Nanoscale

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Nanoscale*, 2025, **17**, 19767

Data-driven optimization of nanoparticle size using the prediction reliability enhancing parameter (PREP)

Seyed Saeid Tayebi, Nate Dowdall, Todd Hoare * and Prashant Mhaskar*

The particle size of a nanoparticle plays a crucial role in regulating its biodistribution, cellular uptake, and transport mechanisms and thus its therapeutic efficacy. However, experimental methods for achieving a desired nanoparticle size and size distribution often require numerous iterations that are both time-consuming and costly. In this study, we address the critical challenge of achieving nanoparticle size control by implementing the Prediction Reliability Enhancing Parameter (PREP), a recently developed data-driven modeling-based product design approach that significantly reduces the number of experimental iterations needed to meet specific design goals. We applied PREP to effectively predict and control particle sizes of two distinct nanoparticle types with different target particle size properties: (1) thermoresponsive covalently-crosslinked microgels fabricated *via* precipitation polymerization with targeted temperature-dependent size properties and (2) physical polyelectrolyte complexes fabricated *via* charge-driven self-assembly with particle sizes and colloidal stabilities suitable for effective circulation. In both cases, PREP enabled efficient and precise size control, achieving target outcomes in only two iterations in each case. These results provide motivation to further utilize PREP in streamlining experimental workflows in various biomaterials optimization challenges.

Received 23rd April 2025,
Accepted 9th August 2025

DOI: 10.1039/d5nr01664a

rsc.li/nanoscale

1. Introduction

Polymer-based nanoparticles have attracted increasing interest in drug delivery and other biomedical applications due to their capacity to encapsulate therapeutic agents, facilitate long-term circulation, traverse tissue barriers, interact with cell surface receptors, and facilitate the delivery of drugs directly into target cells.[1] These features have been leveraged for a range of therapeutic applications including transporting chemotherapeutics to both primary and metastatic cancer sites,[2,3] delivering imaging agents specifically to cells or tissues to aid in accurate disease diagnosis,[4,5] facilitating gene delivery,[4,6] and providing preventative treatments for infectious diseases.[7,8]

The success of each of these applications depends strongly on the size of the nanoparticle,[9] which regulates both the convective transport of nanoparticles due to blood shear and variations in interstitial pressure as well as the potential for nanoparticles to interact with active and passive transport pathways that enable intracellular transport and/or transport across biological barriers such as the blood–brain barrier.[6,10–16] In response, significant effort has been invested in developing strategies to synthesize nanoparticles with precise and uniform sizes across different particle size ranges suitable for different biomedical transport tasks.[1,2,10,14,17,18] Such efforts can be broadly classified into two categories: (1) the assembly of pre-fabricated polymers into particles and (2) the direct synthesis of nanoparticles from monomeric building blocks. In the former case, techniques such as self-assembly, triggered precipitation, and template-assisted synthesis are commonly employed due to their ability to produce nanoparticles with well-defined characteristics.[19–23] Self-assembly, for instance, relies on the spontaneous organization of polymeric building blocks through secondary intermolecular interactions like hydrophobic interactions, hydrogen bonding, electrostatic forces, and π–π stacking, with particle size control enabled by rational tuning of the composition of the building blocks and the solution conditions used.[19,20] However, the inherent dispersity in size and composition among the typical polymeric building blocks for self-assembled nanoparticles can lead to broad particle size distributions, multiple particle populations, and/or the potential for aggregation. In the latter case, emulsion, precipitation, and/or suspension polymerization methods can all be applied to achieve particle size control, with the combination of such templating methods with controlled free radical polymerization strategies (*e.g.* atom transfer

Department of Chemical Engineering, McMaster University, 1280 Main St. W., Hamilton, Ontario, Canada L8S 4L7. E-mail: hoaretr@mcmaster.ca, mhaskar@mcmaster.ca

radical polymerization in emulsion polymerization) particularly beneficial to produce nanoparticles with tunable sizes.[17,18] However, factors such as the variability of the local shear field, variable particle aggregation/nucleation, variability in surfactant or other surface stabilizer performance under different environmental/solvent conditions, and/or localized temperature gradients can result in poor control over nanoparticle size and polydispersity, particularly for methods that do not rely on more complex polymerization pathways and are thus more amenable to practical translation.

Solving these size and stability challenges is challenging based on the frequent interdependence of the key factors that regulate such properties; for example, adjusting one parameter such as monomer concentration, surfactant type/concentration, or reaction temperature can affect polymerization and/or assembly kinetics, the stability of the nanoparticle/solvent interface, and/or particle nucleation kinetics in sometimes unanticipated ways. This interconnectedness makes relying solely on experimental techniques for nanoparticle size optimization both time-consuming and costly, especially without a strategic framework to guide the process.[24–27] In this context, incorporating model-based design techniques that can capture underlying patterns and relationships within the synthesis process offer significant promise to accelerate nanoparticle design. By leveraging model-based computational tools, researchers can plan experimental iterations more efficiently, reducing resource consumption and expediting the development of nanoparticles with desired characteristics.

Modeling approaches for optimizing nanoparticle size can be broadly classified into deterministic and data-driven models. Deterministic models leverage fundamental principles to describe system behavior, offering detailed insights into mechanisms like particle growth and nucleation. Studies have demonstrated the utility of deterministic models in solving reaction–diffusion equations and predicting size distributions under varying conditions.[28–35] However, these models require extensive computational resources, detailed mechanistic knowledge (including measurement of several often hard-to-measure or estimate rate or interaction parameters), and costly validation, making them less practical for complex systems. In contrast, data-driven models bypass the need for detailed mechanistic understanding by uncovering patterns directly from experimental data. These models have been widely used to predict nanoparticle properties such as size and morphology by correlating recipe parameters with outcomes[24,26,36] and have been particularly leveraged in polymerization-based processes to establish correlations between recipe parameters and final nanoparticle size, facilitating predictive particle size control while accounting for radical polymerization kinetics, diffusion rates, and interaction dynamics.[1,27,29,33,36–38]

Among various data-driven modeling techniques such as neural networks and advanced nonlinear regression models,[24,25,27,33] latent variable models (LVM) such as Principal Component Analysis (PCA) and Partial Least Square-Projection to Latent Structure (PLS) have garnered significant attention for their ability to identify a reduced set of latent variables—underlying patterns or structures—that explain most of the system's variability.[39–42] While effective, these methods also pose drawbacks in the context of nanoparticle size optimization given their typical need for large datasets and prediction uncertainty when applied to new data points. Existing literature has proposed uncertainty metrics including Hotelling's T^2 and Squared Prediction Errors (SPE) to address these limitations.[43–50] While these metrics assess the alignment of new data points with the calibration dataset, their interpretations can vary depending on the specific metric used. Recently, we introduced the Prediction Reliability Enhancing Parameter (PREP), a unified metric that enhances predictive reliability by combining multiple model alignment metrics, to address this prediction uncertainty challenge. The PREP method was validated on synthetic datasets and shown to outperform existing methods to identify optimum inputs to achieve target outputs, particularly in cases in which the optimal solution is outside the design space of the original dataset.[51] However, to-date the method has not been validated on an experimental use case.

Herein, we apply the PREP method to optimize nanoparticle size and nanoparticle size distributions in one polymerization-based nanoparticle synthesis use case (the synthesis of dual temperature/pH responsive microgels based on poly(*N*-isopropylacrylamide) (PNIPAM) *via* precipitation polymerization) and in one self-assembly-based nanoparticle synthesis use case (the fabrication of doxorubicin-loaded polyelectrolyte complexes based on sulfated yeast beta glucan and cationic dextran). The first case builds on previous literature from our group and our previous data-driven modeling efforts to optimize the size and colloidal stability of acid-functionalized PNIPAM microgels that have broad utility for drug delivery given their potential for environmentally-responsive reversible swelling responses, their capacity to deform and thus enhance penetration through biological barriers, and their highly hydrated surface properties that can suppress immune system recognition.[40,52–54] The specific target was to match the crosslinking density and the acid content (4–8 mol%) to microgels in the existing dataset while achieving smaller particle sizes that remain stable over time. Specifically, while the pre-existing data set did not include a microgel with a size less than 170 nm that met the crosslink density and acid content criteria, a size of 100 nm was targeted to better exploit the biological penetration properties of the compressible microgels for drug delivery applications. The second case targeted a key challenge around the ionic strength tolerance of polyelectrolyte complexes, which are typically fabricated in water or low ionic strength buffers but often lose colloidal stability when then transferred to the physiological ionic strength conditions typically required for practical clinical use. The specific target was to achieve nanoparticles with diameter <200 nm (target = 170 nm) and a polydispersity index (PDI) as low as possible (target = 0.15), properties most suitable for long-term circulation, that remained colloidally stable under physiological ionic strength. We demonstrate that in both cases the PREP method can achieve the target properties with minimal historical data

following only two iterations, opening the potential to apply PREP more broadly to address nanoparticle design challenges.

2. Preliminaries

2.1 Latent variable models (LVM)

Ordinary least squares (OLS) regression assumes that system outputs are independent; however, this assumption frequently breaks down in real-world industrial applications—such as nanoparticle size control—where variables are inherently interdependent, often resulting in poor model performance. In contrast, latent variable modeling (LVM), while also a linear modeling approach, is well-suited for capturing complex interdependencies by isolating the core independent structures within the dataset. By identifying and operating within an uncorrelated latent space, LVM establishes meaningful connections between system inputs and outputs, particularly in scenarios where data is limited but intervariable dependencies are critical to capture.

Specifically, LVM can either (1) extract correlations within a single block of data—*via* Principal Component Analysis (PCA)—and project the original correlated data into a latent uncorrelated space (referred to as scores) or (2) define relationships between input variables (X) and output variables (Y) by jointly mapping them onto a latent space. In both cases, the resulting scores are represented as linear combinations of the original variables that are orthogonal to one another. The general structure of LVM is illustrated in Fig. 1; for detailed mathematical formulations, and data-blocking configurations, the reader is referred to our prior manuscript.[51]

2.2 Latent variable model inversion (LVMI)

The primary objective of modeling is typically to identify a suitable set of input values that lead to a predetermined set of desired output properties, referred to as $Y_{desirable}$. This process is known as model inversion, and within the framework of LVM it is termed latent variable modeling inversion (LVMI). The outcomes of model inversion depend on the relationship between the number of underlying independent latent factors in the input space (A)—the number of underlying independent factors (or latent variables) driving the input space, rather than merely the number of independent input variables —and the number of output variables (K):

1. If $A < K$, there is no input set X for which $Y_{predicted} = Y_{desirable}$. In this case, model inversion identifies an input X where its $Y_{predicted}$ is as close as possible to $Y_{desirable}$.
2. If $A = K$, there is a single solution for which its $Y_{predicted} = Y_{desirable}$ that can be identified by model inversion.
3. If $A > K$ (the most common case in practice), there are an infinite number of input sets X for which $Y_{predicted} = Y_{desirable}$. In this context, these solutions form a continuous set known as the Null Space (NS) that represents various input combinations that leave the output prediction unchanged.

Solutions derived from LVMI can either match the targeted predetermined value (as in the second and third scenarios) or come as close as possible to the predetermined value (as in the first scenario). While the prediction accuracy for these solutions varies across different samples, the degree of accuracy cannot be confirmed until all the solutions are experimentally tested, which can be a costly and time-consuming process. To address this issue, specific modeling alignment metrics can be computed solely from the input data (X), metrics that are generally classified into three categories:

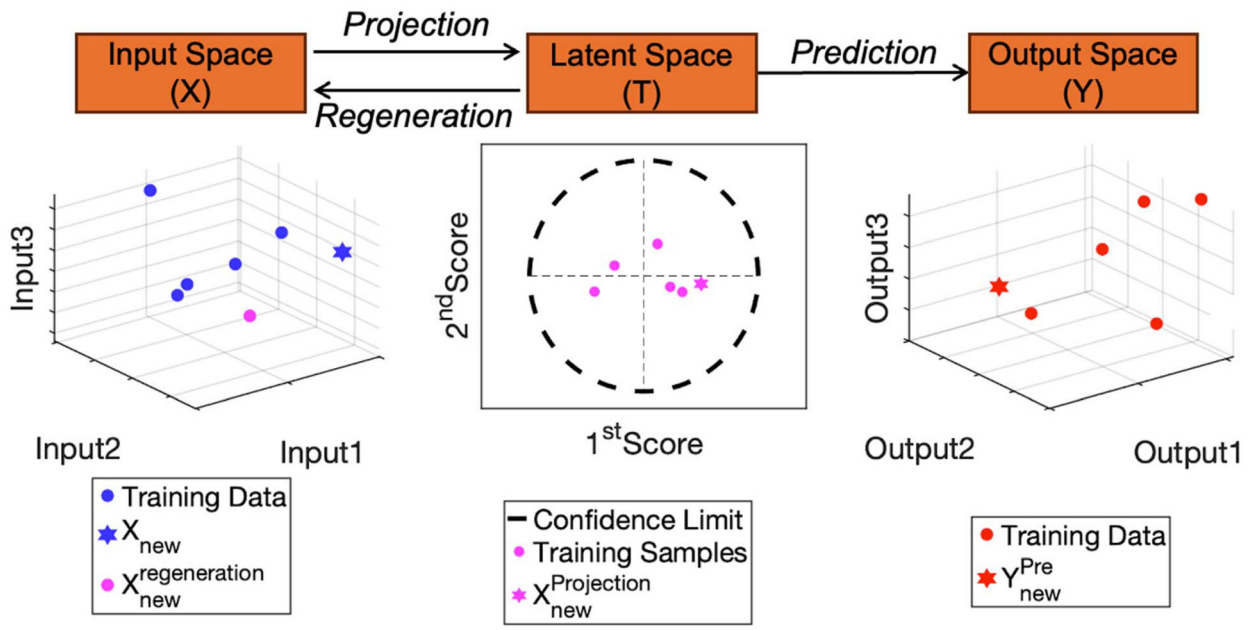

Fig. 1 General latent variable modeling framework.

(a) Hotelling's T^2 metrics measure the distance of a new data point's projection to the latent space from the center of the latent space, indicating how far the new data point deviates from the calibration set.

(b) Squared prediction error (SPE) metrics assess how well the new data point can be reconstructed or regenerated by the model.

(c) Score alignment (H_{PLS} & H_{PCA}) metrics evaluate the similarity of the score structure of the new data point to that of the calibration data, indicating how closely the new sample aligns with the model's learned structure.

Fig. 1 also provides a conceptual summary of the Hotelling T^2 and SPE metrics in which the SPE corresponds to the distance between the $X_{new}^{regenerated}$ and X_{new} in the input space (reflecting how well the model can reconstruct the new sample) and the Hotelling T^2 metric reflects the distance between the latent projection of the new sample and the center of the latent space (capturing how far the sample deviates from the distribution of the calibration set). For the Score Alignment metric (H), when a new sample is projected into a less populated region of the latent space, it reflects a lower resemblance to the calibration data point score structure, resulting in a higher H score (and *vice versa*).

3. Proposed methodology

Although each of the above-mentioned metrics has its own general threshold beyond which model predictions are unlikely to be accurate, there is no single threshold across all metrics that can define a universally reliable range for predictions and thus determine when model predictions can be trusted. Additionally, different expectations may arise depending on which metric is being considered. To address this limitation, the PREP parameter is defined as a linear combination of the metrics, weighted by different coefficients and powers, that are optimized using a validation dataset in which both actual and predicted Y values are available for comparison. The parameters are optimized such that samples with low prediction accuracy are assigned a higher PREP value while samples with higher prediction accuracy are assigned a lower PREP value, allowing the list of potential candidates coming from LVMI to be ranked based on their likelihood of accurate predictions and thus enabling prioritization of those samples that either have the highest chance of success in meeting the target properties or will provide the model with the most new information possible for further model refinement. The general equation for PREP is presented in eqn (1),[51] in which the values of C and P are determined specifically for each dataset through an optimization algorithm.

$$\begin{aligned}\mathrm{PREP} = {} & C_1 \mathrm{hoteling}T2_{\mathrm{pls}}^{P1} + C_2 \mathrm{SPE}_{x.\mathrm{pls}}^{P2} + C_3 \mathrm{hoteling}T2_{\mathrm{pca}}^{P3} \\ & + C_4 \mathrm{SPE}_{\mathrm{pca}}^{P4} + C_5 h_{\mathrm{pls}}^{P5} + C_6 h_{\mathrm{pca}}^{P6}\end{aligned} \quad (1)$$

To implement the PREP method, an initial dataset and a desired target output set are chosen and the k-nearest neighbors (with k being a tuning parameter) to the target output in the output space are identified and used to train both a PLS and a PCA model. The PLS model generates a list of potential design space (PDS) candidates comprised of candidate recipes expected to meet the target output. Model alignment metrics are subsequently calculated for the training data alongside the prediction accuracy, using a jackknife approach in which the PLS model is developed using a subset of the samples and the predicted output is compared to the actual value(s) of the excluded sample(s). The alignment metrics and prediction accuracy of the training dataset are then used to optimize the coefficients and powers of the PREP equation (C and P in eqn (1)), enabling the ranking of PDS samples by assigning a score to each candidate based on its likelihood of accurate prediction. Candidates with the lowest PREP score (indicating high prediction confidence) and the highest PREP score (representing high uncertainty, which can aid model refinement near the target output) are selected for synthesis. If the synthesized samples do not achieve the target, they are added to the dataset, the list of k-nearest neighbors is updated, and the process is repeated iteratively until the desired outcome is obtained. Fig. 2 illustrates the general scheme of the method, with further details available in the original paper.[51]

The PREP method has two key advantages relative to previous methods for assessing prediction accuracy: (1) only a single parameter needs to be evaluated to compare samples, reducing uncertainty and bias in prediction assessment; and (2) the method does not require a large number of data points for practical implementation, with as few as $A + 2$ data points needed in which A represents the number of independent principal components of the system input. Note that while Bayesian and Gaussian process-based approaches can also be applied effectively to similar optimization challenges, they tend to rely on more sample-intensive strategies (*e.g.*, Monte Carlo sampling) and thus often require significantly more data to achieve convergence relative to the PREP method, particularly in complex or high-dimensional settings.[51] Relative to non-linear modeling approaches such as support vector regression, decision trees, and Gaussian process regression that have also performed well for predicting materials properties using relatively smaller sample sizes, PREP offers a key advantage in that it is fundamentally a linear latent variable-based framework, thus reducing the risk of overfitting, making interpretability simpler, and facilitating more robust extrapolation along well-defined latent variable directions (the latter of which is particularly beneficial for inverse design).

4. Experimental case studies

To validate the performance of the PREP method for optimizing and controlling nanoparticle sizes and size distributions, two case studies were performed.

4.1 Case study 1: multi-responsive microgels

Smart microgels that respond to external stimuli such as pH and temperature are typically fabricated *via* a free radical precipitation polymerization by combining a temperature-sensi-

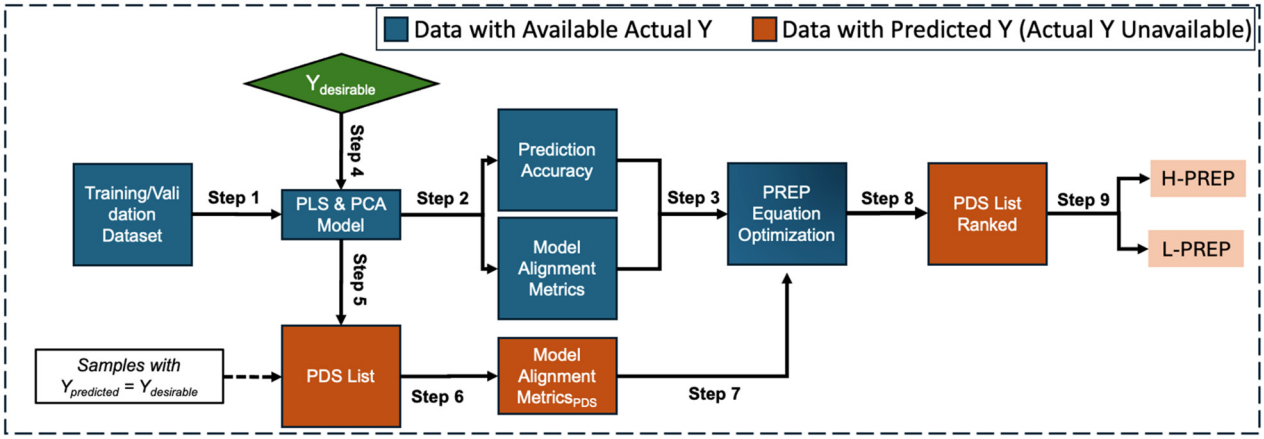

Fig. 2 Schematic illustration of the proposed PREP method. The green box represents the desired target output set. Blue boxes indicate the training and validation data in which actual *Y* values are known and used for optimizing the PREP equation. Orange boxes depict the dataset of potential candidates, for which only *X* values are available. Candidates selected through the PREP method are prioritized for experimental testing.

tive monomer (most typically *N*-isopropylacrylamide, NIPAM), and a pH-responsive comonomer selected among acrylic acid, methacrylic acid, fumaric acid, maleic acid, or vinyl acetic acid.[40] Achieving precise control over microgel size thus requires balancing of the different copolymerization kinetics of the multiple comonomers incorporated, the different water solubilities/hydrophilicities of the different comonomers, and the interactions between any included surfactant with the monomers and the growing copolymers. Our target was to fabricate three microgels with the same crosslinking density and an acid monomer content between 4–8 mol% (sufficient for inducing pH-responsive effects or enabling ligand grafting without compromising the desirable complementary temperature responsiveness[55]) but with as high as possible range in particle size at pH 7.4 and 37 °C. The pre-existing microgel dataset for this project is presented in Table 1. While the dataset already included samples with moderate (~300 nm, Sample 15) and large (~950 nm, Sample 12) sizes that met the design criteria, the smallest microgel that met all the criteria was Sample 4 (diameter ~175 nm), which was relatively close to the moderate size microgel and significantly higher than the ~100 nm particle size previously reported to bypass reticuloendothelial system clearance and pass through the liver sinusoidal fenestrae to promote long-term particle circulation.[56] As such, the optimization objective was to synthesize a 100 nm microgel that would meet this criteria while maintaining the same MBA content as Samples 12 and 15 (160 mg) and an acid content remained within the targeted 4–8 mol% range.

4.1.1 Experimental details

Materials. *N*-Isopropylacrylamide (NIPAM) (Sigma-Aldrich, 97%) was purified by recrystallization with 60 : 40 toluene/hexane mixture. N–N′-Methylene(bis)acrylamide (MBA) (Sigma-Aldrich, 99%), vinylacetic acid (VAA) (Aldrich, 97%), sodium dodecyl sulfate (SDS) (Sigma-Aldrich, 99%), potassium chloride (KCl) (Fisher Chemical, ACS grade), and ammonium persulfate (APS) (Sigma-Aldrich, 98%) were all used as received.

Table 1 Pre-existing microgel formulations and corresponding particle size data. Bolded columns represent the data used as the input (MBA, VAA, SDS) and output (size) variables for the PREP optimization process

Sample ID	NIPAM (g)	MBA (mg)	VAA (mg)	SDS (mg)	APS (mg)	Size[a] (nm)
1	1.6	**160**	**342**	**57**	50	**426**
2	1.6	**160**	**114**	**57**	50	**283**
3	1.6	**160**	**80**	**57**	50	**177**
4[b]	1.6	**160**	**46**	**57**	50	**176**
5	1.6	**205**	**114**	**57**	50	**298**
6	1.6	**114**	**114**	**57**	50	**269**
7	1.6	**80**	**114**	**57**	50	**299**
8	1.6	**46**	**114**	**57**	50	**319**
9	1.6	**160**	**114**	**34**	50	**396**
10	1.6	**160**	**114**	**23**	50	**444**
11	1.6	**160**	**114**	**0**	50	**657**
12[b]	1.6	**160**	**342**	**0**	50	**954**
13	1.6	**173**	**45**	**42**	50	**190**
14	1.6	**244**	**176**	**24**	50	**332**
15[b]	1.6	**160**	**228**	**57**	50	**300**

[a] Sizes correspond to the intensity-averaged effective diameter measured at pH = 7.4 and 37 °C. [b] Represents the best available candidates based on the existing dataset to meet the design criteria of creating a set of microgels with the same crosslinking density/acid content but as different as possible particle sizes.

MilliQ-grade water (>18 Ω resistance) was used for all experiments.

Microgel synthesis. The initial dataset used in this study is summarized in Table 1. For each synthesis recipe, specified amounts of NIPAM, MBA, SDS, and VAA were combined in a 250 mL round-bottom flask containing 150 mL of MilliQ water. The solution was deoxygenated by purging with nitrogen gas for 30 minutes at room temperature before being transferred to an oil bath preheated to 70 °C, with nitrogen purging continued throughout the process. Polymerization was initiated by dissolving 0.05 g of APS in 10 mL MilliQ water and introducing it to the flask using a syringe. The reaction

proceeded under magnetic stirring at 160 rpm for 4 hours at 70 °C. Upon completion, the reaction mixture was cooled to room temperature and dialyzed for six cycles, each lasting 6 hours, to remove residual surfactant and unreacted monomers. The resulting microgel suspension was then lyophilized and stored at ambient conditions.

Particle size measurements. The particle sizes of the microgels were determined using dynamic light scattering (Brookhaven 90Plus) operating at a fixed scattering angle of 90°. Measurements were performed at 37 °C in 10 mM KCl solutions, with the pH adjusted to 7.4 using 0.1 M HCl or NaOH. For each sample, five independent *z*-average particle size measurements were taken, and the average value of the intensity-weighted effective diameter was reported as the particle size. All microgels displayed a unimodal particle size distribution during analysis, such that the effective diameter is representative of the full particle size distribution.

4.1.2 Modeling preparation, integration, and iterations. Since the amounts of NIPAM and APS remained constant across the initial dataset, they were not considered in the model and only the three variables that do change (MBA, VAA, and SDS) were retained. Considering that each of these key variables can affect the kinetics of the polymerization, the nucleation mechanism of new polymer chains, and the maximum size to which the precipitation polymerization proceeds, from a modeling perspective microgel formation is a highly non-linear process and non-linear modeling approaches represent an attractive option. While Artificial Neural Networks (ANNs) are particularly appealing in this context given that they can capture intricate non-linear relationships in the data, ANNs require large amounts of training data to achieve reliable results, a key challenge in product design in which generating new data points is costly and time-consuming. Instead, we implemented an approach of combining a conventional LVMI with an optimization algorithm called Inversion by Optimization (IbO) that utilizes a PLS model to identify solutions in which the predicted outputs ($Y_{\text{predicted}}$) closely match the desired targets ($Y_{\text{desirable}}$) while minimizing certain soft constraints that help ensure statistical validity. The optimization framework enforces key conditions (*e.g.*, MBA = 160 mg and VAA mol% within the specified range) while minimizing PLS Hotelling's T^2 and SPE values. The complete framework is presented in eqn (2).

$$\min_{x^{\text{new}}} \left\{ w_1 \left(\hat{y}^{\text{new}} - y^{\text{des}}\right) \Gamma \left(\hat{y}^{\text{new}} - y^{\text{des}}\right)^T + w_2 \text{Hoteling} T^2_{\text{pls}} + w g_3 \text{SPE}_{x^{\text{new}}} \right\} \tag{2}$$

s.t.; $\hat{y}^{\text{new}} = \tau Q^T$; $\hat{x}^{\text{new}} = \tau P^T$; $\tau = x^{\text{new}} W^*$; where Γ is a $[L \times L]$ diagonal matrix containing the weights assigned to each output variable (emphasizing their relative importance). Given that particle size is the only output variable in this scenario, this term was simplified to $w_1(\hat{y}^{\text{new}} - \hat{y}^{\text{des}})$ in which w_1 represents the weight of each term.

The number of PLS components in such cases is typically determined using data-driven approaches such as cross-validation[57] the eigenvalue-less-than-one rule,[58] or based on experimental knowledge of the dependencies among input variables. In this microgel dataset, the selection was guided by experimental knowledge, as all three input variables—MBA, VAA, and SDS—could be independently manipulated within feasible ranges to synthesize new microgels. Consequently, three PLS components were chosen to sufficiently capture the relationships between the inputs and the output. Using this PLS model, the optimization framework in eqn (2) was applied, resulting in the recipe outlined in Table 2 (IbO 1st itr). The particle size obtained from this recipe (170 nm) was very close to the smallest microgel already available in the dataset. This new recipe was subsequently incorporated into the dataset, and the optimization algorithm was executed again for the next iteration. However, the synthesis of the suggested solution in the second iteration (IbO 2nd itr in Table 2) resulted in aggregation. It is worth noting that the direct model inversion solution was not applicable in this case, as it provided a single answer that failed to meet the required conditions around the VAA content (reaching as low as 2.4 mol%). As such, a more conventional approach did not achieve the targeted particle size, motivating the implementation of the PREP method, which was applied next to overcome these constraints.

The PREP method was implemented by first identifying the list of nearest neighbors; with three latent space components and a single output variable, a minimum of $A + 2 = 5$ nearest neighbors was required. To ensure clarity and avoid any perception that PREP was enhanced by the IbO method and the

Table 2 Measured microgel particle sizes from optimized recipes generated by both the Inversion by Optimization (IbO) method and the PREP method relative to the direct model inversion solution (target size = 100 nm). Bolded columns represent the data used as the input (MBA, VAA, and SDS) and output (size) variables for the PREP optimization process

Sample ID	MBA (mg)	VAA (mg)	SDS (mg)	Size (nm)	Comments
Direct Model Inversion	**158**	**33**	**57**	—	MBA and acid content both too low
IbO 1st itr	**160**	**62**	**65**	**170**	
IbO 2nd itr	**160**	**108**	**74**	—	Sample showed large-scale aggregation
PREP 1st itr (L1)	**160**	**92**	**91**	**144**	
PREP 1st itr (H1)	**160**	**70**	**80**	**151**	
PREP 2nd itr (L2)	**160**	**84**	**134**	**104**	
PREP 2nd itr (H2)	**160**	**101**	**133**	**118**	

similarity of the IbO 1st itr sample to a pre-existing datapoint (Sample 4), the IbO 1st itr sample generated in the initial attempt was excluded from the list of neighbors to ensure that PREP started with the same dataset originally provided to IbO method. Fig. 3 depicts all available datapoints and five nearest neighbors to the target in both the input (a) and output (b) spaces.

Subsequently, PLS and PCA models were constructed using the selected neighbors followed by the creation of the Potential Design Space (PDS). In this case, the number of PLS components exceeded the number of output variables by two, resulting in a two-dimensional null space (*i.e.* for any given $Y_{desirable}$, there exists a two-dimensional surface in the input and latent spaces where all points satisfy $Y_{predicted} = Y_{desirable}$). However, given the imposition of the constraint fixing the MBA content at 160 mg to match the crosslink density of the target microgel with the existing microgels in the series, the number of degrees of freedom was reduced to collapse the null space to a single dimension (*i.e.* a line within the original two-dimensional space), as shown in Fig. 4(i). Further analysis of the points along the blue line revealed that none of the candidates met the 4–8 mol% acid content requirement, necessitating the creation of the Potential Design Space (PDS) using an optimization-based algorithm. The algorithm generated a list of 50 candidates whose predicted outputs ($Y_{predicted}$) were as close as possible to the desired target ($Y_{desirable}$) while still satisfying all specified constraints. It is important to emphasize that the list generated through this optimization process fundamentally differs from the results obtained *via* IbO approach; while the PREP optimization algorithm produces a list of candidates by considering only the input range requirements, IbO yields a single solution by incorporating modeling alignment metrics such as Hotelling's T^2 and Squared Prediction Error (SPE). The new list generated by the implemented optimization algorithm (the PDS) is also shown in Fig. 4(i).

To identify the most relevant candidates for synthesis within the Potential Design Space (PDS), model alignment metrics were calculated for both the nearest neighbor samples and the PDS members and then used together with the prediction accuracy of the nearest neighbor samples to optimize the PREP equation parameters (C and P in eqn (1)). The resulting optimized PREP equation was then applied to rank all PDS candidates, from which two samples corresponding to the lowest (L-PREP) and highest (H-PREP) PREP scores were selected for experimental synthesis. The results of the PREP optimization and the ranking of Potential Design Space (PDS) samples for iteration 1 are presented in Fig. 4 where panel (ii) illustrates the relationship between the prediction accuracy and the PREP score for the validation data points used in optimizing the PREP equation and panel (iii) shows the PDS candidates ranked by their PREP scores; the two selected formulations for synthesis, corresponding to the highest ranked (L-PREP) and lowest ranked (H-PREP) ranked candidates, are also clearly highlighted. As expected, lower prediction accuracy is associated with higher PREP scores, confirming the metric's effectiveness in assessing prediction reliability. The measured particle sizes of the L-PREP and H-PREP recipes, as shown in

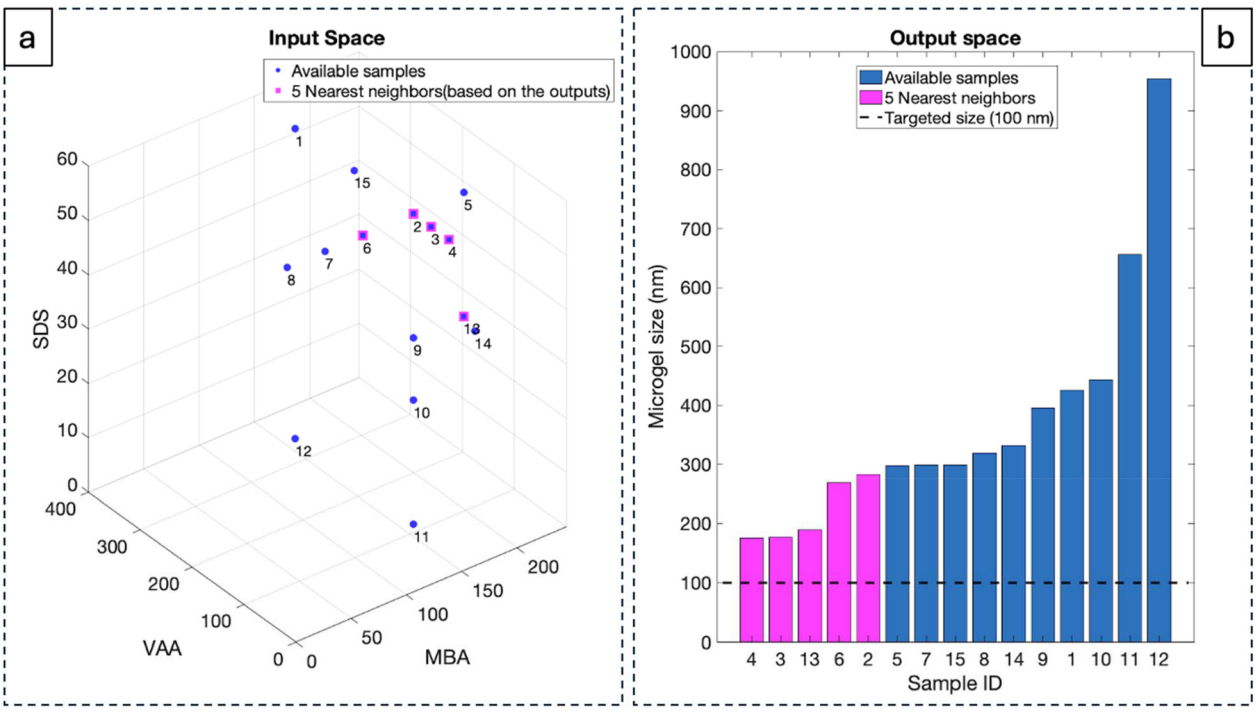

Fig. 3 Visualization of all available datapoints alongside the five nearest neighbors to the target in both the input (a) and output (b) spaces derived from the pre-existing dataset (Table 1).

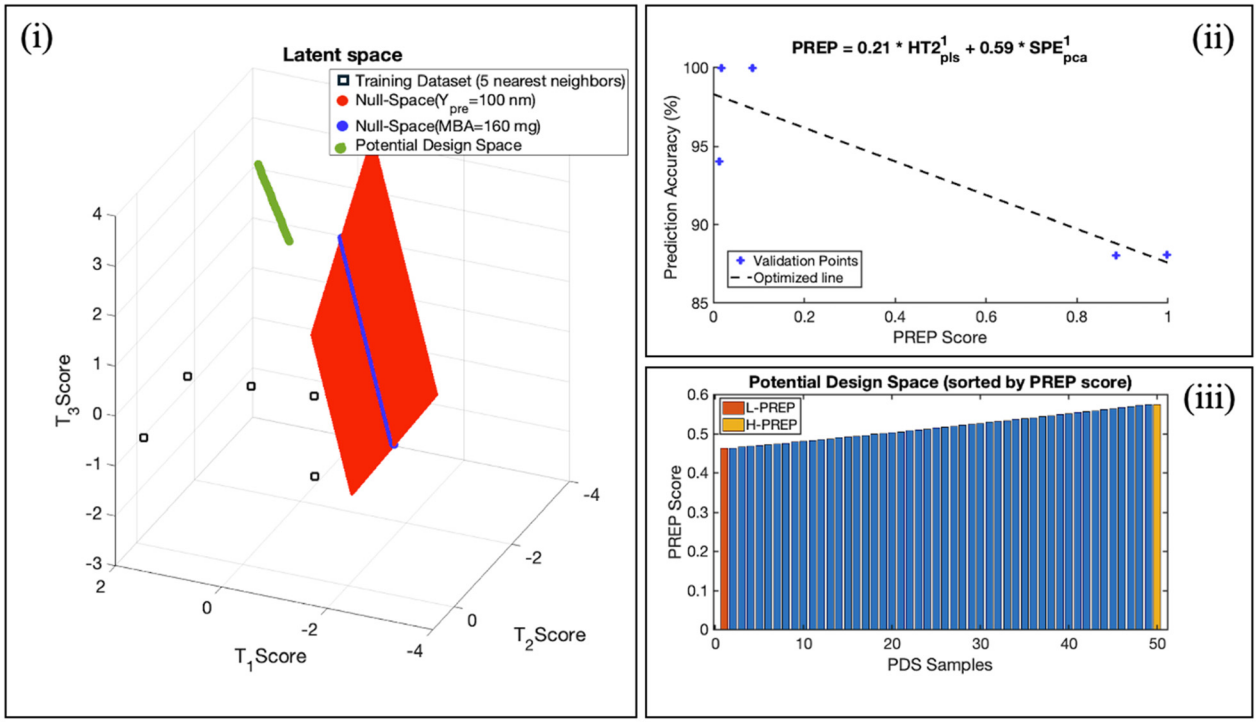

Fig. 4 Results from iteration 1 of the PREP implementation on microgel optimization. Sub-panel (i) represents the visualization of the Potential Design Space (PDS) in the latent space, (ii) shows the outcome of the PREP equation optimization demonstrating the alignment of validation data points along the optimized line (with higher PREP scores corresponding to lower prediction accuracy), and (iii) shows the ranked PDS samples based on their PREP scores with the selected candidates for synthesis (L-PREP – highest expected reliability and H-PREP – highest uncertainty used to enhance model refinement) highlighted.

Table 2, demonstrated that the samples suggested by the PREP method outperformed all existing datapoints in the dataset as well as those proposed by IbO approach. However, since the particle sizes of these samples still did not meet the ~100 nm target size, the newly synthesized samples from this first iteration were added to the dataset, the list of nearest neighbors was updated, and the PREP method was reapplied to generate new synthesis recipes. Note that including the two recipes from the first iteration (and thus removing the two samples from the five nearest neighbors from the first iteration) results in a 40% change in the dataset for the second iteration compared to the first iteration, a key advantage of using a smaller number of samples such that each sample carries disproportionately high weight in reframing the model (*i.e.* adding or replacing even a few samples can substantially alter the dataset, the model parameters, and thus the second iteration predictions).

The updated latent space based on the revised dataset are shown in Fig. 5(i). Note that enforcing all design constraints—particularly the specified acid content range of 4–8 mol%—did not yield a sufficient number of solutions within the actual null space (NS); consequently, the Potential Design Space (PDS) for the second iteration was expanded using the same optimization-based approach as in the first iteration, ensuring that all constraints were satisfied while generating at least 50 candidate datapoints within the PDS. The PREP equation parameters (C and P) were then re-optimized and the resulting equation was re-applied to rank all PDS candidates, with the resulting H-PREP and L-PREP samples identified in Fig. 5(iii) subsequently synthesized. As shown in Table 2, the L-PREP sample demonstrates exceptional proximity to the target particle size, achieving a size of 104 nm. Correspondingly, as shown in Fig. 5 panel (ii), the PLS model developed for the second iteration demonstrates significantly improved accuracy near the target output of 100 nm. Even the lowest-performing validation sample achieved over 97% accuracy—an improvement from 88% in the first iteration—indicating that the PREP method effectively guided the dataset expansion toward the desired region and enhanced model precision around the target.

Table 2 provides a summary of the particle sizes of the synthesized samples suggested by both the PREP and optimization-based methods. The microgel recipes proposed by the PREP method outperformed not only those generated by the optimization-based approach but also all samples in the initial dataset in terms of closeness to the target. The L-PREP and H-PREP samples from the first iteration achieved 75% and 78% accuracy relative to the target (particle sizes = 151 nm and 144 nm, respectively), while the second iteration recipes achieved accuracies of 92% and 98% (118 nm and 104 nm) that surpassed the predefined acceptable threshold of 95% closeness to the target. The PREP method's capacity to deliver an

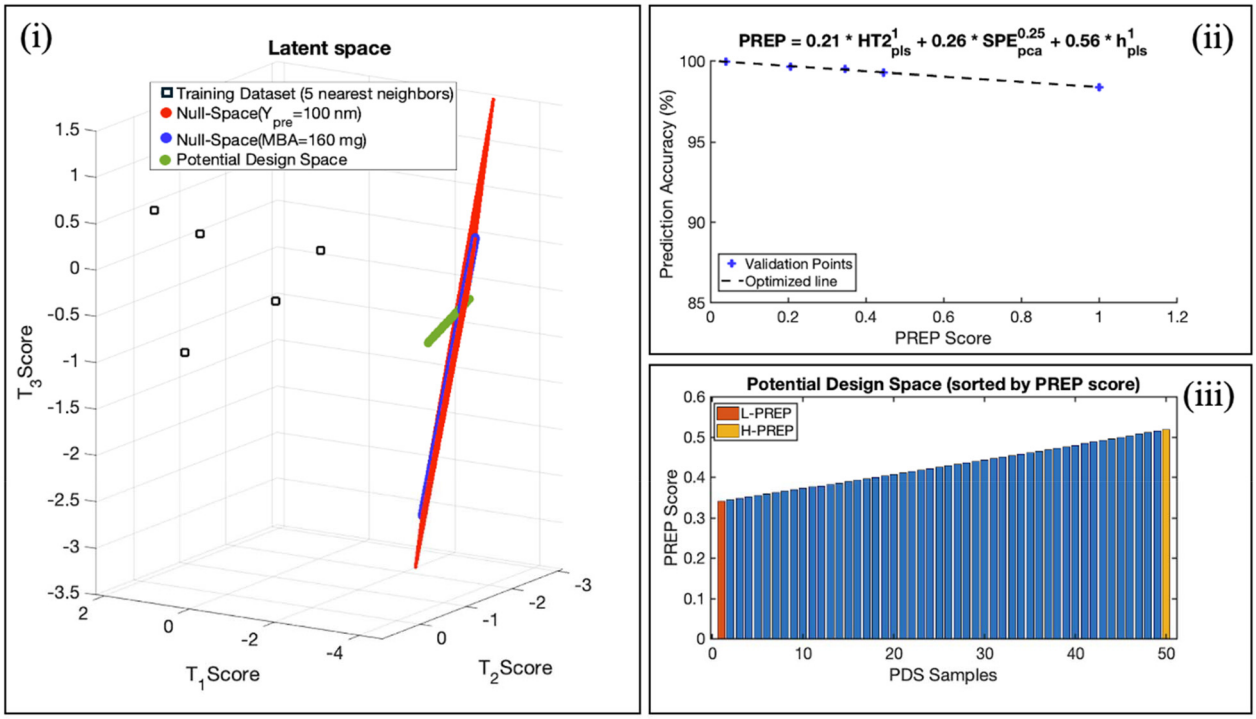

Fig. 5 Results from iteration 2 of the PREP implementation on microgel optimization. Sub-panel (i) represents the visualization of the Potential Design Space (PDS) in the latent space, (ii) shows the outcome of the PREP equation optimization demonstrating the alignment of validation data points along the optimized line (with higher PREP scores corresponding to lower prediction accuracy), and (iii) shows the ranked PDS samples based on their PREP scores with the selected candidates for synthesis (L-PREP – highest expected reliability and H-PREP – highest uncertainty used to enhance model refinement) highlighted.

optimized solution within just two iterations underscores the method's ability to handle dataset expansion rationally, rapidly refine predictions, and adapt to challenging design constraints in a highly non-linear system.

4.2 Case study 2: salt-stable polyelectrolyte complexes

Polyelectrolyte complexation presents several advantages over other nanoparticle fabrication techniques including as rapid self-assembly, relatively simple experimental setup, and the potential to eliminate the use of organic solvents.[59–61] Polyelectrolyte complexes (PECs) are particularly beneficial for delivering ionic therapeutics, which can either be used directly as a building block for nanoparticle assembly (*e.g.* DNA polyplexes[62,63]) or as an additive with tunable release based on the ionic interactions between the charged drug and its counterion polymer.[64,65] However, PECs are particularly sensitive to the high ionic strength of physiological fluids due to their reliance on electrostatic interactions for both intraparticle stabilization and colloidal stability, both of which can be disrupted at high salt concentrations due to charge screening. Thus, identifying PEC formulations with improved stability at high ionic strength without compromising either their favorable size for effective circulation (<200 nm to avoid splenic filtration[1]) or their capacity to load clinically-relevant concentrations of drug is of interest. Given the multiple variables that can influence the size and stability of PECs including the molecular weight and charge ratios of the polyelectrolytes, the pH, the ionic strength, and the drug concentration,[59,66] identifying a formulation that meets both size and stability requirements typically necessitates the fabrication of an extensive library of formulations that lends itself ideally to the implementation of optimization models. The specific case study selected involves the combination of sulfated yeast beta-glucan (GS, anion, a carbohydrate with known immunomodulatory potential to reprogram macrophages away from a pro-fibrotic state toward a pro-inflammatory state[67]) with quaternized dextran (Dex, cation) and the cationic chemotherapeutic drug doxorubicin (DOX), with the combination of the DOX chemotherapeutic loading plus the immunomodulatory properties of GS offering potential benefits for cancer immunotherapy. The target was to achieve initial particle sizes as small as possible and a polydispersity index (PDI) below 0.1 following fabrication in low ionic strength buffer and a final particle size <200 nm (model target: 170 nm) and PDI <0.2 (model target: 0.15) upon transfer of the formed PECs to phosphate buffered saline matching physiological pH and ionic strength.

4.2.1 Experimental details

Materials. Sulfated yeast beta glucan (glucan sulfate, GS) from *S. cerevisiae* was prepared as described by Williams *et al.*[68] (M_n = 13.5 kDa, $Đ$ = 5.5, sulfur degree of substitution = 0.33, charge density = 1.54 ± 0.06 μeq per mg). Cationic

dextran (Dex-GTAC) was prepared *via* functionalization with glycidyltrimethylammonium chloride in the presence of NaOH according to previous methods[69,70] (M_n = 3.7 kDa, $Đ$ = 1.05, nitrogen degree of substitution = 0.50, charge density = 2.09 ± 0.1 μeq per mg). Doxorubicin hydrochloride (DOX, 97.8%) was obtained from Millipore Sigma and used as received. MilliQ-grade water (>18Ω resistance) was used for all experiments. PBS stocks were prepared from PBS tablets (Millipore Sigma) and adjusted to pH 6.5 prior to nanoparticle fabrication. Full-strength PBS (150 mM ionic strength, 10 mM phosphate ions) was denoted as "1× PBS", with all other concentrations used expressed as a fraction of the full-strength concentration.

Polyelectrolyte complex (PEC) fabrication. Polyelectrolyte complexes were prepared using a flash nanoprecipitation method, with the recipes comprising the initial dataset used for optimization summarized in Table 3. GS, Dex-GTAC, and DOX were dissolved in PBS prepared at the ionic strength identified in Table 3, after which 3 mL of the GS solution was loaded into a 6 mL syringe and 3 mL of a 1 : 1 volume ratio of the Dex-GTAC and DOX solutions was loaded into a second 6 mL syringe. The syringes were loaded onto a confined impinging jet mixer and co-jetted over ~2–2.5 seconds into a fresh scintillation vial using a pneumatic plunger. The resulting PEC suspension was left to stir for 10–15 minutes prior to analysis. Note that all formulations followed the same general composition of GS mass ratio > Dex-GTAC mass ratio > DOX mass ratio, maintaining a sulfur : nitrogen ratio greater than 1 in each case.

PEC characterization. PECs were characterized for their size and PDI as a function of time and ionic strength using dynamic light scattering (Brookhaven NanoBrook 90Plus; Long Island, NY, USA; temperature = 25 °C, N = 5 technical replicates). Freshly prepared PECs were 0.2 μm syringe filtered into a polystyrene cuvette prior to analysis. To assess the formulation's stability in physiologically relevant ionic strength, the PECs were diluted (1 : 1 v/v) in concentrated PBS to a final ionic strength corresponding to 1× PBS (~150 mM ionic strength) and analyzed again *via* DLS. The intensity-averaged effective diameter and PDI were reported as the average of 5 technical replicates.

4.2.2 Modeling preparation, integration and iterations. In the available dataset, the PBS ionic strength (expressed as a ratio of the physiological PBS ionic strength), the total polymer concentration, and the GS and Dex-GTAC mass ratios were selected as the system's manipulatable parameters. DOX was not included among the manipulatable variables given that all GS and Dex-GTAC ratios were defined relative to DOX (DOX = 1) in the key input variables used for modeling; as such, the DOX concentration was represented as a normalized variable across all samples. Since the objective was to achieve final particle sizes <200 nm and PDI values <0.2 after exposure to physiological ionic strength solutions, the 1× PBS column from Table 3 was used as the model output. Fig. 6 illustrates how well this target aligns with the existing dataset. While some samples met the size requirement, no sample achieved sufficiently low polydispersity; alternately, other samples met the polydispersity requirement but failed to achieve the target particle size. As such, the optimization approach aimed to identify formulations that satisfied both criteria simultaneously.

Although four input variables were available for manipulation, an additional constraint was imposed to require that samples have a higher GS concentration relative to Dex-GTAC concentration such that the nanoparticle surface is GS-rich (to

Table 3 Initial dataset of PEC formulations. Bolded columns represent the data used as the input variables (assembly solvent as a fraction of full-strength PBS, total precursor concentration added, GS : DOX ratio, and Dex-GTAC : DOX ratio) and output variables (size and PDI in 1× PBS) for the PREP optimization process

Sample ID	Assembly solvent [× PBS]	Total precursor conc. [mg mL^{-1}]	Pre-assembly GS conc. [mg mL^{-1}]	Pre-assembly Dex-GTAC conc. [mg mL^{-1}]	Pre-assembly DOX conc. [mg mL^{-1}]	GS : DOX ratio	Dex-GTAC : DOX ratio	Assembly solvent		1× PBS	
								Size [nm]	PDI	Size [nm]	PDI
1	**0.5**	**0.5**	0.750	0.200	0.050	**15.0**	**4.0**	156	0.11	**208**	**0.11**
2	**0.1**	**0.5**	0.750	0.200	0.050	**15.0**	**4.0**	109	0.13	**362**	**0.04**
3	**0.5**	**0.75**	1.125	0.300	0.075	**15.0**	**4.0**	147	0.14	**229**	**0.09**
4	**0.1**	**0.75**	1.125	0.300	0.075	**15.0**	**4.0**	110	0.14	**357**	**0.08**
5	**0.5**	**1**	1.500	0.400	0.100	**15.0**	**4.0**	161	0.15	**260**	**0.06**
6	**0.1**	**0.25**	0.375	0.100	0.025	**15.0**	**4.0**	133	0.18	**326**	**0.11**
7	**0.5**	**0.5**	0.750	0.188	0.063	**12.0**	**3.0**	146	0.09	**217**	**0.11**
8	**0.1**	**0.5**	0.750	0.188	0.063	**12.0**	**3.0**	124	0.16	**298**	**0.08**
9	**0.1**	**0.75**	1.125	0.281	0.094	**12.0**	**3.0**	123	0.19	**313**	**0.05**
10	**0.5**	**1**	1.500	0.375	0.125	**12.0**	**3.0**	164	0.10	**243**	**0.05**
11	**0.1**	**1**	1.500	0.375	0.125	**12.0**	**3.0**	124	0.20	**744**	**0.25**
12	**0.5**	**0.5**	0.750	0.125	0.125	**6.0**	**1.0**	141	0.10	**170**	**0.21**
13	**0.5**	**0.5**	0.727	0.182	0.091	**8.0**	**2.0**	153	0.03	**409**	**0.12**
14	**0.26**	**0.72**	1.119	0.255	0.067	**16.7**	**3.8**	113	0.08	**142**	**0.28**
15	**0.17**	**0.83**	1.275	0.311	0.074	**17.2**	**4.2**	112	0.07	**150**	**0.26**
16	**0.2**	**0.82**	1.269	0.292	0.079	**16.1**	**3.7**	113	0.11	**137**	**0.21**
17	**0.16**	**0.78**	1.206	0.279	0.075	**16.0**	**3.7**	116	0.08	**141**	**0.23**
18	**0.17**	**0.53**	0.875	0.116	0.068	**12.8**	**1.7**	117	0.22	**142**	**0.31**
19	**0.1**	**0.54**	0.882	0.130	0.068	**12.9**	**1.9**	144	0.23	**171**	**0.21**

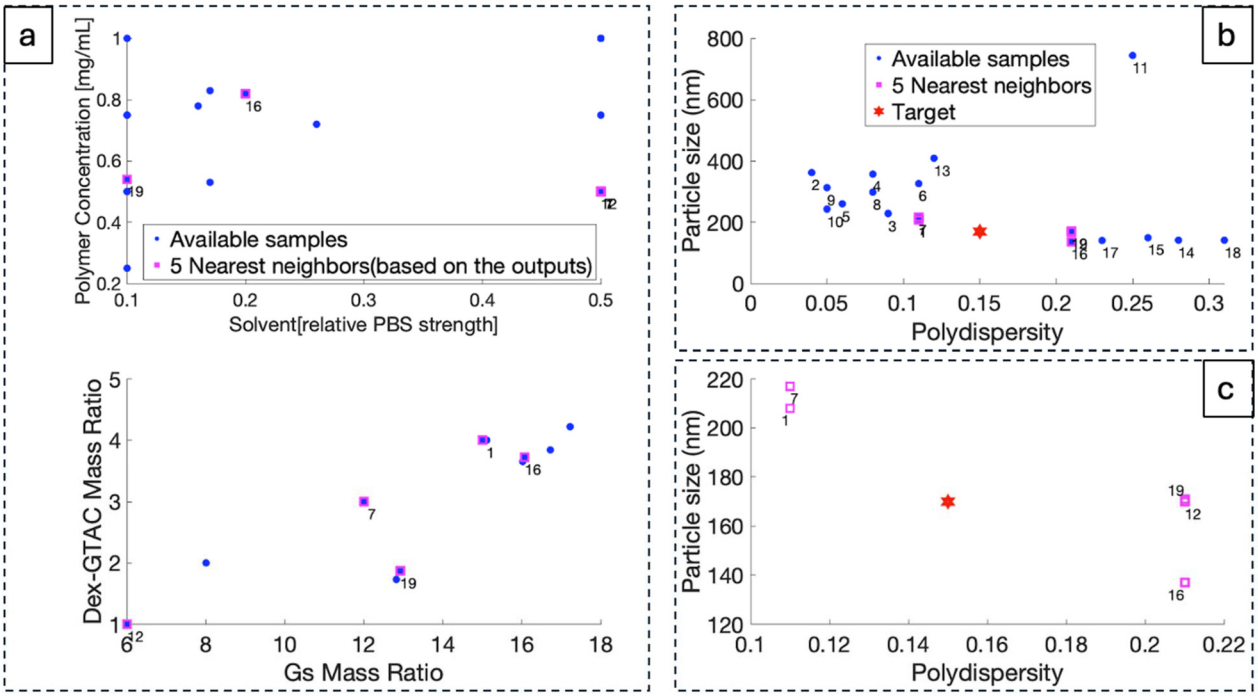

Fig. 6 Visualization of all available data points along with the five nearest neighbors to the target in the input space (a) and output spaces showing all samples (b) and only the nearest neighbors (c) as derived from the pre-existing dataset summarized in Table 3.

promote nanoparticle/macrophage interactions) and the final net charge in the PEC is anionic, key to minimize interactions with proteins in physiological fluids and representing a common design criteria for PECs.[71–73] As a result, the number of truly independent variables was reduced to three, and the number of PLS components was set to three, and the number of nearest neighbors to activate the PREP analysis was A (= 3) + 2 = 5. Fig. 6 illustrates all available data points and highlights the five nearest neighbors to the target in both the input space (a) and the output space (b), with panel (c) representing a zoomed-in version of the area around the target in panel (b).

Next, the PREP method was iteratively applied to the dataset following the same structured sequence of steps described in Case Study 1 for each iteration: developing PLS and PCA models, generating the PDS, optimizing the PREP equation, ranking the PDS, selecting the L-PREP and H-PREP candidates, synthesizing the L-PREP and H-PREP recipes, evaluating whether the target was met, and (if necessary) updating the list of nearest neighbors before repeating the process until satisfactory experimental results were achieved. Given the number of measurable variables and the number of PLS components, the dataset had a one-dimensional null space, *i.e.* there exists a line in the three-dimensional latent space along which variations do not affect the predicted Y. All points on this line, provided they satisfy the constraint GS mass > Dex-GTAC mass, constitute the PDS and were ranked based on their PREP score.

The outcomes of PREP implementation for the first two iterations are presented in Fig. 7. In each sub-figure, panel (i) illustrates the limited portion of the null space (NS) that is spanned by the Potential Design Space (PDS) within the latent space, panel (ii) displays the results of the PREP equation optimization, highlighting the alignment of the validation data points along the optimized trend line according to the calculated PREP scores, and panel (iii) shows the PDS candidates for each iteration ranked by their PREP scores; the two selected candidates for experimental synthesis denoted as L-PREP (low PREP score, high reliability) and H-PREP (high PREP score, high uncertainty) are clearly indicated in the graph and consistently labeled as Lx or Hx where x is the iteration number. The first iteration of the model exhibited relatively poor predictive performance near the target output (Fig. 7(a)), with two of the validation data points yielding prediction accuracy values as low as 60%. However, in the second iteration (Fig. 7(b)), model accuracy improved substantially, with the lowest prediction accuracy among the validation data points showing a prediction accuracy of 85%. Table 4 confirms that the optimization objectives were successfully achieved within just two iterations, yielding a particle with a size of 171 nm (target <200 nm) and a polydispersity index of 0.19 (target <0.2). Nonetheless, two additional iterations (Fig. 8(a) and (b)) were conducted to explore the possibility of further improving the dispersity, leading to the synthesis of a more narrowly dispersed PEC with a particle size of 182 nm and a PDI of 0.15 (Table 4) that precisely matched the model's targeted dispersity value. Note that by the fourth iteration (Fig. 8(b)) even the least accurate validation sample achieved a prediction accuracy above 93%, showing the relevance of the

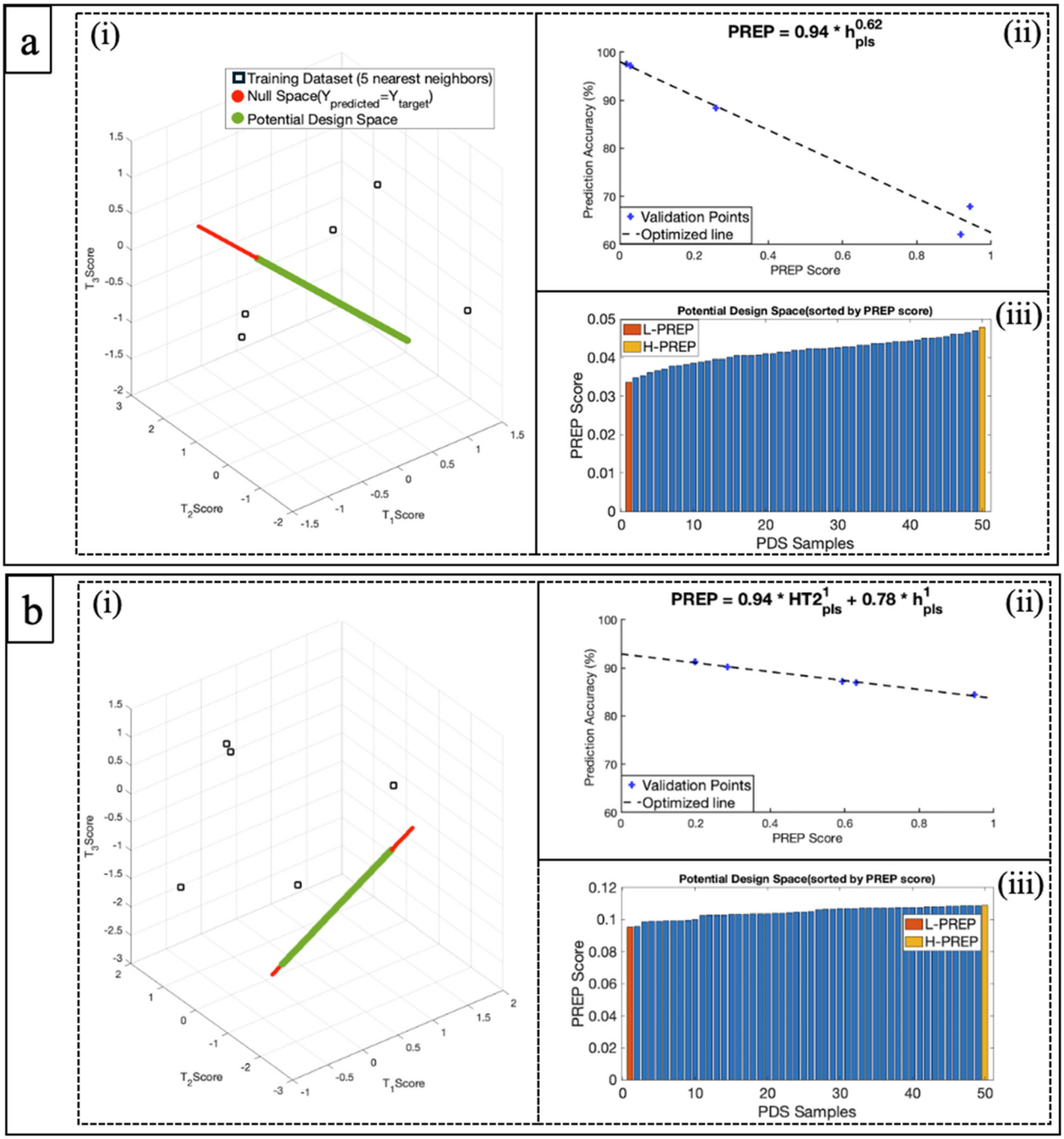

Fig. 7 Results from iteration 1 (a) and iteration 2 (b) of the PREP implementation on PEC optimization. In each sub-panel, (i) represents the visualization of the Potential Design Space (PDS) in the latent space, (ii) shows the outcome of the PREP equation optimization demonstrating the alignment of validation data points along the optimized line (with higher PREP scores corresponding to lower prediction accuracy), and (iii) shows the ranked PDS samples based on their PREP scores with the selected candidates for synthesis (L-PREP – highest expected reliability and H-PREP – highest uncertainty used to enhance model refinement) highlighted.

PREP method to improve model outputs in minimal iterations. It is important to note that conducting the PREP algorithm over another two iterations (Table 4) did not yield further improvements over the best sample obtained in iteration 4 (Sample L4), consistent with the high accuracy of the model already achieved at iteration 4 such that additional iterations did not offer significant further benefits in model prediction accuracy (Fig. S1(a) and S1(b)). This behavior is consistent with the probabilistic nature of the PREP algorithm, which while generally effective in guiding dataset expansion does not guarantee monotonic performance improvement across iterations. As shown in our prior work, the sample rankings based on

Table 4 PEC recipes and particle size results from the iterations generated by PREP model. The sample names correspond to either the H-PREP (H) or L-PREP (L) samples synthesized in each iteration (the number) of the PREP algorithm. Bolded columns represent the data used as the input variables (assembly solvent as a fraction of full-strength PBS, total precursor concentration added, GS : DOX ratio, and Dex-GTAC : DOX ratio) and output variables (size and PDI in 1× PBS) for the PREP optimization process

Sample ID	Assembly solvent [× PBS]	Total precursor conc. [mg mL^{-1}]	Pre-assembly GS conc. [mg mL^{-1}]	Pre-assembly Dex-GTAC conc. [mg mL^{-1}]	Pre-assembly DOX conc. [mg mL^{-1}]	GS : DOX ratio	Dex-GTAC : DOX ratio	Assembly solvent		1× PBS	
								Size [nm]	PDI	Size [nm]	PDI
L1	**0.18**	**0.40**	0.625	0.102	0.073	**8.6**	**1.4**	121	0.23	**178**	**0.34**
H1	**0.13**	**0.86**	1.341	0.309	0.070	**19.1**	**4.4**	105	0.14	**126**	**0.23**
L2[a]	**0.50**	**0.88**	1.257	0.274	0.229	**5.5**	**1.2**	97	0.21	**171**	**0.19**
H2	**0.46**	**0.88**	1.178	0.447	0.135	**8.7**	**3.3**	96	0.06	**131**	**0.24**
L3	**0.30**	**0.83**	1.273	0.306	0.081	**15.8**	**3.8**	94	0.02	**125**	**0.27**
H3	**0.76**	**0.66**	0.924	0.066	0.330	**2.8**	**0.2**	108	0.25	**118**	**0.4**
L4[a]	**0.10**	**0.80**	1.060	0.353	0.186	**5.7**	**1.9**	111	0.02	**182**	**0.15**
H4	**0.13**	**0.94**	1.436	0.368	0.075	**19.1**	**4.9**	93	0.09	**126**	**0.23**
L5	**0.10**	**0.65**	1.000	0.250	0.050	**20.0**	**5.0**	106	0.10	**131**	**0.25**
H5	**0.10**	**0.71**	1.061	0.300	0.060	**17.7**	**5.0**	126	0.11	**166**	**0.20**
L6	**0.59**	**0.65**	1.128	0.120	0.052	**21.6**	**2.3**	108	0.25	**105**	**0.39**
H6	**0.33**	**0.51**	0.862	0.128	0.030	**29.0**	**4.3**	81	0.18	**104**	**0.51**

[a] Best performing samples.

PREP scores do not always correspond directly to prediction accuracy, and in some iterations high PREP score candidates may unexpectedly yield better results than low PREP ones (presumably by exploring less explored parts of the design space that have higher prediction errors but yield superior performance). This highlights the value of PREP's dual-candidate strategy (L-PREP and H-PREP) while also illustrating the convergence limits of the model once optimal regions of the design space have been sufficiently explored. Collectively, these results illustrate PREP's capacity to efficiently converge on an optimal solution within a constrained design space while requiring minimal experimental effort.

Fig. 9 illustrates the outcomes of each iteration alongside the initial nearest neighbors from the pre-existing dataset in the output space, highlighting the proximity of each iteration result to the target. Notably, while the L2 (second iteration L-PREP) sample significantly outperformed all other samples in the dataset (*i.e.* was positioned closer to the target within the output space), the third iteration H-PREP and L-PREP samples both significantly underperformed the initial nearest neighbor samples; however, extending the iterations for one more cycle resulted in the L4 formulation that improved on the performance of L2. This example shows that the aggressiveness of the PREP method in terms of revising the number of nearest neighbor and thus "historical" samples in each iteration can lead to some significant iteration-to-iteration variability but ultimately converges faster on a recipe with target properties. Of note, the optimized L4 recipe resulted in a DOX encapsulation efficiency and loading capacity of 31% and 2.3 wt%, respectively; while this result represents a modest encapsulation efficiency, the loading capacity is significant and the potent nature of DOX (IC_{50} values in the micromolar/nanomolar range[74,75]) is relevant for practical chemotherapeutic use. Furthermore, if additional optimization of the DOX content within these PECs is desirable, the PREP method may be applied to the same system while adding DOX loading as an additional target property.

Relative to the first case study, this case presented additional challenges associated with a greater number of output variables, a lower degree of freedom in the null space (1D compared to 2D in the first case study), and the need to optimize properties that were not intrinsic to the initially synthesized particles but instead emerged after their introduction into a higher ionic strength solution. The successful implementation of PREP in this complex scenario further underscores its potential for handling high-dimensional systems with greater complexity.

5. Discussion

The implementation of the PREP method for nanoparticle size control demonstrates its strong potential as a data-driven optimization tool in scenarios in which existing datasets are limited in their coverage of the desired output space. One of the most notable strengths of PREP observed in this work is its ability to extrapolate beyond the bounds of the original dataset while preserving the fundamental correlations inherent to the system. This capability is especially valuable in nanoparticle design, in which the empirical design space defined by available experimental data may not sufficiently explore the parameter space associated with more challenging design targets. For example, in Case Study 1, despite no sample in the initial dataset having a size below 170 nm within the targeted crosslinker/acid concentrations, PREP successfully leveraged the underlying statistical structure of the data to suggest a formulation that resulted in a microgel significantly smaller than any previously observed sample. A similar advantage was

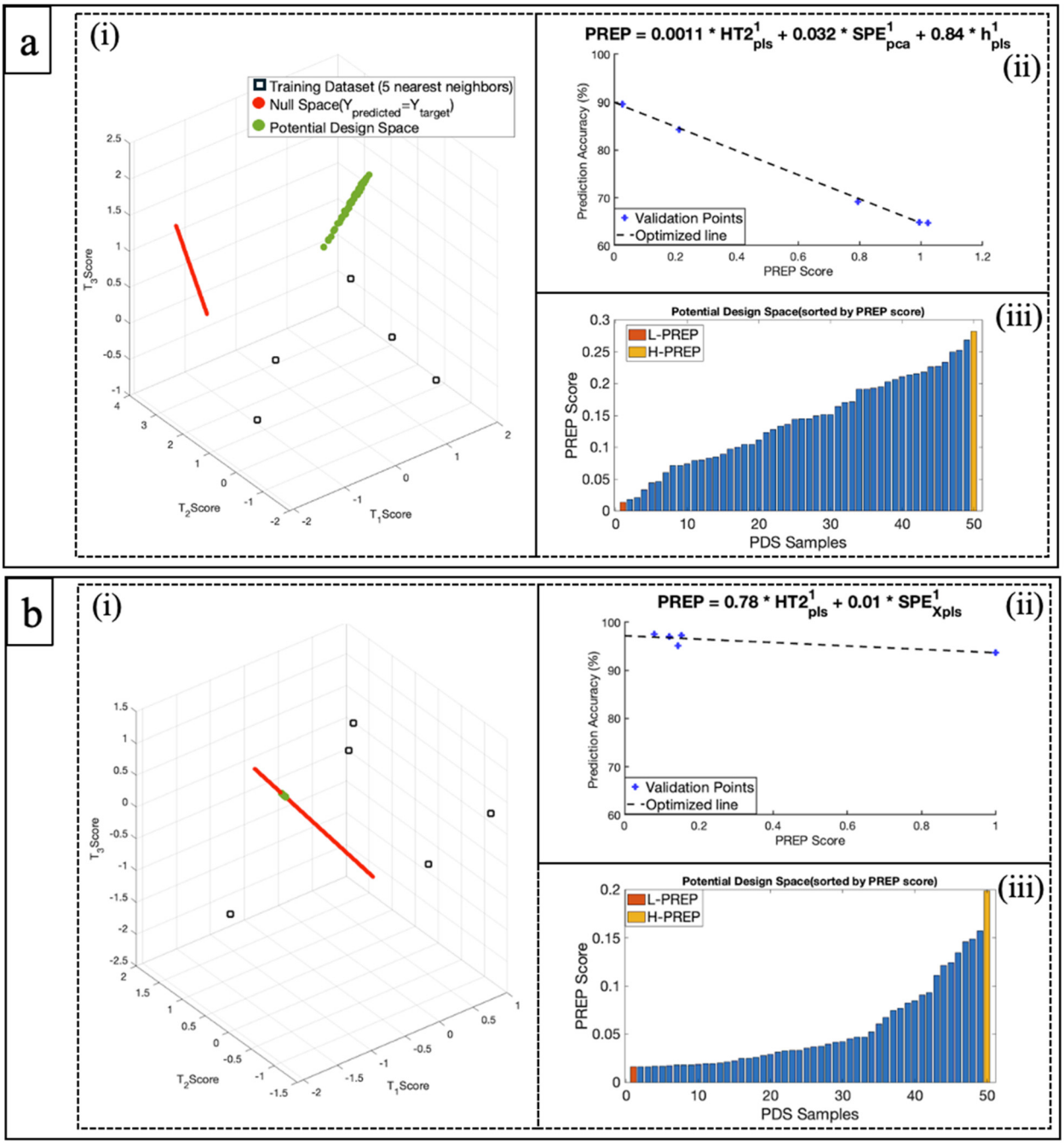

Fig. 8 Results from iteration 3 (a) and iteration 4 (b) of the PREP implementation on PEC optimization. In each sub-panel, (i) represents the visualization of the Potential Design Space (PDS) in the latent space, (ii) shows the outcome of the PREP equation optimization demonstrating the alignment of validation data points along the optimized line (with higher PREP scores corresponding to lower prediction accuracy), and (iii) shows the ranked PDS samples based on their PREP scores with the selected candidates for synthesis (L-PREP – highest expected reliability and H-PREP – highest uncertainty used to enhance model refinement) highlighted.

observed in Case Study 2, in which the initial dataset included samples that met one of the low particle size or low polydispersity index design criteria but not both; PREP was able to identify and prioritize formulations that bridged this gap, producing nanoparticles that simultaneously satisfied both the size and polydispersity targets in only two iterations. As such, the PREP method has clear utility not just in optimizing within known boundaries but also in directing the evolution of the dataset toward previously unexplored but desirable regions of the output space. In particular, while previously published work has focused primarily on forward modeling approaches (*i.e.* developing models to predict particle size or other pro-

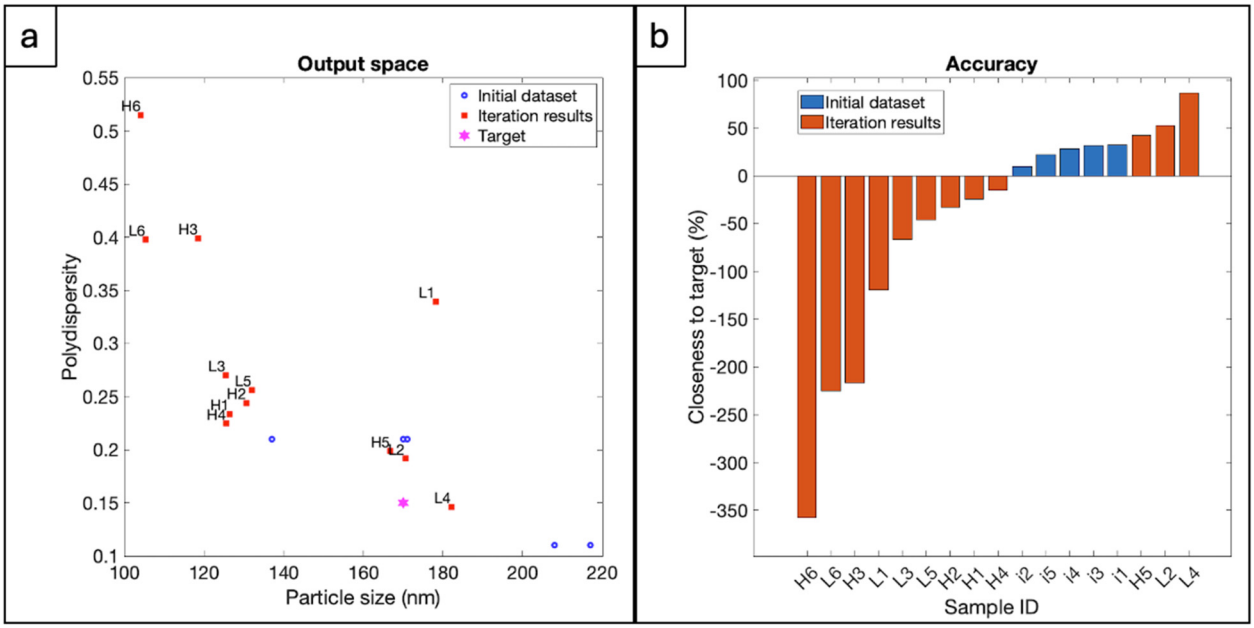

Fig. 9 Assessment of iteration results relative to the target particle size and polydispersity expressed relative to (a) the actual output space and (b) the proximity of each datapoint to the target size.

perties based on known input variables), the PREP method offers improved predictive performance when inverting the problem (*i.e.* suggesting new formulations dissimilar to the training data but that can yield specific desired outputs), particularly under data-limited conditions.

The iterative feedback structure of PREP is also highly advantageous in that it allows the PREP method to rapidly incorporate new data and revise its predictions, offering an efficient means of dataset expansion with each iteration contributing meaningful directional insight. These results suggest that PREP is particularly well-suited to systems in which the relationships among input variables are complex, the output space is multidimensional, and the design goals are not fully represented in the initial data. More specifically, the second case study presented additional challenges due to a higher number of output variables, reduced flexibility in the null space, and the need to optimize properties that emerged only after the particles were introduced into physiological conditions, all challenges that were successfully navigated by the PREP algorithm.

The success of PREP in these studies highlights its potential as a transformative tool for nanoparticle design and optimization. By leveraging data-driven modeling, PREP offers a systematic approach to refining synthesis protocols, reducing resource-intensive trial-and-error processes, and ensuring precise control over key material properties. Note that while the case studies described here in focus only on particle size optimization for two types of systems (covalently-crosslinked microgels and polyelectrolyte complexes), we expect the underlying PREP framework to be broadly applicable optimizing the size or other property of other types of nanoparticle systems in which the experimental design variables (inputs) and measured properties (outputs) can be organized into well-defined multivariate X and Y blocks respectively. Moving forward, the application of PREP to datasets with an even higher degree of input and output complexity remains an open avenue for exploration, presenting opportunities to further extend its impact across a broader range of nanoparticle engineering challenges.

6. Conclusions

The Prediction Reliability Enhancing Parameter (PREP) method was successfully applied to streamline the synthesis and optimization of nanoparticles with precise size and size distribution characteristics. Across two distinct case studies involving very different types of nanoparticles and nanoparticle fabrication methods (precipitation polymerization of dual pH- and temperature-responsive microgels and physical self-assembly of polyelectrolyte complex nanoparticles), PREP effectively achieved target size and/or polydispersity properties in just two iterations while achieving highly accurate results under complex design constraints, overcoming the limitations of traditional approaches that consistently failed to reach the desired size. As such, the application of PREP offers significant potential to address other types of nanoparticle optimization challenges and other complex materials design challenges, leveraging its demonstrated reliability in high-dimensional optimization problems.

Conflicts of interest

There are no conflicts to declare.

Data availability

The majority of the data supporting this study are presented in the main text. The full dataset is available from the corresponding author upon reasonable request or as required by the journal.

PREP optimization results for the second case study (PEC) including outcomes from Iterations 5 and 6. See DOI: **https://doi.org/10.1039/d5nr01664a**.

Acknowledgements

The Natural Sciences and Engineering Research Council of Canada (NSERC, Discovery grant RGPIN-2017-06455 and CREATE grant 555324), the McMaster Advanced Control Consortium (MACC), and the Canada Research Chairs program (to both T. H. and P. M.) are gratefully acknowledged for funding this work.

References

1 J. W. Hickey, *et al.*, Control of polymeric nanoparticle size to improve therapeutic delivery, *J. Controlled Release*, 2015, **219**, 536–547.
2 N. Sun, T. Wang and S. Zhang, Radionuclide-labelled nanoparticles for cancer combination therapy: a review, *J. Nanobiotechnol.*, 2024, **22**, 728.
3 S. Gavas, S. Quazi and T. M. Karpiński, Nanoparticles for cancer therapy: current progress and challenges, *Nanoscale Res. Lett.*, 2021, **16**(1), 173.
4 K.-R. Kim, *et al.*, Sentinel lymph node imaging by a fluorescently labeled DNA tetrahedron, *Biomaterials*, 2013, **34**(21), 5226–5235.
5 S. T. Proulx, *et al.*, Use of a PEG-conjugated bright near-infrared dye for functional imaging of rerouting of tumor lymphatic drainage after sentinel lymph node metastasis, *Biomaterials*, 2013, **34**(21), 5128–5137.
6 L. Y. Chou, K. Zagorovsky and W. C. Chan, DNA assembly of nanoparticle superstructures for controlled biological delivery and elimination, *Nat. Nanotechnol.*, 2014, **9**(2), 148–155.
7 V. B. Patravale, A. A. Date and A. B. Jindal, *Nanomedicines for the Prevention and Treatment of Infectious Diseases*, Springer, 2023, vol. 56.
8 E. Sharifi, *et al.*, Nanostructures for prevention, diagnosis, and treatment of viral respiratory infections: from influenza virus to SARS-CoV-2 variants, *J. Nanobiotechnol.*, 2023, **21**(1), 199.
9 I. M. Adjei, C. Peetla and V. Labhasetwar, Heterogeneity in nanoparticles influences biodistribution and targeting, *Nanomedicine*, 2014, **9**(2), 267–278.
10 J.-W. Yoo, N. Doshi and S. Mitragotri, Adaptive micro and nanoparticles: temporal control over carrier properties to facilitate drug delivery, *Adv. Drug Delivery Rev.*, 2011, **63**(14–15), 1247–1256.
11 L. Wu, J. Zhang and W. Watanabe, Physical and chemical stability of drug nanoparticles, *Adv. Drug Delivery Rev.*, 2011, **63**(6), 456–469.
12 C. Wong, *et al.*, Multistage nanoparticle delivery system for deep penetration into tumor tissue, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**(6), 2426–2431.
13 K. Y. Win and S.-S. Feng, Effects of particle size and surface coating on cellular uptake of polymeric nanoparticles for oral delivery of anticancer drugs, *Biomaterials*, 2005, **26**(15), 2713–2722.
14 J.-M. Williford, *et al.*, Shape control in engineering of polymeric nanoparticles for therapeutic delivery, *Biomater. Sci.*, 2015, **3**(7), 894–907.
15 S. H. Choi, S. H. Lee and T. G. Park, Temperature-sensitive pluronic/poly (ethylenimine) nanocapsules for thermally triggered disruption of intracellular endosomal compartment, *Biomacromolecules*, 2006, **7**(6), 1864–1870.
16 A. Albanese, *et al.*, Tumour-on-a-chip provides an optical window into nanoparticle tissue transport, *Nat. Commun.*, 2013, **4**(1), 2718.
17 X. Wang, *et al.*, Size control synthesis of melanin-like polydopamine nanoparticles by tuning radicals, *Polym. Chem.*, 2019, **10**(30), 4194–4200.
18 Z. Song, *et al.*, Size control of copper nanodrugs through emulsion atom transfer radical polymerization, *Polym. Chem.*, 2024, **15**(17), 1777–1785.
19 A. C. Mendes, *et al.*, Self–assembly in nature: using the principles of nature to create complex nanobiomaterials, *Wiley Interdiscip. Rev.: Nanomed. Nanobiotechnol.*, 2013, **5**(6), 582–612.
20 M. Grzelczak, *et al.*, Directed self-assembly of nanoparticles, *ACS Nano*, 2010, **4**(7), 3591–3605.
21 P. M. Valencia, *et al.*, Microfluidic platform for combinatorial synthesis and optimization of targeted nanoparticles for cancer therapy, *ACS Nano*, 2013, **7**(12), 10671–10680.
22 D. Liu, *et al.*, A versatile and robust microfluidic platform toward high throughput synthesis of homogeneous nanoparticles with tunable properties, *Adv. Mater.*, 2015, **27**(14), 2298–2304.
23 R. Karnik, *et al.*, Microfluidic platform for controlled synthesis of polymeric nanoparticles, *Nano Lett.*, 2008, **8**(9), 2906–2912.
24 R. Saswade, *et al.*, Data-driven Model Predictive Control of Nanoparticle Production in Modular Reactors. in *2024 European Control Conference (ECC)*. 2024, IEEE.
25 E. D. Koronaki, *et al.*, Nonlinear manifold learning determines microgel size from Raman spectroscopy, *AIChE J.*, 2024, **70**(10), e18494.
26 S. Dong, *et al.*, Gaussian processes modeling for the prediction of polymeric nanoparticle formulation design to enhance encapsulation efficiency and therapeutic efficacy, *Drug Delivery Transl. Res.*, 2024, 1–17.
27 N. Sahai, M. Gogoi and N. Ahmad, Mathematical modeling and simulations for developing nanoparticle-based cancer

drug delivery systems: a review, *Curr. Pathobiol. Rep.*, 2021, **9**, 1–8.
28 S. Keßler, K. Drese and F. Schmid, Simulating copolymeric nanoparticle assembly in the co-solvent method: How mixing rates control final particle sizes and morphologies, *Polymer*, 2017, **126**, 9–18.
29 S. Keßler, F. Schmid and K. Drese, Modeling size controlled nanoparticle precipitation with the co-solvency method by spinodal decomposition, *Soft Matter*, 2016, **12**(34), 7231–7240.
30 R. Stepanyan, *et al.*, Controlled nanoparticle formation by diffusion limited coalescence, *Phys. Rev. Lett.*, 2012, **109**(13), 138301.
31 P. López-Domínguez, *et al.*, Precise Modeling of the Particle Size Distribution in Emulsion Polymerization: Numerical and Experimental Studies for Model Validation under Ab Initio Conditions, *Polymers*, 2023, **15**(22), 4467.
32 A. A. Lahiq and S. M. Alshahrani, State-of-the-art review on various mathematical approaches towards solving population balanced equations in pharmaceutical crystallization process, *Arabian J. Chem.*, 2023, **16**(8), 104929.
33 P. Dogra, *et al.*, Mathematical modeling in cancer nanomedicine: a review, *Biomed. Microdevices*, 2019, **21**, 1–23.
34 N. Sheibat–Othman, *et al.*, Is modeling the PSD in emulsion polymerization a finished problem? An overview, *Macromol. React. Eng.*, 2017, **11**(5), 1600059.
35 D. Suzuki, *et al.*, Machine-learning-assisted prediction of the size of microgels prepared by aqueous precipitation polymerization, *Chem. Commun.*, 2024, **60**(93), 13678–13681.
36 E. Mikayilov, *et al.*, Role of computational modeling in the design and development of nanotechnology-based Drug Delivery systems, *Chem. Biochem. Eng. Q.*, 2024, **38**(2), 97–110.
37 T. Al Najjar, N. K. Allam and E. N. El Sawy, Anionic/nonionic surfactants for controlled synthesis of highly concentrated sub-50 nm polystyrene spheres, *Nanoscale Adv.*, 2021, **3**(19), 5626–5635.
38 M. A. Ahmed, J. Erdőssy and V. Horváth, The role of the initiator system in the synthesis of acidic multifunctional nanoparticles designed for molecular imprinting of proteins, *Period. Polytech., Chem. Eng.*, 2021, **65**(1), 28–41.
39 D. Palací-López, *et al.*, Improved formulation of the latent variable model inversion-based optimization problem for quality by design applications, *J. Chemom.*, 2020, **34**(6), e3230.
40 S. S. Tayebi, *et al.*, Predicting the Volume Phase Transition Temperature of Multi-Responsive Poly (N-isopropylacrylamide)-Based Microgels Using a Cluster-Based Partial Least Squares Modeling Approach, *ACS Appl. Polym. Mater.*, 2022, **4**(12), 9160–9175.
41 F. Yacoub and J. F. MacGregor, Product optimization and control in the latent variable space of nonlinear PLS models, *Chemom. Intell. Lab. Syst.*, 2004, **70**(1), 63–74.
42 L. Zhang and S. Garcia-Munoz, A comparison of different methods to estimate prediction uncertainty using Partial Least Squares (PLS): a practitioner's perspective, *Chemom. Intell. Lab. Syst.*, 2009, **97**(2), 152–158.
43 M. C. Denham, Prediction intervals in partial least squares, *J. Chemom., A*, 1997, **11**(1), 39–52.
44 S. Serneels, P. Lemberge and P. J. Van Espen, Calculation of PLS prediction intervals using efficient recursive relations for the Jacobian matrix, *J. Chemom.. A*, 2004, **18**(2), 76–80.
45 N. K. M. Faber, Uncertainty estimation for multivariate regression coefficients, *Chemom. Intell. Lab. Syst.*, 2002, **64**(2), 169–179.
46 I. S. Helland, Partial least squares regression and statistical models, *Scand. J. Stat.*, 1990, 97–114.
47 K. Faber and B. R. Kowalski, Prediction error in least squares regression: Further critique on the deviation used in The Unscrambler, *Chemom. Intell. Lab. Syst.*, 1996, **34**(2), 283–292.
48 S. Van Huffel and J. Vandewalle, The partial total least squares algorithm, *J. Comput. Appl. Math.*, 1988, **21**(3), 333–341.
49 A. Phatak, P. Reilly and A. Penlidis, An approach to interval estimation in partial least squares regression, *Anal. Chim. Acta*, 1993, **277**(2), 495–501.
50 B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*, Chapman and Hall/CRC, 1994.
51 S. S. Tayebi, T. Hoare and P. Mhaskar, Fast-Tracking Design Space Identification with the Prediction Reliability Enhancing Parameter (PREP), *Comput. Chem. Eng.*, 2025, 109159.
52 S. Petrusic, *et al.*, Properties and drug release profile of poly (N-isopropylacrylamide) microgels functionalized with maleic anhydride and alginate, *J. Mater. Sci.*, 2013, **48**, 7935–7948.
53 S. Campora, *et al.*, Functionalized poly (N-isopropylacrylamide)-based microgels in tumor targeting and drug delivery, *Gels*, 2021, **7**(4), 203.
54 A. Das, *et al.*, Poly (N–isopropylacrylamide) and its copolymers: a review on recent advances in the areas of sensing and biosensing, *Adv. Funct. Mater.*, 2024, **34**(37), 2402432.
55 T. Hoare and R. Pelton, Highly pH and temperature responsive microgels functionalized with vinylacetic acid, *Macromolecules*, 2004, **37**(7), 2544–2550.
56 Z. H. Mok, The effect of particle size on drug bioavailability in various parts of the body, *Pharm. Sci. Adv.*, 2024, **2**, 100031.
57 D. W. Osten, Selection of optimal regression models via cross–validation, *J. Chemom.*, 1988, **2**(1), 39–48.
58 N. Cliff, The eigenvalues-greater-than-one rule and the reliability of components, *Psychol. Bull.*, 1988, **103**(2), 276.
59 V. S. Meka, *et al.*, A comprehensive review on polyelectrolyte complexes, *Drug Discovery Today*, 2017, **22**(11), 1697–1706.
60 P. Sarika and N. R. James, Polyelectrolyte complex nanoparticles from cationised gelatin and sodium alginate for curcumin delivery, *Carbohydr. Polym.*, 2016, **148**, 354–361.
61 N. P. Birch and J. D. Schiffman, Characterization of self-assembled polyelectrolyte complex nanoparticles formed

from chitosan and pectin, *Langmuir*, 2014, **30**(12), 3441–3447.
62 W. Liu, *et al.*, An investigation on the physicochemical properties of chitosan/DNA polyelectrolyte complexes, *Biomaterials*, 2005, **26**(15), 2705–2711.
63 D. Oupický, *et al.*, DNA delivery systems based on complexes of DNA with synthetic polycations and their copolymers, *J. Controlled Release*, 2000, **65**(1–2), 149–171.
64 L. Zhang, *et al.*, Preparation of polyelectrolyte complex nanoparticles of chitosan and poly (2-acry1amido-2-methylpropanesulfonic acid) for doxorubicin release, *Mater. Sci. Eng., C*, 2016, **58**, 724–729.
65 Y. Luo, *et al.*, Preparation, characterization and drug release behavior of polyion complex micelles, *Int. J. Pharm.*, 2009, **374**(1–2), 139–144.
66 C. Schatz, *et al.*, Formation and properties of positively charged colloids based on polyelectrolyte complexes of biopolymers, *Langmuir*, 2004, **20**(18), 7766–7778.
67 H. Yuan, *et al.*, Effect of the modifications on the physicochemical and biological properties of β-glucan—A critical review, *Molecules*, 2019, **25**(1), 57.
68 D. L. Williams, *et al.*, Development of a water-soluble, sulfated (1→ 3)-β-D-glucan biological response modifier derived from Saccharomyces cerevisiae, *Carbohydr. Res.*, 1992, **235**, 247–257.
69 J. J. Thomas, M. Rekha and C. P. Sharma, Dextran–glycidyltrimethylammonium chloride conjugate/DNA nanoplex: A potential non-viral and haemocompatible gene delivery system, *Int. J. Pharm.*, 2010, **389**(1–2), 195–206.
70 J. Bendoraitiene, *et al.*, Peculiarities of starch cationization with glycidyltrimethylammonium chloride, *Starch/Staerke*, 2006, **58**(12), 623–631.
71 T. Mohan, *et al.*, Highly protein repellent and antiadhesive polysaccharide biomaterial coating for urinary catheter applications, *ACS Biomater. Sci. Eng.*, 2019, **5**(11), 5825–5832.
72 P. Mocchiutti, *et al.*, Cationic and anionic polyelectrolyte complexes of xylan and chitosan. Interaction with lignocellulosic surfaces, *Carbohydr. Polym.*, 2016, **150**, 89–98.
73 A. Kodiyan, *et al.*, Surface modification with alginate-derived polymers for stable, protein-repellent, long-circulating gold nanoparticles, *ACS Nano*, 2012, **6**(6), 4796–4805.
74 R. D. Olson, *et al.*, Doxorubicin cardiotoxicity may be caused by its metabolite, doxorubicinol, *Proc. Natl. Acad. Sci. U. S. A.*, 1988, **85**(10), 3585–3589.
75 K. Ashikawa, *et al.*, Evidence that activation of nuclear factor-κB is essential for the cytotoxic effects of doxorubicin and its analogues, *Biochem. Pharmacol.*, 2004, **67**(2), 353–364.