X. Ma, P. Macchiarini, H. Meng, H. Möhwald, P. Mulvaney, A. E. Nel, S. Nie, P. Nordlander,

T. Okano, J. Oliveira, T. H. Park, R. M. Penner, M. Prato, V. Puntes, V. M. Rotello, A. Samarakoon, R. E. Schaak, Y. Shen, S. Sjöqvist, A. G. Skirtach, M. G. Soliman, M. M. Stevens, H.-W. Sung, B. Z. Tang, R. Tietze, B. N. Udugama, J. S. VanEpps, T. Weil, P. S. Weiss, I. Willner, Y. Wu, L. Yang, Z. Yue, Q. Zhang, Q. Zhang, X.-E. Zhang, Y. Zhao, X. Zhou and W. J. Parak, Diverse Applications of Nanomedicine, *ACS Nano*, 2017, **11**(3), 2313–2381, DOI: 10.1021/acsnano.6b06040.

2 M. Murar, L. Albertazzi and S. Pujals, Advanced Optical Imaging-Guided Nanotheranostics towards Personalized Cancer Drug Delivery, *Nanomaterials*, 2022, **12**(3), 399, DOI: 10.3390/nano12030399.

3 C. Feng, Y. Li, B. E. Ferdows, D. N. Patel, J. Ouyang, Z. Tang, N. Kong, E. Chen and W. Tao, Emerging Vaccine Nanotechnology: From Defense against Infection to Sniping Cancer, *Acta Pharm. Sin. B*, 2022, **12**(5), 2206–2223, DOI: 10.1016/j.apsb.2021.12.021.

4 J. Nam, S. Son, K. S. Park, W. Zou, L. D. Shea and J. J. Moon, Cancer Nanomedicine for Combination Cancer Immunotherapy, *Nat. Rev. Mater.*, 2019, **4**(6), 398–414, DOI: 10.1038/s41578-019-0108-1.

5 M. J. Mitchell, M. M. Billingsley, R. M. Haley, M. E. Wechsler, N. A. Peppas and R. Langer, Engineering Precision Nanoparticles for Drug Delivery, *Nat. Rev. Drug Discovery*, 2021, **20**(2), 101–124, DOI: 10.1038/s41573-020-0090-8.

6 R. T. Stiepel, E. Duggan, C. J. Batty and K. M. Ainslie, Micro and Nanotechnologies: The Little Formulations That Could, *Bioeng. Transl. Med.*, 2023, **8**(2), e10421, DOI: 10.1002/btm2.10421.

7 G. Yamankurt, E. J. Berns, A. Xue, A. Lee, N. Bagheri, M. Mrksich and C. A. Mirkin, Exploration of the Nanomedicine-Design Space with High-Throughput Screening and Machine Learning, *Nat. Biomed. Eng.*, 2019, **3**(4), 318–327, DOI: 10.1038/s41551-019-0351-1.

8 P. M. Valencia, E. M. Pridgen, M. Rhee, R. Langer, O. C. Farokhzad and R. Karnik, Microfluidic Platform for Combinatorial Synthesis and Optimization of Targeted Nanoparticles for Cancer Therapy, *ACS Nano*, 2013, **7**(12), 10671–10680, DOI: 10.1021/nn403370e.

9 E. Blanco, H. Shen and M. Ferrari, Principles of Nanoparticle Design for Overcoming Biological Barriers to Drug Delivery, *Nat. Biotechnol.*, 2015, **33**(9), 941–951, DOI: 10.1038/nbt.3330.

10 W. Poon, B. R. Kingston, B. Ouyang, W. Ngo and W. C. W. Chan, A Framework for Designing Delivery Systems, *Nat. Nanotechnol.*, 2020, **15**(10), 819–829, DOI: 10.1038/s41565-020-0759-5.

11 P. M. Valencia, O. C. Farokhzad, R. Karnik and R. Langer, Microfluidic Technologies for Accelerating the Clinical Translation of Nanoparticles, *Nat. Nanotechnol.*, 2012, **7**(10), 623–629, DOI: 10.1038/nnano.2012.168.

12 S. J. Shepherd, D. Issadore and M. J. Mitchell, Microfluidic Formulation of Nanoparticles for Biomedical Applications, *Biomaterials*, 2021, **274**, 120826, DOI: 10.1016/j.biomaterials.2021.120826.

13 Y. Liu, G. Yang, D. Zou, Y. Hui, K. Nigam, A. P. J. Middelberg and C.-X. Zhao, Formulation of Nanoparticles Using Mixing-Induced Nanoprecipitation for Drug Delivery, *Ind. Eng. Chem. Res.*, 2020, **59**(9), 4134–4149, DOI: 10.1021/acs.iecr.9b04747.

14 R. Haase, E. Fazeli, D. Legland, M. Doube, S. Culley, I. Belevich, E. Jokitalo, M. Schorb, A. Klemm and C. Tischer, A Hitchhiker's Guide through the Bio-Image Analysis Software Universe, *FEBS Lett.*, 2022, **596**(19), 2472–2485, DOI: 10.1002/1873-3468.14451.

15 M. Mattiazzi Usaj, E. B. Styles, A. J. Verster, H. Friesen, C. Boone and B. J. Andrews, High-Content Screening for Quantitative Cell Biology, *Trends Cell Biol.*, 2016, **26**(8), 598–611, DOI: 10.1016/j.tcb.2016.03.008.

16 D. J. Brayden, S.-A. Cryan, K. A. Dawson, P. J. O'Brien and J. C. Simpson, High-Content Analysis for Drug Delivery and Nanoparticle Applications, *Drug Discovery Today*, 2015, **20**(8), 942–957, DOI: 10.1016/j.drudis.2015.04.001.

17 B. Yang, C. J. Richards, T. B. Gandek, I. de Boer, I. Aguirre-Zuazo, E. Niemeijer and C. Åberg, Following Nanoparticle Uptake by Cells Using High-Throughput Microscopy and the Deep-Learning Based Cell Identification Algorithm Cellpose, *Front. nanotechnol.*, 2023, **5**, DOI: 10.3389/fnano.2023.1181362.

18 M. B. Cutrona and J. C. Simpson, A High-Throughput Automated Confocal Microscopy Platform for Quantitative Phenotyping of Nanoparticle Uptake and Transport in Spheroids, *Small*, 2019, **15**(37), 1902033, DOI: 10.1002/smll.201902033.

19 Y. Rui, D. R. Wilson, S. Y. Tzeng, H. M. Yamagata, D. Sudhakar, M. Conge, C. A. Berlinicke, D. J. Zack, A. Tuesca and J. J. Green, High-Throughput and High-Content Bioassay Enables Tuning of Polyester Nanoparticles for Cellular Uptake, Endosomal Escape, and Systemic in Vivo Delivery of mRNA, *Sci. Adv.*, 2022, **8**(1), DOI: 10.1126/sciadv.abk2855.

20 S. Kelly, M. H. Byrne, S. J. Quinn and J. C. Simpson, Multiparametric Nanoparticle-Induced Toxicity Readouts with Single Cell Resolution in HepG2 Multicellular Tumour Spheroids, *Nanoscale*, 2021, **13**(41), 17615–17628, DOI: 10.1039/D1NR04460E.

21 X. Chen and H. Lv, Intelligent Control of Nanoparticle Synthesis on Microfluidic Chips with Machine Learning, *NPG Asia Mater.*, 2022, **14**(1), 69, DOI: 10.1038/s41427-022-00416-1.

22 H. Tao, T. Wu, M. Aldeghi, T. C. Wu, A. Aspuru-Guzik and E. Kumacheva, Nanoparticle Synthesis Assisted by Machine Learning, *Nat. Rev. Mater.*, 2021, **6**(8), 701–716, DOI: 10.1038/s41578-021-00337-5.

23 F. Mekki-Berrada, Z. Ren, T. Huang, W. K. Wong, F. Zheng, J. Xie, I. P. S. Tian, S. Jayavelu, Z. Mahfoud, D. Bash, K. Hippalgaonkar, S. Khan, T. Buonassisi, Q. Li and X. Wang, Two-Step Machine Learning Enables Optimized Nanoparticle Synthesis, *npj Comput. Mater.*, 2021, **7**(1), 55, DOI: 10.1038/s41524-021-00520-w.

24 K. Abdel-Latif, R. W. Epps, F. Bateni, S. Han, K. G. Reyes and M. Abolhasani, Self-Driven Multistep Quantum Dot

Synthesis Enabled by Autonomous Robotic Experimentation in Flow, *Adv. Intell. Syst.*, 2021, **3**(2), 2000245, DOI: 10.1002/aisy.202000245.

25 O. Voznyy, L. Levina, J. Z. Fan, M. Askerka, A. Jain, M.-J. Choi, O. Ouellette, P. Todorović, L. K. Sagar and E. H. Sargent, Machine Learning Accelerates Discovery of Optimal Colloidal Quantum Dot Synthesis, *ACS Nano*, 2019, **13**(10), 11122–11128, DOI: 10.1021/acsnano.9b03864.

26 D. Van Tilborg, H. Brinkmann, E. Criscuolo, L. Rossen, R. Özçelik and F. Grisoni, Deep Learning for Low-Data Drug Discovery: Hurdles and Opportunities, preprint, *Chemrxiv*, 2024, DOI: 10.26434/chemrxiv-2024-w0wvl.

27 Z. Bao, J. Bufton, R. J. Hickman, A. Aspuru-Guzik, P. Bannigan and C. Allen, Revolutionizing Drug Formulation Development: The Increasing Impact of Machine Learning, *Adv. Drug Delivery Rev.*, 2023, **202**, 115108, DOI: 10.1016/j.addr.2023.115108.

28 A. Krause, A. Singh and C. Guestrin, Near-Optimal Sensor Placements in Gaussian Processes: Theory, Efficient Algorithms and Empirical Studies, *J. Mach. Learn. Res.*, 2008, **9**(2), 235–284.

29 D. Reker and G. Schneider, Active-Learning Strategies in Computer-Assisted Drug Discovery, *Drug Discovery Today*, 2015, **20**(4), 458–465, DOI: 10.1016/j.drudis.2014.12.004.

30 R. Karnik, F. Gu, P. Basto, C. Cannizzaro, L. Dean, W. Kyei-Manu, R. Langer and O. C. Farokhzad, Microfluidic Platform for Controlled Synthesis of Polymeric Nanoparticles, *Nano Lett.*, 2008, **8**(9), 2906–2912, DOI: 10.1021/nl801736q.

31 A. G. Mares, G. Pacassoni, J. S. Marti, S. Pujals and L. Albertazzi, Formulation of Tunable Size PLGA-PEG Nanoparticles for Drug Delivery Using Microfluidic Technology, *PLoS One*, 2021, **16**(6), e0251821, DOI: 10.1371/journal.pone.0251821.

32 C. McQuin, A. Goodman, V. Chernyshev, L. Kamentsky, B. A. Cimini, K. W. Karhohs, M. Doan, L. Ding, S. M. Rafelski, D. Thirstrup, W. Wiegraebe, S. Singh, T. Becker, J. C. Caicedo and A. E. Carpenter, CellProfiler 3.0: Next-Generation Image Processing for Biology, *PLoS Biol.*, 2018, **16**(7), e2005970, DOI: 10.1371/journal.pbio.2005970.

33 A. Alijagic, N. Scherbak, O. Kotlyar, P. Karlsson, X. Wang, I. Odnevall, O. Benada, A. Amiryousefi, L. Andersson, A. Persson, J. Felth, H. Andersson, M. Larsson, A. Hedbrant, S. Salihovic, T. Hyötyläinen, D. Repsilber, E. Särndahl and M. Engwall, A Novel Nanosafety Approach Using Cell Painting, Metabolomics, and Lipidomics Captures the Cellular and Molecular Phenotypes Induced by the Unintentionally Formed Metal-Based (Nano) Particles, *Cells*, 2023, **12**(2), 281, DOI: 10.3390/cells12020281.

34 B. Settles, *Active Learning Literature Survey; Technical Report*, University of Wisconsin-Madison Department of Computer Sciences, 2009, https://minds.wisconsin.edu/handle/1793/60660 (accessed 2023-08-04).

35 A. Graves, Practical Variational Inference for Neural Networks, In *Advances in Neural Information Processing Systems*, ed. J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira and K. Q. Weinberger, Curran Associates, Inc., 2011, vol. 24.

36 Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan and J. Snoek, Can You Trust Your Model' s Uncertainty? Evaluating Predictive Uncertainty under Dataset Shift, In *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2019, vol. 32.

37 A. Wu, S. Nowozin, E. Meeds, R. E. Turner, J. M. Hernández-Lobato and A. L. Gaunt, *Deterministic Variational Inference for Robust Bayesian Neural Networks*, 2019.

38 T. A. Meyer, C. Ramirez, M. J. Tamasi and A. J. Gormley, A User's Guide to Machine Learning for Polymeric Biomaterials, *ACS Polym. Au*, 2023, **3**(2), 141–157, DOI: 10.1021/acspolymersau.2c00037.

39 Z. Bao, F. Yung, R. J. Hickman, A. Aspuru-Guzik, P. Bannigan and C. Allen, Data-Driven Development of an Oral Lipid-Based Nanoparticle Formulation of a Hydrophobic Drug, *Drug Delivery Transl. Res.*, 2023, DOI: 10.1007/s13346-023-01491-9.

40 J.-M. Rabanel, V. Adibnia, S. F. Tehrani, S. Sanche, P. Hildgen, X. Banquy and C. Ramassamy, Nanoparticle Heterogeneity: An Emerging Structural Parameter Influencing Particle Fate in Biological Media?, *Nanoscale*, 2019, **11**(2), 383–406, DOI: 10.1039/C8NR04916E.

41 L. V. Jospin, H. Laga, F. Boussaid, W. Buntine and M. Bennamoun, Hands-On Bayesian Neural Networks—A Tutorial for Deep Learning Users, *IEEE Comput. Intell. Mag.*, 2022, **17**(2), 29–48, DOI: 10.1109/MCI.2022.3155327.

42 D. P. Kingma and J. Ba, A Method for Stochastic Optimization, *arXiv*, Preprint, 2014, arXiv:1412.6980, DOI: 10.48550/ARXIV.1412.6980.

43 C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

44 T. Chen and C. Guestrin, XGBoost: A Scalable Tree Boosting System, In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, San Francisco California USA, 2016, pp. 785–794, doi: DOI: 10.1145/2939672.2939785.

45 F. Zhdanov, Diverse Mini-Batch Active Learning, *arXiv*, Preprint, 2019, arXiv:1901.05954v1, DOI: 10.48550/ARXIV.1901.05954.

46 A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai and S. Chintala, PyTorch: An Imperative Style, High-Performance Deep Learning Library, In *Advances in Neural Information Processing Systems*, ed. H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox and R. Garnett, Curran Associates, Inc., 2019, vol. 32.

47 E. Bingham, J. P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, T. Karaletsos, R. Singh, P. Szerlip, P. Horsfall and N. D. P. Goodman, Deep Universal Probabilistic Programming, *J. Mach. Learn. Res.*, 2019, 1–6.

48 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman,

G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, *Scikit-Learn: Machine Learning in Python, arXiv*, Preprint, 2012, arXiv:1201.0490v4, DOI: **10.48550/ARXIV.1201.0490**.

49 Team, *R. D. C. R.: A Language and Environment for Statistical Computing*, 2010.

50 H. Wickham, Data Analysis, In *ggplot2: Elegant Graphics for Data Analysis*, ed. H. Wickham, Use R.!, Springer International Publishing, Cham, 2016, pp. 189–201, DOI: **10.1007/978-3-319-24277-4_9**.