Rapid Model-Based Recipe Design from Limited-Sample Datasets: Experimental Validation in Nanoparticle and Microgel Systems

Rapid Model-Based Recipe Design from Limited-Sample Datasets: Experimental Validation in Nanoparticle and Microgel Systems

By

Seyed Saeid Tayebi

A Thesis Submitted to the School of Graduate Studies in the Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy

Doctor of Philosophy (2025) McMaster University

(Chemical Engineering) Hamilton, Ontario, Canada

TITLE: Rapid Model-Based Recipe Design from Limited-Sample Datasets: Experimental Validation in Nanoparticle and Microgel Systems

AUTHOR: Seyed Saeid Tayebi

(McMaster University, Hamilton, ON)

SUPERVISORS: Dr. Prashant Mhaskar & Dr. Todd Hoare

NUMBER OF PAGES: 161

“Love and compassion are necessities, not luxuries. Without them, humanity cannot survive.”
— *Dalai Lama*

Abstract

In data-constrained experimental domains such as nanoparticle engineering, microgel synthesis, and pharmaceutical formulation, researchers frequently face the challenge of modeling systems governed by complex and highly nonlinear relationships among variables. These applications often involve limited datasets due to the cost, time, and resource demands of generating new samples, making conventional trial-and-error approaches inefficient. As a result, there is a growing need for data-driven methodologies that can reliably predict product behavior and guide recipe design using minimal experimental input. This thesis presents a series of strategies to enhance prediction accuracy and reliability in small datasets through localized modeling, quantitative model reliability assessment, and guided expansion of the available dataset. The first contribution involves coupling Latent Variable Modeling (LVM) with clustering to create local Partial Least Squares (PLS) models tailored to subsets of similar samples. This combination simplifies the underlying data structure by reducing multicollinearity via projection into latent space and grouping structurally similar data, thereby improving prediction fidelity. The framework was validated using the prediction of the Volume Phase Transition Temperature (VPTT) of dual-responsive microgels—a property influenced by several formulation variables—with results that showed significantly improved prediction accuracy. Building on these advances, the second contribution focuses on the inverse problem of design space identification—determining input configurations that are most likely to yield desired output properties. To do this robustly, the Prediction Reliability Enhancing Parameter (PREP) is introduced, a novel metric that unifies multiple LVM alignment diagnostics including Hotelling T^2, Squared Prediction Error (SPE), and score alignment factors into a single predictive reliability score. PREP is calibrated in a data-driven, case-specific manner and facilitates the ranking of candidate formulations by their expected predictive reliability. Extensive validation on simulated datasets has demonstrated that PREP significantly accelerates the identification of optimal solutions, particularly under highly nonlinear conditions and limited data regimes. PREP was subsequently deployed across real experimental case studies involving the formulation of nanoparticles and microgels. In one study, a microgel with a tightly constrained particle size of ~100 nm was successfully designed from an initial dataset spanning sizes of 170–900 nm, with PREP delivering a near-target solution in minimal iterations while competing design approaches failed under the applied constraints. In another case, PREP enabled the identification of polyelectrolyte complexes with particle sizes below 200 nm and polydispersity indices under 0.2, again demonstrating superior efficiency and accuracy relative to conventional approaches. Overall, this work offers a practical and scalable pathway for predictive modeling and recipe design in settings constrained by data scarcity and high experimental costs. The methodologies developed—particularly the integration of local LVM models and the PREP-based design space identification—can be broadly applied to other high-value domains requiring precision formulation and optimization such as drug delivery, nanomedicine, and advanced materials development.

Lay Abstract

In many real-world applications—from pharmaceuticals to materials science—researchers face the challenge of finding the right “recipe” to produce a product with desired properties. Each ingredient in a formulation affects the final outcome, but degree to which different ingredients affect the outcome—and how changes in one ingredient interact with the other ingredients—are often unknown and difficult to predict. A natural solution is to try to learn these patterns by experimenting with different recipes. However, testing many combinations is usually expensive, time-consuming, and sometimes even impossible, creating a serious problem of data scarcity. This thesis proposes a new approach to make the most of the limited data available. Instead of relying on the global datasets, the method begins by grouping existing recipes into clusters based on their similarities. It then carefully analyzes each group to uncover internal relationships between ingredients and outcomes. Building on this, the thesis introduces a new scoring system that helps determine when the predictions made by a model can be trusted—and when they cannot. By learning more from fewer experiments and making smarter suggestions for future recipes, this research offers a pathway to faster, more efficient product development in complex systems in which trial-and-error is not a viable option.

ACKNOWLEDGEMENTS

I would like to begin by expressing my deepest gratitude to my supervisors, Dr. Prashant Mhaskar and Dr. Todd Hoare, whose mentorship has profoundly shaped not only the trajectory of my PhD, but the person I have become over the past four years. From the earliest stages of my journey—when everything was unfamiliar and I was prone to making big, sometimes odd mistakes—they consistently offered patience, support, and encouragement. Not once did they make me feel that my missteps were anything other than part of the learning process. The environment they created was remarkably calm and pressure-free; in fact, what should have been the most stressful period of my life became one of the most peaceful and creatively fulfilling. Somehow, they each understood me better than I understood myself. Through their insight and quiet guidance, I experienced growth not only in my academic capabilities, but also in my confidence, independence, and character. I will forever carry with me the lessons and values they instilled, and no matter where life takes me, they will always remain at the top of the list of the most impactful people in my life.

I would also like to extend my sincere thanks to my supervisory committee member, Dr. Paul McNicholas, whose critical insights have been an invaluable part of my research journey. Dr. McNicholas consistently viewed my work through a fresh lens, offering perspectives and feedback that I had not considered and truly needed. His ability to pinpoint weaknesses or blind spots in my approach—things I often could not see myself—challenged me to think more deeply and rigorously. His questions were always sharp, constructive, and grounded in a desire to help me grow as a researcher. I am especially grateful for the way he brought clarity to complex aspects of my work and helped me better understand the broader implications of my decisions. His involvement not only strengthened the quality of my thesis but also sharpened my ability to evaluate my own work critically and thoughtfully.

Next, I would like to thank my wife, whose unwavering support has been the quiet force behind everything I have accomplished. I could never begin to list even a fraction of what she has done for me. One of my greatest hopes is that one day I can repay even the smallest part of her countless sacrifices. Her presence has made every challenge more manageable, and her belief in me—especially during times when I struggled to believe in myself—has been one of the greatest gifts of my life. Over the past ten years, since the day I met her, I can clearly see how my life has steadily evolved into its best form. She is, without question, my better half—by far. I know that no words will ever be enough to express the depth of my gratitude, but I hope she knows that everything good I've built is because of her love and strength. I also owe heartfelt thanks to two more members of our household—our fluffy cats, Fandogh and Farrokh. Their calming purrs and warm presence often helped me untangle the most complex ideas—on the strict condition that I pet them exactly

the right way. They asked for nothing in return... other than fish, treats, occasional bits of chicken, and anything else vaguely edible that they could persuade me to surrender.

I am endlessly grateful to my parents, Maryam and Hasan, whose unconditional love has been the foundation of everything I am. No matter where I am or how I'm doing, their love remains constant—a steady, grounding force in my life. They are, quite simply, the Pillars of Life for me. Nothing I achieve holds meaning unless I know it brings joy to them—and I know they feel the same way about me. I am also incredibly thankful for my two beautiful sisters, Susan and Sajedeh, who are among the greatest gifts life has given me. Their presence has always filled my world with laughter, comfort, and an unshakable sense of belonging. Thanks to them, I've never truly felt alone. Being far away from them has been the most painful part of this journey, and I deeply miss the days we could simply be together every day. But distance has only deepened my love and appreciation for them, and I carry them with me in everything I do. I also want to express my heartfelt thanks to my brother-in-law, Amir, who has stepped into a role I once believed would be my responsibility—caring for my parents—and has done so with incredible heart and commitment. He is one of the main reasons I can endure the distance, and I sometimes wonder if he was the best gift life gave to my sister or to me (though I think we both lucked out in the deal!).

I would also like to sincerely thank my in-laws, Mohammad and Azar, who supported me with open hearts even when there was no expectation for them to do so. Their unconditional care and belief in me gave me the freedom and security to pursue my dreams−something that would not have been possible without their generosity and love.

Finally, I would like to thank the broader academic communities that have supported and inspired me throughout this journey. I am grateful to Dr. Mhaskar's group, Dr. Hoare's lab, and the MACC group, all of whom have offered thoughtful discussions, technical insights, and a collaborative spirit that enriched my research experience. I am especially thankful to Dr. Hoare for introducing me to the CREATE program, which became a turning point in my academic path. Through CREATE, I was exposed to new disciplines, professional development opportunities, and career-building tools that reshaped how I think about my future. It opened windows I hadn't even known existed and gave me the clarity and confidence to start carving out a roadmap toward the career I hope to build.

Table of Contents

1 Introduction 1

1.1 Background and Motivation 1

1.2 Thesis outline 6

1.3 References 10

2 Predicting the Volume Phase Transition Temperature of Multi-Responsive Poly(N-isopropylacrylamide)-Based Microgels Using a Cluster-Based Partial Least Squares Modeling Approach 14

2.1 Introduction 15

2.2 Experimental 19

2.2.1 Materials 19
2.2.2 Microgel Synthesis 19
2.2.3 Swelling Profile Measurement 20

2.3 Modelling Frameworks 21

2.3.1 Principal Component Analysis (PCA) 21
2.3.2 Partial Least Squares (PLS) 22
2.3.3 Clustering Procedure (K-means Algorithm) 22

2.4 Methodology Development, Results and Discussion 23

2.4.1 Data Pre-Processing 23
2.4.2 Clustering-Based PLS Modelling 24
2.4.3 Suggested Data Arrangements 27
2.4.4 Swelling Profile Prediction 30
2.4.5 VPTT Prediction 32
2.4.6 Predicting the Properties of New Microgels 42

2.5 Conclusions 44

2.6 References 45

2.7 Supporting Information 49

2.7.1 Training Dataset Formulations, Summary of the Models' Performance and supplementary Figures 49
2.7.2 PCA and PLS Modelling Supplementary Discussion 52
2.7.3 Methods and Implementation 62
2.7.4 Swelling Profile Prediction Quality 65

3 Fast-Tracking Design Space Identification with the Prediction Reliability Enhancing Parameter (PREP) 69

3.1 Introduction 71

3.2 Preliminaries 77

3.2.1 Principal Component Analysis (PCA) 77
3.2.2 Partial Least Square-Projection to Latent Structure (PLS) 77
3.2.3 Latent Variable Model Inversion (LVMI) 78

3.3 Proposed Methodology 81

3.3.1 Involved Model Alignment Parameters and Initial Blocks Creation 82
3.3.2 Optimization Framework 87
3.3.3 PREP Implementation for Design Space Candidates 89
3.3.4 Refinement of Design Space Candidates 90
3.3.5 Iterative Execution of PREP Methodology 91

3.4 Model Assessment 92

3.4.1 First Simulated Dataset....95
3.4.2 Second Simulated Dataset....100
3.5 Conclusion....106
3.6 References....108
3.7 Supporting Information....110
3.7.1 Third Simulated Dataset....110
3.7.2 Fourth Simulated Dataset....112
3.7.3 Fifth Simulated Dataset....115
• *Fifth Simulated Dataset Results Summary*....115
4 Data-Driven Optimization of Nanoparticle Size Using the Prediction Reliability Enhancing Parameter (PREP)....118
4.1 Introduction....120
4.2 Preliminaries....123
4.2.1 Latent Variable Models (LVM)....123
4.2.2 Latent Variable Model Inversion (LVMI)....125
4.3 Proposed Methodology....126
4.4 Experimental Case Studies....129
4.4.1 Case Study 1: Multi-Responsive Microgels....129
4.4.2 Case Study 2: Salt-Stable Polyelectrolyte Complexes....138
4.5 Discussion....149
4.6 Conclusions....151
4.7 References....152
4.8 Supporting information of last chapter....156
5 Conclusions and Recommendations for Future Works....158
5.1 Conclusions....158
5.2 Future Works....160

List of Figures

Figure 2-1: Data arrangements for PLS modelling of microgel swelling: (A) no clustering is applied(one PLS model is applied on the whole observation set), (B) observations are clustered based on X (microgel recipe), with one PLS model applied per cluster, and (C) observations are clustered based on Y (microgel swelling response), with one PLS model applied per cluster.....25

Figure 2-2: Data distribution among different clusters based on both microgel recipes (A) and microgel swelling profiles (B). Organization of microgel observations into clusters as classified based on (C) [VAA] and (D) Summative Size Parameter values among all cluster members. The number of clusters required to separate the observations in the latent variable space was determined by the model under both policies.....28

Figure 2-3: Five data arrangements used in this work from Type 1 to Type 5 are represented by (A) to (E) respectively (K represents the number of input variables and m denotes the number of output variables). using these different data arrangements to generate non-biased comparisons between the different data arrangement types.....30

Figure 2-4: Experimentally observed (y-axis) versus model-predicted (x-axis) values of microgel particle sizes across all pH temperature values tested for each observation in the dataset using different data arrangements (rows) and different data clustering policies (columns).....31

Figure 2-5: Prediction error of the model for predicting the microgel volume phase transition temperature at pH=4 for each individual microgel in the dataset using (A) no clustering, (B) recipe (X-based) clustering, (C) swelling (Y-based) clustering, and (D) combined clustering, in each case using the best-performing data arrangement for each clustering policy. Observations highlighted in green are correctly predicted to either have or not have a discrete phase transition while observations highlighted in red are ones for which the existence of VPTT was mis-predicted; (E, F) comparison of VPTT prediction accuracies achieved using each clustering policy in A-D based on (E) the percentage of microgels for which the experimental VPTT is predicted within <3°C and (F) the average and worst-case prediction errors observed.....34

Figure 2-6: Recipe score plot for the prediction of the volume phase transition temperature of dual pH/temperature-responsive microgels at pH 4 using Combined clustering in which the green points represent

samples for which the VPTT was predicted within 3°C while the red points represent samples for which VPTT was either poorly predicted or mis-predicted.....36
Figure 2-7: 1st (A) and 2nd (B) PCA Loading vectors corresponding to the recipe score plot.....36
Figure 2-8: Prediction error of the model for the microgel volume phase transition temperature at pH10 using (A) no clustering, (B) recipe (X-based) clustering, (C) swelling (Y-based) clustering, and (D) combined clustering, in each case using the best-performing data arrangement for each clustering policy. Observations highlighted in green are correctly predicted to either have or not have a discrete phase transition while observations highlighted in red are ones for which the existence of VPTT was mis-predicted; (E, F) comparison of VPTT prediction accuracies achieved using each clustering policy in A-D based on (E) the percentage of microgels for which the experimental VPTT is predicted within <6°C and the Transition Prediction Parameter and (F) the average and worst-case prediction errors observed. 39
Figure 2-9: Recipe score plot for the prediction of the volume phase transition temperature of dual pH/temperature-responsive microgels at pH 10 using X (recipe-based) clustering in which the green points represent samples for which the VPTT was predicted within 6°C while the red points represent samples for which VPTT was either poorly predicted or mis-predicted.....41
Figure 2-10: (A,B) Microgel swelling and VPTT prediction quality of two new microgels (New Obs 1, A and New Obs 2, B) using the optimal data processing approaches identified at each pH: (A) pH = 4 using the combined clustering mode and (B) pH = 10 using recipe-based clustering; (C,D) Score plots for pH 4 (C) and pH 10 (D) showing that both microgels lie within the well-predicted range consistent with the good correlation between actual vs. measured sizes and VPTT values. 44
Figure 3-1: PCA (a) and PLS (b) blocking configurations 78
Figure 3-2: Schematic representation of the proposed PREP method. Blue boxes denote the training/validation data for which the actual Y values are known and used for optimizing the PREP equation. Orange boxes represent the dataset of potential candidates for which only X values are available, and the candidates must be ranked, with the candidates selected via the PREP method to be experimentally tested. .82

Figure 3-3: A schematic of an ideal PREP equation optimization showing what is valued within the optimization algorithm for the cost function. 88
Figure 3-4: Schematic representation of the simulated dataset implementation for TDS identification and comparison of previously reported optimization techniques relative to PREP.....93
Figure 3-5: Distributions of the first simulated dataset and targets: (a) Selected targets 1 (T1), 2 (T2), and 3 (T3) along with corresponding nearest neighbors for each target; (b-d) TDS for T1 (b), T2 (c), and T3 (d) in the original space; (e) Projection of each TDS from the second row into the latent space. 96
Figure 3-6: Performance of different methods in reaching (a) T1, (b) T2, and (c) T3 using the first simulated dataset. The first row displays the results from the Original Model Inversion, the second row displays the results from the Tomba method, and the third row illustrates the results from the proposed PREP method. ..98
Figure 3-7: PREP iteration results from first to last iteration for T2 using the first simulated dataset. (a) the outcome of the PREP optimization and its alignment in each iteration; (b) the projections of the TDS and PDS into the PLS latent space in each iteration, focusing only on the first and second scores (the third score is excluded for simplicity); (c) comparison of actual Y values, predicted Y values, and T2 values for all PDS samples sorted by PREP scores: (top row) actual Y vs. predicted Y; (bottom row) actual Y versus T2. 99
Figure 3-8: Comparison of the performance of different methods (original model inversion, Tomba, and PREP) in achieving the same targets shown in Figure 3-5 using various calibration datasets for the first simulated dataset.....100
Figure 3-9: Distributions of the second simulated dataset and targets: (a) Selected targets 1 (T1), 2 (T2), and 3 (T3) along with corresponding nearest neighbors for each target; (b-d) TDS for T1 (b), T2 (c), and T3 (d) in the original space; (e) Projection of each TDS from the second row into the latent space. 101
Figure 3-10: Performance of different methods in reaching (a) T1, (b) T2, and (c) T3 using the second simulated dataset. The first row displays the results from the Original Model Inversion, the second row displays the results from the Tomba method, and the third row illustrates the results from the proposed PREP method.....103

Figure 3-11: PREP iteration results from first to last iteration for T2 using the second simulated dataset: (a) the outcome of the PREP optimization and its alignment; (b) the projections of the TDS and PDS into the PLS latent space in each iteration, focusing PREP iteration results from first to last iteration for T2 using the second simulated dataset: (a) the outcome of the PREP optimization and its alignment; (b) the projections of the TDS and PDS into the PLS latent space in each iteration, focusing only on the first and second scores (the third score is excluded for simplicity); (c) comparison of actual Y values, predicted Y values, and T2 values for all PDS samples sorted by PREP scores: (top row) actual Y vs. predicted Y; (bottom row) actual Y versus T2. 104
Figure 3-12: Comparison of the performance of different methods (original model inversion, Tomba, and PREP) in achieving the same targets shown in Figure 3-9 using various calibration datasets for the second simulated dataset. 105
Figure 4-1: General latent variable modeling framework 124
Figure 4-2: 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. 128
Figure 4-3: 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 4-1). 134
Figure 4-4: 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 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.....136
Figure 4-5: 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. 137
Figure 4-6: 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 4-3.....143
Figure 4-7: 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 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.....145
Figure 4-8: 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. 146

Figure 4-9: 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.....149

List of tables

Table 2-1: Mole fractions of all recipe components used to synthesize the microgel library.....19

Table 2-2: VPTT prediction quality within the well-predicted area for both pH=4 and pH=10 42

Table 3-1: R^2 values from PLS conducted on the entire dataset compared to using the 5 nearest neighbors corresponding to each target for the first simulated dataset.....96

Table 3-2: R^2 values from PLS conducted on the entire dataset compared to using the 5 nearest neighbors corresponding to each target for the second simulated dataset.....101

Table 4-1: 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.....131

Table 4-2: 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.....138

Table 4-3: 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. 141

Table 4-4: 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.....147

Chapter 1

1 Introduction

1.1 Background and Motivation

In industrial product development, particularly in areas such as pharmaceutical formulation, nanomaterials, and process optimization, finding the ideal formulation or process condition is a critical and often complex task. Typically, what is available are a limited number of tried-and-tested options, each with a corresponding response or outcome. The primary challenge is to adjust the manipulable parameters, either to achieve a predetermined response or to enhance the consistency and reliability of the outcome. However, these relationships are frequently too intricate and nonlinear to be addressed manually. Consequently, the application of data-driven modeling approaches becomes indispensable in such contexts [1-6].

A fundamental issue in this area is the uncertainty inherent in predicting outcomes, especially when dealing with new, unseen data points. In most industrial settings, as the system complexity increases, a larger dataset is usually required to refine the model. However, collecting additional experimental data in such frameworks is often resource-intensive. The challenge lies in finding ways to maximize the efficiency of data utilization while minimizing the number of required samples. This necessitates careful and strategic expansion of the dataset to ensure that each new data point contributes meaningfully to the model's predictive power, without overburdening resources [7-11].

In these circumstances, the first step is to develop a model that offers reliable predictions with a reasonable level of accuracy. Even if the model can identify formulations or input sets that are less likely to succeed, it can still offer significant value by guiding future experimentation. While this provides some level of insight, such models must be refined to enhance their predictive reliability. In particular, as these

models are used in reverse to suggest the input configurations that will generate products with desired properties, the reliability of these models is crucial [12-14]. The next step is to enhance the model's ability to not only predict but to recommend optimal input sets with high confidence.

It is common in many industrial applications that the number of manipulable input variables often exceeds the number of output variables. This discrepancy carries significant implications both in practice and within the modeling framework. In practice, for any given target in the response space, there is not just one solution, but rather several potential solutions, collectively referred to as the Design Space (DS). The process of identifying this Design Space is commonly known as Design Space Identification [7, 15-18]. The ability to identify the most optimal solution is critical, as doing so can lead to considerable savings in terms of time, material costs, and overall resource utilization. From a modeling perspective, this mismatch between the number of inputs and outputs results in a scenario in which multiple input configurations can lead to identical predictions of the target output. This flexibility creates an opportunity to explore a range of input combinations but also introduces the challenge of determining which of these solutions is the most reliable and resource-efficient to pursue [7, 19]. This situation is akin to the "null space" concept, where the system's complexity and multiple input variables result in several potential solutions that produce the same response. However, not all of these predictions are made with the same level of accuracy [20]. Some input configurations will have predictions that are more reliable than others. It is crucial to identify these configurations, as they can provide greater certainty in model predictions and help prioritize those formulations or input sets for further exploration. In this context, the ability to assess the uncertainty of model predictions and evaluate their reliability across different formulations is paramount. A model that can identify high-confidence predictions is essential, as it allows for more efficient experimental design and ensures that resources are directed toward the most promising candidates.

Given the importance of identifying high-confidence predictions and ensuring reliable experimental design, it becomes evident that advanced modeling techniques are necessary to manage the complexity of

high-dimensional datasets. One such technique is Latent Variable Modeling (LVM) approaches such as Principal Component Analysis (PCA) and Partial Least Squares (PLS), which offer a powerful approach to handle the inherent complexity in these systems [21-23]. By identifying underlying factors that influence the system's behavior, LVMs reduce the dimensionality of the data and provide a more manageable framework for prediction. This characteristic aligns particularly well with the goals of design space identification, as it enables the modeling of complex systems by identifying and isolating key factors that influence the outcomes. LVMs facilitate the reduction of complexity by transforming the data from its original, highly correlated space into a latent space where variables are more independent. This process simplifies the modeling task, allowing the focus to shift to understanding the relationship between the input (X) and output (Y) variables in their more manageable, internally uncorrelated forms [24].

Getting back to the original purpose of using such data-driven modeling techniques, which is to extract as many patterns as possible from the data and then suggest new samples to expand the dataset purposefully toward a predetermined target, it is worth noting that compared to traditional approaches like Design of Experiments (DOE), which also aim to explore the Knowledge Space (KS) and identify regions that ensure consistent product quality, LVMs offer a distinct advantage [25, 26]. DOE methods often require large numbers of experimental samples to account for a wide range of input variables and process conditions, which can become impractical especially when experimental resources are limited [7-10, 27]. LVMs, on the other hand, are adept at handling multicollinearity in high-dimensional spaces, making them ideal for situations where the relationship between inputs and outputs is complex and not easily captured by simpler methods[28-31].

Despite their widespread use in various industries, particularly for handling complex, high-dimensional datasets, LVMs face a significant challenge: the lack of established and validated methods for quantifying prediction uncertainty, especially when applying these models to new, unseen observations [7, 8, 13, 20]. This gap becomes even more pronounced in datasets with limited samples, where the challenge of prediction reliability is critical.

One major approach to addressing prediction reliability in design space identification is to estimate model prediction uncertainty. This line of research seeks to enhance model precision and define prediction intervals—the range within which the actual outcome is expected to fall. Several studies have attempted to estimate prediction uncertainty by calculating either the variance in predicted outcomes or the variance in regression coefficients, which, in turn, affects the variance of predictions [20, 32].

Prediction uncertainty has traditionally been estimated using three main approaches: ordinary least squares (OLS)-based approximation techniques, linearization methods, and re-sampling strategies. OLS-based methods estimate the prediction interval by measuring the distance between new observations and the center of the input space, with larger distances leading to higher estimated uncertainty and wider prediction intervals [20, 29, 33]. Another common approach involves linearization, in which the regression coefficients are approximated using a first-order Taylor series expansion to capture how changes in output affect the coefficients. This method can estimate the variance of model parameters through matrix differential calculus [20, 28, 30, 34, 35]. Additionally, re-sampling methods like bootstrapping and jackknifing create new datasets by perturbing the original samples or residuals, allowing researchers to assess the distribution of prediction intervals [36, 37].

Probabilistic design space characterization methods—particularly Bayesian-based approaches—have gained considerable attention for their ability to quantify uncertainty and define design spaces in terms of feasibility probabilities [9, 38-43]. These methods are particularly valued in contexts such as regulatory compliance and risk-sensitive optimization, in which characterizing confidence in predictions is critical[39, 44, 45]. However, because their core focus is on mapping feasibility across the entire design space through probability estimation, they often rely on feasibility thresholds that can introduce conservatism and exclude promising but underrepresented regions[46]. While effective in many scenarios, these methods are not inherently designed to guide iterative sample selection or to prioritize candidates for efficient dataset expansion.

As an alternative to probabilistic frameworks, black-box modeling approaches such as Gaussian process regression, neural networks, or ensemble machine learning have been widely used for experimental optimization[47-51]. These models emphasize predictive flexibility and can perform well in capturing nonlinear input–output relationships [52, 53]. However, they are typically decoupled from the structure of the calibration dataset and do not leverage latent-variable diagnostics such as Hotelling's T^2 or SPE. As a result, they lack built-in mechanisms to assess prediction reliability or guide sampling in a structured way [54]. This can limit their practical utility in resource-constrained experimental workflows, in which interpretability and confidence in predictions are critical[55, 56].

Geometrical methods represent another class of design space identification tools, focusing on delineating feasible regions through constraint surface approximations in the input space[57-60]. While they provide a global view of where feasible solutions may reside, these methods do not prioritize which candidates within the feasible region should be explored first [61]. This absence of strategic ranking makes them less suitable for stepwise experimental design, especially when only a limited number of formulations can be tested. Furthermore, they often assume that the true design space lies entirely within the current knowledge space, without offering mechanisms for targeted exploration beyond its boundaries.

This thesis addresses the challenge of predicting reliable outcomes in systems with complex correlations and limited sample datasets, focusing on enhancing model prediction accuracy and providing a robust tool to assess the reliability of predictions for unseen data points. The core problem lies in evaluating and improving data-driven models applied to such datasets, where the goal is not only to refine prediction accuracy but also to identify which data points are most reliable for expanding the dataset. This optimization of the dataset is crucial, as it ensures that resources are directed toward the most informative data to improve both the model's precision and experimental efficiency. The ability to trust model predictions, even in resource-constrained environments, is essential for advancing predictive modeling in industrial applications, particularly when the dataset expansion is critical for achieving a desired outcome. The approach of this thesis is thus built around two key contributions: improving model precision through

dataset manipulation and developing a metric to quantify the level of prediction uncertainty, which together form the foundation for more efficient and reliable design space identification in complex systems. In particular, the PREP framework that represents the central contribution of this work was specifically developed to prioritize efficient discovery of feasible solutions by ranking candidates based on prediction reliability rather than mapping the entire design space, leveraging latent-variable structures embedded in the dataset to select samples most likely to succeed in experimental validation. Unlike probabilistic, black-box, or geometrical approaches, PREP is not designed to define the entire feasible region or maximize global uncertainty coverage; instead, its goal is to accelerate convergence to reliable solutions and promote intelligent expansion of the knowledge space. This distinction makes PREP especially valuable in cases where data are limited, experiments are costly, and targeted progress is more critical than exhaustive design space coverage.

1.2 Thesis outline

To address the problem outlined earlier, this thesis presents a structured approach across its chapters. Each section builds on the previous work, beginning with the development of methods to enhance prediction accuracy followed by tools to assess and improve the reliability of these predictions. The following sections provide a detailed exploration of the key contributions, methodologies, and experimental validations that underpin this research.

The first contribution of this thesis (Chapter 2) focuses on improving prediction accuracy by employing a combination of clustering and Partial Least Squares (PLS) modeling. The underlying principle here is that in systems governed by highly complex rules, the behavior of the system tends to be less complicated when observed locally as opposed to globally. In practice, this means that complex systems can be divided into smaller and more manageable subsets or clusters, within which the relationships between input and output variables are simpler. To leverage this principle, the dataset is divided into distinct clusters and separate PLS models are developed for each cluster. This approach reduces the overall size of the calibration dataset while simultaneously increasing its consistency. By focusing on more consistent

data points within each cluster, the model is better able to capture underlying patterns by first identifying the cluster it belongs to and then applying the corresponding cluster-specific model for prediction.

In the clustering process, two main approaches can be considered: clustering based on input (X) similarity and clustering based on output (Y) similarity. Clustering by X is simpler, as the input data for an unseen sample are always available and thus finding the appropriate cluster is straightforward. However, clustering based on output similarity (Y) can potentially be more effective in capturing the underlying patterns of the system, as it groups together samples that exhibit similar response behaviors. The challenge with clustering by Y, however, is that for unseen samples the output Y is not known, making it impossible to directly assign a cluster based on output similarity.

To overcome this challenge, a novel clustering mechanism that incorporates both the X and Y spaces was introduced. This method takes into account the relationships in both the input and output spaces, allowing for the clustering process to be applied even to unseen data points. By utilizing both spaces, the clustering approach can identify similar patterns more effectively, providing a robust solution for predicting outcomes with limited data.

This method was tested in the context of predicting the properties of multi-responsive microgels—complex systems in which various parameters influence the final properties. The clustering-based approach demonstrated enhanced prediction accuracy in this scenario; furthermore, it revealed situations in which the model's coverage within certain clusters was insufficient. This knowledge was particularly valuable in product design applications, as being able to identify clusters with unreliable models helps flag potentially inappropriate candidate formulations early in the process. Overall, this contribution demonstrates the effectiveness of combining clustering with PLS modeling to improve prediction accuracy in complex systems with limited data. The clustering mechanism not only enhances the model's predictive power but also provides valuable insights into the model's reliability, helping to guide experimental design and optimize resource allocation.

Building on the insights from the first contribution, the second contribution (Chapter 3) addresses the need for a more nuanced evaluation of prediction reliability, especially when local models (despite having high coverage) may still provide poor predictions for some new data points. This limitation highlights the challenge of reliably assessing model performance, even when the model's coverage appears sufficient. To resolve this issue, a new numerical metric was developed to quantify the reliability of predictions. This metric, called the Prediction Reliability Enhancing Parameter (PREP), integrates several existing metrics from Latent Variable Modeling (LVM), such as Hotelling's T^2, Squared Prediction Error (SPE), and Score Alignment, into a single composite score.

The advantage of PREP is that it addresses the issue of conflicting conclusions that might arise when using individual metrics. For example, one metric may suggest high reliability, while another may indicate uncertainty, making it difficult to make informed decisions based on model predictions. PREP resolves this problem by combining these parameters into a unified score, allowing for a more balanced evaluation of model performance. This comprehensive approach ensures that the model does not prematurely dismiss samples that, despite performing well on certain metrics, may still hold value for further exploration. Thus, PREP enhances the model's ability to assess which data points are most likely to contribute to the accurate identification of design spaces and which ones are necessary to improve model performance.

To effectively implement PREP, the process begins by generating a list of potential design space members, which includes samples for which the model prediction either matches or closely approximates a predetermined output target. Using the PREP equation, the model assigns a composite score to each member, with the coefficients and powers of the equation determined based on the nature of the available dataset. Once this list is established, all potential solutions are ranked according to their PREP scores. Two distinct options emerge from this ranking: the new data point with the lowest PREP score has a high likelihood of accurate prediction, while the new data point with highest PREP score provides insight into areas where the model's predictions need refinement. In this context, the PREP-based iterative process is

designed to expand the dataset toward the inclusion of True Design Space (TDS) members—those input sets that reliably produce the desired output properties. This iterative approach optimizes the model's predictive capability by continually refining the dataset with highly informative data points via a stepwise dataset expansion that helps guide the exploration of design spaces more efficiently.

Comparative evaluations of PREP against other commonly used methods applied to synthetic datasets of varying complexity showed that PREP is significantly more resource-efficient. By identifying the right solutions with fewer iterations, PREP reduces the number of experimental samples needed to pinpoint optimal formulations. These results highlight PREP's ability to streamline the process of design space identification, making it an invaluable tool in situations where resources and experimental data are limited.

The final contribution (Chapter 4) shifts the focus from simulated datasets to real experimental cases. In this phase, PREP was implemented across two highly complex and nonlinear systems—precipitation polymerization (to synthesize covalently crosslinked microgels) and polyelectrolyte complex formation (to fabricate ionically crosslinked nanoparticle coacervates). Although both case studies were experimental, they posed fundamentally different optimization challenges. In the first case, the objective was to achieve a particle size well outside the range covered by the initial dataset while adhering to strict hard constraints on the formulation space. These limitations significantly reduced the viable search area, rendering other methods incapable of even proposing experimentally feasible samples. PREP, however, was able to identify viable formulations and reached the target rapidly, demonstrating both efficiency and robustness. In the second case, constraints were again present in the formulation space but the optimization objective was not to achieve a fixed target value but rather to expand the output space coverage into previously unexplored regions. This required PREP to strategically learn from the available data and iteratively direct sampling toward promising areas, thereby driving dataset expansion in a purposeful manner toward achieving previously unachievable particle properties. Together, these case studies effectively validated PREP's capacity to solve real-world, constraint-heavy design problems in

which data scarcity and accuracy requirements coexist, confirming its value in experimental design space identification.

In summary, this thesis presents a coherent framework that advances predictive modeling and design space identification in data-scarce, nonlinear experimental systems. Each chapter contributes a key element toward addressing this overarching challenge: Chapter 2 improves prediction accuracy by localizing the calibration dataset through clustering, Chapter 3 enhances the stability and reliability of localized models using a novel scoring metric, and Chapter 4 demonstrates the practical validity of these methods through complex experimental case studies. Notably, the proposed approach consistently identified feasible and high-performing solutions without relying on infeasible intermediate steps—a limitation that compromised the effectiveness of competing methods. Collectively, these contributions offer a robust and generalizable strategy for accelerating experimental discovery in constrained design environments.

1.3 References

1. Torres, J.M.G.T., et al., *Designing multi-responsive polymers using latent variable methods.* Polymer, 2014. **55**(2): p. 505-516.
2. MacGregor, J.F., K. Muteki, and T. Ueda, *On the rapid development of new products through empirical modeling with diverse data-bases*, in *Computer Aided Chemical Engineering*. 2006, Elsevier. p. 701-706.
3. Dogra, P., et al., *Mathematical modeling in cancer nanomedicine: a review.* Biomedical microdevices, 2019. **21**: p. 1-23.
4. Sheibat-Othman, N., et al., *Is modeling the PSD in emulsion polymerization a finished problem? An overview.* Macromolecular Reaction Engineering, 2017. **11**(5): p. 1600059.
5. Yu, L.X., *Pharmaceutical quality by design: product and process development, understanding, and control.* Pharmaceutical research, 2008. **25**(4): p. 781-791.
6. Giordano, A., A.A. Barresi, and D. Fissore, *On the use of mathematical models to build the design space for the primary drying phase of a pharmaceutical lyophilization process.* Journal of Pharmaceutical Sciences, 2011. **100**(1): p. 311-324.
7. Tomba, E., M. Barolo, and S. García-Muñoz, *General framework for latent variable model inversion for the design and manufacturing of new products.* Industrial & engineering chemistry research, 2012. **51**(39): p. 12886-12900.
8. Tomba, E., et al., *Latent variable modeling to assist the implementation of Quality-by-Design paradigms in pharmaceutical development and manufacturing: A review.* International journal of pharmaceutics, 2013. **457**(1): p. 283-297.

9. Stockdale, G.W. and A. Cheng, *Finding design space and a reliable operating region using a multivariate Bayesian approach with experimental design.* Quality Technology & Quantitative Management, 2009. **6**(4): p. 391-408.
10. N. Politis, S., et al., *Design of experiments (DoE) in pharmaceutical development.* Drug development and industrial pharmacy, 2017. **43**(6): p. 889-901.
11. Djuris, J. and Z. Djuric, *Modeling in the quality by design environment: Regulatory requirements and recommendations for design space and control strategy appointment.* International journal of pharmaceutics, 2017. **533**(2): p. 346-356.
12. Debevec, V., S. Srčič, and M. Horvat, *Scientific, statistical, practical, and regulatory considerations in design space development.* Drug development and industrial pharmacy, 2018. **44**(3): p. 349-364.
13. Destro, F. and M. Barolo, *A review on the modernization of pharmaceutical development and manufacturing–Trends, perspectives, and the role of mathematical modeling.* International Journal of Pharmaceutics, 2022. **620**: p. 121715.
14. Hasan, M.R., et al., *Application of mathematical modeling and computational tools in the modern drug design and development process.* Molecules, 2022. **27**(13): p. 4169.
15. Jaeckle, C.M. and J.F. MacGregor, *Industrial applications of product design through the inversion of latent variable models.* Chemometrics and intelligent laboratory systems, 2000. **50**(2): p. 199-210.
16. Facco, P., et al., *Bracketing the design space within the knowledge space in pharmaceutical product development.* Industrial & Engineering Chemistry Research, 2015. **54**(18): p. 5128-5138.
17. Paris, A., C. Duchesne, and É. Poulin, *Establishing multivariate specification regions for incoming raw materials using projection to latent structure models: comparison between direct mapping and model inversion.* Frontiers in Analytical Science, 2021: p. 7.
18. Chatterjee, S., C.M. Moore, and M.M. Nasr, *An overview of the role of mathematical models in implementation of quality by design paradigm for drug development and manufacture.* Comprehensive quality by design for pharmaceutical product development and manufacture, 2017: p. 9-24.
19. MacGregor, J.F., et al., *Data-based latent variable methods for process analysis, monitoring and control.* Computers & chemical engineering, 2005. **29**(6): p. 1217-1223.
20. Zhang, L. and S. Garcia-Munoz, *A comparison of different methods to estimate prediction uncertainty using Partial Least Squares (PLS): a practitioner's perspective.* Chemometrics and intelligent laboratory systems, 2009. **97**(2): p. 152-158.
21. Lorber, A., L.E. Wangen, and B.R. Kowalski, *A theoretical foundation for the PLS algorithm.* Journal of Chemometrics, 1987. **1**(1): p. 19-31.
22. Miyashita, Y., et al., *Comments on the NIPALS algorithm.* Journal of chemometrics, 1990. **4**(1): p. 97-100.
23. Kresta, J., T. Marlin, and J. MacGregor, *Development of inferential process models using PLS.* Computers & Chemical Engineering, 1994. **18**(7): p. 597-611.
24. Geladi, P. and B.R. Kowalski, *Partial least-squares regression: a tutorial.* Analytica chimica acta, 1986. **185**: p. 1-17.
25. Kucherenko, S., et al., *Computationally efficient identification of probabilistic design spaces through application of metamodeling and adaptive sampling.* Computers & Chemical Engineering, 2020. **132**: p. 106608.
26. Louchard, G., *Probabilistic analysis of adaptative sampling.* Random Structures & Algorithms, 1997. **10**(1-2): p. 157-168.
27. Bae, S., N.H. Kim, and S.-g. Jang, *Reliability-based design optimization under sampling uncertainty: shifting design versus shaping uncertainty.* Structural and Multidisciplinary Optimization, 2018. **57**(5): p. 1845-1855.
28. Phatak, A., P. Reilly, and A. Penlidis, *An approach to interval estimation in partial least squares regression.* Analytica chimica acta, 1993. **277**(2): p. 495-501.

29. Faber, K. and B.R. Kowalski, *Prediction error in least squares regression: Further critique on the deviation used in The Unscrambler.* Chemometrics and Intelligent Laboratory Systems, 1996. **34**(2): p. 283-292.
30. Denham, M.C., *Prediction intervals in partial least squares.* Journal of Chemometrics: A Journal of the Chemometrics Society, 1997. **11**(1): p. 39-52.
31. MacGregor, J.F., et al., *Data-based latent variable methods for process analysis, monitoring and control.* Computers & Chemical Engineering, 2005. **29**(6): p. 1217-1223.
32. Bano, G., et al., *Uncertainty back-propagation in PLS model inversion for design space determination in pharmaceutical product development.* Computers & Chemical Engineering, 2017. **101**: p. 110-124.
33. Van Huffel, S. and J. Vandewalle, *The partial total least squares algorithm.* Journal of computational and applied mathematics, 1988. **21**(3): p. 333-341.
34. Serneels, S., P. Lemberge, and P.J. Van Espen, *Calculation of PLS prediction intervals using efficient recursive relations for the Jacobian matrix.* Journal of Chemometrics: A Journal of the Chemometrics Society, 2004. **18**(2): p. 76-80.
35. Helland, I.S., *Partial least squares regression and statistical models.* Scandinavian journal of statistics, 1990: p. 97-114.
36. Faber, N.K.M., *Uncertainty estimation for multivariate regression coefficients.* Chemometrics and intelligent laboratory systems, 2002. **64**(2): p. 169-179.
37. Efron, B. and R.J. Tibshirani, *An introduction to the bootstrap*. 1994: Chapman and Hall/CRC.
38. Laky, D., et al., *An optimization-based framework to define the probabilistic design space of pharmaceutical processes with model uncertainty.* Processes, 2019. **7**(2): p. 96.
39. Hua, S., G. Qu, and S.S. Bhattacharyya. *Exploring the probabilistic design space of multimedia systems*. in *14th IEEE International Workshop on Rapid Systems Prototyping, 2003. Proceedings.* 2003. IEEE.
40. Bano, G., et al., *Probabilistic Design space determination in pharmaceutical product development: A Bayesian/latent variable approach.* AIChE Journal, 2018. **64**(7): p. 2438-2449.
41. Peterson, J.J. and M. Yahyah, *A Bayesian design space approach to robustness and system suitability for pharmaceutical assays and other processes.* Statistics in Biopharmaceutical Research, 2009. **1**(4): p. 441-449.
42. Bano, G., et al., *A novel and systematic approach to identify the design space of pharmaceutical processes.* Computers & Chemical Engineering, 2018. **115**: p. 309-322.
43. Peterson, J.J., *A Bayesian approach to the ICH Q8 definition of design space.* Journal of biopharmaceutical statistics, 2008. **18**(5): p. 959-975.
44. Yerramilli, S., et al., *Fully bayesian inference for latent variable gaussian process models.* SIAM/ASA Journal on Uncertainty Quantification, 2023. **11**(4): p. 1357-1381.
45. Seeger, M., *Bayesian Gaussian process models: PAC-Bayesian generalisation error bounds and sparse approximations.* 2003.
46. Ko, J. and D. Fox, *GP-BayesFilters: Bayesian filtering using Gaussian process prediction and observation models.* Autonomous Robots, 2009. **27**(1): p. 75-90.
47. Demis, P., S. Kucherenko, and O.V. Klymenko, *Design Space Approximation with Gaussian Processes*, in *Computer Aided Chemical Engineering*. 2021, Elsevier. p. 905-911.
48. Xing, W., et al., *Shared-gaussian process: Learning interpretable shared hidden structure across data spaces for design space analysis and exploration.* Journal of Mechanical Design, 2020. **142**(8): p. 081707.
49. Frigola, R., et al., *Bayesian inference and learning in Gaussian process state-space models with particle MCMC.* Advances in neural information processing systems, 2013. **26**.
50. Hebbal, A., et al., *Bayesian optimization using deep Gaussian processes with applications to aerospace system design.* Optimization and Engineering, 2021. **22**(1): p. 321-361.
51. Snoek, J., H. Larochelle, and R.P. Adams, *Practical bayesian optimization of machine learning algorithms.* Advances in neural information processing systems, 2012. **25**.

52. Frazier, P.I., *A tutorial on Bayesian optimization.* arXiv preprint arXiv:1807.02811, 2018.
53. Shahriari, B., et al., *Taking the human out of the loop: A review of Bayesian optimization.* Proceedings of the IEEE, 2015. **104**(1): p. 148-175.
54. Gramacy, R.B., H.K. Lee, and W.G. Macready. *Parameter space exploration with Gaussian process trees*. in *Proceedings of the twenty-first international conference on Machine learning*. 2004.
55. Lookman, T., et al., *Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design.* npj Computational Materials, 2019. **5**(1): p. 21.
56. Liu, Y., et al., *Machine Learning-Based Methods for Materials Inverse Design: A Review.* Computers, Materials & Continua, 2025. **82**(2).
57. Peterson, J.J., *A posterior predictive approach to multiple response surface optimization.* Journal of Quality Technology, 2004. **36**(2): p. 139-153.
58. Burges, C.J., *Geometric methods for feature extraction and dimensional reduction-a guided tour*, in *Data mining and knowledge discovery handbook*. 2010, Springer. p. 53-82.
59. Danhaive, R. and C.T. Mueller, *Design subspace learning: Structural design space exploration using performance-conditioned generative modeling.* Automation in Construction, 2021. **127**: p. 103664.
60. Karabutov, N., *Geometrical frameworks in identification problem.* Intelligent control and automation, 2021. **12**(2): p. 17-43.
61. Rogers, A. and M.G. Ierapetritou, *Mathematical tools for the quantitative definition of a design space*, in *Process Simulation and Data Modeling in Solid Oral Drug Development and Manufacture*. 2016, Springer. p. 225-279.

Chapter 2:

2 Predicting the Volume Phase Transition Temperature of Multi-Responsive Poly(N-isopropylacrylamide)-Based Microgels Using a Cluster-Based Partial Least Squares Modeling Approach

The contents of this chapter have been published in *ACS Applied Polymer Materials* .

Seyed Saeid Tayebi,† Elizabeth Keane,‡ Nahieli Preciado Rivera,† Todd Hoare,† and Prashant Mhaskar†

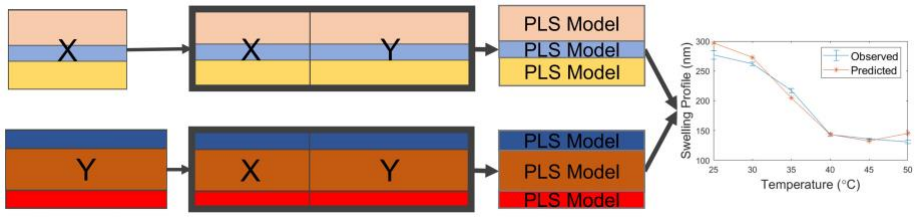
†*Department of Chemical Engineering, McMaster University, Hamilton, Ontario, Canada, L8S 4L7*

‡*Department of Materials Science Engineering, McMaster University, Hamilton, Ontario, Canada, L8S 4L7*

Authorship Contribution

Seyed Saeid Tayebi: Conceptualized the idea, developed the methodology, wrote the code, tested the models, experimentally synthesized the samples, wrote the manuscript draft, and addressed comments during the revision process. Prashant Mhaskar: Contributed to the conceptualization of the modeling approach, provided supervision, reviewed and contributed to the writing of the manuscript, and managed project administration and funding. Todd Hoare: Contributed to the conceptualization of the experimental approach, provided supervision, reviewed and contributed to the writing of the manuscript, and managed project administration and funding. Elizabeth Keane: Conducted preliminary studies exploring the potential application of Latent Variable Modeling (LVM) to this system. Nahieli Preciado Rivera: Provided training to Seyed Saeid Tayebi and Elizabeth Keane in sample preparation and contributed to the experimental synthesis of the samples.

Abstract



Despite the various potential applications of dual pH/temperature-responsive microgels, the multiple (and often interacting) physical and chemical factors that influence the volume phase transition temperature (VPTT) in such microgels make it challenging to directly design a microgel with a particular targeted swelling response. Herein, we address this challenge by designing and implementing a data-driven model that can predict a microgel swelling profile and subsequently VPTT based only on the microgel recipe. A clustering-based adaptation of partial least squares (PLS) modelling is developed and subsequently applied to a data library of pH 4 (fully protonated) and pH 10 (fully ionized) swelling responses of 32 pH/temperature-responsive poly(N-isopropylacrylamide) microgels functionalized with various carboxylic acidfunctionalized comonomers. We demonstrate that the best-performing clustering and data arrangement strategies can predict the VPTT of the microgels within 1.0°C at pH 4 and 2.4°C at pH 10, an accuracy similar to the uncertainty estimates from the experimental transition temperature data (0.6°C at pH 4 and 2.2°C at pH 10). Such an approach thus paves the way for faster customization of a microgel swelling profile as needed for a target application.

Keywords: pH-Temperature responsive microgels , VPTT , latent variable modelling, PLS PCA

2.1 Introduction

Smart microgels have attracted increasing interest given their ability to reversibly swell and de-swell in response to multiple stimuli including temperature, pH, light, ionic strength, and magnetic fields. The stimulus-specific swelling response of a smart microgel is thus considered a key characteristic property and dictates the implementation of such materials in various applications including drug delivery, oil

recovery, bio-sensing, separations, smart windows, cosmetics, and others [1-5]. Poly(N-isopropylacrylamide) (PNIPAM)-based microgels have attracted particular attention given that they exhibit a discontinuous volume phase transition temperature (VPTT) of 32°C in aqueous solution, close to physiological temperature [6, 7]. The resulting deswelling/swelling transitions observed upon increasing/decreasing the solution temperature, coupled with the corresponding changes in transparency, pore size, and interfacial hydrophobicity, directly lead to the broad applicability of these microgels [8, 9].

The incorporation of other functional comonomers into PNIPAM microgels can create a multi-responsive microgel in which the swelling/de-swelling transitions can be regulated by two or more external stimuli[10, 11]. The nature of the dual stimulus swelling response can be adjusted by changing the type/amount of comonomer used, the distribution of the comonomer within the microgel, and/or other polymerization conditions [10, 12-14]. Alternately, post-polymerization modification of a functionalized microgel can be used to introduce a functional group often not compatible with the precipitation-based free radical process typically used for microgel synthesis [15, 16].

Coupling temperature sensitivity with pH-responsiveness in microgels can unlock applications in sensing [17], drug delivery [18],catalysis [19-21], and tissue engineering or cell scaffolding [22]. The specific pH and magnitude of the pH-induced transition can be controlled by incorporating anionic or cationic functional comonomers with different pK_a values [11, 23]. However, given that both pH-induced ionization and temperature-induced volume phase transitions alter the hydrophilic/hydrophobic balance of the microgel, the pH and temperature transitions do not occur fully independently; pH ionization significantly increases the temperature at which the VPTT can occur, while the thermal transition can significantly enhance the magnitude of pH swelling that is possible for a given pH change/microgel composition [24]. In this context, while the dual pH/temperature-responsive swelling responses of a multi-functional microgel can offer benefits in various applications (e.g. targeting drug delivery to cancer tumors characterized by lower pH values and higher temperatures than normal tissues [9, 25-28]), *a priori* identification of a multi-functional microgel recipe that can exhibit a targeted swelling/de-swelling profile

with a specific VPTT is challenging since the swelling response of the microgel to one stimulus directly affects its response to the second stimulus [10, 12, 26, 29].

One approach to predicting the VPTT in such microgels would be to utilize first principles approaches that leverage existing thermodynamic and kinetic frameworks. However, the significant nanoscale heterogeneity in both crosslinker and functional monomer distributions as well as the occurrence of microphase separation dependent on those local distributions makes the first principles prediction of microgel swelling responses much more challenging than it is for bulk hydrogels[13, 30]. Kinetic modelling approaches have been used to link radial functional group distributions in microgels with the copolymerization kinetics of the constituent comonomers [31][32], with the local compositional knowledge then used in conjunction with Flory-Huggins equilibrium swelling theory to predict relative swelling responses in multi-functional microgels [32, 33]. However, the implementation of such models requires either measuring or estimating a wide range of parameters including copolymerization ratios between all monomers/cross-linkers used, parameters that can be complex to measure (particularly as the number of comonomers in a microgel is increased, as is often the case in multi-responsive microgels) and often require dedicated experimental setups and/or equipment [31, 34, 35]. Furthermore, the occurrence of local microphase separation and/or the quantitative effects of pH changes on the Flory-Huggins parameter are challenging to experimentally assess, both of which are required for accurate prediction of phase transition behaviors in microgels.

This diversity of challenges associated with the use of first-principal models has raised interest in utilizing purely data-driven modelling techniques for predicting microgel swelling responses. In this regard, principal component analysis (PCA) and partial least squares (PLS) have gained traction for tackling materials-based optimization problems. PCA arranges data in one block and reduces dimensionality by placing a higher weight on variables with higher variance. In contrast, PLS arranges data into two blocks and prioritizes input variables that have a major impact on the output responses. Such datadriven models have previously been applied to a variety of materials design challenges, including to design polyethylene

blown film resins with superior performance [36], develop soft sensors to predict the melt flow index of a polymerization reactor's output [37], and apply ATR-FTIR-based methods to estimate the viscosity of polymer solutions [38]. Closer to the context of smart microgels, PLS modelling approaches have been applied to optimize the cloud point, molecular weight, and percentage yield of dual thermo/photoresponsive polymers that combine the temperature-responsiveness of PNIPAM with the light-responsiveness of cinnamate functional groups [39]. MacGregor et al. [40, 41] have also developed a variation of PLS modelling called multi-block PLS that takes parallel blocks of data (including previously synthesized recipes and the raw material properties) into account to accelerate the design and production of new industrial products. In light of the success of such previous efforts, we hypothesize that data-driven modelling can address the challenge of predicting swelling responses and thus VPTT values for multi-responsive microgels.

Herein, we develop and validate a method to predict the swelling profiles and VPTT values of multi-responsive microgels using a partial least squares modelling technique that incorporates a clustering step in which the data is first clustered into groups with either like polymerization recipes or like swelling responses before building the PLS models. In particular, to avoid the the necessity of knowing the swelling profile to cluster recipes based on swelling responses, a new algorithm denoted as 'combined clustering mode' is developed. Using a validation dataset of fully protonated (pH 4) and fully ionized (pH 10) swelling profiles for 32 different carboxylic acid-functionalized PNIPAM-based microgels, the resulting model is demonstrated to accurately predict both the thermal phase transition temperature as well as the absolute microgel particle sizes as a function of pH and/or temperature (or, at minimum, explicitly identify recipes that cannot be well predicted). This approach thus enables the accelerated design of microgel recipes that can achieve new types of swelling profiles targeted to specific applications without requiring lengthy trial-and-error syntheses or any specific fundamental or phenomenological insight into the dynamics or microstructure of microgels, both of which become prohibitively complex when multiple functional comonomers are added to achieve customized swelling profiles.

2.2 Experimental

2.2.1 Materials

N-isopropylacrylamide (NIPAM) (Sigma-Aldrich, 97%) was purified by recrystallization with 60:40 toluene:hexane mixture. N-N'-methylene(bis)acrylamide (MBA) (SigmaAldrich, 99%), acrylic acid (AA) (Aldrich, 99%), methylacrylic acid (MAA) (Aldrich, 99%), vinylacetic acid (VAA) (Aldrich, 97%), fumaric acid (FA) (Sigma-Aldrich, 99%), maleic acid (MA) (Sigma-Aldrich, 99%), 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. MilliQ-grade water (>18Ω resistance) was used for all experiments.

2.2.2 Microgel Synthesis

The raw recipes used to fabricate each microgel in the dataset are shown in Supporting Information Table S1, while the resulting mole fractions of each monomer component of the microgels fabricated (the input data used to build the models) are shown in Table 2-1. For any given recipe, the required amounts of NIPAM, MBA, SDS and functional monomer(s) were added into a 250 mL round-bottom flask with 150 mL of MilliQ water. The mixture was purged under nitrogen for 30 minutes at room temperature before being transferred to the oil bath at 70°C, with continuous nitrogen purging maintained. To initiate the polymerization, 0.05 g APS was mixed with 10 mL of MilliQ water and delivered to the flask via a syringe. The reaction was run for 4 hours at 70°C under magnetic mixing (magnet size 4cm) at 160 RPM. Subsequently, the reaction mixture was cooled and subjected to 6 x 6 hour cycles of dialysis to remove surfactant and any unreacted monomers. The resulting microgel suspension was lyophilized to dryness and stored at room temperature.

Table 2-1: Mole fractions of all recipe components used to synthesize the microgel library

Sam ID	NIPAM	MBA	AA	MAA	FA	MA	VA	SDS
1	0.870	0.064	0.055	0	0	0	0	0.012
2	0.875	0.064	0	0	0.049	0	0	0.012

3	0.731	0.054	0	0	0	0	0.206	0.010
4	0.784	0.058	0	0	0	0	0.147	0.011
5	0.847	0.062	0	0	0	0	0.079	0.012
6	0.867	0.064	0	0	0	0	0.057	0.012
7	0.889	0.065	0	0	0	0	0.033	0.012
8	0.832	0.078	0	0	0	0	0.078	0.012
9	0.862	0.045	0	0	0	0	0.081	0.012
10	0.874	0.032	0	0	0	0	0.082	0.012
11	0.886	0.019	0	0	0	0	0.083	0.012
12	0.851	0.062	0	0	0	0	0.080	0.007
13	0.853	0.063	0	0	0	0	0.080	0
14	0.857	0.063	0	0	0	0	0.080	0
15	0.834	0.061	0.093	0	0	0	0	0.012
16	0.864	0.063	0	0	0.060	0	0	0.012
17	0.857	0.063	0	0.080	0	0	0	0
18	0.738	0.054	0	0	0	0	0.208	0
19	0.866	0.070	0	0	0.059	0	0	0.005
20	0.753	0.090	0	0	0.147	0	0	0.011
21	0.833	0.083	0	0	0.025	0	0.047	0.008
22	0.898	0.067	0	0	0	0	0.030	0.005
23	0.784	0.089	0	0.051	0.071	0	0	0.005
24	0.891	0.068	0	0	0	0	0.031	0.010
25	0.784	0.088	0	0	0.121	0	0	0.007
26	0.892	0.071	0	0.008	0	0	0.025	0
27	0.888	0.070	0	0	0	0	0.033	0.009
28	0.792	0.089	0	0	0	0	0.114	0
29	0.732	0.054	0	0	0	0.204	0	0.010
30	0.849	0.062	0	0.076	0	0	0	0.012
31	0.932	0.068	0	0	0	0	0	0
32	0.756	0.090	0	0	0.114	0	0.036	0

2.2.3 Swelling Profile Measurement

The particle sizes of each microgel as a function of pH and temperature were measured using dynamic light scattering (Brookhaven 90Plus) operating at a scattering angle of 90 degrees. Particle size was measured at 6 different temperatures (25°C, 30°C, 35°C, 40°C, 45°C, 50°C) in 10 mM KCl solutions adjusted using 0.1 M HCl or NaOH to either pH 4 (fully protonated state) or pH 10 (fully ionized state). Five independent z-average particle size measurements were collected at each pH/temperature tested, with the average of the repeat measurements reported as the microgel particle size. Note that all tested

microgels exhibited a unimodal particle size distribution at both pH values tested. Particle sizes at various temperatures were then used to calculate VPTT values for the synthesized microgels at each pH value by fitting a sigmoidal curve to the experimental data, with the mid-point of the curve identified as the volume phase transition temperature (see Figure S1 for an illustration of how the transition temperature was calculated for one of the synthesized samples). Error bars associated with the VPTT measurements were estimated by fitting sigmoidal curves to the top and bottom of the error bar ranges of each individual particle size measurement and subtracting the maximum and minimum VPTT estimate stemming from these fits.

2.3 Modelling Frameworks

The authors developed all modelling code and implemented the code in MATLAB R2021b.

2.3.1 Principal Component Analysis (PCA)

PCA works with only one block of data and thus does not capture any relationships between variables in different blocks of data. Nonetheless, it is commonly used as a pre-processing tool in which the correlation between the predictors among the block of data can be captured and, ultimately, the number of variables can be reduced to facilitate faster and more reliable modelling [42, 43].

The directions in which the observations are distributed among the data space play a pivotal role in dimensionality reduction. These directions can be captured and then explained using a linear combination of the original predictor variables with weighted coefficients (i.e. PCA components). Since this linear combination of the original predictors indicates the correlation among the predictor variables, if such directions are ordered based on the variance of the original block, it is possible to select the first two, three or more directions that explain more variance than the other directions based on the level of correlation among the columns of the dataset and work with them as opposed to working with the original predictor variables.

2.3.2 Partial Least Squares (PLS)

PLS is analogous to applying two PCA analyses on two blocks (X and Y) simultaneously; as such, opposed to PCA, PLS aims to capture relationships between more than one block of data. Unlike with PCA modelling, in which explaining the maximum variance in each block is the primary goal, PLS modelling instead aims to arrange the data in a way that maximizes the correlation between the calculated scores in X and Y [44]; in other words, instead of covering the maximum variance in the X and Y spaces, PLS instead aims to capture the maximum variance in the Y block that can be explained by the variances in the X block [43, 45].

2.3.3 Clustering Procedure (K-means Algorithm)

Clustering is a non-supervised classification approach that uses only the observation block of data when there is no label vector or response matrix available. Clustering is one of the few data mining tools that is applicable for data pre-processing as it does not require training and works based on two key ideas [46, 47]:

a) The members of one cluster must be as similar to each other as possible; and
b) The members of one cluster must be as different from those of the other clusters as possible.

A wide range of procedures including partitioning methods, hierarchical clustering methods, densitybased clustering methods, and grid-based methods have been introduced to identify the similarity or differences among the observations and thus enable optimal clustering of the samples [48-51]. The Lioyd algorithm is used in this work given that it is a partition method that can be applied easily using the K-means technique [47, 52], which involves the iterative application of three steps until convergence is achieved:

1. Randomly initializing the center point of the clusters;
2. Distributing samples among the clusters by clustering each sample to its nearest center; and
3. Updating the centers' position using the average values of each cluster's members.

The last two steps are repeated until there are no changes in the members of each cluster [52-54]. Figure S2 schematically illustrates the implementation of K-means on an cluster-based dataset to determine the right clustering index, showing how fast the repetition of the last two stages in each iteration enables clustering of the data with high precision.

2.4 Methodology Development, Results and Discussion

2.4.1 Data Pre-Processing

Prior to analyzing the data, the available dataset was pre-processed to identify outliers within the dataset. The mean-centered and scaled values of the data reported in Table 2-1 were first calculated, values that are compiled in Table S2; a positive value for a variable indicates a value that is higher than the average value of that variable among all observations while a negative value indicates the opposite. A PLS model was then applied on the dataset to correlate the X (recipe) block to the Y (swelling) block, after which the dimensionally-reduced blocks of data were used to estimate the original blocks of data and the squared prediction error (SPE) (the difference between each predicted observation value and its experimental value) was compared based on the SPE 95% confidence limit of the corresponding block calculated using eq. (1):

$$SPE_{lim} = \frac{\nu}{2m}\chi^2_{(\frac{2m^2}{\nu},0.95)} \quad (eq.1)$$

Here, m and v are the SPE mean and variance calculated for all observations in each block and χ^2 is the chi-square distribution with ($\frac{2m}{\nu}^2$) degrees of freedom and a significance level of 95% [55]. Observations with higher SPE than SPE_{lim}(either in the recipe space or in the swelling space) were classified as outliers given that the covariance among the outlier samples differs from rest of the dataset and would thus significantly skew the predictive power of the model. Four recipes (as shown in Figure S3) were identified as outliers following this analysis and thus were removed from subsequent analysis. Observations 31 and 32 are both large microgels prepared with either a very small comonomer fraction

(observation 31) or a high concentration of cross-linker (observation 32). Both these conditions have been previously noted to lead to an increased probability of microgel aggregation, a phenomenon comfirmed experimentally by the observation that the swollen state (25°C) particle size at pH 4 exceeds that at pH 10 for both samples; such a trend is not physically realistic in the absence of microgel aggregation given the key role of functional monomer ionization in driving microgel swelling responses [1]. Alternately, observations 29 (prepared with the highest concentration of MA) and 30 (prepared with a high concentration of MAA) were both microgels containing higher concentrations of functional comonomers for which there are few observations in our dataset. In these two cases, the model is provided with less relevant "learning" data for accurately predicting the recipe formulation of these microgels; in this case, unlike with observations 31 and 32, adding more microgels to the dataset that include these comonomers at higher concentrations may lead to the ultimate inclusion of these samples within the model, although that is outside of the scope of the current work. The remaining 28 microgels that passed the pre-screening test were thus used as the dataset for all subsequent analyses.

2.4.2 Clustering-Based PLS Modelling

To account for potential differences in the swelling responses when different comonomers and/or different mole fractions of functional comonomers are used to prepare microgels, an adaptation of PLS-based modeling is developed to cluster the dataset prior to applying the PLS model. Three clustering policies were evaluated:

a. No clustering, using the X block as the model input and the Y block as the model output directly (Figure 2-1(A)).
b. Recipe-based clustering in which clustering was applied on the input (recipe) variables (X) and PLS was then performed on each set of observations categorized in the same cluster based on similar microgel recipes (Figure 2-1(B)).
c. Swelling response-based clustering in which clustering was applied on the output (swelling) variables (Y) and PLS was then performed on each set of observations categorized in the same cluster based on the microgel swelling profiles. While this clustering approach has the obvious

disadvantage of requiring data specific to the swelling profiles for a given microgel recipe to be available to perform clustering, we will address this challenge later in this manuscript that enables the practical implementation of this policy (Figure 2-1(C)).

To cluster either the X or Y block, a PCA model is first applied to the targeted block to extract the main trends in the dataset (including up to four first PCA scores), after which the K-means strategy was used to classify which observations should be in which cluster. Columns of the PCA score matrix that are the most

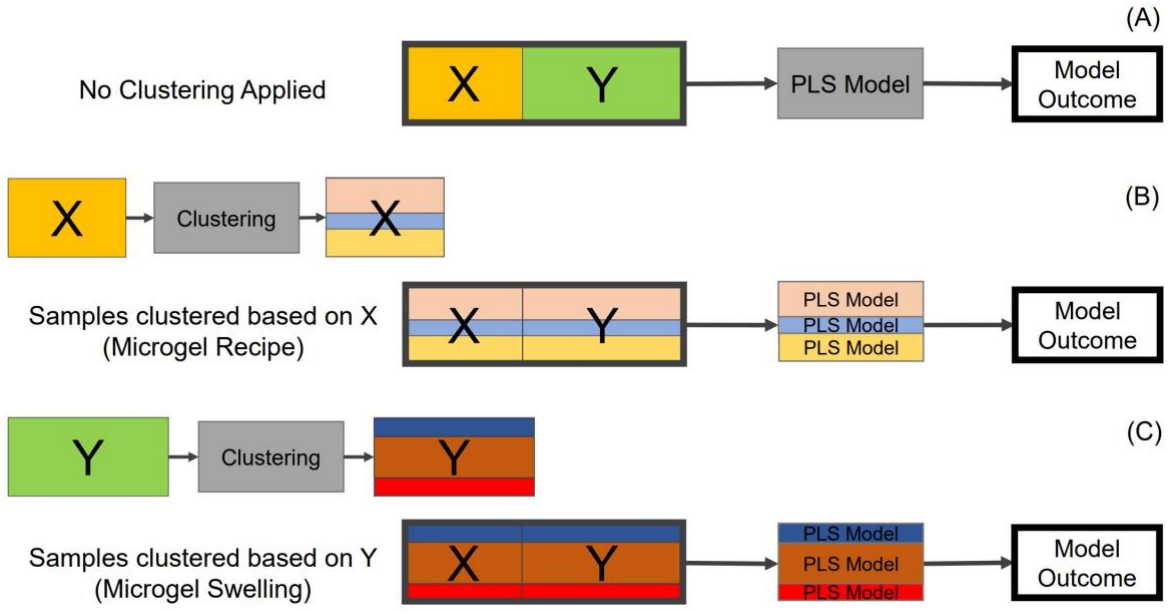


Figure 2-1: Data arrangements for PLS modelling of microgel swelling: (A) no clustering is applied(one PLS model is applied on the whole observation set), (B) observations are clustered based on X (microgel recipe), with one PLS model applied per cluster, and (C) observations are clustered based on Y (microgel swelling response), with one PLS model applied per cluster

discriminatory among the given blocks of data were then selected to optimize the utility of the clustering approach. Figure 2-2(A and B) shows that the 1st and 2nd score vectors are sufficient to provide a linearly separable dataset either in the recipe space (X) or in the swelling profile space (Y) among the different clusters; as such, only the two first scores from the PCA analysis need to be used to inform clustering in this work. Of note, given that the K-means final clustering index can be affected by the initial guesses of the centers at the first stage, each clustering process was repeated 20 times and only repeated indexing

patterns occurring in at least 85 percent of the runs were selected for further analysis. The number of clusters resulting from this analysis was managed by first setting a maximum of 10 clusters for each policy but coding the model such that if any cluster includes only one sample that cluster should be omitted, the maximum number of clusters should be reduced by one, and K-means should be re-applied on the targeted block. By applying this condition, the number of clusters required to discriminate differences in both the X and Y spaces was equal to 3 in 100 percent of cases. Figure 2-2(A and B) shows the distribution of these clusters in the

corresponding score plot.

For the recipe based clustering policy, the presence or absence of vinylacetic acid (VAA) in the recipe plays a pivotal role in determining the host cluster (i.e. the cluster into which a particular observation is classified). Cluster 3 primarily contains high VAA content microgels while Cluster 2 contains predominantly microgels prepared without any VAA comonomer (Figure 2-2(C)). The apparently different swelling profiles of the VAA-containing microgels are consistent with the unique reactivity of VAA among all the comonomers tested, as VAA reacts primarily as a chain transfer agent rather than a comonomer and is thus localized primarily at chain ends toward the outer periphery of the microgel [24]. However, while this correlation between VAA content and hosting cluster shows how the model can reflect fundamental microgel properties (e.g. the chemistry of the comonomers), other recipe components can change the optimal hosting cluster even for VAA copolymer microgels, particularly when lower VAA contents are used. Thus, while clustering based only on recipe components (i.e. whether or not VAA is used as a comonomer) would improve prediction quality versus not clustering at all, better predictions can be drawn from the recipe score plot information that takes into account all the interacting contributions of different recipe components on the swelling response. Better qualitative correlations can be drawn between the overall size of the microgel across the full range of the swelling profiles observed and the identified home cluster of that microgel. Specifically, if a representative Summative Size Parameter is computed according to eq. (2) below:

$$Summative\ Size\ Parameter = \sum D_i \quad (eq.2)$$

where D_i is the observed particle size for each sample at each measured temperature and pH value (which were the same for each observation to enable direct comparisons), the observations are unambiguously sorted into clusters based on the particle size across the full swelling profile (Figure 2-2(D)). As such, the cluster into which each microgel is classified can be qualitatively correlated with the microgel recipe or (more rigorously) the size of the microgels across the full dataset.

2.4.3 Suggested Data Arrangements

In addition to clustering observations with like properties together to improve predictability over a broad range of potential microgel recipes/swelling responses, different data arrangement strategies were also investigated to further improve the model predictions. In the available dataset for each sample, the X (recipe) block contains 8 concentration variables including N-isopropylacrylamide (NIPAM, the temperature responsive monomer), N,N'-methylene(bis)acrylamide (MBA, the crosslinker), sodium dodecyl sulfate (SDS, the surfactant used to control particle size), and five different pH-responsive functional monomers (acrylic acid (AA), methacrylic acid (MAA), fumaric acid (FA), maleic acid (MA), and vinylacetic acid (VAA) while the Y (swelling response) block includes 12 values corresponding to the equilibrium particle sizes at six temperatures (25°C, 30°C, 35°C, 40°C , 45°C, 50°C) at two different pH values (pH 4, the fully protonated state, and pH 10, the fully ionized state). A general view of both input (recipes) and output (swelling profile) variables in this data arrangement is shown schematically in Figure 2-3(A).

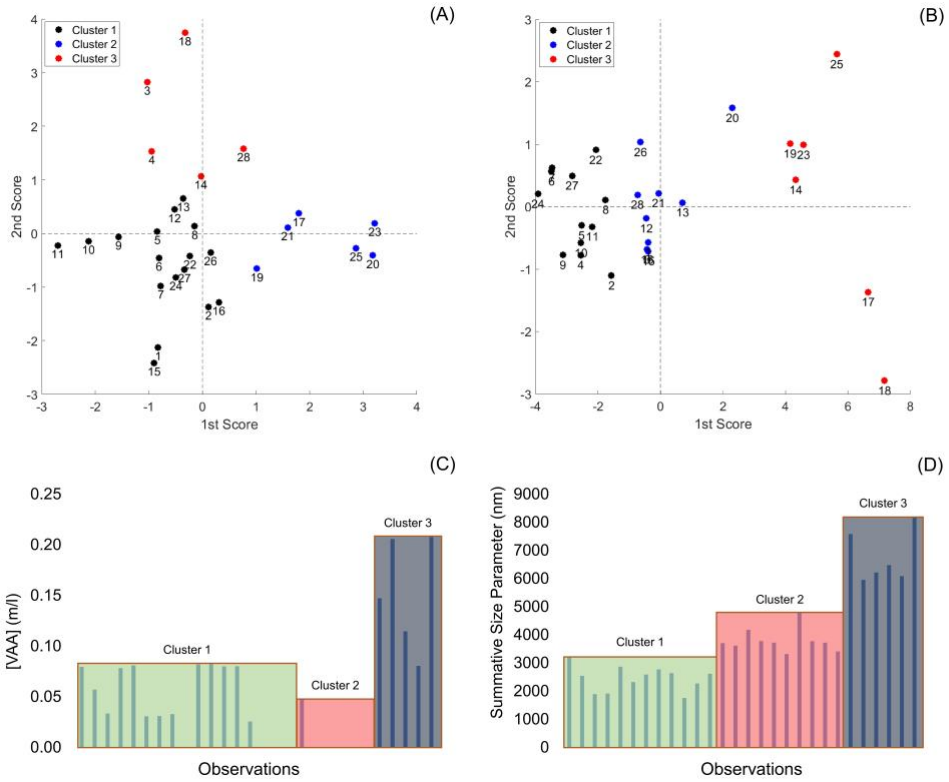


Figure 2-2: Data distribution among different clusters based on both microgel recipes (A) and microgel swelling profiles (B). Organization of microgel observations into clusters as classified based on (C) [VAA] and (D) Summative Size Parameter values among all cluster members. The number of clusters required to separate the observations in the latent variable space was determined by the model under both policies.

Considering the Y data all in a single block (a Type 1 arrangement, as per Figure 2-3(A), neither temperature nor pH was explicitly included in the input or output blocks of the variables considered, making it impossible for the model to predict swelling responses at temperatures or pH values other than those represented directly within the dataset. To incorporate these variables more directly in the analysis and permit the explicit prediction of temperature and/or pH-responsive swelling profiles at any pH/temperature value and/or improve the prediction accuracy of swelling profiles at specific measured pH/temperature values, four additional data arrangement formats were proposed:

- The Type 2 data arrangement divides the Y block into two sections corresponding to the two pH values at which the swelling profiles were recorded, with PLS modelling subsequently applied on each section separately (8 input variables, 6 output variables per PLS model - Figure 2-3(B). This data arrangement directly assesses the potential benefits of pH-based separation of the dataset

on prediction accuracy but does not allow for the prediction of the swelling profile of microgels at different pHs and temperatures beyond those directly tested.

- The Type 3 data arrangement adds an extra column to the input block corresponding to the temperature at which each particle size was observed (Figure 2-3(C)). In this case, the output block holds only 2 values representing the particle sizes measured at pH = 4 and pH = 10 at each given temperature, with the number of observations correspondingly increased by a factor of six to account for the six measurement temperatures used. Such a data arrangement would make it possible to predict microgel particle size at any desired temperature (not limited to the six observation temperatures) but cannot predict swelling at different pH values.
- The Type 4 data arrangement breaks the Y block of the Type 3 data arrangement into two parts based on the pH while keeping the X block the same (Figure 2-3(D)). In this arrangement, the input block includes 9 values (the 8 recipe variables plus temperature) but the number of variables in the output block is only 1 (the particle size at a specific temperature/pH value). The two parts are then combined to cover the whole original observation space (i.e. both pH values). This arrangement allows for any potential benefits (if any) of the pH-based separation of the Y block of the Type 3 arrangement to be realized but cannot predict pH swelling responses at pH values other than 4 and 10.
- The Type 5 data arrangement expands the input block to 10 variables in which the two last columns include the corresponding temperature and pH values in addition to the 8 recipe variables; the Y block in this case thus includes only one value (the particle size at each specific pH/temperature value) (Figure 2-3(E)). In this case, the number of observations is increased by a factor of 12 to cover the entire dataset (i.e. one observation for each individual pH/temperature combination at which the particle size was measured). In this arrangement, if adequate experimental data is available, it would be possible to estimate microgel particle size at any temperature or pH value, not just those at which particle size was specifically measured.

Note that while the Type 3, 4, and 5 data arrangements can in principle predict the swelling profile at temperature and/or pH values other than those explicitly measured, to assess the benefits of different data arrangements for improving prediction accuracy we will limit our analysis only to those explicitly measured pH and temperature values such that comparisons with Type 1 and Type 2 data arrangements can be unambiguously made. Refer to the SI section 8.3 for more details on how PLS models were applied

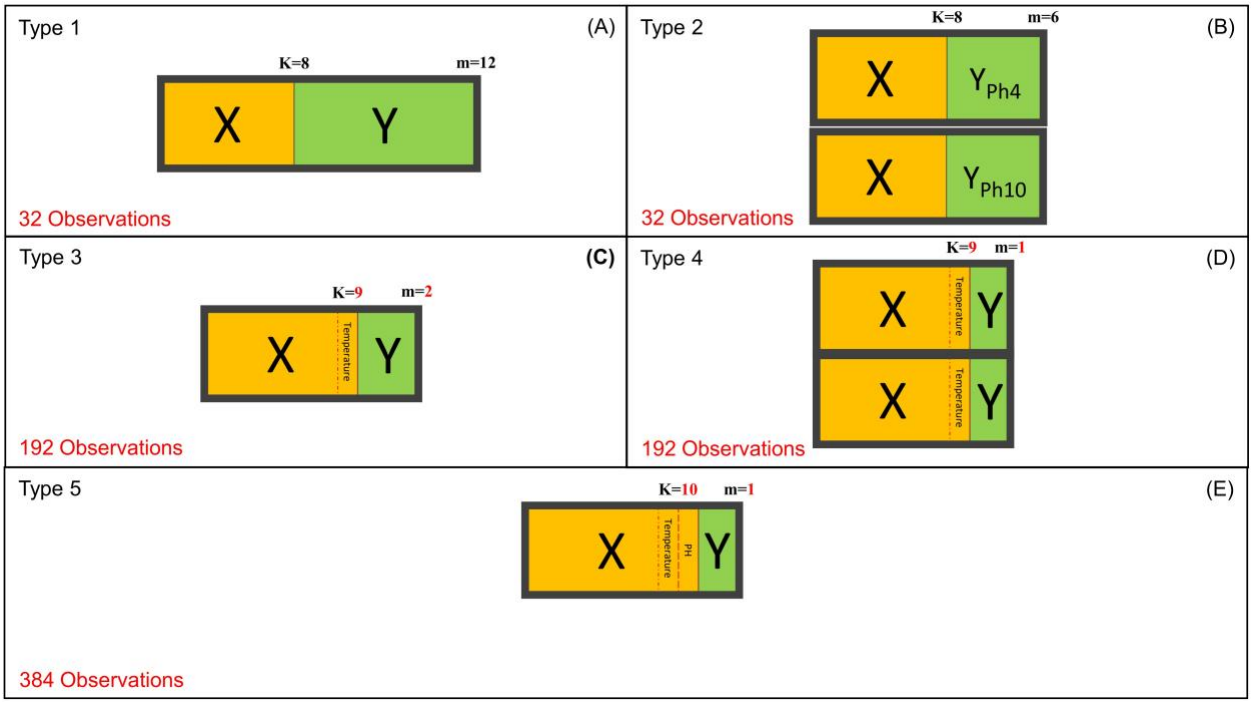


Figure 2-3: Five data arrangements used in this work from Type 1 to Type 5 are represented by (A) to (E) respectively (K represents the number of input variables and m denotes the number of output variables). using these different data arrangements to generate non-biased comparisons between the different data arrangement types.

2.4.4 Swelling Profile Prediction

The observed (experimental) versus fitted particle size plots shown in Figure 2-4 allow for the quantitative assessment of the impacts of the different clustering policies and/or data arrangements on the accuracy of microgel swelling predictions. Each row represents the impact of different clustering policies under the same data arrangement policy while each column represents the impact of different data

arrangement policies under the same clustering policy. As observed in the Figure 2-4, applying clustering significantly better concentrates the results around the bisector of the plot, meaning that the predicted particle size values were closer to the experimentally observed values. Moreover, Y-based clustering gives a slightly better prediction of particle size relative to X-based clustering under the same data arrangement format, as indicated both by visual observation as well as the R^2 and MSE parameters for all 15 cases (also tabulated in each subpanel of Figure 2-4).

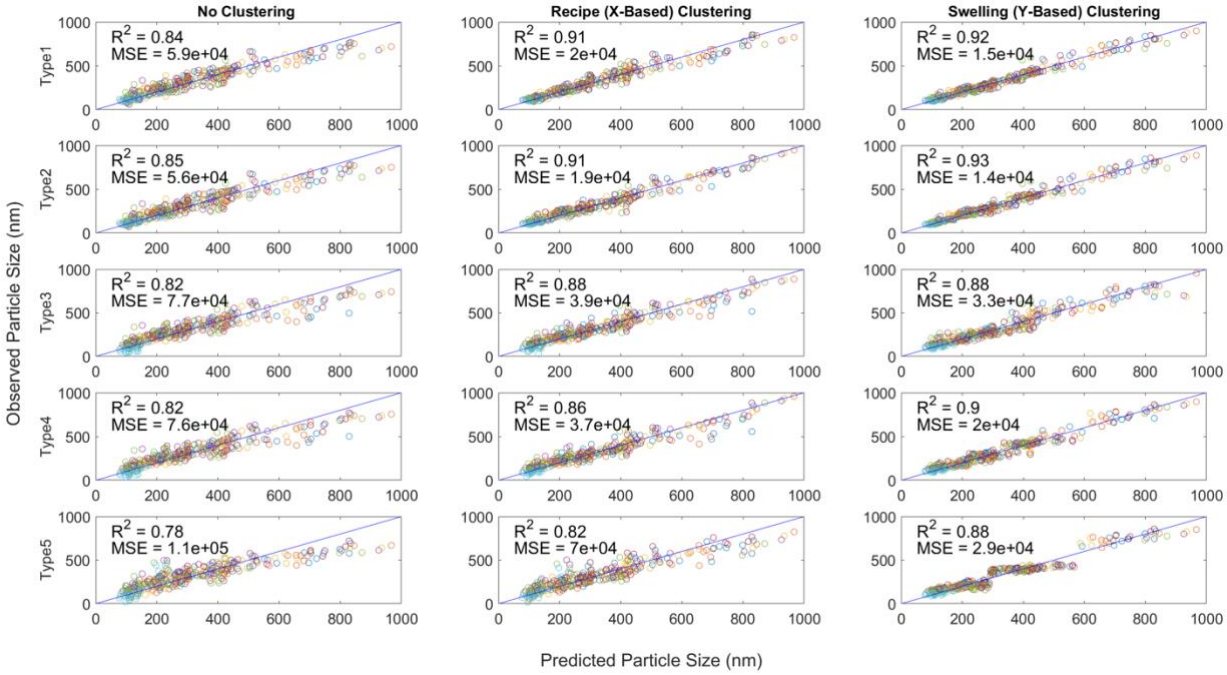


Figure 2-4: Experimentally observed (y-axis) versus model-predicted (x-axis) values of microgel particle sizes across all pH temperature values tested for each observation in the dataset using different data arrangements (rows) and different data clustering policies (columns)

Experimentally observed (y-axis) versus model-predicted (x-axis) values of microgel particle sizes across all pH temperature values tested for each observation in the dataset using different data arrangements (rows) and different data clustering policies (columns).

2.4.5 VPTT Prediction

Following the demonstrated potential of the model to predict particle size at different pH/temperature conditions, the clustering policies and data arrangement formats were next evaluated in terms of their ability to predict microgel volume phase transition temperatures. Note that the optimal clustering policies and/or data arrangements for the VPTT prediction may be different from those identified for the particle size predictions given that not all particle sizes as a function of temperature are equally important for estimating the VPTT; in particular, the VPTT prediction model emphasizes the accuracy of the particle size measurements just above and just below the VPTT whereas points in the plateau regions (although important to predict for the swelling model) are less important. In this context, and in light of the very different shapes of the particle size versus temperature curves of microgels in the fully protonated regime at pH 4 (in which the microgels have more similar transition temperature responses and consistently show upper and lower temperature range plateaus corresponding to the fully swollen and fully collapsed state) and the fully ionized regime at pH 10 (in which some microgels exhibit no phase transition response whatsoever), the capability of the model to predict the transition temperature was investigated separately for each pH value.

- *Transition Temperature Prediction at pH=4*

Figure 2-5(A to C) show the VPTT prediction error stemming from the various clustering strategies for all microgels in the dataset at pH=4, using the best data arrangement format under each of the clustering policies as identified in Table S5. The reported ’Transition prediction parameter’ in Table S5 refers to the percentage of observations for which the model can accurately predict the existence of a discrete transition temperature rather than a more continuous or non-existent transition within the tested temperature range. Good estimates of the VPTT of the microgels can be achieved using Y-based Clustering strategy, enabling a 96 percent accuracy in predicting the presence of a discrete phase transition and an average VPTT prediction error of just 1.8°C; furthermore, only 4 of the 28 total

microgels in the dataset had a more than 4°C difference between their actual and predicted transition temperature values.

While Y (swelling profile)-based clustering provided the best predictive potential, its direct implementation requires knowledge of the actual swelling profile values to find the right hosting cluster - information that is not available for a new recipe without actually synthesizing and analyzing the microgel. To achieve a *priori* predictions of VPTT without requiring the particle size data in advance, a combination between recipe and swelling-based clustering named Combined Clustering was developed. Using this strategy, the swelling profile of a new microgel is first predicted using the recipe-based clustering approach, enabling the initial identification of potential hosting clusters for a given new observation. Subsequently, a new swelling profile is predicted using swelling-based clustering and the hosting clusters for the already predicted swelling profile values are determined in order to ensure that the hosting cluster for the predicted swelling profile matches the cluster that was used to predict the swelling profile. If the host clusters match, the value is reported; if the host clusters do not match, a new swelling profile is predicted using the current hosting clusters and swelling-based PLS models and the process is iterated until the hosting clusters match. Figure S4 provides a schematic of how the Combined Clustering approach can predict the correct Y-based cluster for a new microgel observation without requiring the actual swelling data for a given observation.

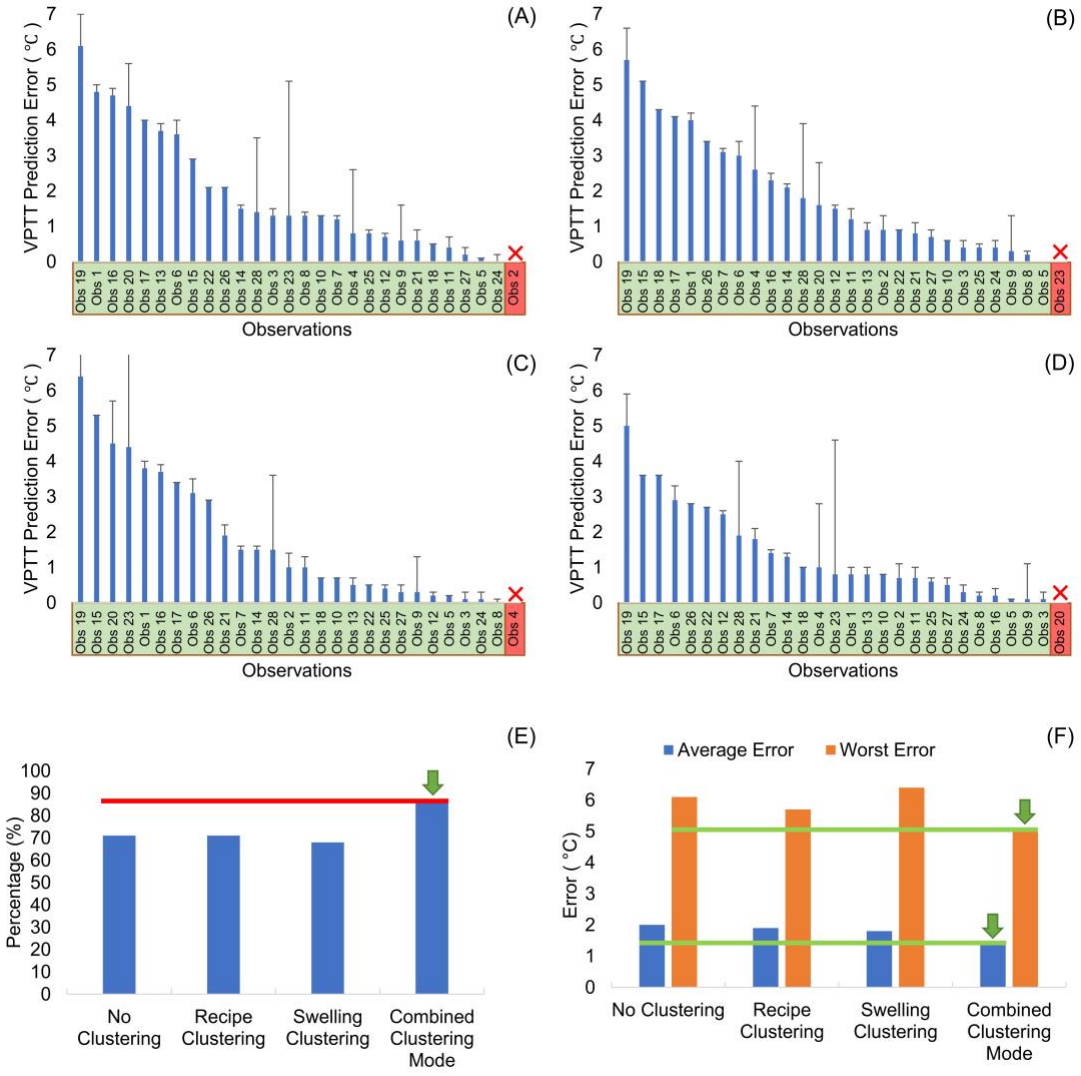


Figure 2-5: Prediction error of the model for predicting the microgel volume phase transition temperature at pH=4 for each individual microgel in the dataset using (A) no clustering, (B) recipe (X-based) clustering, (C) swelling (Y-based) clustering, and (D) combined clustering, in each case using the best-performing data arrangement for each clustering policy. Observations highlighted in green are correctly predicted to either have or not have a discrete phase transition while observations highlighted in red are ones for which the existence of VPTT was mis-predicted; (E, F) comparison of VPTT prediction accuracies achieved using each clustering policy in A-D based on (E) the percentage of microgels for which the experimental VPTT is predicted within <3°C and (F) the average and worst-case prediction errors observed.

Figure 2-5(D) shows the prediction accuracy using Combined Clustering policy to predict VPTT values at pH 4. As shown, using this approach allows for significantly more accurate predictions of the microgel transition temperature than can be achieved by clustering the data directly using the swelling data. The average error per observation using Combined Clustering is only 1.4 °C (close to the mean $\approx 0.6C$ experimental error associated with estimating the VPTT from the fitting procedure described), with 93

percent of the microgels in the dataset having VPTT values predicted within an accuracy of 3.6°C; indeed, the worst predicted VPTT value is <5°C away from the experimental VPTT of that microgel.

To summarize the fitting results, Figure 2-5(E) shows to the percentage of observations for which the VPTT is predicted as close as 3°C using each of the clustering policies in (A) to (D) while Figure 2-5(F) provides a visual comparison of the average VPTT prediction error and the worst case VPTT prediction error achieved with each clustering policy used. Combined Clustering achieves both improved average and worst-case prediction errors relative to the other reported clustering options; however, this clustering approach cannot accurately predict VPTT values for four microgels in the dataset (i.e. the VPTT prediction error is >3°C or the existence of a VPTT is mis-predicted). As such, for practical use of the model, it would be beneficial if we could at least predict whether an accurate VPTT prediction is likely to be possible or not with a given microgel recipe. To make this assessment, the recipe score plot of the full dataset indicated that samples that give good VPTT predictions (VPTT prediction error <3°C, Figure 2-6 green dots) can be spatially distinguished from samples with poor VPTT predictions(Figure 2-6 red dots). In this context, we can (based only on the microgel recipe) explicitly identify whether a new recipe would fall within the "well-predicted area", in which the average and worst prediction errors are 1.0°C and 2.9°C respectively and 100 percent accuracy is expected in the Transition Prediction Parameter, or the "poorly-predicted area", in which the model cannot be expected to provide a valid prediction. As such, while not every microgel VPTT can be predicted accurately depending on how similar a new recipe is to those in the existing dataset, we can predict with high confidence whether or not a VPTT prediction from the model is *expected* to be accurate based on the location of a new recipe within the latent variable space. Note that such information would be very beneficial for the next stage where a recipe is needed to be determined to yield a particular target profile.

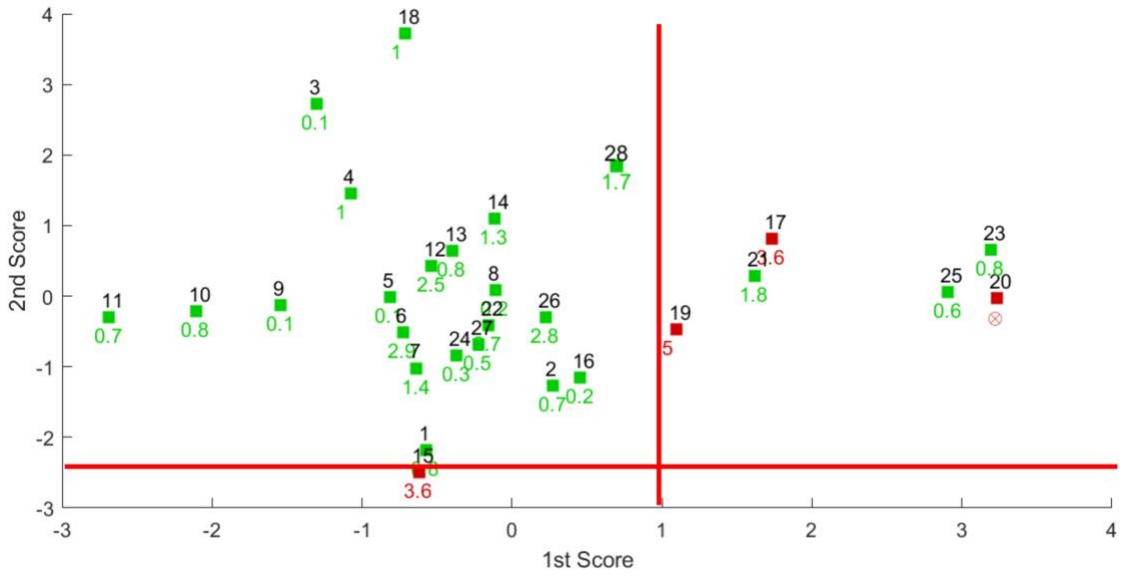


Figure 2-6: Recipe score plot for the prediction of the volume phase transition temperature of dual pH/temperature-responsive microgels at pH 4 using Combined clustering in which the green points represent samples for which the VPTT was predicted within 3°C while the red points represent samples for which VPTT was either poorly predicted or mis-predicted.

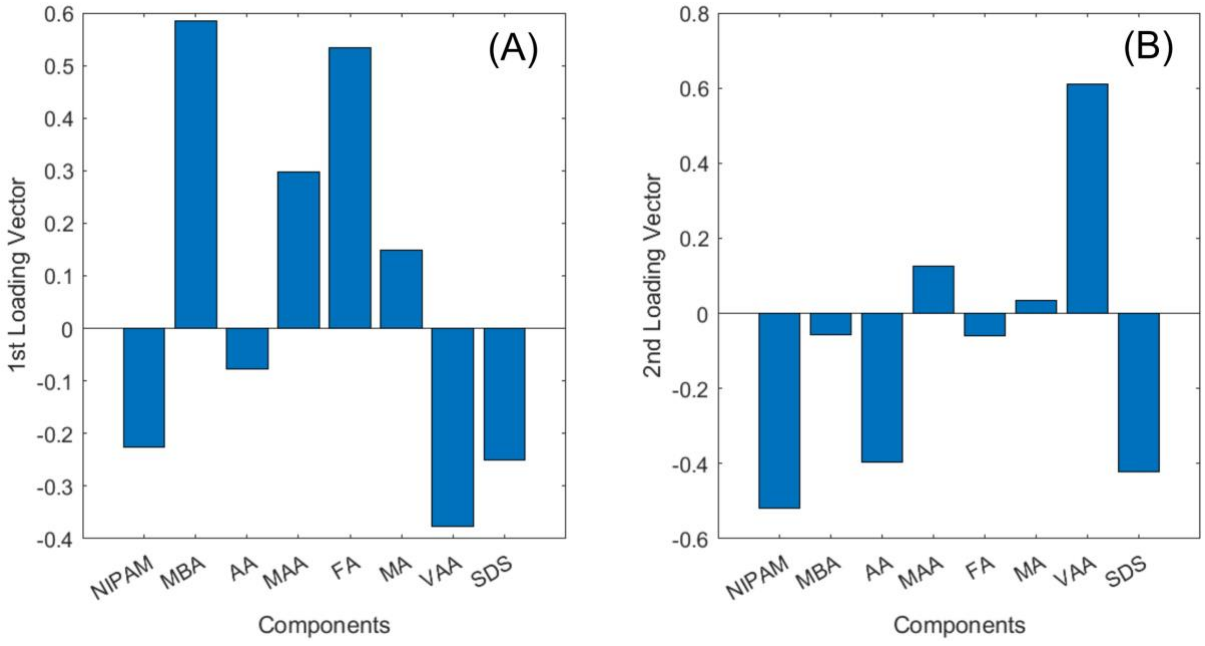


Figure 2-7: 1st (A) and 2nd (B) PCA Loading vectors corresponding to the recipe score plot

Referring to the score plot in Figure 2-6, microgels with a high 1st score or low 2nd score are significantly more likely to have poorly predicted VPTT values using the Combined Clustering policy. Based on the first and second loading vectors of the PCA model in the recipe space (Figure 2-7(A and B)), these ranges correspond to samples with low VAA contents based on the large negative coefficient for VAA in the first loading vector and the large positive coefficient for VAA in the second loading vector. Correspondingly, looking at the highlighted observations in Table S4 and their recipes in Table 2-1, Table S1 or Table S2, each sample that is poorly predicted does not contain vinylacetic acid as a comonomer. As such, the score plot analysis and the qualitative assessment of the dataset yield the same overall conclusions about what types of microgel swelling profiles are likely to be well-predicted using the model.

• *Transition Temperature Prediction at pH=10*

Due to the ionization of the acid-containing pH-responsive monomer residues at pH=10, the volume phase transition temperature is at minimum increased or, depending on the number and distribution of the functional monomers present, fully suppressed. As such, not only predicting the magnitude of the VPTT but also whether or not a sample will show a discrete VPTT value over a defined temperature range (in this work, from 25°C to 50°C) is important for predicting microgel swelling responses. As shown in Figure 2-8(E) and Table S5, unlike at pH 4, the Transition Prediction Parameter shows that even in the best case (recipe-based X-clustering) the presence or absence of a discrete phase transition is predicted 82 percent of the time; however, among the 12 samples that do show discrete VPTT values between 25°C and 50°C (as per Table S5), 11 of them were correctly predicted to have a discrete VPTT and 10 of those 11 microgels had estimated VPTT values within 6°C of the measured VPTT. Note that, due to the much higher uncertainty associated with VPTT experimental measurements at pH 10 (estimated as $\approx 2.2C$) owing to the much less discrete nature of the transitions in the ionized state relative to those observed at pH 4, prediction accuracies within 6°C are practically useful for microgel design at pH 10. In contrast, the Y-based clustering approach predicts no transition even though one exists in 6/12 cases (Table S5, True

Detection column) and fewer samples with <6°C prediction errors (Figure 2-8(E)) compared to X-based clustering, even when the Combined Clustering approach is used. We attribute this observation to the different variance structure in the pH 10 dataset relative to the pH 4 dataset, emphasizing the importance of considering all potential clustering/data arrangement approaches to ensure optimal predictions under each condition.

Similar to pH=4, in this case (pH=10), there are also 6 observations for which the existence of a discrete transition temperature either was mis-predicted or an actual VPTT was poorly predicted (i.e. the VPTT prediction error was >6°C) using the best-performing clustering policy (X-based). To assess if these samples could be explicitly identified by the model, the score plot for the VPTT prediction at pH 10 is shown in Figure 2-9; properly predicted samples are highlighted in green and poorly predicted samples are highlighted in red.

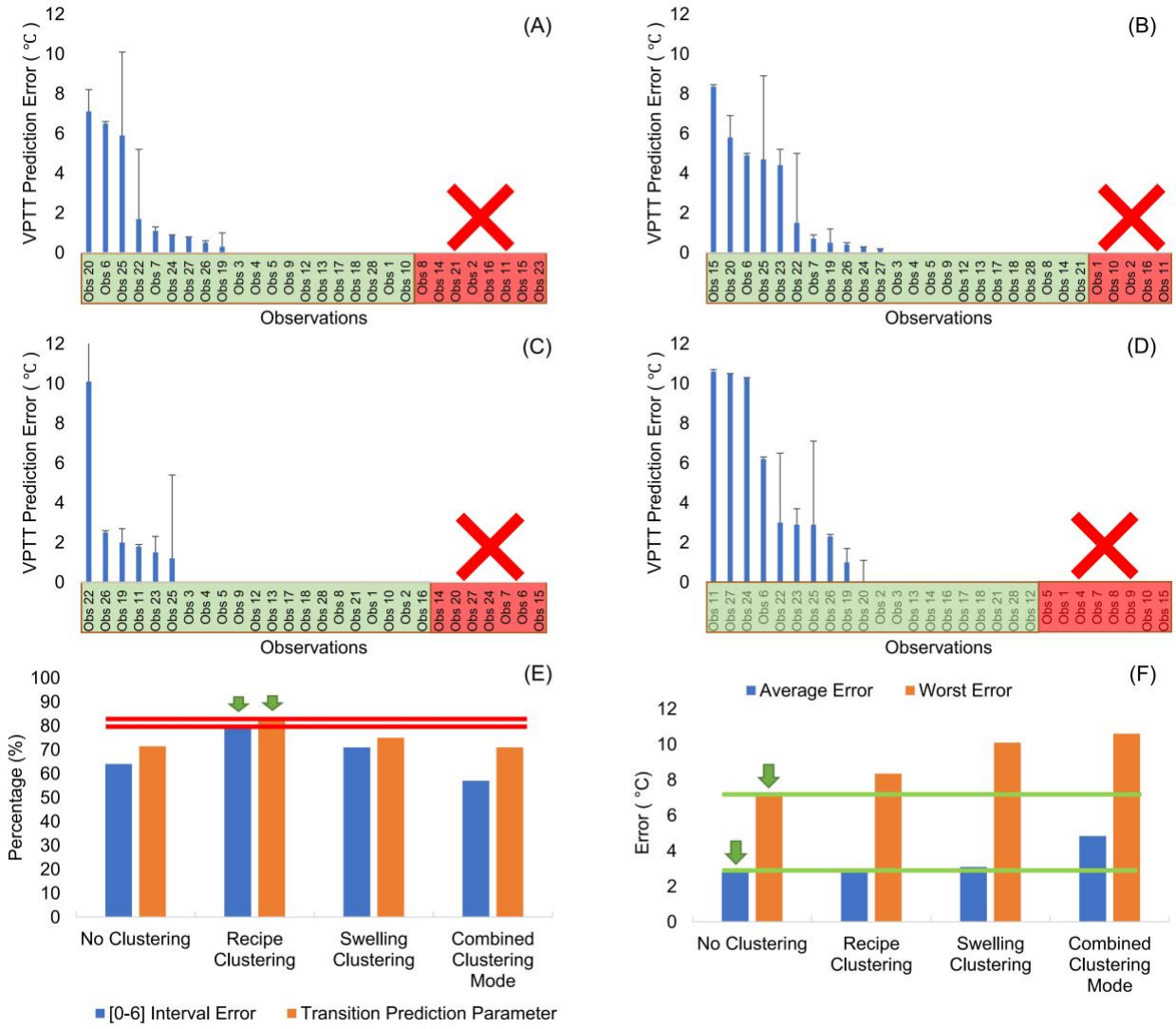


Figure 2-8: Prediction error of the model for the microgel volume phase transition temperature at pH10 using (A) no clustering, (B) recipe (X-based) clustering, (C) swelling (Y-based) clustering, and (D) combined clustering, in each case using the best-performing data arrangement for each clustering policy. Observations highlighted in green are correctly predicted to either have or not have a discrete phase transition while observations highlighted in red are ones for which the existence of VPTT was mis-predicted; (E, F) comparison of VPTT prediction accuracies achieved using each clustering policy in A-D based on (E) the percentage of microgels for which the experimental VPTT is predicted within <6°C and the Transition Prediction Parameter and (F) the average and worst-case prediction errors observed.

Although the space in which the recipe-based clustering model is expected to accurately predict the transition temperature is quite wide, microgel recipes exhibiting very low 1st scores and/or very low 2nd scores tend to be more poorly predicted. Based on the loading vector values in Figure 2-7(A and B), the parameters for NIPAM, AA, and SDS are all both negative in both loading vectors, implying that the

swelling properties of less functionalized and smaller microgels are not predicted as accurately at pH 10. Correspondingly, the majority of the poorly predicted samples were microgels with moderate-high functional monomer contents prepared using higher SDS concentrations that result in microgels with smaller particle sizes. Note that the SDS content cannot alone predict a poorer model prediction; for example, despite having the exact high identical amount of SDS in their recipes, samples 5, 6, 7, 8 and 9 all showed good VPTT predictions corresponding with their position in the well-predicted region in the score plot. The relatively low number of microgels with similar recipes in the training data set is likely the cause of the less accurate predictions for such microgels and could be resolved by further expanding the training data set using more recipes with similar compositions. However, given that all well-predicted samples are in one quadrant and all (and only) the poorly-predicted samples are in the other quadrants, a Transition Prediction Parameter of 100 percent and average and worst-case VPTT prediction errors equal to 2.3°C and 5.8°C respectively can be achieved within the well-predicted range. In this context, at pH 10, the statistical analysis can not only well-predict the VPTT behavior of most microgels but also unambiguously identify whether or not a model VPTT prediction is likely to be accurate.

Overall, the performance of the model in predicting the volume phase transition properties of dual pH/temperature microgels is summarized in Table 2-2. While the data-driven modelling approaches developed herein cannot unambiguously predict the swelling profile or VPTT value for any new microgel tested, the existence of a discrete phase transition for an observation (97 percent at pH 4 and 82 percent at pH 10) and the resulting volume phase transition temperature (100 percent within 5°C at pH 4 and 83 percent within 6°C at pH 10) can be accurately predicted for the vast majority of microgel recipes tested. Moreover, if the predictive power of the best-performing models (Combined Clustering for pH 4 and recipe-based clustering for pH 10) are considered only over their well-predicted areas as identified by PLS analysis, the existence of a discrete phase transition for an observation can be predicted with 100 percent accuracy at both pH values and VPTT prediction errors of <3°C at pH 4 and <6°C at pH 10 can be achieved. As such, even for samples that cannot be well-predicted, PCA score plot analysis can

explicitly categorize whether or not a reliable VPTT prediction can be expected based only on the microgel recipe, allowing for immediate flagging of potentially inaccurate particle size or VPTT estimates within the context of a microgel design framework. Expanding the "training set" of available microgel data in areas in which few observations were present in the current dataset (e.g. higher monomer loadings of MAA, MA, or FA and/or higher SDS contents) would likely further increase the ratio of samples that appear in the well-predicted area relative to the poorly predicted area, albeit with the drawback of requiring significant additional microgel synthesis and characterization work that statistical modelling approaches like this typically aim to avoid. The capacity of the model presented to make functional predictions as well as flag potential poor predictions based only on the recipe used to prepare the microgel should also in principle (upon inversion) enable the identification of a microgel recipe with a specific targeted swelling profile as long as that profile is within the "well-predicted" solution space of the model. This capacity offers the potential to significantly accelerate the rate of microgel development for target applications that require precise and specific swelling/deswelling profiles at multiple pH values.

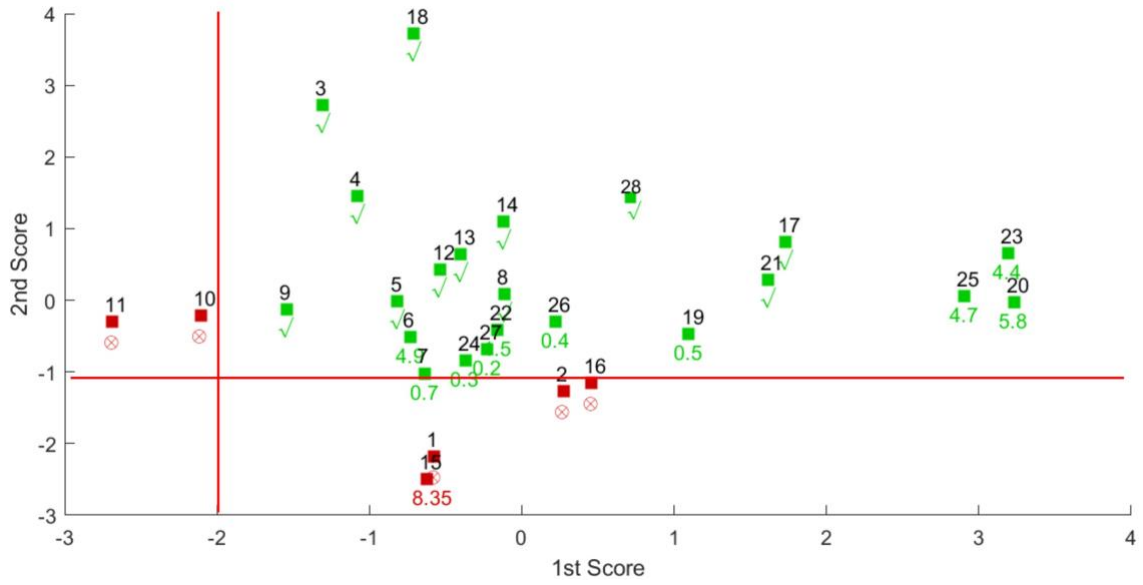


Figure 2-9: Recipe score plot for the prediction of the volume phase transition temperature of dual pH/temperature-responsive microgels at pH 10 using X (recipe-based) clustering in which the green points represent samples for which the VPTT was predicted within 6°C while the red points represent samples for which VPTT was either poorly predicted or mis-predicted

Table 2-2: VPTT prediction quality within the well-predicted area for both pH=4 and pH=10

	Clustering Policy	Type	Transition Prediction Parameter	[0-3] Error Interval pH4 [0-6] Error Interval pH10	Average Error	The Worst Error
pH 4	Combined Clustering Mode	Type 1	100%	100%	1.0	2.9
pH 10	Recipe Clustering	Type 1	100%	100%	2.4	5.8

2.4.6 Predicting the Properties of New Microgels

Finally, to assess whether the model can predict the swelling response of a microgel it has not previously seen, we synthesized two new microgels (New Obs 1 and New Obs 2) using recipes incorporating mid-high mole fractions of three functional monomers (AA, FA, and VAA) into a single microgel; the model had not been trained with any microgels prepared with this combination of comonomers (see Table S3 for the new microgel recipe information). Figure 2-10(A and B) show the observed versus predicted particle size versus temperature plots at both pH 4 and pH 10 for both microgels using the optimum clustering and data arrangement policies identified at both pH values (combined clustering for pH 4, recipe-based clustering at pH 10). The high correlation between the actual and predicted particle sizes confirming that the model can accurately predict the swelling results for a microgel recipe distinct from those trained by the model. Correspondingly, the score plots Figure 2-10(C and D) show that both new microgels lie within the well-predicted range that would suggest a good prediction is possible. The samples did not exhibit any VPTT at pH 10 in the temperature range of 25°C to 50°C, as the model correctly predicted; at pH 4, the predicted VPTT values lied within 0.6 °C and 0.4 °C of the actual measured VPTT values for New Obs 1 and New Obs 2 respectively. As such, the model can accurately predict both the swelling profile and the VPTT value of a new microgel not used to derive the model.

While we demonstrate that our data-driven technique can predict the particle size versus temperature swelling responses of a diverse range of microgels (or at minimum identify that a particular prediction is likely to be poor), the developed technique also has some limitations. In particular, model predictions in any data-driven model are only as good as the data used to train the model. Ideally, all new recipes would be included in the model's training dataset (as we have done herein using our jackknife training

approach); only after this requirement is satisfied can the location of the new recipes in the recipe score plot be confidently interpreted as a measure of predictability. That being said, the two new microgels synthesized to test our model prediction were not processed in this way and still gave good predictions, which would be expected provided the recipes were not radically different than those already in the training set. Expanding the training set over a broader range of microgel compositions (e.g. the mid/high AA content/high SDS content microgels that were poorly predicted at pH 10) would reduce the need to perform such re-training for each sample and expand the well-predicted range. In addition, while the current methodology can be used on any new microgel dataset that has been synthesised under various reaction conditions, the model will not necessarily hold true for samples that were produced under different reaction conditions beyond the base temperature, stirring speed, and reaction time used consistently across our training data set. Although the model can easily be modified to include variables like temperature and mixing as additional X (process) parameters, a sufficient number of samples must be synthesised under various conditions to allow the model to be adequately trained to give good predictions.

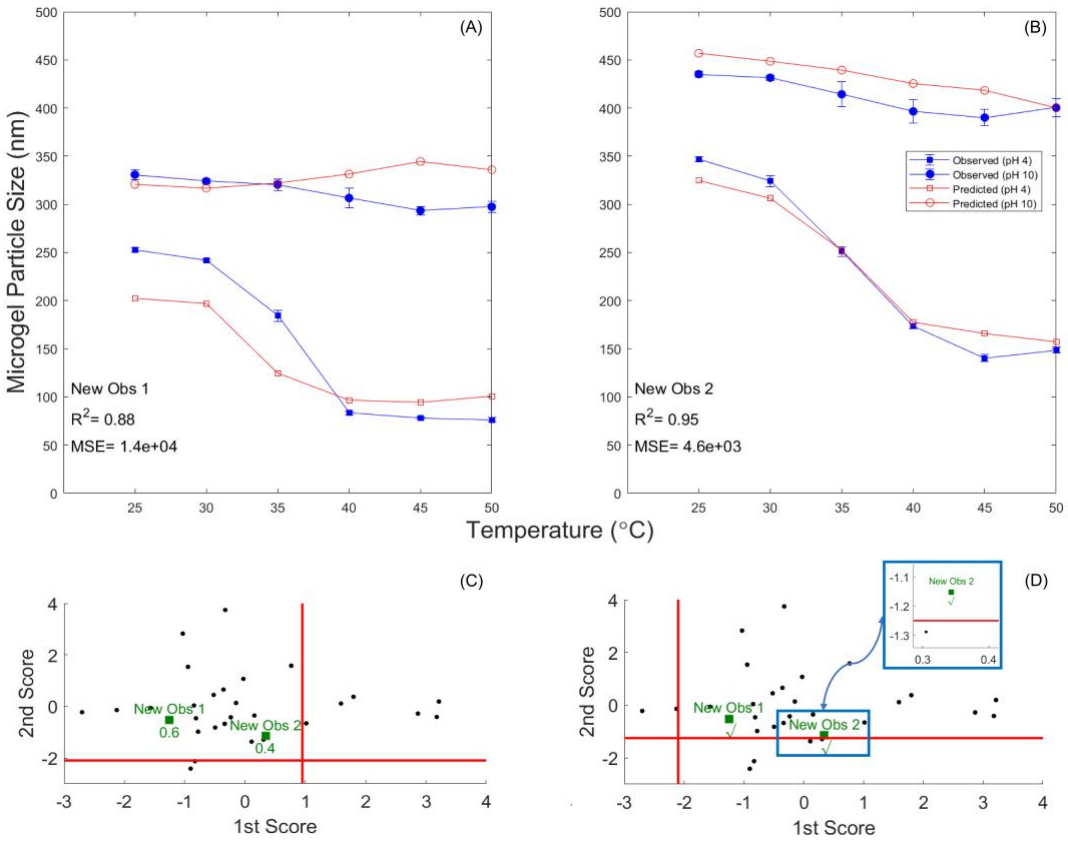


Figure 2-10: (A,B) Microgel swelling and VPTT prediction quality of two new microgels (New Obs 1, A and New Obs 2, B) using the optimal data processing approaches identified at each pH: (A) pH = 4 using the combined clustering mode and (B) pH = 10 using recipe-based clustering; (C,D) Score plots for pH 4 (C) and pH 10 (D) showing that both microgels lie within the well-predicted range consistent with the good correlation between actual vs. measured sizes and VPTT values.

2.5 Conclusions

In this work, a coupling of partial least squares modelling technique with data clustering was introduced and developed to predict the swelling responses and volume phase transition temperature of dual pH/temperature-responsive microgels at both pH 4 (fully protonated state) and at pH 10 (fully ionized state). By first clustering data either in the recipe space or in the swelling profile space followed by developing PLS based models for each individual cluster, improved predictions of the VPTT can be achieved. Specifically, clustering based on the swelling profile provides better predictions of transition temperatures at pH 4 while clustering based on the microgel recipe provides better predictions at pH 10. A new version of swelling-based clustering called Combined Clustering was also developed that enables the implementation of swelling profile-based clustering without the prior need to know the swelling response

of the microgel, leading to improved property predictions at pH 4. The optimized cluster-based PLS technique developed successfully predicts the microgels' transition temperature within an mean prediction error of 1.0°C at pH=4 and 2.3°C at pH=10, values that are close to the error bars of experimentally measured transition temperatures (0.6°C and 2.2°C at pH 4 and pH 10, respectively). Furthermore, although not every microgel swelling response in the dataset can be accurately predicted, it is possible to *a priori* predict whether an accurate VPTT prediction can be made for a new microgel recipe using PCA score plot analysis. This ability to predict microgel swelling profiles (and assess how accurate any prediction is likely to be) has the potential to eliminate the need for using the slow trial-and-error approach now used to design a microgel with a specific pH or temperature swelling profile, significantly accelerating the design of new functional multi-responsive microgels with targeted swelling responses for specific applications.

Supporting Information

"Training dataset recipes and summary of the models' performance, PCA and PLS modelling supplementary discussion, Methods and implementation clarification, and Swelling Profile prediction quality for all dataset

(PDF)"

Acknowledgements

The Natural Sciences and Engineering Research Council of Canada (NSERC, Discovery Grant RGPIN-2017-06455), The McMaster Advanced Control Consortium (MACC), and the Canada Research Chairs program (to both TH and PM) are gratefully acknowledged for funding this work.

2.6 References

1. Gichinga, M.G., et al., *Miniemulsion polymers as solid support for transition metal catalysts.* Polymer, 2010. **51**(3): p. 606-615.
2. Kiser, P.F., G. Wilson, and D. Needham, *A synthetic mimic of the secretory granule for drug delivery.* Nature, 1998. **394**(6692): p. 459-462.
3. Kuentz, A., et al., *Stable microdispersions and microgels based on acrylic polymers, method for obtaining them and compositions, particularly cosmetic compositions, containing them*. 1998, Google Patents.
4. Mei, Y., et al., *Catalytic activity of palladium nanoparticles encapsulated in spherical polyelectrolyte brushes and core− shell microgels.* Chemistry of Materials, 2007. **19**(5): p. 1062-1069.
5. Smeets, N.M. and T. Hoare, *Designing responsive microgels for drug delivery applications.* Journal of Polymer Science Part A: Polymer Chemistry, 2013. **51**(14): p. 3027-3043.
6. Heskins, M. and J.E. Guillet, *Solution properties of poly (N-isopropylacrylamide).* Journal of Macromolecular Science—Chemistry, 1968. **2**(8): p. 1441-1455.
7. Li, G., et al., *Nanoscale mechanical properties of core–shell-like Poly-NIPAm microgel particles: effect of temperature and cross-linking density.* The Journal of Physical Chemistry B, 2021. **125**(34): p. 9860-9869.
8. Meyer-Kirschner, J., et al., *Monitoring Microgel Synthesis by Copolymerization of N-isopropylacrylamide and N-vinylcaprolactam via In-Line Raman Spectroscopy and Indirect Hard Modeling.* Macromolecular Reaction Engineering, 2018. **12**(3): p. 1700067.
9. Muratalin, M., et al., *Study of N-isopropylacrylamide-based microgel particles as a potential drug delivery agents.* Colloids and Surfaces A: Physicochemical and Engineering Aspects, 2017. **532**: p. 8-17.
10. Hoare, T., *Multi-responsive microgels: synthesis, characterization, and applications*. 2007: Library and Archives Canada= Bibliotheque et Archives Canada: Ottawa.
11. Zhang, L., M.W. Spears Jr, and L.A. Lyon, *Tunable swelling and rolling of microgel membranes.* Langmuir, 2014. **30**(26): p. 7628-7634.
12. Khan, A., et al., *Effect of Experimental Variables on the Physicochemical Characteristics of Multi-Responsive Cellulose Based Polymer Microgels.* Russian Journal of Physical Chemistry A, 2020. **94**(7): p. 1503-1514.
13. Hoare, T. and R. Pelton, *Functionalized microgel swelling: comparing theory and experiment.* The Journal of Physical Chemistry B, 2007. **111**(41): p. 11895-11906.
14. Sheikholeslami, P., et al., *Semi-batch control over functional group distributions in thermoresponsive microgels.* Colloid and Polymer Science, 2012. **290**: p. 1181-1192.
15. Windbiel, J.T. and A. Llevot, *Microgel preparation by miniemulsion polymerization of passerini multicomponent reaction derived acrylate monomers.* Macromolecular Chemistry and Physics, 2021. **222**(24): p. 2100328.
16. Abdelaty, M.S. and D. Kuckling, *Altering of lower critical solution temperature of environmentally responsive poly (N-isopropylacrylamide-co-acrylic acid-co-vanillin acrylate) affected by acrylic acid, vanillin acrylate, and post-polymerization modification.* Colloid and Polymer Science, 2021. **299**: p. 1617-1629.
17. Zhang, Y., Y. Guan, and S. Zhou, *Synthesis and volume phase transitions of glucose-sensitive microgels.* Biomacromolecules, 2006. **7**(11): p. 3196-3201.
18. Gorelikov, I., L.M. Field, and E. Kumacheva, *Hybrid microgels photoresponsive in the near-infrared spectral range.* Journal of the American Chemical Society, 2004. **126**(49): p. 15938-15939.
19. Khan, S.R., et al., *Synthesis, characterization, and silver nanoparticles fabrication in N-isopropylacrylamide-based polymer microgels for rapid degradation of p-nitrophenol.* Journal of dispersion science and technology, 2013. **34**(10): p. 1324-1333.

20. Farooqi, Z.H., et al., *Catalytic reduction of 2-nitroaniline in aqueous medium using silver nanoparticles functionalized polymer microgels.* Journal of Inorganic and Organometallic Polymers and Materials, 2015. **25**: p. 1554-1568.
21. Farooqi, Z.H., et al., *Engineering of silver nanoparticle fabricated poly (N-isopropylacrylamide-co-acrylic acid) microgels for rapid catalytic reduction of nitrobenzene.* Journal of Polymer Engineering, 2016. **36**(1): p. 87-96.
22. Bakaic, E., et al., *pH-Ionizable in Situ Gelling Poly (oligo ethylene glycol methacrylate)-Based Hydrogels: The Role of Internal Network Structures in Controlling Macroscopic Properties.* Macromolecules, 2017. **50**(19): p. 7687-7698.
23. Supasuteekul, C., et al., *A study of hydrogel composites containing pH-responsive doubly crosslinked microgels.* Soft Matter, 2012. **8**(27): p. 7234-7242.
24. Hoare, T. and R. Pelton, *Highly pH and temperature responsive microgels functionalized with vinylacetic acid.* Macromolecules, 2004. **37**(7): p. 2544-2550.
25. Malmsten, M., H. Bysell, and P. Hansson, *Biomacromolecules in microgels—Opportunities and challenges for drug delivery.* Current Opinion in Colloid & Interface Science, 2010. **15**(6): p. 435-444.
26. Lei, B., et al., *Double security drug delivery system DDS constructed by multi-responsive (pH/redox/US) microgel.* Colloids and Surfaces B: Biointerfaces, 2020. **193**: p. 111022.
27. Teng, D., et al., *Glucosamine-carrying temperature-and pH-sensitive microgels: Preparation, characterization, and in vitro drug release studies.* Journal of colloid and interface science, 2008. **322**(1): p. 333-341.
28. Qi, X., et al., *Fabrication and characterization of a novel anticancer drug delivery system: salecan/poly (methacrylic acid) semi-interpenetrating polymer network hydrogel.* ACS Biomaterials Science & Engineering, 2015. **1**(12): p. 1287-1299.
29. Hoare, T. and D. McLean, *Multi-Component Kinetic Modeling for Controlling Local Compositions in Thermosensitive Polymers.* Macromolecular theory and simulations, 2006. **15**(8): p. 619-632.
30. Wu, J., G. Huang, and Z. Hu, *Interparticle potential and the phase behavior of temperature-sensitive microgel dispersions.* Macromolecules, 2003. **36**(2): p. 440-448.
31. Hoare, T. and D. McLean, *Kinetic prediction of functional group distributions in thermosensitive microgels.* The Journal of Physical Chemistry B, 2006. **110**(41): p. 20327-20336.
32. Schneider, S., et al., *Model-based design and synthesis of ferrocene containing microgels.* Polymer Chemistry, 2020. **11**(2): p. 315-325.
33. Jung, F., et al., *Model-based prediction of the hydrodynamic radius of collapsed microgels and experimental validation.* Chemical Engineering Journal, 2019. **378**: p. 121740.
34. Camerin, F., et al., *Modelling realistic microgels in an explicit solvent.* Scientific reports, 2018. **8**(1): p. 14426.
35. Janssen, F.A., et al., *Kinetic modeling of precipitation terpolymerization for functional microgels*, in *Computer Aided Chemical Engineering*. 2018, Elsevier. p. 109-114.
36. Fiscus, D.M. *Developing Structure–Process–Property Relationships Using Multivariate Analysis*. in *Macromolecular Symposia*. 2020. Wiley Online Library.
37. Sharmin, R., et al., *Inferential sensors for estimation of polymer quality parameters: Industrial application of a PLS-based soft sensor for a LDPE plant.* Chemical Engineering Science, 2006. **61**(19): p. 6372-6384.
38. Mohammadi, M., M.K. Khorrami, and H. Ghasemzadeh, *ATR-FTIR spectroscopy and chemometric techniques for determination of polymer solution viscosity in the presence of SiO2 nanoparticle and salinity.* Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2019. **220**: p. 117049.
39. Torres, J.M.G.T., et al., *Designing multi-responsive polymers using latent variable methods.* Polymer, 2014. **55**(2): p. 505-516.

40. MacGregor, J.F., K. Muteki, and T. Ueda, *On the rapid development of new products through empirical modeling with diverse data-bases*, in *Computer Aided Chemical Engineering*. 2006, Elsevier. p. 701-706.
41. Muteki, K. and J.F. MacGregor, *Multi-block PLS modeling for L-shape data structures with applications to mixture modeling.* Chemometrics and Intelligent Laboratory Systems, 2007. **85**(2): p. 186-194.
42. MacGregor, J.F., et al., *Data-based latent variable methods for process analysis, monitoring and control.* Computers & chemical engineering, 2005. **29**(6): p. 1217-1223.
43. Kresta, J., T. Marlin, and J. MacGregor, *Development of inferential process models using PLS.* Computers & Chemical Engineering, 1994. **18**(7): p. 597-611.
44. Geladi, P. and B.R. Kowalski, *Partial least-squares regression: a tutorial.* Analytica chimica acta, 1986. **185**: p. 1-17.
45. Haenlein, M. and A.M. Kaplan, *A beginner's guide to partial least squares analysis.* Understanding statistics, 2004. **3**(4): p. 283-297.
46. Syakur, M.A., et al. *Integration k-means clustering method and elbow method for identification of the best customer profile cluster*. in *IOP conference series: materials science and engineering*. 2018. IOP Publishing.
47. Sinaga, K.P. and M.-S. Yang, *Unsupervised K-means clustering algorithm.* IEEE access, 2020. **8**: p. 80716-80727.
48. Mahmud, M.S., et al., *A survey of data partitioning and sampling methods to support big data analysis.* Big Data Mining and Analytics, 2020. **3**(2): p. 85-101.
49. Jafarzadegan, M., F. Safi-Esfahani, and Z. Beheshti, *Combining hierarchical clustering approaches using the PCA method.* Expert Systems with Applications, 2019. **137**: p. 1-10.
50. Wu, J.M.-T., et al., *The density-based clustering method for privacy-preserving data mining.* 2019.
51. Wu, T.-Y., et al., *A grid-based swarm intelligence algorithm for privacy-preserving data mining.* Applied Sciences, 2019. **9**(4): p. 774.
52. Irawan, Y., *Implementation Of Data Mining For Determining Majors Using K-Means Algorithm In Students Of SMA Negeri 1 Pangkalan Kerinci.* Journal of Applied Engineering and Technological Science (JAETS), 2019. **1**(1): p. 17-29.
53. Hossain, M.Z., et al., *A dynamic K-means clustering for data mining.* Indonesian Journal of Electrical engineering and computer science, 2019. **13**(2): p. 521-526.
54. Lu, W., *Improved K-means clustering algorithm for big data mining under Hadoop parallel framework.* Journal of Grid Computing, 2020. **18**(2): p. 239-250.
55. Paris, A., C. Duchesne, and É. Poulin, *Establishing multivariate specification regions for incoming raw materials using projection to latent structure models: comparison between direct mapping and model inversion.* Frontiers in Analytical Science, 2021. **1**: p. 729732.

2.7 Supporting Information

2.7.1 Training Dataset Formulations, Summary of the Models' Performance and supplementary Figures

Table S1 and Table S2 show the raw (in grams) and scaled value recipe formulations of each of the microgels used to build the training dataset for all the models. The four microgels excluded based on the pre-screening analysis described in the manuscript are highlighted in rows 29 to 32. Table S4 also provides quality of VPTT prediction at pH 4 (High = <1.5°C error, Moderate = 1.5°C <VPPT error < 3°C ; Poor = >3°C error) and pH 10 (High = <3°C error, Moderate = 3°C < VPTT error < 6°C ; Poor = >6°C error) as well as their hosting cluster based on either their recipes or swelling profiles. Finally, the VPTT prediction quality at both pH values among all policies for all samples is tabulated in Table S5 with the best-performing data arrangements under each clustering policy highlighted. The true detection column in Table S5, refers to the portion of observations for which there was a discrete VPTT and the model has correctly guessed identified such (this value does not include those portion of observations for which there was not an actual VPTT and the model correctly guessed that). For reference, Figure S1 provides a visual representation into how the VPTT was calculated based on the available swelling data and the corresponding error estimates made from the uncertainties in the experimental particle size measurements. Figure S2, Figure S3, and Figure S4 also provide supporting information pertaining to section 3.3, section 4.1, and section 4.5.1 respectively.

Table S1: Raw formulation (in grams) of all recipes prior to any pre-processing

Sam ID	NIPAM	MBA	AA	MAA	FA	MA	VAA	SDS
1	1.600	0.160	0.064	0	0	0	0	0.056
2	1.600	0.159	0	0	0.092	0	0	0.056
3	1.600	0.161	0	0	0	0	0.343	0.056
4	1.600	0.161	0	0	0	0	0.228	0.057
5	1.600	0.160	0	0	0	0	0.114	0.058
6	1.600	0.161	0	0	0	0	0.080	0.056
7	1.600	0.159	0	0	0	0	0.045	0.055
8	1.600	0.204	0	0	0	0	0.114	0.059
9	1.600	0.114	0	0	0	0	0.114	0.057

10	1.600	0.080	0	0	0	0	0.114	0.056
11	1.600	0.047	0	0	0	0	0.114	0.055
12	1.600	0.159	0	0	0	0	0.114	0.034
13	1.600	0.161	0	0	0	0	0.114	0.024
14	1.600	0.160	0	0	0	0	0.114	0
15	1.600	0.159	0.114	0	0	0	0	0.059
16	1.600	0.159	0	0	0.114	0	0	0.057
17	1.600	0.160	0	0.114	0	0	0	0
18	1.600	0.160	0	0	0	0	0.343	0
19	1.600	0.176	0	0	0.112	0	0	0.024
20	1.600	0.261	0	0	0.320	0	0	0.060
21	1.600	0.217	0	0	0.049	0.006	0.069	0.039
22	1.600	0.163	0	0	0	0	0.041	0.023
23	1.600	0.247	0	0.079	0.149	0	0	0.026
24	1.600	0.166	0	0	0	0	0.042	0.046
25	1.600	0.245	0	0	0.253	0	0	0.036
26	1.600	0.174	0	0.011	0	0	0.034	0.023
27	1.600	0.172	0	0	0	0	0.045	0.041
28	1.600	0.245	0	0	0	0	0.175	0.026
29	1.600	0.161	0	0	0	0.457	0	0.056
30	1.600	0.159	0	0.109	0	0	0	0.058
31	1.600	0.159	0	0	0	0	0	0
32	1.600	0.259	0	0	0.248	0	0.058	0.025

Table S2: Mean-centered and scaled values of the dataset for all 32 samples prior to any pre-processing

Sam ID	NIPAM	MBA	AA	MAA	FA	MA	VAA	SDS
1	0.587	-0.074	2.450	-0.281	-0.486	-0.189	-0.956	0.916
2	0.698	-0.049	-0.263	-0.281	0.759	-0.189	-0.956	0.933
3	-2.273	-0.711	-0.263	-0.281	-0.486	-0.189	2.538	0.444
4	-1.166	-0.464	-0.263	-0.281	-0.486	-0.189	1.544	0.627
5	0.116	-0.179	-0.263	-0.281	-0.486	-0.189	0.393	0.838
6	0.542	-0.084	-0.263	-0.281	-0.486	-0.189	0.011	0.909
7	0.988	0.015	-0.263	-0.281	-0.486	-0.189	-0.389	0.982
8	-0.187	0.843	-0.263	-0.281	-0.486	-0.189	0.370	0.787
9	0.431	-1.238	-0.263	-0.281	-0.486	-0.189	0.418	0.889
10	0.675	-2.058	-0.263	-0.281	-0.486	-0.189	0.437	0.931
11	0.925	-2.901	-0.263	-0.281	-0.486	-0.189	0.456	0.972
12	0.199	-0.160	-0.263	-0.281	-0.486	-0.189	0.400	-0.307
13	0.241	-0.151	-0.263	-0.281	-0.486	-0.189	0.403	-0.883

14	0.325	-0.133	-0.263	-0.281	-0.486	-0.189	0.410	-2.043
15	-0.149	-0.238	4.382	-0.281	-0.486	-0.189	-0.956	0.795
16	0.481	-0.098	-0.263	-0.281	1.051	-0.189	-0.956	0.899
17	0.325	-0.133	-0.263	4.272	-0.486	-0.189	-0.956	-2.043
18	-2.117	-0.676	-0.263	-0.281	-0.486	-0.189	2.574	-2.043
19	0.515	0.293	-0.263	-0.281	1.028	-0.189	-0.956	-0.804
20	-1.816	1.556	-0.263	-0.281	3.269	-0.189	-0.956	0.516
21	-0.172	1.151	-0.263	-0.281	0.164	5.103	-0.149	-0.161
22	1.163	0.115	-0.263	-0.281	-0.486	-0.189	-0.439	-0.764
23	-1.172	1.488	-0.263	2.600	1.332	-0.189	-0.956	-0.776
24	1.026	0.208	-0.263	-0.281	-0.486	-0.189	-0.434	0.419
25	-1.180	1.464	-0.263	-0.281	2.610	-0.189	-0.956	-0.372
26	1.039	0.378	-0.263	0.146	-0.486	-0.189	-0.525	-0.919
27	0.962	0.343	-0.263	-0.281	-0.486	-0.189	-0.400	0.170
28	-1.007	1.491	-0.263	-0.281	-0.486	-0.189	0.989	-0.911
29	-2.239	-0.703	-0.263	-0.281	-0.486	321.782	-0.956	0.449
30	0.171	-0.166	-0.263	4.052	-0.486	-0.189	-0.956	0.848
31	1.867	0.211	-0.263	-0.281	-0.486	-0.189	-0.956	-2.043
32	-1.761	1.550	-0.263	-0.281	2.439	-0.189	-0.348	-0.912

Table S3: New (test) microgel recipes

	Sam ID	NIPAM	MBA	AA	MAA	FA	MA	VAA	SDS
Mole Fraction	New Obs 1	0.7961	0.0477	0.0522	0	0.0136	0	0.0775	0.0129
	New Obs 2	0.8454	0.0638	0.0307	0	0.0465	0	0.0056	0.0080
in gram	New Obs 1	1.6	0.13	0.07	0	0.03	0	0.12	0.07
	New Obs 2	1.6	0.16	0.04	0	0.09	0	0.01	0.04
Centered-Scaled	New Obs 1	0.090	-0.069	1.263	-0.281	0.702	-0.189	-0.861	-0.104
	New Obs 2	-0.925	-1.077	2.331	-0.281	-0.139	-0.189	0.361	1.098

2.7.2 PCA and PLS Modelling Supplementary Discussion

• *PCA*

PCA aims to identify the directions in which the observations are distributed and order these directions based on the variance of the original block of data that each of these directions covers. These directions can be explained using a linear combination of the original variables with weighted coefficients. The directions are referred to as the PCA components or score matrix (denoted by t) and the weighted coefficients used to calculated them are referred to as the PCA loading matrix (denoted by p,Figure S5). In PCA, if only the first direction (which explains the most variation within the original block) is chosen for dimensionality reduction, it means that in the new space there is only one variable for each observation (i.e. the 1st score) that has been calculated through eq. (S1); if the two directions with the highest variance are chosen, there are two variables representing each observation in the new space called the 1st and 2nd scores (and so on). The general structure of PCA modelling is shown in Figure S5.

The scores are computed as follows:

$$t_{1st}^{i} = X_i \times p_{1st} \quad (eq.S1)$$

where X_i and p_{1st} represent the i_{th} observation and the first direction with highest variance respectively.

Moving backward from score space to the original space will give $X\hat{}$ (eq. (S2)) which may not equal the original X if the number of selected directions is less than the number of original predictor variables. The distance between the actual and the estimated values is calculated using eq. (S3). It should be noted that in eq. (S2), T and P are used to refer to the loading and score matrices; as such, t_{ith} and p_{ith} in eq. (S1) represent

Table S4:Quality of VPTT prediction at pH 4 (High = <1.5°C error, Moderate = 1.5°C <VPTT error <3°C ; Poor = >3°C error or mis-predicted) and pH 10 (High = <3°C error, Moderate = 3°C <VPTT error <6°C ; Poor = >6°C error or mis-predicted) as well as their hosting cluster based on either X (recipe-based) or Y (swelling profiles) clustering policies

Sam ID	VPTT Prediction Quality [pH4]	VPTT Prediction Quality [pH10]	Hosting Cluster Recipe-Based	Hosting Cluster Swelling-Based
1	High	Poor	1	2
2	High	Poor	1	1
3	High	High	3	2
4	High	High	3	1
5	High	High	1	1
6	Moderate	Moderate	1	1
7	High	High	1	1
8	High	High	1	1
9	High	High	1	1
10	High	Poor	1	1
11	High	Poor	1	1
12	Moderate	High	1	2
13	High	High	1	2
14	High	High	3	3
15	Poor	Poor	1	2
16	High	Poor	1	2
17	Poor	High	2	3
18	High	High	3	3
19	Poor	High	2	3
20	Poor	Moderate	2	2
21	Moderate	High	2	2
22	Moderate	High	1	1
23	High	Moderate	2	3
24	High	High	1	1
25	High	Moderate	2	3
26	Moderate	High	1	2
27	High	High	1	1
28	Moderate	High	3	2
New Obs 1	High	High	1	1
New Obs 2	High	High	1	1

Table S5: Summary of the VPTT prediction quality using different clustering policies and data arrangement strategies

		Type	Transition Prediction Parameter	[0-3] Interval Error pH4 [0-6] Interval Error pH10	True detection	R^2	Average Error	The Worst Error
pH 4	No Clustering	1	96%	71%	96%	93%	1.9	6.1
		2	100%	70%	100%	92%	2.2	6.2

		3	75%	64%	75%	96%	1.3	3.6
		4	75%	64%	75%	96%	1.3	3.7
		5	43%	32%	43%	95%	1.6	4.3
	Recipe Clustering	1	96%	68%	96%	94%	2.0	6.1
		2	96%	71%	96%	94%	1.9	5.7
		3	71%	54%	71%	94%	2.1	5.2
		4	75%	54%	75%	94%	2.1	5.2
		5	43%	32%	43%	95%	1.6	3.7
	Swelling Clustering	1	89%	71%	89%	95%	1.5	6.1
		2	96%	68%	96%	95%	1.8	6.4
		3	89%	57%	89%	92%	2.8	9.1
		4	71%	54%	71%	93%	2.4	8.4
		5	50%	32%	50%	93%	2.5	7.2
	Combined Clustering Mode	1	96%	86%	96%	96%	1.4	5.0
		2	100%	75%	100%	95%	1.8	5.3
		3	75%	46%	75%	90%	3.3	10.3
		4	71%	29%	71%	89%	3.8	9.9
		5	36%	11%	36%	89%	3.5	10.0
pH 10	No Clustering	1	64%	64%	75%	94%	2.3	5.9
		2	71%	64%	75%	92%	2.70	7.1
		3	57%		0%		0.0	0.0
		4	57%		0%		0.0	0.0
		5	57%		0%		0.0	0.0
	Recipe Clustering	1	82%	79%	92%	93%	2.80	8.4
		2	71%	68%	83%	92%	3.1	8.3
		3	54%	54%	0%		0.0	0.0
		4	57%	57%	0%		0.0	0.0
		5	54%	54%	0%		0.0	0.0
	Swelling Clustering	1	61%	61%	67%	96%	1.7	6.0
		2	64%	64%	67%	97%	1.1	4.7
		3	54%	43%	50%	85%	5.7	10.9
		4	75%	71%	50%	91%	3.2	10.1
		5	71%	67%	50%	96%	3.1	10.2
	Combined Clustering Mode	1	57%	50%	58%	95%	1.9	6.0
		2	61%	53%	67%	95%	1.8	6.1
		3	46%	35%	25%	81%	7.2	6.9
		4	53%	39%	42%	81%	6.9	11.1
		5	71%	57%	83%	87%	4.8	10.6

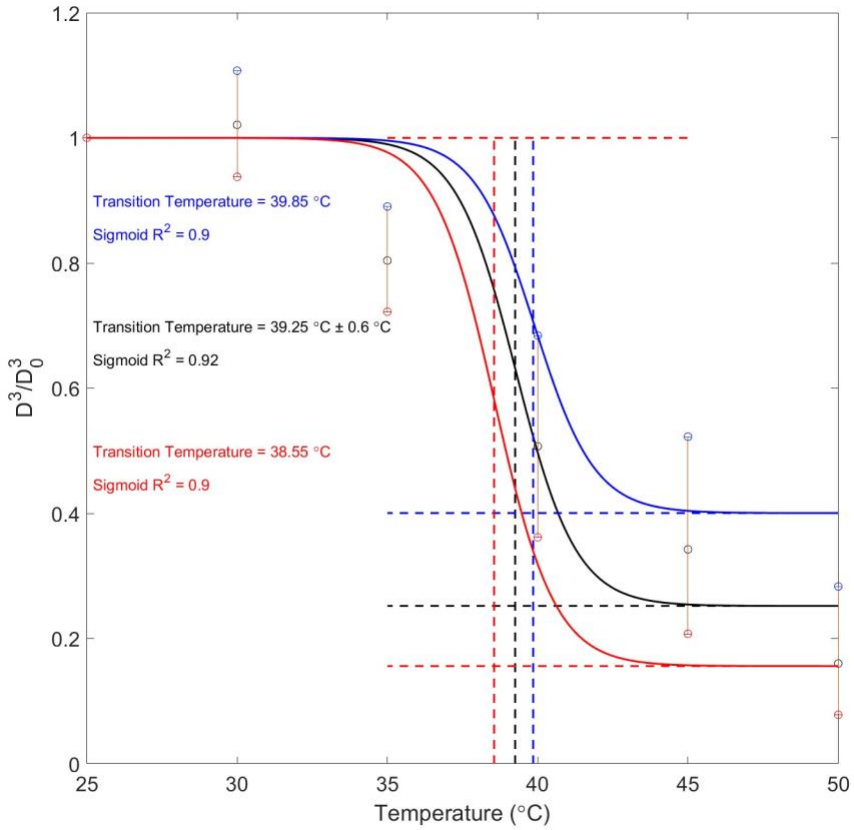


Figure S1: Method for estimating VPTT values from experimental swelling profiles D^3/D_0^3, where D denotes the measured z-average microgel diameter at the specific temperature tested and D_0 denotes the particle size diameter at 25°C (fully swollen state), by fitting a sigmoidal curve to the data (black line). Error bars on the experimental VPTT data were estimated based on the difference in the VPTT values measured according to the blue curve (fit based on the upper boundary of error bars of each observation point) and the red curve (fit based on the lower boundary of the error bars of each observation point the

i_{th} columns of the corresponding matrices.

$$\hat{X} = T \times P^T \quad (eq.S2)$$

$$E = X - \hat{X} \quad (eq.S3)$$

The R^2 coverage of an applied PCA is calculated through eq. (S4), with the equation reflecting that using more directions (components) will promote broader coverage of the original dataset while at the same time increasing the chance of over-fitting the data.

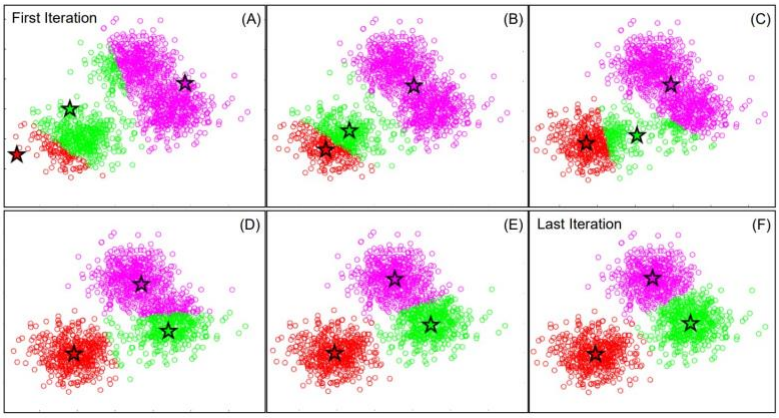


Figure S2: A K-means clustering example spanning from the first iteration (A) (randomly selected centers) to the last iteration (F) (fully clustered data)

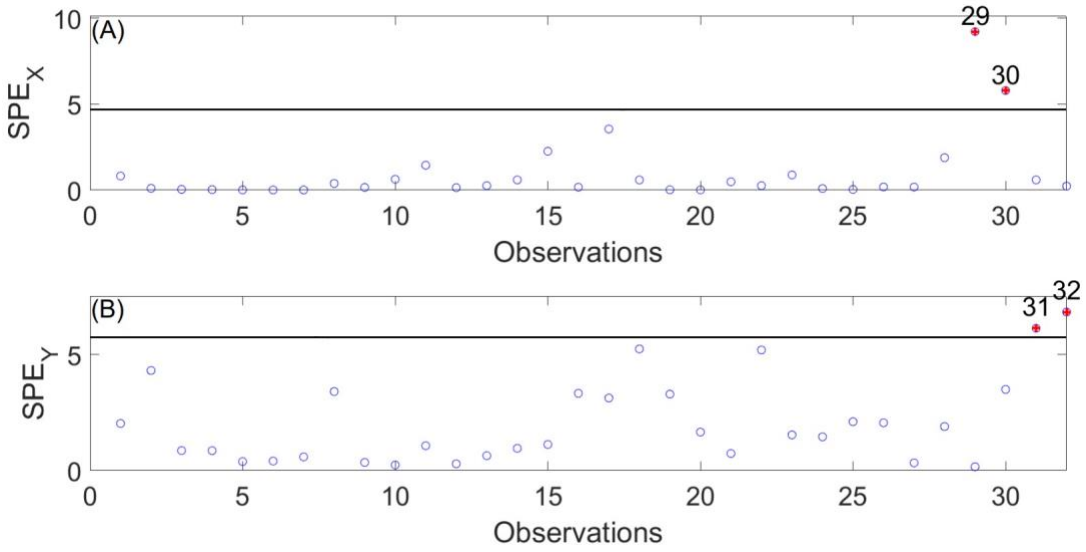


Figure S3: Squared prediction error (SPE) for all observations in (A) the X (recipe) and (B) the Y (swelling) space. The solid black line represents the SPE_{lim} (threshold) in each block, with the red-crossed samples showing the excluded observations as a result of the pre-screening algorithm.

$$R^2 = 1 - \frac{var(E)}{var(X)} \quad (eq.S4)$$

• *PCA NIPALS Algorithm*

Non-linear Iterative PArtial Least Squares (NIPALS) is one of the most powerful algorithms utilized for calculating the PCA loading (P) and score (T) matrices. This method calculates the components of PCA

one by one based on their importance in representing the original X matrix [1, 2]. To calculate each component, only the part of data that has not been described using the previous components is used. The NIPALS algorithm is quite flexible and also easy to apply on any data space, particularly when some of the measurements are missing [2, 3]. This algorithm includes the following steps for each component [1, 4]:

a) Mean-centering and unit-scaling the data in each column (based on predictors)

b) Choosing an arbitrary column for the t_{ith} score (this randomly selected column can be any vector yet at least one element of this vector needs to be not equal to zero. The most commonly used vector in this step is a randomly selected column of $X_{(i-1)th}$).

c) Regressing each column of $X_{(i-1)th}$ onto t_{ith}

d) Assigning the best-fit slope as the *ith* element of p_{ith}

e) Normalizing the p_{ith} to unit value

f) Regressing all rows of $X_{(i-1)th}$ onto p_{ith} to update each element of $t_{ith(new)}$

g) Repeating steps 3 to 6 until there are no subsequent changes in the calculated t_{ith}

h) Recording p_{ith} and t_{ith} in the *ith* column of the P and T matrices respectively.

i) Deflating the $X_{(i-1)th}$ matrix (i.e. the E calculated through eq. (S3) is replaced as $X_{(i)th}$ and the new components are calculated for the new X)

Step (i) practically means that E corresponds to the variance that has not been described with the already calculated components, such that for the following component only this part of data should be processed. A general summary of regressions and how the NIPALS algorithm deals with finding loading and score vectors is depicted in Figure S6. A note should be made in Figure S6 that in each iteration normalizing P_{ith}' is performed after part (A) and prior to part (B).

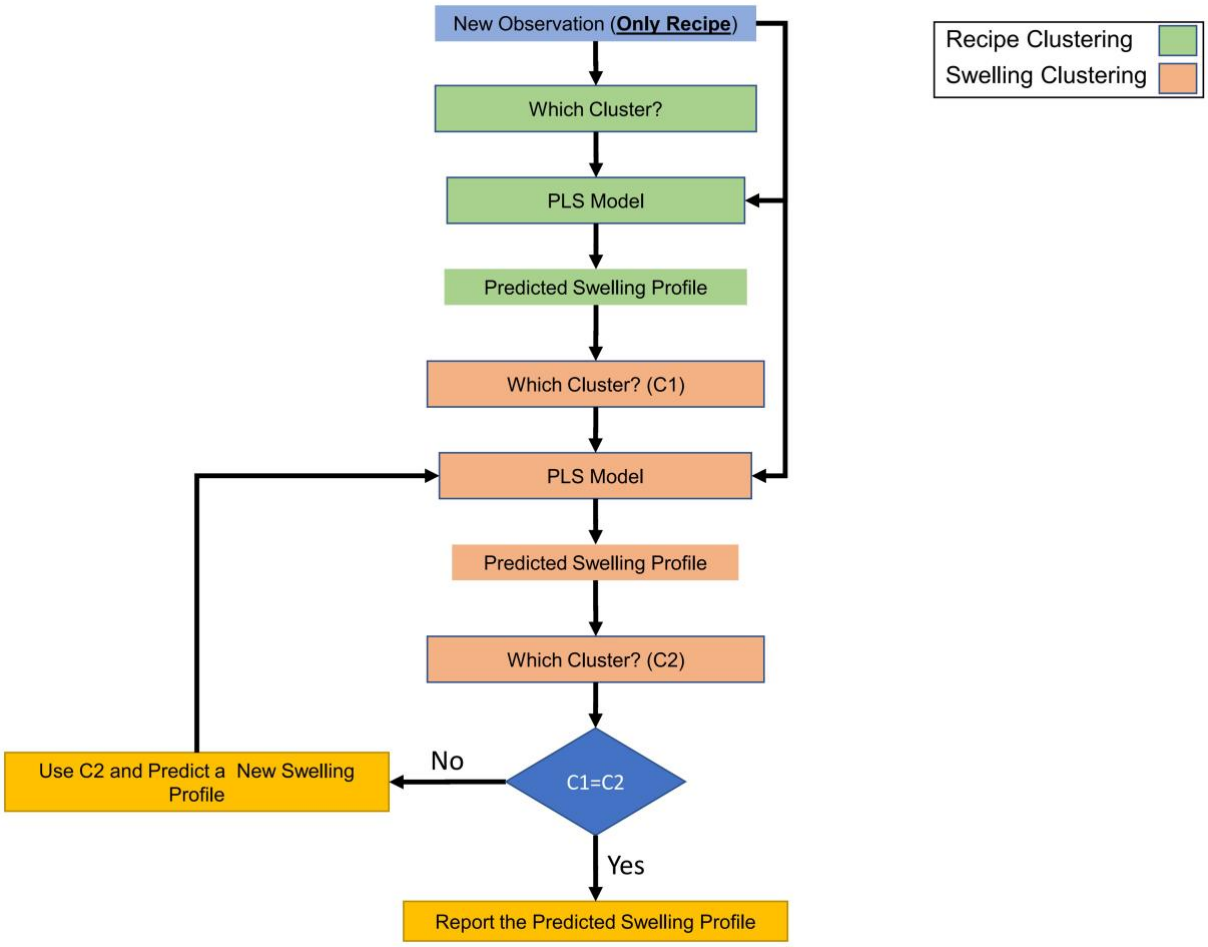


Figure S4: Schematic of the heuristic used to classify a new microgel observation into the correct Y-based swelling cluster based only on the microgel recipe using the developed Combined Clustering policy

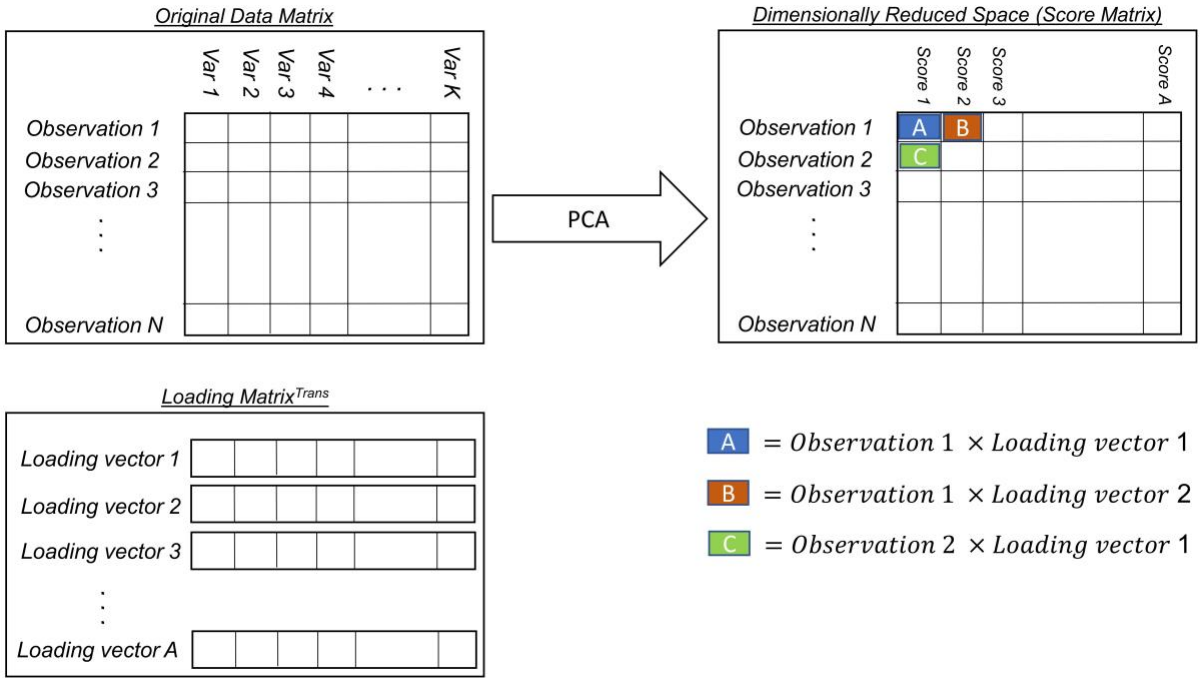


Figure S5: PCA general structure

• *PLS*

PLS NIPALS Algorithm

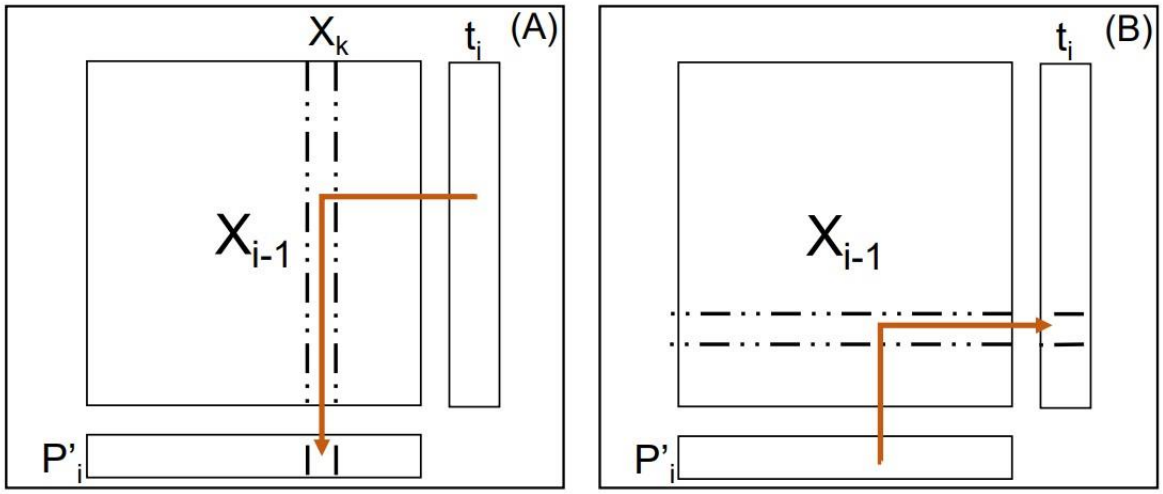


Figure S6: A schematic depiction of the NIPALS algorithm to calculate the loading and score vectors in PCA. The orange arrow shows the direction of the regression.

The general concept of PLS was described in section 3.2. The NIPALS algorithm is also one of the most practical options for performing PLS [5, 6] and calculating its all blocks. Figure S7 shows a standard scheme of a NIPALS-PLS structure for calculating the i_{th} component where t and u are score vectors and c and p are the loading vectors corresponding to the X and Y blocks respectively. In addition to already mentioned vectors, PLS includes another vector (denoted by w) that plays a pivotal role in relating the Y block variance to that of X block. The advantages of using NIPALS for PLS modelling are similar to those discussed for PCA; however, the steps through which it works are different [7]. The NIPALS algorithm for PLS modelling is implemented as follows:

a) Mean-centering and unit-scaling data in both the X and Y blocks based on their columns.

b) Choosing an arbitrary column for u or simply setting u equal to randomly selected column of Y.

c) Regressing columns of X onto u and assigning the resulting slopes as the elements of the w vector

d) Normalizing the w vector to unit-scale

e) Regressing rows of X onto w to update the elements of the t vector

f) Regressing columns of Y onto t and assigning the resulting slopes as the elements of the c vector

g) Updating the u vector by regressing rows of Y onto c

h) Checking if convergence (i.e. no changes in the recently calculated u with the u from previous iteration) is reached. If yes, proceed to step (i); if no, go back to (c) to iterate.

i) Calculating the loading vector of X (p) by regressing the columns of X onto t

j) Deflating the already described parts of both X and Y blocks using eq. (S5) and eq. (S6)

$$E = X - t \times p^T \quad (eq.S5)$$

$$F = Y - t \times c^t \quad (eq.S6)$$

k) Recording the t, u, c, w, p vectors as the corresponding columns of T, U, C, W, P matrices

l) Setting the new X and Y equal to E and F and repeating all steps to calculate the next component (this process should be iterated as many as the number of required components)

In PLS, blocks of data are assigned to either the X or Y section based on which block's information is available (or more likely to be available in the future) and which block's information is needed in the future [7]. In addition, while the X block has two loading vectors (p and w), the w vector is used to calculate the corresponding t vector while the p vector is used to calculate $X\hat{}$ matrix (for more information, refer to [5, 8]).

PLS and PCA both need to be trained using an available dataset to develop loading and score matrices. For PLS, once it is trained, these matrices are used to estimate the corresponding responses for a new observation. To do so, the new observation should first be mean-centered and unit-scaled using the X block's average and standard deviation (Std) values; following, eqs. (S7) to (S9) can be utilized to predict the corresponding y_{fit} values. The predicted y_{fit} values must then be de-normalized using the training Y block's average and Std values [5, 8].

$$t^{new} = x^{new} \times W^{*} \quad (eq.S7)$$

$$y^{fit} = t^{new} \times C^{T} \quad (eq.S8)$$

where W^{*} is:

$$W^{*} = \frac{W}{(P^{T} \times W)} \quad (eq.S9)$$

A schematic of how the PLS model was applied in this work is shown in Figure S7.

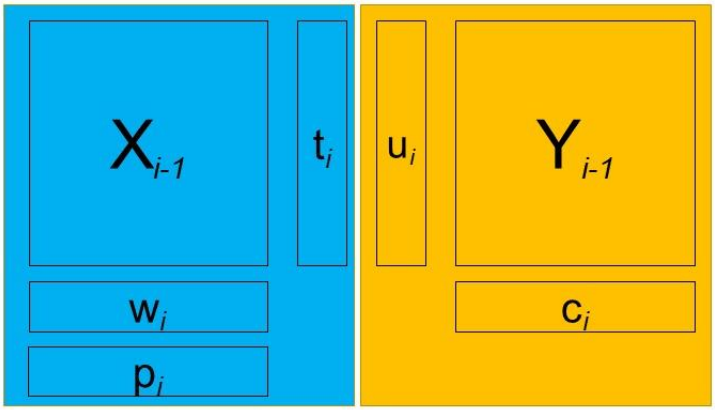


Figure S7: PLS model components showing two PCA applied on two blocks simultaneously

2.7.3 Methods and Implementation

Based on the provided information, there were 15 possible dataset IDs (i.e. five possible data arrangements under three possible clustering policies) that needed to be evaluated to determine the best approach to achieve the maximum predictive power for the available dataset.

In this work, the Non-linear Iterative Partial Least-Squares (NIPALS) algorithm was employed to apply PLS on all introduced datasets. The NIPALS algorithm is a very flexible and straightforward way of applying PLS, as it calculates the PLS components one by one and thus makes it possible to set the number of required components (i.e. linear combinations of variables) to cover the whole X and Y space during the modelling process.

The number of components in this paper was not pre-determined but rather was identified based on the NIPALS algorithm for each case analyzed, allowing for flexibility as the number of observations and number of variables was changed using different arrangements and clustering policies and thus minimizing the risk that the model was over-fit in different scenarios. Since the NIPALS algorithm calculates the components in order, the eigenvalue-greater-than-one rule [9] was used to increase the number of components considered only if the variance explained with the recently calculated t is greater than one; as such, if a specific component does not improve predictions, it is excluded from the model. In

this context, comparisons of the outcomes of different models are significantly less biased and over-fitting of the models is either avoided or at least much significantly less likely to happen.

To be able to rigorously compare the output results of all 15 cases mentioned above, a systematic procedure unbiased by the clustering policies, data arrangement formats, or even the selection of the training and testing observations was required. Comparisons among all suggested data arrangements through evaluating R^2, MSE, or other error parameters are highly influenced by the number of output and input variables; as such, the modelling outputs under all introduced data arrangements and clustering policies were unfolded to the original format (8 inputs and 12 outputs) and then calculated and reported accordingly. In addition, considering the limited number of available observations (i.e. microgel recipes) in the dataset, comparisons of results under different clustering policies and/or data arrangements can be affected by how the observations were grouped in either the testing or training dataset. More specifically, if a certain group of observations is excluded for testing and the rest for training from the beginning, the final comparison between different arrangements and clustering policies can be biased given that some observations may be well-covered by certain arrangement formats and/or clustering policies but not others; in such a case, the model's final coverage will show a non-realistic privilege for some dataset arrangements over the others. To further avoid any bias in the comparisons between modelling frameworks, we employed a Jackknife approach in which we repeated the process of applying PLS models on the training data (and its evaluation) as many times as the number of observations under each case. In each iteration, one observation from the dataset was excluded from the "training" data and used as a "test" data point – at the end of the process, each individual microgel has been used as the "test" data at least once. This approach allows all the microgel swelling responses/VPTT values to be predicted by the corresponding models without seeing the observations ahead of time. Figure S8 provides a graphical representation of our procedure on how these steps were consistently used in this work.

a. Taking the original dataset in its original format (8 input variables, 12 output variables)

b. Applying pre-processing on the data to exclude samples that are not physically realistic (based on the
heuristics described in the manuscript)

c. Determining the applied clustering policy (no clustering, X-based clustering, or Y-based clustering)

d. Excluding observation 1 as the testing observation and using the rest of the observations as the training dataset

e. Creating all 5 introduced data arrangement types of both the training and testing observations for PLS modelling.

f. Applying K-means clustering on the right block of data based on the selected policy in step (c)

g. Applying the PLS model (or models, depending on the clustering policy) on each of the data arrangements and recording the model predictions for the training dataset

h. Applying the corresponding testing dataset to the corresponding PLS model (or models) and recording the model predictions for the testing observation

i. Unfolding the predicted values for the output to the original format (12 values for each observation) and calculating the model predictive power for different data arrangements (R^2, MSE, predicted values, etc.) for both the training dataset and testing observations

j. Returning to step (d), choosing the next observation as the testing observation (assigning the rest of the observations to the training dataset), and repeating steps (e) to (i).

k. Returning to step (c) to change the clustering policy mode and repeating steps (d) to (j).

l. Visualizing the models' outcomes for all cases and making conclusions about the relative performance
of the different modelling approaches.

2.7.4 Swelling Profile Prediction Quality

Figure S9, depicts the observed versus actual temperature-induced deswelling profiles at pH 4 (fully protonated state) and pH 10 (fully ionized state) for each microgel used in the training dataset, including relevant quality of fit metrics for the overall model predictability of each of the swelling profiles. The best-fit curves shown are calculated based on the best result among all three clustering policies and five data arrangements tested. Figure S10 shows the frequency at which each clustering (A) and data arrangement policy (B) yields the best fit to the experimental data.

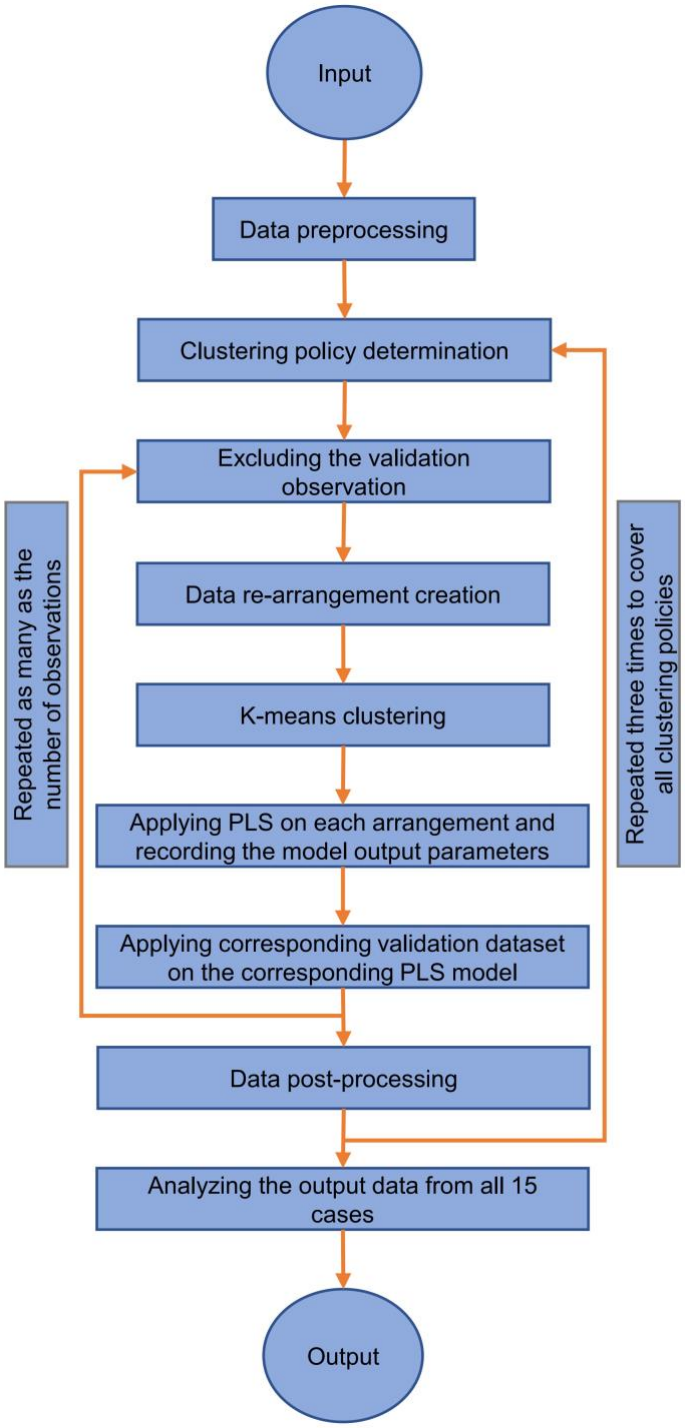


Figure S8: Summary of the heuristic used in this work to evaluate and compare all potential models incorporating different data arrangement and clustering policies

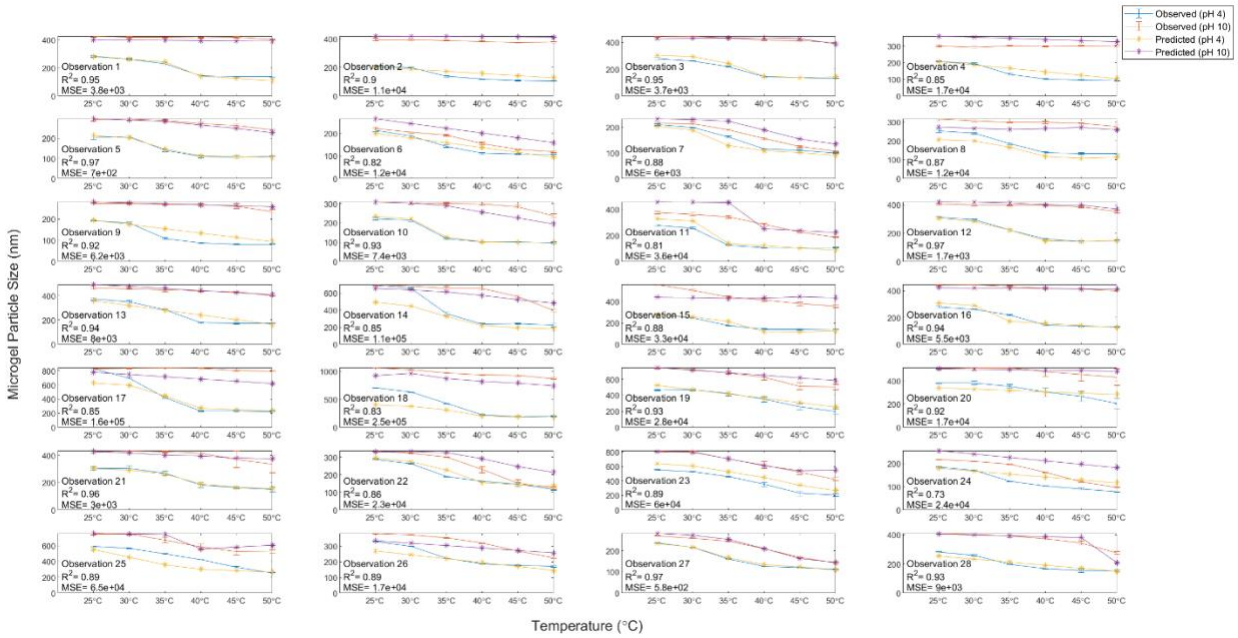


Figure S9: Observed temperature-induced deswelling profiles versus the best predicted deswelling profile for each observation among all clustering policies and data arrangements tested at both pH 4 (fully protonated state) and pH 10 (fully ionized state)

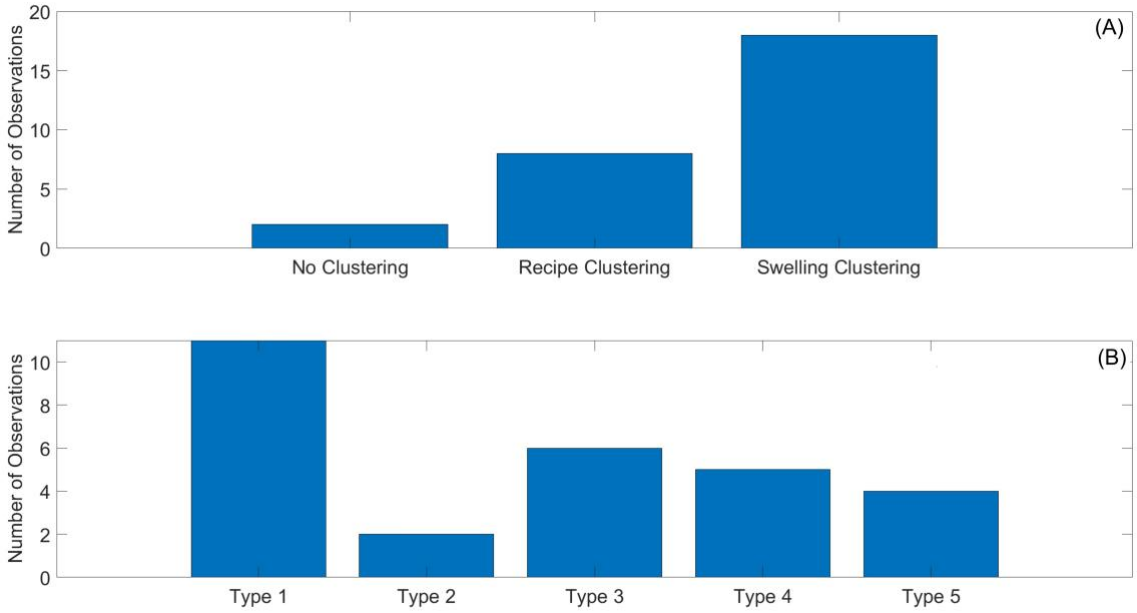


Figure S10: The number of observations for which each (A) clustering policy (B) data arrangement policy provides the most accurate prediction of the swelling profiles

References

1. Risvik, H., *Principal component analysis (PCA) & NIPALS algorithm.* It. lut. fi, 2007.

2. Wu, W., D. Massart, and S. De Jong, *The kernel PCA algorithms for wide data. Part I: theory and algorithms.* Chemometrics and Intelligent Laboratory Systems, 1997. **36**(2): p. 165-172.
3. Howley, T., et al. *The effect of principal component analysis on machine learning accuracy with high dimensional spectral data*. in *International Conference on Innovative Techniques and Applications of Artificial Intelligence*. 2005. Springer.
4. Miyashita, Y., et al., *Comments on the NIPALS algorithm.* Journal of chemometrics, 1990. **4**(1): p. 97-100.
5. Lorber, A., L.E. Wangen, and B.R. Kowalski, *A theoretical foundation for the PLS algorithm.* Journal of Chemometrics, 1987. **1**(1): p. 19-31.
6. Dayal, B.S. and J.F. MacGregor, *Improved PLS algorithms.* Journal of Chemometrics: A Journal of the Chemometrics Society, 1997. **11**(1): p. 73-85.
7. Kresta, J., T. Marlin, and J. MacGregor, *Development of inferential process models using PLS.* Computers & Chemical Engineering, 1994. **18**(7): p. 597-611.
8. Abdi, H., et al., *Computational toxicology.* 2013.
9. Kaiser, H.F., *The application of electronic computers to factor analysis.* Educational and psychological measurement, 1960. **20**(1): p. 141-151.

Chapter2:

3 Fast-Tracking Design Space Identification with the Prediction Reliability Enhancing Parameter (PREP)

The contents of this chapter have been published in Computers & Chemical Engineering Journal.

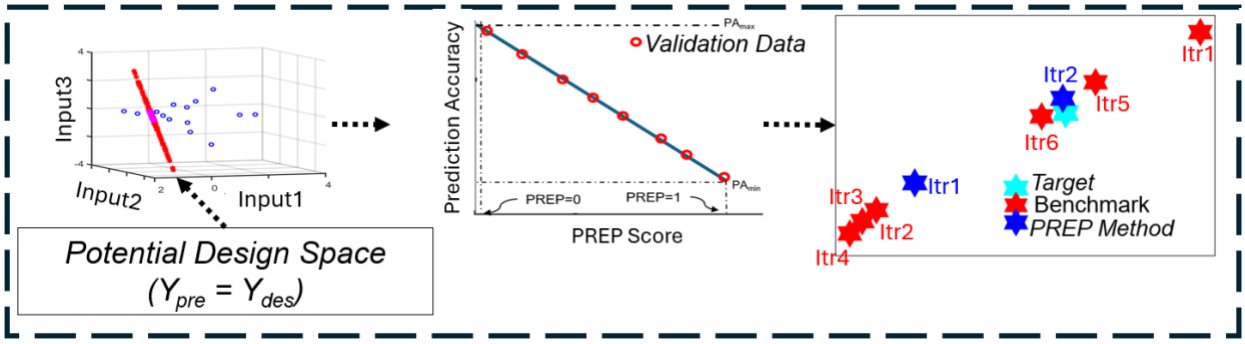
Seyed Saeid Tayebi, Todd Hoare, Prashant Mhaskar

Department of Chemical Engineering, McMaster University, Hamilton, Ontario L8S 4L7, Canada

Authorship Contribution

Seyed Saeid Tayebi: Conceptualized the idea, developed the methodology, wrote the code, wrote the manuscript draft, and addressed comments during the revision process. Prashant Mhaskar: Contributed to the conceptualization of the methodology, provided supervision, reviewed and contributed to the writing of the manuscript, and managed project administration and funding. Todd Hoare: Contributed to the conceptualization of the methodology, provided supervision, reviewed and contributed to the writing of the manuscript, and managed project administration and funding.

Abstract



In industrial product development, latent variable modeling tools are widely used to address challenges like multicollinearity and small sample sizes. However, these methods are often limited by prediction uncertainty, particularly when identifying optimal operating conditions or formulations to achieve desired product characteristics. This study introduces a methodology that leverages latent variable modeling alignment metrics, including partial least squares and principal components analysis Hotelling T^2, Sum of Squared Prediction Errors (SPE), and score alignment metrics (h_{PLS} and h_{PCA}), to quantify and enhance prediction reliability. These metrics are integrated into a Prediction Reliability Enhancing Parameter (PREP), a quantitative measure designed to identify recipes with higher reliability relative to the general model uncertainty. Using an iterative optimization-based algorithm, the methodology expands the Knowledge Space (KS) to efficiently determine the True Design Space (TDS), even when the TDS lies outside the KS. Validation with simulated nonlinear datasets demonstrates that the PREP approach achieves desired targets with significantly fewer iterations compared to conventional methods, particularly in cases in which the data are highly non-linear. The PREP approach thus provides a practical and effective solution for improving prediction reliability in complex, data-driven product design, offering enhanced accuracy and flexibility in identifying optimal formulations or operating conditions.

Keywords: Design space identification, Accelerated product design, Prediction uncertainty, Reliability assessment, Model validation, Latent variable modeling

3.1 Introduction

Predicting reliable outcomes in product development and formulation optimization is a crucial challenge, especially pertinent for most cases in which a first principles model may be difficult to build and/or maintain and data-driven modeling approaches are required. A major issue in such models is the uncertainty in predictions for new unseen data points, which can hinder the identification of optimal solutions and design spaces. In particular, standard error metrics like the Standard Error of Prediction (SEP) and Standard Error of Calibration (SEC) often fail to accurately reflect the reliability of predictions, as they are calculated based on samples the model has already encountered [1]. This lack of reliability assessment makes it difficult to confidently trust model predictions when trying to determine critical design spaces for desired outputs e.g. the regions within the input space that are expected to yield consistent quality in the final product properties. This concept plays a crucial role in the Quality by Design (QbD) framework, as outlined by the ICH Q8 guideline [2] that defines the design space as **"**the multidimensional combination and interaction of input variables (e.g., material attributes) and process parameters that have been demonstrated to provide assurance of quality**"** [2, 3].

The fundamental principle of latent variable models (LVMs) such as Principal Component Analysis (PCA) and Partial Least Squares (PLS)—that the number of underlying factors influencing a system is much smaller than the number of measurements taken—aligns well with the goals of design space determination and has led to the widespread application of LVMs in this field. Relative to Design of Experiments (DOE) methods that can also be useful in exploring the KS and identifying regions that ensure desired quality control, LVMs avoid the limitations of DOE methods in dealing with a large number of input variables or process conditions that often result in the need for an impractically high number of experimental samples [1, 3-5]. However, despite the use of latent variable modeling techniques in a range of industries based on their ability to handle multicollinearity, there is no established and validated method for assessing prediction uncertainty for new observations within the latent variable modeling framework [1, 3, 6, 7].

One major approach to addressing prediction reliability in design space identification focuses on estimating model prediction uncertainty. Researchers focusing on prediction uncertainty estimation aim to improve model precision and delineate prediction intervals, the range within which the actual outcome is expected to fall. Equation 1 shows the general format of a linear modeling approach in which model prediction uncertainty estimation methods attempt to calculate either the variance in $\hat{y}$ directly or the variance in the regression coefficients (β) that subsequently leads to a variance in $\hat{y}$.

$$\hat{y} = \beta . x \ (eq. 1)$$

where, β $[I \times L]^{\mathbb{R}}$ is the regression coefficients, x$[N \times I]^{\mathbb{R}}$ is the predictor variable, and $\hat{y}$ $[N \times K]^{\mathbb{R}}$is the response prediction for y. In this notation, N refers to the number of data points, I represents the number of predictor variables, and K indicates the number of response variables. Although it is generally assumed that the variance in prediction accuracy is a function of the t-statistic, a critical aspect of this approach involves estimating the standard deviation of the prediction error (S) for new observations, as shown in Equation 2.

$$CI = \widehat{y_0} \pm t_{\frac{\alpha}{2},N-df} . S \ (eq. 2)$$

where $\widehat{y_0}$ is the prediction for a new observation, N is the number of data points in the calibration dataset, df is the degrees of freedom used by the model, and α is the significance level for the interval (i.e. 100(1-α)% is the confidence interval) [7, 8].

Estimations of this prediction error are typically conducted using one of three main approaches: approximation techniques based on Ordinary Least Squares (OLS) expressions, methods involving linearization, and re-sampling strategies [7]. OLS-type expressions primarily rely on the distance of a new observation from the center of the input space used to train the model as an indicator for estimating the prediction interval; the greater this distance, the higher the estimated uncertainty and the wider the prediction interval. Prediction uncertainty has also been attributed to three primary factors: variability in the estimated model parameters, the unexplained variance in the response variable y, and inaccuracies in

the measurements of the predictor variables [9, 10]. Linearization-based methods, in contrast, recognize that the regression coefficients (β) depend nonlinearly on the outputs (y) by applying a first-order Taylor series expansion to β with respect to y around the current model output y_0. The resulting approximation is $\beta \approx \beta(y_0) + J(y - y_0)$, where J is the Jacobian matrix that captures how changes in y affect β. The variance of the model parameters can then be estimated using this Jacobian using methods like matrix differential calculus or inductive algorithms [7, 11-14]. Re-sampling methods, with bootstrapping and jackknifing being the most common, involve creating new datasets from the existing dataset by applying perturbations to either the samples or their residuals, thus providing a general expectation of the distribution of the prediction interval across the input space. Various adaptations of this technique have been developed to assess uncertainty in the model parameters that are then used to estimate the prediction interval [15, 16].

Probabilistic design space characterization methods [17-19], such as Bayesian-based approaches[20], also offer a structured strategy to quantify uncertainty and define design spaces in terms of feasibility probability, making them particularly valuable in scenarios in which risk assessment and regulatory compliance are key considerations. While these methods have demonstrated strong performance (especially in low-dimensional problems) and have been shown to be competitive with flexibility-based optimization techniques, they typically rely on extensive sampling strategies such as Monte Carlo methods; additionally, while probabilistic approaches provide a valuable means of defining design spaces under uncertainty, their reliance on feasibility probability thresholds may introduce conservatism, potentially limiting operational flexibility. As a result of these limitations, none of the existing probabilistic design space characterization methods can be robustly implemented across diverse datasets of interest [7], highlighting the need for alternative strategies for efficient and adaptable design space identification.

A common scenario in design space determination arises when the number of input variables exceeds the number of output variables that must be kept within an acceptable range. In such cases, there exists a

region within the input space, known as the null space, in which adjustments can be made with minimal or no impact on the output variables [1, 21-23]. The primary objective in design space determination is to find this null space, which provides the flexibility to make changes without affecting the desired output values. Numerous studies have sought to calculate this null space by accounting for prediction uncertainty and propagating these uncertainties back to the input space [1, 8, 19, 21-29]. Despite these efforts, the result has typically been the identification of a region with a high probability of containing the True Design Space (TDS). However, even within this region, there are multiple candidate areas and further prioritization is needed to efficiently pinpoint the exact design space.

Tomba [1] introduced a framework comprising four major scenarios for design space determination, each addressing different situations: (1) unconstrained inputs with specific targeted output values; (2) unconstrained inputs with a targeted output ranges; (3) constrained inputs with specific targeted output values; and (4) constrained inputs with a targeted output ranges. Various optimization algorithms were then tailored to each case that emphasized the importance of maintaining low Hotelling's T^2 and SPE (Squared Prediction Error) for all potential formulations or recipes. While these frameworks can address challenges around design space determination when the design space is within the range of the training dataset or closely resembles the calibration data points, they also have certain limitations. Biases within the algorithms often make it challenging to deviate from the already defined Knowledge Space (KS), which can slow or hinder the effective expansion of the KS. Active adjustments are required to the weights of soft and hard constraints to ensure that subsequent iterations yield solutions that differ from previous ones or from existing observations in the calibration dataset used for the PLS-regression model. Furthermore, the validity of this approach relies on the assumption that the design space to be identified must resemble the calibration data points, despite there being no explicit or standardized definition of the degree of similarity required to justify initiating the search for a new target sample within the same Knowledge Space [1, 26, 27]. Finally, none of these methods addresses scenarios in which the actual Design Space is entirely separate from and does not overlap with the established Knowledge Space. The

present manuscript presents a technique to address real-world experimental scenarios, where researchers often begin with an existing dataset of randomly distributed samples. In Latent Variable Modeling (LVM), the KS is typically defined based on the existing dataset and is determined using a 95% confidence region around the variation of the original data points in the latent space. This confidence region extends beyond the observed data, but it does not necessarily encompass all potential viable solutions. In some cases, the true Design Space lies beyond this predefined KS, meaning that models trained on the available data remain valid only within the KS and may not reliably predict outcomes outside of it. Therefore, in such cases, it is essential to guide the expansion of the KS in the correct direction to ensure that it eventually includes the true DS. Importantly, this expansion should be achieved in as few iterations as possible to optimize resource efficiency and minimize experimental costs.

Our proposed method to address this challenge aims to enhance prediction reliability in latent variable modeling via a more efficient identification of the design space, thus optimizing experimental resource use by minimizing the number of samples required to reach the target. PREP iteratively expands the dataset in a rational manner, significantly reducing the number of experimental iterations compared to conventional methods and ultimately lowering material and time consumption. Rather than depending solely on individual LVM alignment metrics (e.g., Hotelling T^2 and SPE), we introduce a composite parameter referred to as the Prediction Reliability Enhancing Parameter (PREP) that combines multiple monitoring metrics including not only Hotelling T^2 and SPE but also other relevant parameters. The covariance similarity between these metrics in the calibration data points and new unseen data points serves as a robust indicator for accepting or rejecting model predictions. Our approach iteratively refines the Potential Design Space (PDS) by focusing on subsets of data in each iteration, selecting candidates that closely resemble well-predicted samples to ensure higher prediction accuracy and improving model performance over time to offer a more consistent estimation of prediction uncertainty across various observations. The proposed method is illustrated through simulations on multiple types of nonlinear datasets, showing superior performance and fewer iterations compared to other available methods.

It should be noted that the choice of Tomba's framework for benchmarking design space identification in this work was made based on the similarity of Tomba's method and our method in terms of relying on first identifying a member of the Design Space (DS) and then expanding upon that foundation to define the region spanned by the DS. We acknowledge that alternative approaches exist, including geometrical methods that explicitly define the entire DS region by approximating constraint boundaries, probabilistic methods that determine DS feasibility by estimating probability distributions within the input space, and black-box methods that identify the DS through iterative exploration and function optimization. However, once a geometrical or probabilistic method defines the DS region, an additional step is still required to select specific candidates for synthesis. PREP could thus serve as a tool for ranking these candidates based on prediction reliability. Conversely, black-box approaches do not leverage dataset-inherent structure and often rely on global function optimization rather than structured latent-variable-driven exploration, in contrast to our (and Tomba's) method. Since our method is fundamentally aligned with approaches that prioritize dataset-driven sample selection and refinement, we chose to benchmark it against Tomba et al.'s framework that shares this data-driven philosophy, ensuring a more meaningful and relevant comparison.

The remainder of this article is organized as follows. In Section 2, preliminary information is provided. This includes an overview of the PLS (Partial Least Squares) and PCA (Principal Component Analysis) frameworks, as well as a discussion of the available model alignment metrics like Hotelling T^2, SPE (Squared Prediction Error), and latent variable score alignment metrics (h_{PLS}, h_{PCA}), and their relevance in assessing model prediction reliability. Section 3 introduces the proposed methodology, detailing how the Prediction Reliability Enhancing Parameter (PREP) is used to identify candidates for which model predictions are expected to be more reliable. In Section 4, we describe the generation of a simulated dataset and the assessment of its degree of nonlinearity, followed by an examination of two different datasets with four different target outcomes to be achieved using our method and the above-mentioned

scenarios using Tomba's work. Section 5 presents the results and compares the performance of different methods, and finally, Section 6 draws the conclusions.

3.2 Preliminaries

3.2.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) reduces a dataset's dimensionality by identifying principal components—linear combinations of the original variables (P $[I \times A]^{\mathbb{R}}$) Here, A represents the number of principal components or the dimensionality of the latent space used to approximate the original I-dimensional space—that capture key variance in the data. These principal components are uncorrelated, enabling efficient analysis of interdependent data while retaining all essential information. The resulting reduced representation (T $[N \times A]^{\mathbb{R}}$) simplifies the dataset and preserves the main structure by capturing the most critical variance. The main mathematical equations of PCA are given in Equation 3, while its data blocking configuration is illustrated in Figure 3-1(a).

$$X_{\text{Original data}} = T.P^T + E \quad (eq.3)$$

where the E $[N \times I]^{\mathbb{R}}$ is the residual between X and its representation using the already trained PCA.

3.2.2 Partial Least Square-Projection to Latent Structure (PLS)

Partial Least Squares (PLS) is a predictive modeling method that analyzes the relationships between two data blocks, typically input (X) and output (Y), and maximizes the correlation between independent components derived from both blocks to highlight how variations in X drive changes in Y. PLS strategies generate T scores for the input data (X) and U $[N \times A]^{\mathbb{R}}$ scores for the output data (Y). Using the coefficients P, the data can be projected from the original X space to the score space (T). Conversely, W* $[I \times A]^{\mathbb{R}}$ coefficients allow the data to be projected back from the score space to the original X space, effectively capturing and reconstructing the key relationships and interdependencies between the input

and output datasets [30-32]. The key equations used to describe a PLS model are given in Equation 4, with the schematic of a PLS model with its blocking configuration shown in Figure 3-1(b).

$$X_{\text{Original data}} = T.P^T + E$$
$$Y_{\text{Original data}} = T.Q^T + F$$
$$T = X.W^* \qquad (eq.4)$$
$$t_{\text{new}} = x_{\text{new}}.W^* \rightarrow y_{\text{new}} = t_{\text{new}}.Q^T$$
$$\beta_{\text{pls}} = W^*.Q^T;\ y_{\text{new}} = x_{\text{new}}.\beta_{\text{pls}}$$

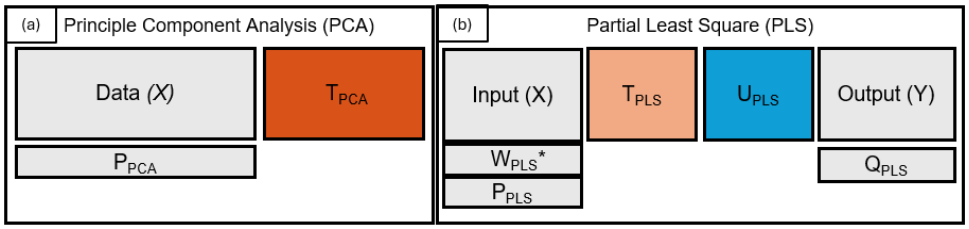


Figure 3-1: PCA (a) and PLS (b) blocking configurations

3.2.3 Latent Variable Model Inversion (LVMI)

PCA or PLS models are predictive tools that use input data (X) to estimate outputs (Y). In product design the inverse use of these models is often needed, by which a product designer would start with a desired output ($Y_{desirable}$) and would seek to determine the input values that can achieve it. Three scenarios arise based on the number of components (A) and the number of output variables (K):

1. If A < K: No exact input produces $Y_{desirable}$, but the model inversion finds an input where $Y_{predicted}$ is closest to $Y_{desirable}$.
2. If A = K: There is a single solution where $Y_{predicted}$ equals $Y_{desirable}$ that can be identified by the model inversion.
3. If A > K (most common): Multiple inputs yield $Y_{predicted} = Y_{desirable}$, with the Null Space (NS) offering infinite input solutions without altering the output prediction.

Equations 5 to 7 outline how potential input solutions can be determined in each case based on the PLS model blocks [1, 21-23, 26].

$(A<k)$: $\tau_{des}=(Q^T.Q)^{-1}.Q^T.y^{des}$ (eq.5)

$(A=k)$: $\tau_{des}=Q^T.y^{des}$ (eq.6)

$(A>k)$: $\tau_{des}=(Q^T.(Q^T.Q)^{-1}.y^{des})+\Delta\tau_{NS}$ (eq.7)

where τ_{des} $[1 \times A]^{\mathbb{R}}$ represents the $t_{pls-scores}$ of the potential input set yielding $Y_{desirable}$ $[1 \times K]^{\mathbb{R}}$ and $\Delta\tau_{NS}$ $[l \times A]^{\mathbb{R}}$ represents an infinite number of l points within the null-space where the variation in latent space has no effect on $Y_{Predicted}$.

To proceed, it is essential to distinguish between the Design Space (DS), Knowledge Space (KS), and Null Space (NS). The KS represents the portion of the input space that has been explored and utilized for model development, while the DS consists of input combinations that yield acceptable product outcomes based on predefined quality criteria. In some cases, the DS is a subset of the KS; in others, the DS is located entirely or partially outside the KS due to limitations in initial sampling or modeling constraints, requiring rational KS expansion to ensure the DS is properly identified. Finally, the NS refers to input variations that do not affect the output, providing flexibility within the DS without altering the desired system response [33-38]. DS can exist without NS if there is a range of acceptable outputs (Y) encompassing inputs that meet this range. However, if both an acceptable output range and NS are present, DS becomes multidimensional in that it incorporates NS across all acceptable Y values. In manufacturing, aligning calculated NS with actual DS enables greater flexibility in input adjustments without changing product properties. This alignment is straightforward in linear systems but requires iterative expansion in nonlinear systems to cover regions likely containing the DS, ensuring the calculated NS matches the actual DS. Direct model inversion may not always be efficient to achieve this goal, particularly with constraints on inputs (X) or the presence of NS that allow multiple input solutions to achieve the same Y target. Exploring solutions along the NS may reveal more efficient or flexible options.

Tomba's framework introduced four scenarios for identifying potential solutions [1]. The first scenario assumes no constraints on the X variables, with the targeted output thus fully determined and direct model

inversion being appropriate. In the second scenario, there are no constraints on X but the output is acceptable within a range, requiring exploration of the solution space. Scenarios three and four introduce constraints on some of the X variables in which Y is either fully or partially defined, making direct model inversion less practical and necessitating iterative approaches to find feasible solutions within the constrained design space. We specifically focus on the first scenario, as it is expected to yield the best result when there are no constraints, and the fourth scenario, which addresses a more general situation. In this fourth scenario, once the region with a high likelihood of containing the design space is identified, the optimization framework described in Equation 8 is suggested to find the most optimal solution based on existing data to be experimentally tested and added to the dataset for the next iteration (provided the target has not already been achieved).

$$\min_{x^{new}} \left\{ g_1 \left(\hat{y}^{new}\text{-}y^{des}\right)\Gamma\left(\hat{y}^{new}\text{-}y^{des}\right)^T + g_2 \text{ Hoteling } T^2_{pls} + g_3 \text{ SPE}_{x^{new}} + g_4 \, d^{-1}_{(\tau,T)} \right\} \quad \text{eq.8}$$

$$\text{s.t.}$$
$$\hat{y}^{new} = \tau.Q^T$$
$$\hat{x}^{new} = \tau.P^T$$
$$\tau = x^{new}.W^*$$
$$d^{-1}_{(\tau,T)} = \frac{1}{\min[(\tau\text{-}\tau_n)^T.\Lambda^{-1}.(\tau\text{-}\tau_n)] + \text{Const}} \quad \forall n \in \{1,2,\ldots,n\}$$

where Γ is a $[K \times K]^{\mathbb{R}}$ diagonal matrix containing the weights assigned to each output variable (emphasizing their relative importance) and g_i represents the weight of each term. It should be noted that any hard constraints can be applied to the suggested x^{new}, its corresponding $\hat{y}^{new}$, or its calculated PLS SPE and Hotelling T^2, allowing for acceptance or rejection of any x^{new} from the outset. The last term in Equation 8 introduces a penalty if the new iteration's suggestion is too close to either an existing dataset member or a previously suggested solution from earlier iterations. This approach, adapted from [28], aims to prevent the optimization process from becoming trapped in local minima and repeatedly suggesting solutions similar to prior iterations or existing calibration samples. The weight of this penalty is

dynamically adjusted based on the proximity of the existing sample to the target value, such that as the suggestion approaches the target the importance of avoiding similarity to existing data diminishes. The goal is to find candidates that not only satisfy the desired output values (the first term of Equation 8) but also minimize the expected prediction uncertainty (the second and third terms of Equation 8), using PLS Hotelling T^2 and PLS SPE (Squared Prediction Error) as metrics to assess the validity of a trained model for new input data. PLS Hotelling T^2 (Equation 9) measures the extent to which a new observation lies within the statistical boundaries of the original data used to calibrate the model (providing a measure of how 'in line' the new data is with the model's existing data structure) while, PLS SPE (Equation 10) calculates the residual error of a new observation relative to the model (indicating how well the new data point can be represented by the model); in Equations 9 and 10, $t_{a,i}$ is the a^{th} score value of the i^{th} observation and S_a is the standard deviation of the a^{th} column of T_{pls}. As such, lower values of Hotelling T^2 and SPE indicate that the new data align well with the model's calibration set and the prediction is likely to be more reliable.

$$\mathrm{SPE}_{\mathrm{pls,x}} = e_i e_i^T \; where \; e_i = x_i - \tau_i . \mathrm{P}_{\mathrm{pls}} \; (eq. 9)$$

$$\text{Hoteling } \mathrm{T}^2_{\mathrm{pls}} = \sum_{a=1}^{A} \left(\frac{t_{\mathrm{pls,a,i}}}{s_a}\right)^2 \; (eq. 10)$$

3.3 Proposed Methodology

While PLS Hotelling T^2 and SPE provide useful insights into how new observations align with an existing model, they focus more on the compatibility of the data with the model rather than offering a direct pathway for optimizing towards specific desired outputs. To address this limitation, we propose a novel methodology that not only considers prediction alignment but also provides a systematic approach to refining the design space and enhancing the prediction accuracy for desired outcomes. In our proposed method, we define an experimental space where the TDS is expected to reside, generate candidate recipes with predicted outputs matching the desired target, and rank them using a new metric to prioritize which

recipes are most likely to expand the dataset toward the TDS, a process summarized schematically in Figure 3-2.

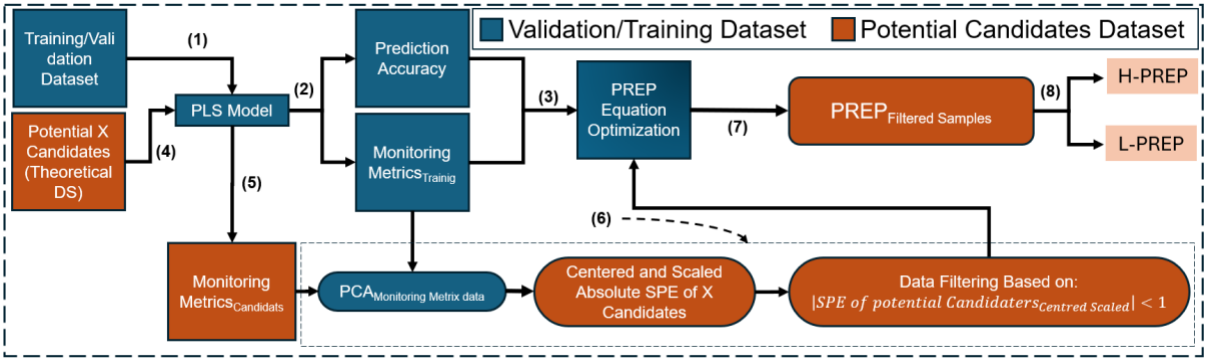


Figure 3-2: Schematic representation of the proposed PREP method. Blue boxes denote the training/validation data for which the actual Y values are known and used for optimizing the PREP equation. Orange boxes represent the dataset of potential candidates for which only X values are available, and the candidates must be ranked, with the candidates selected via the PREP method to be experimentally tested.

This process is repeated iteratively to refine the design space and achieve the target with a predetermined level of accuracy. Our approach keeps the rules used in traditional methods that emphasize keeping low Hotelling's T^2 and SPE values while iteratively generating new datapoints by focusing on the covariance similarity between different model alignment metrics of the well-predicted samples among the calibration (or validation) dataset and that of potential future solutions.

3.3.1 Involved Model Alignment Parameters and Initial Blocks Creation

To implement this process, a standard PLS model is first developed using the k-nearest neighbors of the calibration dataset corresponding to the targeted $Y_{desirable}$ (Step 1 in Figure 3-2). This approach reduces the dataset size while increasing the likelihood of capturing a locally linear (or at least more linear) structure near the area of interest. Typically, k is set between 5 and 10 samples, but in cases with a higher number of input variables, selecting more neighbors may be necessary to maintain model reliability. Additionally, the number of nearest neighbors chosen should not limit the number of PLS components required to adequately explain the model, which is chosen based on the effectiveness of regenerating the X data; this is particularly crucial since the model will be inverted to generate potential X values from targeted Y outputs [1, 21]. Next, two additional pieces of information are required: (1) a matrix of available model

alignment parameters (calculated using only the X data and the trained model) from either the training or validation dataset that we call the Monitoring Metrics Matrix; and (2) a metric that assesses the prediction accuracy of the model on these datasets (Step 2 in Figure 3-2). In our method, we utilize the following model alignment parameters, all of which assess model accuracy from a different perspective:

• *PLS Hotelling T^2*

- *Concept*: As previously mentioned, the PLS Hotelling T^2 score is a multivariate metric that assesses how far a new data point is from the center (mean) of the PLS model's distribution of data in the latent space.
- *Importance*: It captures whether a new data point is "typical" or too extreme relative to the training set in terms of its position in the latent variable space.

• *PLS SPE*

- *Concept*: The PLS SPE score measures how much variation in the new data point is not captured by the PLS model. While PLS seeks to explain the maximum covariance between X and Y, SPE quantifies how much of the original X data is not accounted for by the model.
- *Importance:* it shows how well the PLS model represents the new data point.

• *PCA Hotelling T^2*

- *Concept*: Like PLS Hotelling T^2, PCA Hotelling T^2 measures how far a new data point is from the center of the data's distribution in the PCA space. In this case, PCA does not consider the relationship between X and Y but simply reduces the dimensionality based on variance in X (Equation 11).

$$\text{Hoteling } \mathrm{T}^2_{\mathrm{pca}} = \sum_{a=1}^{A} \left(\frac{t_{\mathrm{pca},a,i}}{S_a}\right)^2 \quad (eq.11)$$

where $t_{pca,a,i}$ is the a[th] score value of the i[th] observation and S_a is the standard deviation of the a[th] column of T_{pca}

- *Importance*: It helps assess whether the new data point is consistent with the variance structure found in the training X data.

• *PCA SPE*

- *Concept*: Like PLS SPE, PCA SPE is the squared error of the reconstruction of the new data point in the PCA model, measuring how much of the variance in the new data point cannot be captured by the principal components found in the training data (Equation 12)

$$SPE_{pca,x} = e_i e_i^T \; where \; e_i = x_i - \tau_{pca,i}.\mathrm{P_{pca}} \; \; where \; \tau_{\mathrm{pca},i} = x_i.\mathrm{P_{pca}^T} \; \; (eq.12)$$

- *Importance:* If PCA SPE is high, it suggests that the new data point is atypical compared to the training data in terms of its X features, potentially indicating poor model performance.

• *PLS Alignment Score (h_{PLS})*

- *Concept*: The h_{PLS} statistic is a quadratic form of the new data's score vector ($t_{pls,new}$) transformed by the PLS score covariance matrix. This metric evaluates how well the new data's latent variables fit within the structure of the existing PLS model (Equation 13).

$$\mathrm{h_{PLS,x_{new}}} = \mathrm{t_{pls_{xnew}}}(T_{pls}T_{pls}^T)^{-1}\,\mathrm{t_{pls_{xnew}}}^{\mathrm{T}} \; (eq.13)$$

- *Importance*: It captures the alignment between the new data and the established PLS latent space. A high h_{PLS} value suggests that the new data are unusual or do not align well with the trained PLS model, signaling a potential reliability issue with predictions.

• *PCA Alignment Score (h_{PCA})*

- *Concept*: Like h_{PLS} but calculated in the PCA space, the h_{PCA} value reflects how well the new data fits the principal components found in the training X data (i.e. how close the new data is to the principal subspace of the training data) (Equation 14).

$$\mathrm{h_{PCA,x_{new}}} = \mathrm{t_{pca_{xnew}}}(\mathrm{T_{pca}T_{pca}^T})^{-1}\,\mathrm{t_{pca_{xnew}}}^{\mathrm{T}} \; (eq.14)$$

- *Importance:* h_{PCA} indicates whether the new data align with the training data's variance structure, helping to identify outliers that might not be well-represented by the PCA model.

The h_{PLS} parameter thus evaluates new data based on the relationship between X and Y (i.e. how well the new data align with the model's explanation of the covariance between X and Y) while h_{PCA} only looks at the structure in X without any regard to the response variable Y. It should be emphasized that the score alignment metrics (h) and the Hotelling T^2 capture different aspects related to potential variability in a particular dataset. Hotelling T^2 focuses on how *far* the point is from typical data while h_{PLS}/h_{PCA} focus on the *alignment* of the new sample with the latent structure; stated in a different way, Hotelling T^2 functions as a geometric measure akin to the Mahalanobis distance while h_{PLS} and h_{PCA} evaluate the fit and projection consistency within the model's latent space. As such, each chosen parameter evaluates the data from a different perspective: PLS Hotelling T^2 and SPE assess the consistency of the X and Y relationship, PCA Hotelling T^2 and SPE check the variance structure of X, and h_{PLS} and h_{PCA} provide a more detailed assessment of fit within the defined latent spaces. Together, these metrics offer a comprehensive evaluation to ensure all aspects of the data are thoroughly considered both for well-predicted samples and for a new unseen observation.

The final step in preparing this block of model alignment matrices involves normalization to ensure that subsequent steps are not influenced by differences in the magnitude of each data column. To achieve this, the 95% confidence limits for both the PLS and PCA SPE and Hotelling T^2 are calculated for the first four columns of data. The data are then divided by these values, allowing samples exceeding these limits to be considered with a higher chance of lower prediction accuracy. For the alignment factors h_{PLS} and h_{PCA}, each column's maximum value is determined, and all data points in that column are scaled by dividing them by this maximum. This normalization prepares the dataset for subsequent analysis steps. After preparing the initial data as described above, a numerical metric was developed to represent the prediction accuracy for each member of the validation dataset and thus assess the prediction quality of a new sample recipe. To achieve this, Equation 15 is utilized:

$$Prediction\ Accuracy_{single\ sample} = \frac{\sum_l^L w_i r_l}{L}$$

$$r_l = 1 - \frac{e_l}{\text{normalizer}_\text{l}} = \frac{\left|y_{l,} - \hat{y}_l\right|}{\frac{Y_l - \min(Y_l)}{N}} \qquad (eq.15)$$

Here, L represents the number of output variables, N denotes the number of samples in the calibration dataset, l signifies the l^{th} output variable, and Y_i refers to the l^{th} column of the output data used for model training. The weighting factor w_i allows for the prioritization of specific output variables as desired. Instead of normalizing the Y values initially, we use a normalization factor to ensure that prediction accuracy is minimally affected by the scale of the i^{th} column of Y. This approach allows the discrepancy between the actual and predicted values to be reported relative to the scale of each variable independently; in contrast, normalizing the data upfront could introduce biases, as errors for values near zero would be disproportionately higher compared to those near one. This metric provides a single prediction accuracy value for Y data with multiple columns and can yield negative values for poor predictions, with the range of the prediction accuracy metric spanning from $-\infty$ (poorest prediction) to 1 (perfect prediction).

To gather these required blocks of data, there are two possible scenarios:

1. *When the dataset is sufficiently large*: In this case, the data can be partitioned into two sets: one for developing the PLS model and the other for validation. For the validation set, all the aforementioned model alignment parameters are calculated and correlated with their prediction accuracy using the actual Y values and Equation (15).
2. *When the dataset is small*: If the dataset is too small to set aside a portion for validation from the outset, the final results can be highly sensitive to the choice of validation data. In such cases, a jackknife approach is recommended in which one observation is left out at a time, the trained model is used to predict its Y value, and the prediction accuracy is determined. Ultimately, a PLS model trained on the full dataset (with k-nearest neighbors) would be used in this case to generate potential candidates.

3.3.2 Optimization Framework

Next, a correlation structure must be developed to elucidate the relationship between all model alignment values and their respective prediction accuracies (Step 3 in Figure 3-2). Identifying this structure requires the use of an optimization algorithm to determine the optimal coefficients and powers for the PREP, as specified in Equation 16. The objective is to configure the algorithm such that samples with high prediction accuracy are assigned lower PREP scores whereas samples with lower prediction accuracy are assigned higher PREP scores. The optimization framework is presented in Equation 17, with the schematic for the ideal case depicted in Figure 3-3.

$$PREP = c_1 \text{hotelingT2}_{\text{pls}}^{p1} + c_2 \text{SPE}_{\text{x,pls}}^{p2} + c_3 \text{hotelingT2}_{\text{pca}}^{p3} + c_4 \text{SPE}_{\text{pca}}^{p4} + c_5 \text{h}_{\text{pls}}^{p5} + c_6 \text{h}_{\text{pca}}^{p6} \qquad (eq.16)$$

$$\min_{c_i p_i} (Data\ Linearly\ Distribution\ Forces + PREM\ Range\ Forcer) \qquad (eq.17)$$

$$s.t.$$
$$c_i \in (0, \alpha = 1]$$
$$p_i \in (0, \beta = 1]$$

In Equation 17, the cost function of the optimization framework consists of several components that can be grouped into two categories:

1. **Data Linear Distribution Forcers:**

   These terms aim to align the PREP scores with the predictive accuracy of the model, aiming to ensure that samples with higher predictive reliability (i.e., those better captured by the model) are assigned lower PREP scores while less reliable predictions receive higher PREP scores. This alignment ensures that PREP scores reflect how well the model captures each sample's behavior, reinforcing the correlation between the PREP coefficient and the underlying predictive accuracy.

2. **PREP Range Forcers:**

   These components ensure that the PREP scores remain within a standardized range (0 to 1) across the dataset, maintaining consistency between the training and validation phases and ensuring

comparability of PREP scores across different datasets and iterations. Ideally, the optimal sample should have a PREP score near 0 (indicating high reliability) while predictions with higher uncertainty should have PREP scores closer to 1.

The ranges for the parameters c and p are defined to ensure they yield only positive values (in this work they are selected in the range of 0 to 1), aligning with the expectation that higher values of the model alignment metrics correspond to poorer predictions. Figure 3-3 provides a schematic representation of the ideal results from the optimized PREP equation, illustrating the considerations taken during the optimization process.

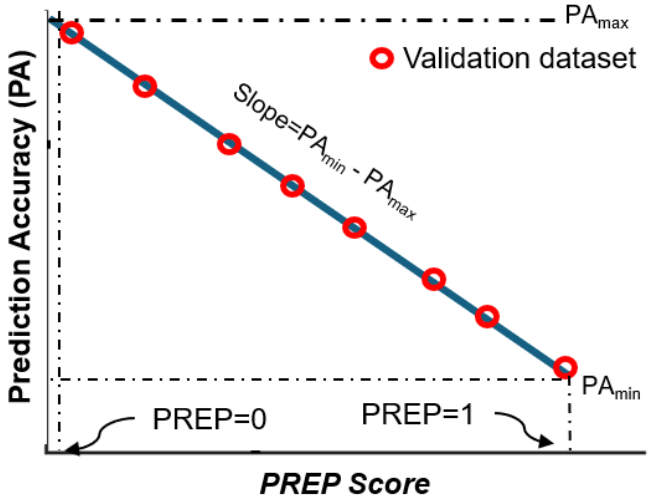


Figure 3-3: A schematic of an ideal PREP equation optimization showing what is valued within the optimization algorithm for the cost function.

Considering the nature of the optimization process, a population-based algorithm was also needed to generate a list of to-be-optimized parameters and manipulate them iteration to iteration to find the best possible values. Among the two of the most practical algorithms in this field - Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) - we chose to use the PSO approach because it offers results comparable to GA but was simpler and significantly faster to implement. In the PSO algorithm, a "particle" (potential solution) adjusts its position within the search space by considering three factors: (1) the best position it has ever been (its personal best), (2) the best position any particle in the group has been (the global best), and (3) its current direction of movement. With each iteration, the particle adjusts

its velocity and position based on these three criteria, hoping to balance exploration (finding new areas of the search space) with exploitation (refining known good areas). Over time, these particles "explore" the solution space and converge toward the best solution, assuming that their collective behavior will lead to an optimal or near-optimal result.

For new candidate samples, PREP scores may sometimes exceed 1 since certain alignment metrics (such as Hotelling T^2) can be significantly higher for future observations compared to the normalized range established during the initial model training. It is essential to emphasize that the PREP method intentionally selects two samples at each iteration: one with the lowest PREP score (L-PREP) and one with the highest PREP score (H-PREP). This dual selection serves complementary purposes. L-PREP is chosen for its expected high prediction accuracy, as it closely aligns with key monitoring metrics; in contrast, H-PREP plays a crucial role in expanding the Knowledge Space (KS) by exposing the model to conditions it has yet to encounter, improving the model's ability to achieve the target output. Additionally, in certain cases—particularly when the target lies far outside the range of available data or even the closest samples—PLS models may struggle to achieve reliable prediction accuracy, with the even best accuracy among the validation dataset being low. In such scenarios, the risk of PREP failure is higher, making it even more crucial to select H-PREP samples to promote KS expansion and increase the likelihood of reaching or moving closer to the target value. Overall, the PREP approach is designed to enhance the model's generalization and adaptability while at the same time expediting the process to achieve a faster identification of the design space.

3.3.3 PREP Implementation for Design Space Candidates

Following the optimization step, a PREP equation will be developed with coefficients and powers specifically tailored to the dataset of interest. This equation undergoes validation during the optimization process, using validation datapoints from the training dataset. As a result, the PREP score provides a reliable indication of the expected accuracy for samples not directly included in the calibration of the PLS model. In the subsequent step, it is necessary to define the region most likely to encompass TDS, which

can be represented as a theoretical null space where $Y_{predicted}=Y_{desirable}$ excluding any prediction uncertainty (Step 4 in Figure 3-2). When such a null space exists, it theoretically allows for the generation of an infinite number of samples depending on the step size used to move from one candidate to the next. To make this step practical, we generate a limited list of N candidates, with N set to 200 when the null space lies within the bounds of the X data and to 20 when it falls outside the bounds and thus requires optimization to generate feasible candidates. The goal of this optimization is to ensure that $Y_{predicted}$ is as close as possible to $Y_{desirable}$ while maintaining sufficient differentiation among the candidates. Once this region or list of candidates is established, the trained PLS model can be employed to calculate all relevant model alignment metrics for these candidates (Step 5 in Figure 3-2). Using the optimized PREP equation, the PREP scores for each candidate are calculated and the candidates are ranked based on these scores; candidate recipes corresponding to the L- PREP and H-PREP scores are then selected for synthesis to initiate the iterative process (Steps 7 and 8 in Figure 3-2).

3.3.4 Refinement of Design Space Candidates

In our methodology, we ensure that the model alignment parameters of potential candidates are similar to those of the validation dataset. While previously methods merely check if candidate recipes fit within the same structure as the validation dataset—as assessed by SPE and Hotelling T^2—we perform Principal Component Analysis (PCA) on the various model alignment metrics from the training/validation dataset and use the resulting model to calculate the SPE for new candidates based on their corresponding model alignment metrics (Step 6 in Figure 3-2). We apply absolute center scaling to the SPE values of all candidates and retain only those with a scaled SPE value less than 1, enabling overly aggressive candidates to be filtered out while ensuring that the statistical structure of the candidates aligns with that of the validation dataset. Candidate recipes within this range are considered more representative of normal X candidates for the given target, reducing the risk of selecting candidates with potentially misleading PREP scores that could skew the iterative process in an undesirable direction.

3.3.5 Iterative Execution of PREP Methodology

To summarize the proposed PREP methodology, the following steps outline the iterative usage of PREP to achieve optimized model alignment and accurate predictions:

1. Select k-nearest neighbors to the target and train PLS and PCA models.
2. Calculate monitoring metrics and prediction accuracies for validation/training data points
3. Optimize PREP equations using alignment metrics and prediction accuracies.
4. Generate Potential Design Space (PDS) candidates with valid X and close proximity between $Y_{predicted}$ and $Y_{desirable}$.
5. Calculate Monitoring Metrics for all PDS candidates.
6. Refine PDS based on the metric alignment.
7. Rank the refined PDS candidates based on PREP scores.
8. Select H-PREP and L-PREP samples corresponding to the highest and lowest PREP scores from the ranking.
9. Synthesize and characterize the selected samples
10. If the target is unmet, update the k-nearest neighbors and iterate until a solution is identified.

The authors developed and implemented all modeling codes using MATLAB R2024b. The simulations were executed on a system with an 11th Gen Intel(R) Core (TM) i7-1165G7 @ 2.80GHz processor and 16 GB RAM. Each iteration, involving a PSO-based optimization (500 iterations, 100 initial particles, repeated five times), required ~30 seconds; as such, the method's computational cost is moderate consistent with its primary objective being experimental efficiency rather than purely computational speed. While higher data dimensionality may require more nearest neighbors in the optimization process, the number of decision variables in the PREP equation remains fixed, thus limiting the scalability impact in higher dimensions.

3.4 Model Assessment

To robustly evaluate the effectiveness of our method, we developed a series of simulated datasets with a tunable level of nonlinearity in which the underlying data structure was known, allowing us to iterate towards a predetermined target while monitoring performance. Specifically, since the aim of the simulated dataset analysis was to assess how quickly our method could identify a member of TDS, we required both a relatively complex dataset and the equation that generated it; this setup enabled us to define the complete set of potential process outputs, establish rational target selection, and provide insight on the actual outcomes of the method's suggestions, guiding it towards the correct direction in the TDS.

Five datasets of varying complexities were generated to evaluate the new PREP method. The two datasets presented in the main body of the paper were chosen to represent different levels of nonlinearity, with the first dataset exhibiting a lower level of nonlinearity generated using trigonometric functions applied to the input variables, and the second dataset introducing a higher level of nonlinearity through the combination of trigonometric functions and complex power-law expressions featuring interdependent exponents. The remaining datasets, discussed in the Supplementary Information, include a third dataset that employs nested power-law terms with variable-dependent exponents, resulting in a significantly higher degree of nonlinearity; a fourth dataset that incorporates trigonometric, exponential, and power-law functions, adding periodicity, oscillations, and greater complexity through interdependencies between input variables; and a fifth dataset that combines sinusoidal, logarithmic, and square-root terms, introducing moderate nonlinearity while maintaining a simpler structure compared to the other datasets. Together, these datasets ensure a comprehensive representation of the PREP method's applicability across a range of complexities.

For dataset creation, we imposed limitations on the X data to better mimic real-world scenarios and provide a realistic range for system outputs. The numbers of input and output variables were set to 3 and 2, respectively, to enable visual representation of model performance during each iteration, more clearly demonstrate the method's ability to handle the multivariate nature of the system, and (by selecting more

inputs than outputs) increase the likelihood of a null space being present to allow for the creation of a theoretical design space that can be rationally assessed without relying solely on potential solution candidates.

It is worth noting that, although the entire dataset is calculated for visualization purposes, only 30 samples were randomly generated in each case to reflect real-world scenarios in which sample preparation is often either expensive or time-consuming. For the present case, the number of input variables is 3, limiting the maximum number of PLS components to 3. To allow for model computation during the iteration (while leaving out one data point for validation), at least five data points are needed to be kept as nearest neighbors. From the initial 30 samples, we therefore selected k=5 nearest neighbors to the targeted output for model development and identifying the design space corresponding to predefined targets. After each iteration, the set of 5 nearest neighbors is updated. If the newly synthesized sample is among the 5 nearest neighbors, it is automatically selected for the next iteration; however, if it performs worse than the others, the new sample is added to the 5 neighbor set and one of the previous members is randomly excluded. This approach ensures that iterations do not get stuck in a loop, continuing to explore alternative solutions if the previous suggestion was suboptimal. Figure 3-4 provides a schematic overview of how this simulated dataset was utilized during the iterations to achieve the final target.

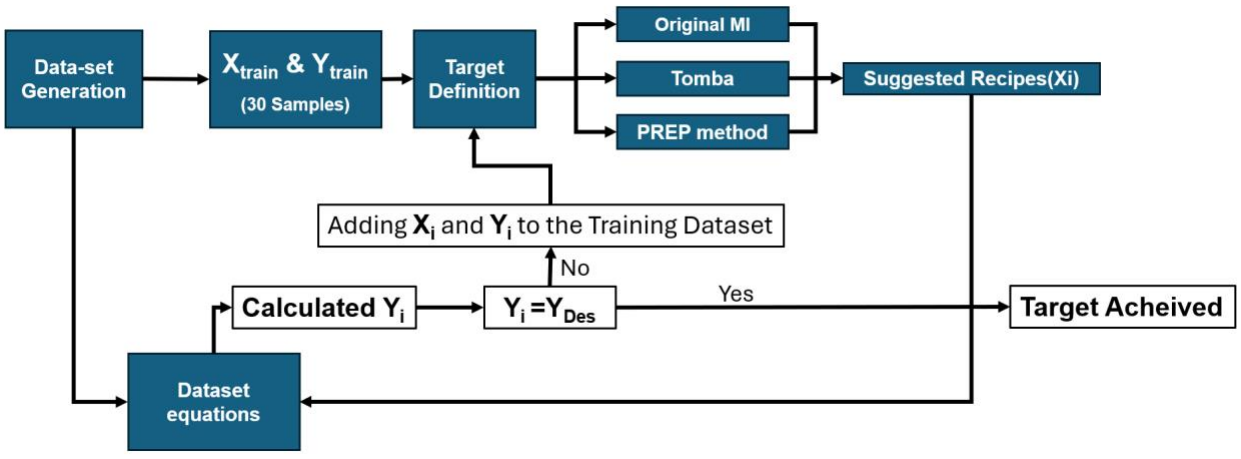


Figure 3-4: Schematic representation of the simulated dataset implementation for TDS identification and comparison of previously reported optimization techniques relative to PREP.

It is important to note that interactions between variables are inherently captured through the PLS modeling framework, which identifies latent variables that explain the maximum variance in both the X and Y data. As PREP is implemented iteratively, PLS continuously refines the mapping between input and output variables, effectively adapting to variable dependencies in the data structure. This enables PREP to capture complex nonlinear relationships within the iterative modeling approach without requiring explicit interaction terms.

To comprehensively assess the performance of the PREP method, each of the datasets is utilized and three different targets are selected for each dataset; the effectiveness of the PREP method relative to conventional model inversion (MI) as well as the method reported by Tomba (scenario 4) is then evaluated based on the number of iterations each method requires to reach the desired target, defined herein as PA > 95%. It is important to note that all datasets impose a constraint on the X values to remain within the range of 0 to 2, making the direct application of the original model inversion impractical; however, to maintain the utility of this method for comparison, for all targets and at every iteration any suggested X value below 0 is set to 0 and any value above 2 is set to 2.

It is important to emphasize that, unlike the MI and Tomba approaches, the PREP method selects two samples at each iteration: L-PREP (which is associated with high expected prediction accuracy) and H-PREP (which presents a significant level of uncertainty in proximity to the target). This uncertainty may facilitate the model's ability to identify critical areas for enhancement. In addition, a potential advantage arises if, during the final iteration, both L-PREP and H-PREP samples demonstrate comparably high accuracy, an outcome that could signify that the TDS has been comprehensively captured by the PDS in the final iteration. Such a scenario would not only affirm the model’s robust predictive capabilities in relation to the target output but also create opportunities for further optimization of solutions by incorporating secondary considerations such as time, energy, or cost-effectiveness into the recipe identification process. However, to ascertain that the TDS is entirely represented, it may be necessary to synthesize samples with PREP scores that more closely align with the average of the distribution. This

supplementary step would serve to validate whether the existing candidates adequately encompass the entire design space or if additional exploration is warranted.

3.4.1 First Simulated Dataset

The first dataset used in this study was generated using the following equation:

For X data:

$$X_{raw} = rand(N, 3)_{[0\ 2]} = [X_1, X_2, X_3]$$

For Y data:

$$Y_1 = \sin\left(\frac{\pi}{6}X_1\right) * \sin\left(\frac{\pi}{6}X_2\right) * \sin\left(\frac{\pi}{6}X_3\right)$$

$$Y_2 = \cos\left(\frac{\pi}{6}X_1\right) * \cos\left(\frac{\pi}{6}X_2\right) * \cos\left(\frac{\pi}{6}X_3\right)$$

$$Y = [Y_1, Y_2]$$

From this function, a training dataset of 30 randomly generated samples was created and three selected target points were selected with a focus on regions that are underrepresented (T1 and T2) and one that lies completely outside the range (T3) (Figure 3-5(a)). These specific areas are of particular interest because they align with the primary objective of the PREP method: to address regions where the model has less coverage by leveraging covariance similarities between the monitoring metrics of PDS candidates and the calibration dataset. To ensure consistency and eliminate sources of random uncertainty, the same dataset and target points were used across all three methods tested (Original Model Inversion (MI), Tomba, and PREP).

Table 3-1 presents the PLS X_R^2 and Y_R^2 values for the entire dataset together with the results for the three separate PLS models developed using the five nearest neighbors to each target. A significant increase in Y_R^2 is observed when using the five nearest samples compared to the PLS model trained on the full set of 30 samples, consistent with a smaller number of samples that are closer to each other being easier to

describe with a linear model compared to the entire dataset that presents higher degrees of variation. Figure 3-5(b to d) present the TDS (all combinations of X data columns for which the actual Y falls within a 95% accuracy range of the targeted Y) in the original X space, providing a quantitative overview of how closely each target's TDS aligns with the calibration dataset and the acceptable X space range (between 0 and 2); Figure 3-5(e) displays the projection of this TDS in the PLS latent space utilizing the entire dataset to more clearly indicate that T2 and T3 are near the 95% confidence limit of the PLS model or entirely outside of it. Collectively, Figure 3-5(b to e) show that the distribution of TDS and its projection resemble a curve rather than a straight line, confirming the nonlinearity present in this dataset.

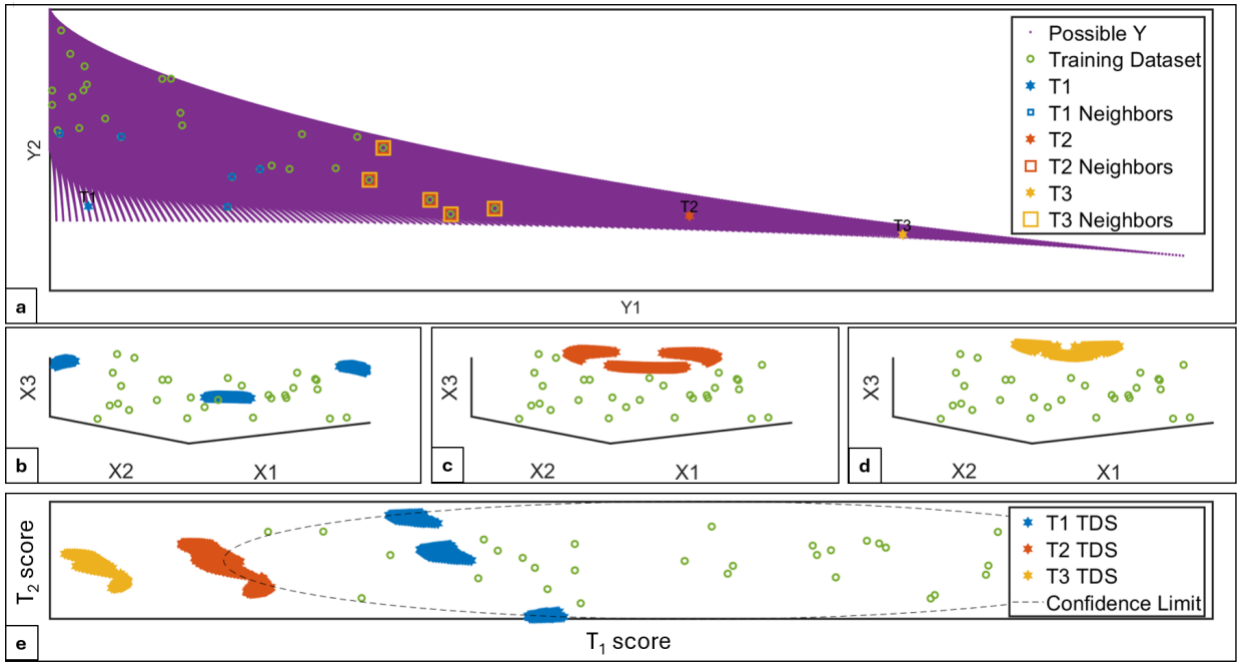


Figure 3-5: Distributions of the first simulated dataset and targets: (a) Selected targets 1 (T1), 2 (T2), and 3 (T3) along with corresponding nearest neighbors for each target; (b-d) TDS for T1 (b), T2 (c), and T3 (d) in the original space; (e) Projection of each TDS from the second row into the latent space.

Table 3-1: R^2 values from PLS conducted on the entire dataset compared to using the 5 nearest neighbors corresponding to each target for the first simulated dataset.

PLS Num Components	**Entire Dataset**		**T1 Neighbors**		**T2 Neighbors**		**T3 Neighbors**	
	R^2X (%)	R^2Y (%)	R^2X (%)	R^2Y (%)	R^2X (%)	R^2Y (%)	R^2X (%)	R^2Y (%)
1	26	82	20	85	49	40	49	40
2	58	86	53	94	95	54	95	54
3	10	88	100	95	100	99	100	99

The results for all targets across all methods are presented in Figure 3-6. For target 1, which was anticipated to be the easiest among the three targets to achieve based on its TDS projection into the latent space, the Original MI and Tomba methods reached 95% closeness to the targeted values in 5 and 4 iterations respectively while the PREP method also achieved the same level of closeness in 4 iterations. However, for targets T2 and T3 that were expected to more clearly benefit from dataset expansion considering the projection of their corresponding TDS in the latent space, the Original MI failed to approach the target and Tomba method made only minimal progress toward the target over 13 iterations for T2 and 17 iterations for T3. This slow progress is consistent with the heavy reliance of these methods on the calibration dataset and these models' more conservative approach regarding dataset expansion. In contrast, the PREP method effectively considered the similarity of all PDS candidates to well-predicted validation samples, enabling it to quickly move toward dataset expansion and thus hit the targeted formulations in 4 steps for T2 and 2 steps for T3. Interestingly, PREP reached the more challenging T3 target even faster than T2; we attribute this result to the closeness of the calibration dataset to the targeted values in T2 that caused the generated PDS to be more similar to the existing data, potentially introducing bias in the lower PREP scores assigned to samples closely resembling the calibration dataset that was not present in T3 given its distance from the entire training dataset.

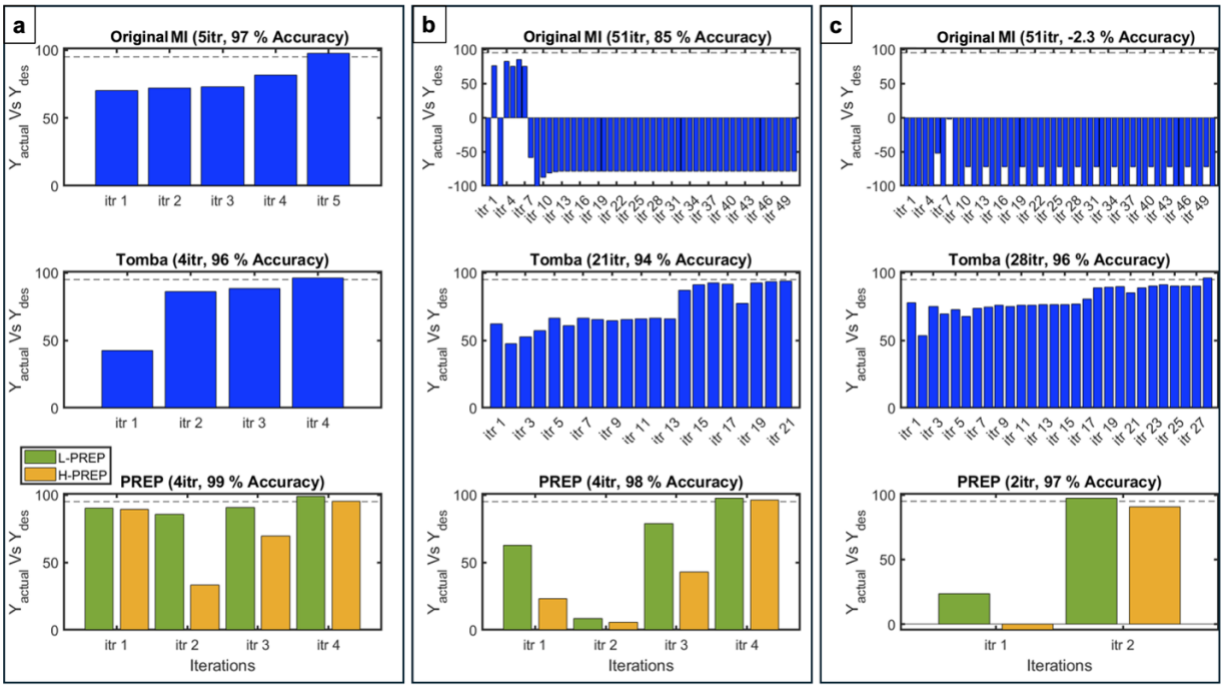


Figure 3-6: Performance of different methods in reaching (a) T1, (b) T2, and (c) T3 using the first simulated dataset. The first row displays the results from the Original Model Inversion, the second row displays the results from the Tomba method, and the third row illustrates the results from the proposed PREP method.

Figure 3-7 provides more insight into how the PREP method expanded the dataset to include T2. The top row shows how well the validation dataset aligns with the line over all four iterations required to hit the T2 coupled with the range of PREP scores for the PDS in each iteration, while the second row displays the projections of the PDS and TDS into the latent space highlighting how these projections changed over the iterations (only the first and second PLS scores are plotted here for clearer visualization, with the third score omitted for simplicity). Note that while the number of TDS samples remains constant throughout the iterations, the occupied area of these samples in the latent space may initially be smaller if the model is far from accurately representing these samples; over the course of the iterations, the area spanned by the TDS samples in the latent space is significantly increased, consistent with the PREP method systematically approaching this space. The third and fourth rows in Figure 3-7 present the actual Y values of the PDS plotted against their prediction accuracy, sorted based on their PREP scores (third row) and the actual Y values of the PDS versus T2 sorted based on their PREP scores (fourth row). The primary objective of PREP is to achieve good performance in the third row, where $Y_{predicted}$ is expected to closely

match Y_{actual}. In cases in which there is a null space, the last two rows are expected to show similar results for all candidates, with $Y_{predicted}$ =T2. However, in situations where samples with $Y_{predicted}$ =T2 do not satisfy the constraints on X, the PDS consists of samples whose X values are within the limits and whose $Y_{predicted}$ values are as close as possible to T2, a direct result of the optimization algorithm used for which the cost function is based on the proximity of $Y_{predicted}$ to T2 to ensure that the solutions differ from those already suggested.

As shown in the first iteration, the pls model in the first iteration is entirely invalid regarding $Y_{desirable}$, with an average Y_{actual} versus $Y_{desirable}$ accuracy being less than 0.4. However, by the last iteration, nearly all PDS samples fall within the TDS of the targeted output, demonstrating the effectiveness of the dataset expansion enabled by PREP and its validity in the area of interest. The general trend in the fourth row demonstrates a decrease in prediction accuracy (PA) as PREP scores increase, confirming that PREP effectively captures the underlying characteristics of the data to enable effective ranking of samples based on their prediction accuracy.

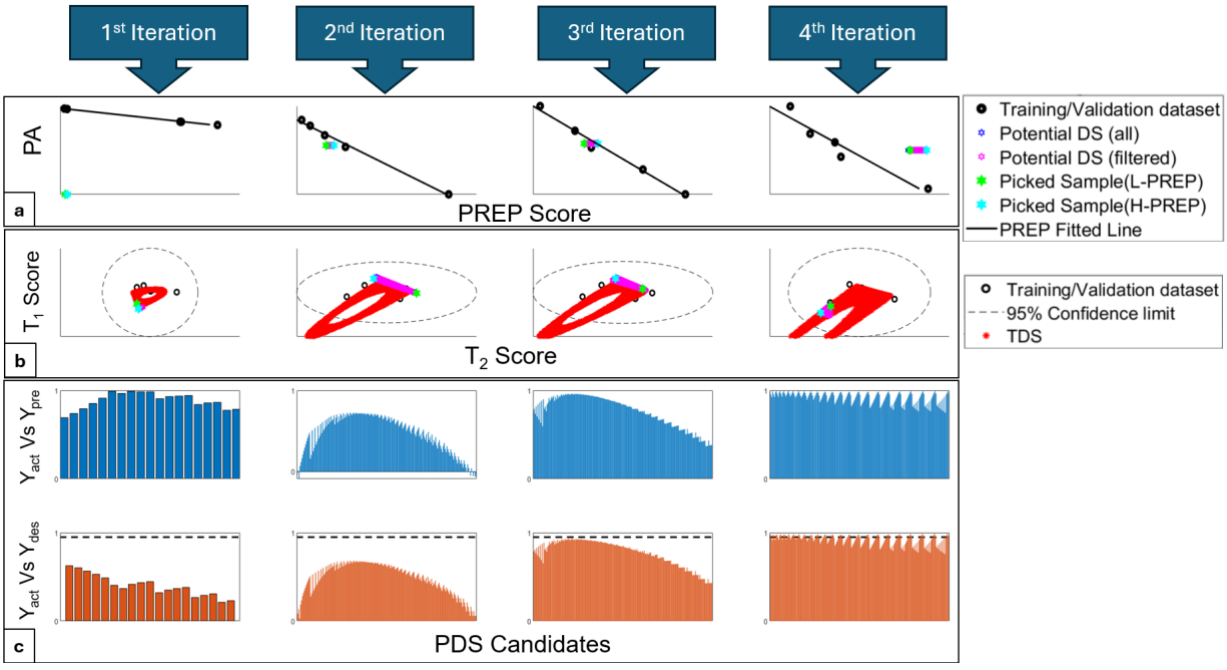


Figure 3-7: PREP iteration results from first to last iteration for T2 using the first simulated dataset. (a) the outcome of the PREP optimization and its alignment in each iteration; (b) the projections of the TDS and PDS into the PLS latent space in each iteration, focusing only on the first and second scores (the third score is excluded for simplicity);

(c) comparison of actual Y values, predicted Y values, and T2 values for all PDS samples sorted by PREP scores: (top row) actual Y vs. predicted Y; (bottom row) actual Y versus T2.

To assess the robustness of our method against varying initial datasets, we generated the simulated dataset 30 times, aiming to reach the same targets. For each case, we measured the number of iterations required for each method to achieve the targets (T1, T2, T3, as shown in Figure 3-5). Figure 3-8 presents a comparison of the number of iterations taken by each method to reach the same targets, highlighting the strong robustness of the PREP method with differing initial datasets.

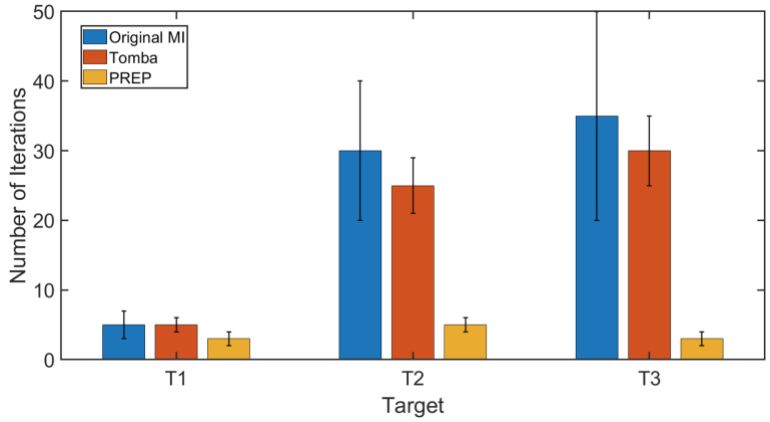


Figure 3-8: Comparison of the performance of different methods (original model inversion, Tomba, and PREP) in achieving the same targets shown in Figure 3-5 using various calibration datasets for the first simulated dataset.

3.4.2 Second Simulated Dataset

For the second dataset, which introduces slightly more complexity, the same X equation is used along with the following equation for Y data generation:

$$\text{part}_1 = \sin\left(\frac{\pi}{6}X_1\right) * \sin\left(\frac{\pi}{6}X_2\right) * \sin\left(\frac{\pi}{6}X_3\right)$$

$$\text{part}_2 = \cos\left(\frac{\pi}{6}X_1\right) * \cos\left(\frac{\pi}{6}X_2\right) * \cos\left(\frac{\pi}{6}X_3\right)$$

$$part_3 = (0.4X_1 + 0.5X_2)^{\alpha {X_3}^{\beta}} + (0.3X_1 + 0.7X_3)^{\alpha {X_2}^{\beta}} + (2.7X_2 + 3.3X_3)^{\alpha {X_1}^{\beta}} \quad eq.20$$

$$part_4 = (0.2X_1 + 0.8X_2)^{\beta {X_3}^{\alpha}} + (0.6X_1 + 0.4X_3)^{\beta {X_2}^{\alpha}} + (3.2X_2 + 2.8X_3)^{\beta {X_1}^{\alpha}}$$

$$Y_{final} = [part_3 \ , \quad \frac{part_4}{part_1.part_2}]$$

where $\alpha = 4$ and $\beta = 0.2$.

For this dataset, which was designed to exhibit a higher degree of nonlinearity due to the inclusion of highly complex mathematical relationships such as combinations of trigonometric functions and power-law expressions with interdependent exponents, similar results were observed. Figure 3-9 illustrates the three targets along with their TDS and projections in the latent space, providing a sense of the difficulty associated with each target and (based on the distribution of the TDS) illustrating the expected nonlinearity of the dataset. Table 3-2 again shows the benefit of considering only the five nearest neighbors for each target rather than the full dataset to enhance the accuracy of the PLS model by better capturing the variance of the data in the vicinity of the targets.

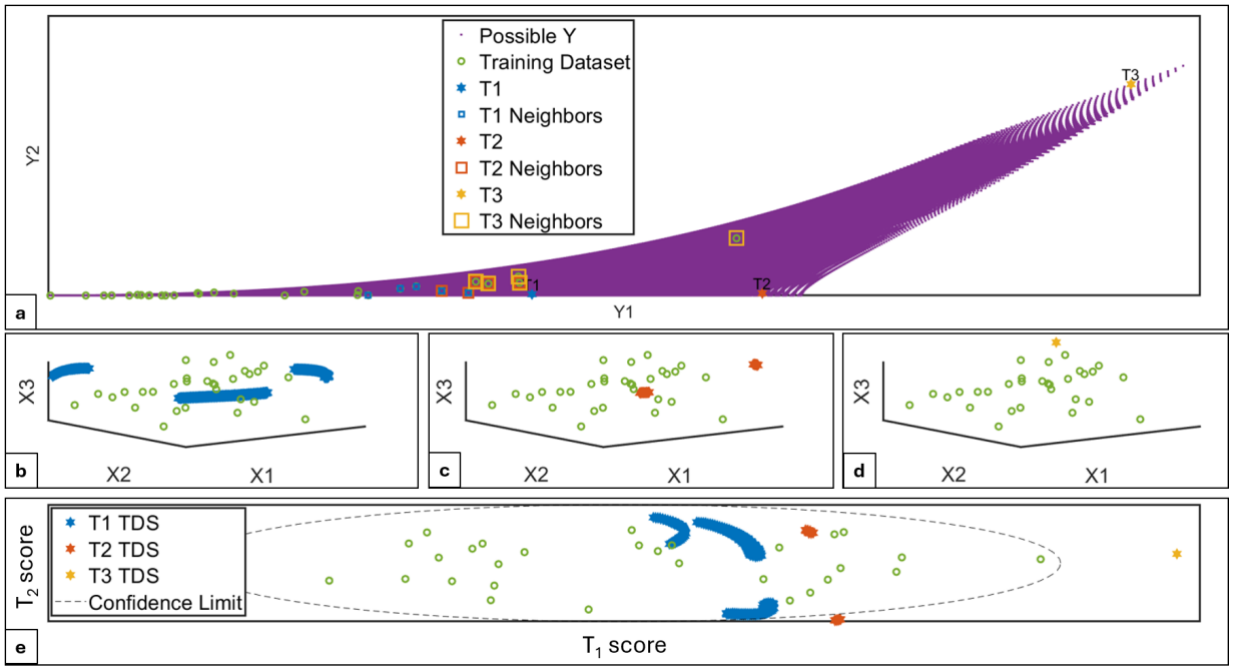


Figure 3-9: Distributions of the second simulated dataset and targets: (a) Selected targets 1 (T1), 2 (T2), and 3 (T3) along with corresponding nearest neighbors for each target; (b-d) TDS for T1 (b), T2 (c), and T3 (d) in the original space; (e) Projection of each TDS from the second row into the latent space.

Table 3-2: R^2 values from PLS conducted on the entire dataset compared to using the 5 nearest neighbors corresponding to each target for the second simulated dataset.

PLS Num Components	Entire Dataset		T1 Neighbors		T2 Neighbors		T3 Neighbors	
	R^2X (%)	R^2Y (%)	R^2X (%)	R^2Y (%)	R^2X (%)	R^2Y (%)	R^2X (%)	R^2Y (%)
1	40.0	61.1	78.0	35.1	71.2	66.6	51.4	69.7
2	72.5	62.0	93.8	79.5	99.6	90.1	94.2	89.9
3	100.0	62.1	100.0	99.4	100.0	96.0	100.0	99.7

The results for each of the three targets are presented in Figure 3-10. As with the first dataset, PREP reaches each target within half or less of the number of iterations required with the Original MI and Tomba methods. Of particular note, for T2 the PREP method hit the target within 4 iterations while the Original MI required 10 iterations and the Tomba method required 15 iterations. Several notable observations can however be made about the differences in the performance of each method in this second simulated dataset compared to the first simulated dataset. Notably, for T3 that was significantly outside the range of the original training dataset, the Original Model Inversion (MI) ultimately reached the target in 10 iterations even though it failed to reach the target in the first simulated dataset that was more linear. This discrepancy arises because the average values for the TDS for T3 in the first simulated dataset were [1.75, 1.75, 1.75] while the same values for T3 in the second simulated dataset were [1.94, 1.99, 2]. We manually set the Original MI suggestions to the (0-2) boundaries whenever the suggestions exceeded those limits, as we wanted to keep this method applicable in examples for which there are constraints on X. This manual adjustment ultimately benefited the Original MI in certain cases in which TDS approached the boundaries, enabling the Original MI to reach T3 in the second simulated dataset faster based on its ability to reset suggestions closer to the target values. In contrast, the other two methods (Tomba and PREP) are constrained to search and generate PDS within the range of 0 to 2, offering them no guidance or bias toward the edges of these limits. Another noteworthy point is that Tomba's method exhibits a more conservative trajectory in approaching the target compared to PREP, consistent with the smaller increments in prediction accuracy (PA) achieved by Tomba's method relative to PREP. This fundamental difference allows PREP to take a more ambitious approach, enabling it to progress toward the target at a faster pace.

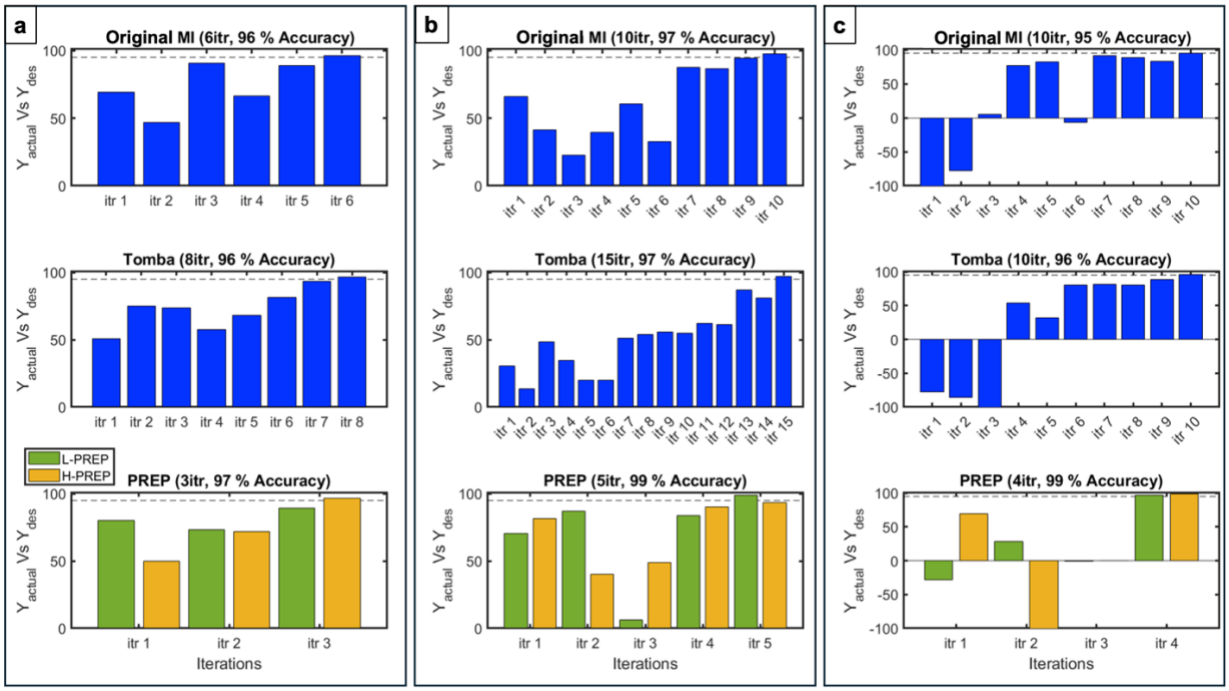


Figure 3-10: Performance of different methods in reaching (a) T1, (b) T2, and (c) T3 using the second simulated dataset. The first row displays the results from the Original Model Inversion, the second row displays the results from the Tomba method, and the third row illustrates the results from the proposed PREP method.

Figure 3-11 provides an overview of how the suggestions made by the PREP method to achieve target T2 facilitated the expansion of the dataset from the first iteration, during which the model predictions were largely invalid around the desired area, to the last iteration, at which point almost all PDS were members of the TDS; achieving such performance with such a non-linear model represents a significant success for the PREP method. Another noteworthy point is that unlike with the first dataset, the strategy of organizing PDS samples by their PREP scores illustrates a advantage of the PREP iterations- at each iteration, the method suggests samples that are either closer to the target, or through explorations, improve the model accuracy. In some instances, the suggestions intended to improved model accuracy also end up moving closer to the desired target. For example, while the second, fourth, and final iterations demonstrated the expected trend of declining PA with rising PREP scores, in the third iteration the PA actually improved with higher PREP scores Thus, in this context, the H-PREP emerged as the optimal choice, underscoring the advantages of considering both outer limits of the PREP algorithm predictions for future iterations.

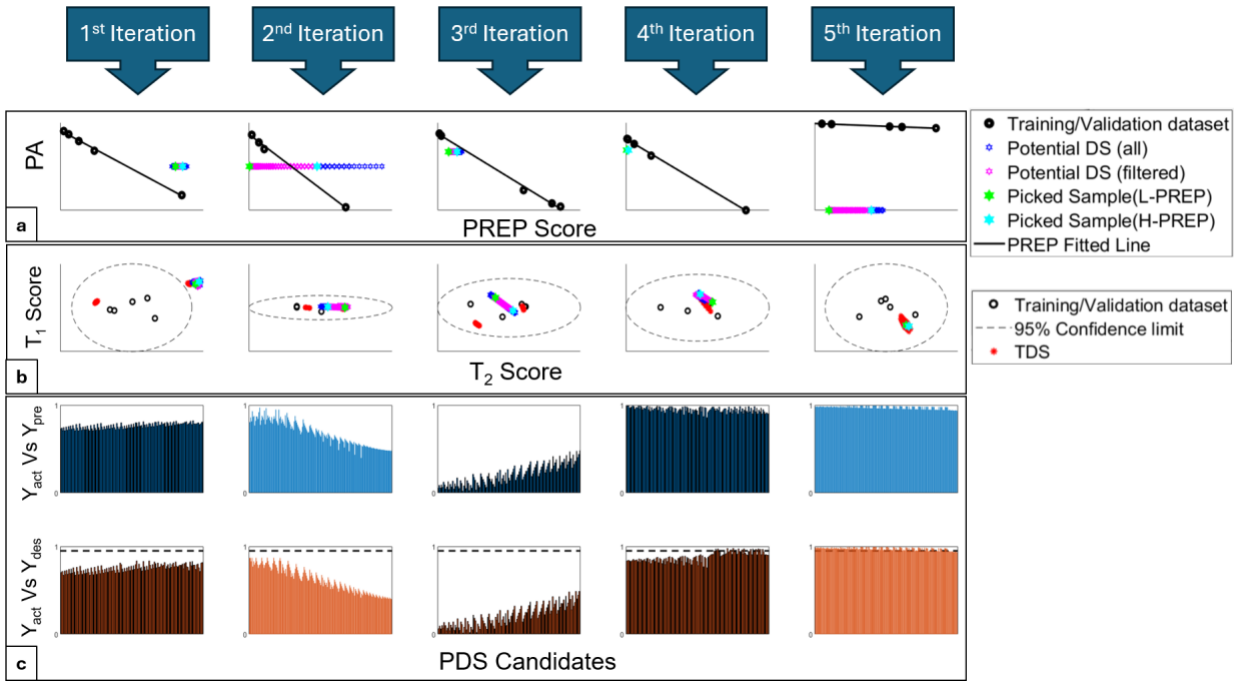


Figure 3-11: PREP iteration results from first to last iteration for T2 using the second simulated dataset: (a) the outcome of the PREP optimization and its alignment; (b) the projections of the TDS and PDS into the PLS latent space in each iteration, focusing PREP iteration results from first to last iteration for T2 using the second simulated dataset: (a) the outcome of the PREP optimization and its alignment; (b) the projections of the TDS and PDS into the PLS latent space in each iteration, focusing only on the first and second scores (the third score is excluded for simplicity); (c) comparison of actual Y values, predicted Y values, and T2 values for all PDS samples sorted by PREP scores: (top row) actual Y vs. predicted Y; (bottom row) actual Y versus T2.

To assess the robustness of the PREP approach, we again randomly generated 30 different datasets from the second simulated dataset and re-ran each method to assess how many iterations were required to achieve the desired target, with the results presented in Figure 3-12. The PREP method again demonstrated greater robustness and a faster identification of the Design Space compared to the other two methods with all three targets, again confirming the ability of PREP to accelerate optimal recipe identification over a range of different initial datasets.

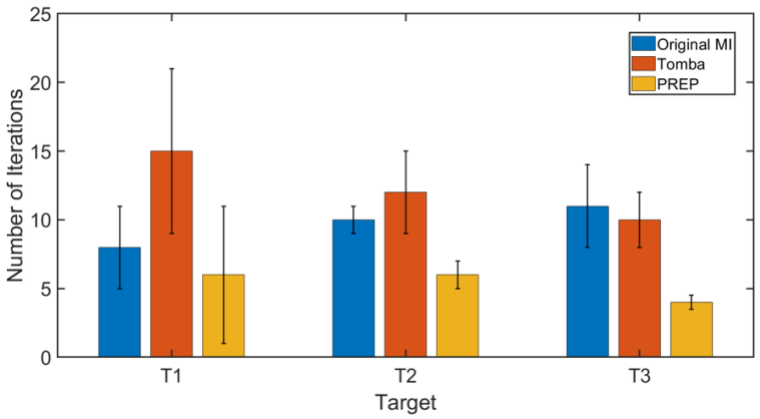


Figure 3-12: Comparison of the performance of different methods (original model inversion, Tomba, and PREP) in achieving the same targets shown in Figure 3-9 using various calibration datasets for the second simulated dataset.

The Supplementary Information highlights the similarly improved performance of the PREP method across three additional simulated datasets designed to represent varying levels of nonlinearity and complexity and thus challenge the robustness of the methods: (1) a highly nonlinear dataset involving nested power-law terms with interdependent exponents (Equations S1, Figures S1–S3); (2) a dataset combining trigonometric, exponential, and power-law terms with interdependencies among variables (Equations S2, Figures S4–S7); and (3) a moderately complex dataset featuring sinusoidal, logarithmic, and square-root functions (Equations S3, Figures S8–S9). PREP consistently required significantly fewer iterations compared to Original MI and Tomba, achieving targets in an average of 3 to 11 iterations whereas Tomba required 10 to 25 iterations and Original MI required 16 to 30 iterations. Furthermore, PREP demonstrated greater robustness, as evidenced by its lower variability in iteration counts across all analyses. These results highlight the efficiency and reliability of PREP in identifying target recipes across datasets with diverse complexities.

Collectively, the results from all simulations suggest that PREP demonstrates superior performance compared to previously reported methods due to its ability to adaptively prioritize solutions that align closely with the desired target properties. However, a potential downside of the PREP method's aggressive approach is that there is a larger risk of moving further away from the target when the optimal solution lies close to existing datapoints. While none of the case studies explored in this study (including

those with datapoints near the target) showed reduced performance due to this risk, this factor should remain a consideration for future applications. It should also be noted that the PREP method requires making two samples (L-PREP and H-PREP) per iteration whereas the other methods require only one sample be made per iteration. However, in most cases, making two samples in parallel is significantly faster than performing multiple iterative cycles that require individual synthesis-modeling steps; furthermore, even if the performance of PREP was assessed based on the number of samples instead of the number of cycles, it would still significantly outperform the other methods in most cases.

In practical applications, PREP could be particularly valuable for industrial challenges in which efficient formulation optimization and product design require targeted dataset expansion, such as pharmaceutical development, drug formulation, and complex chemical processes in which highly nonlinear relationships and limited fundamental understanding make traditional modeling approaches impractical. By guiding dataset expansion toward capturing the actual design space via resource-efficient data collection, PREP enhances experimental efficiency while ensuring reliable identification of optimal operating conditions. The number of input dimensions that PREP can handle is in principle unrestricted, as its nearest-neighbor selection is based on the number of latent variables plus two; however, the method's effectiveness in very high-dimensional datasets remains an open area for future experimental validation. Moreover, PREP is especially valuable in scenarios in which sample preparation is either costly or time-consuming and the number of available datapoints is thus very limited. In such cases, PREP has the potential to expand the dataset iteratively starting with as few as five samples (as shown in all the case studies presented in this paper) with a minimal number of additional samples, guiding the process rationally toward regions of the design space that encompass the desired target properties.

3.5 Conclusion

The proposed PREP method represents a new approach to identify samples for which model predictions are expected to be more reliable than those of alternative candidates, specifically aimed at accelerating the identification of the Design Space. By applying this approach iteratively, we achieved significantly faster

convergence to the actual solution compared to existing methods. The method's efficiency was particularly advantageous in scenarios in which rapid DS identification is critical and/or sample preparation incurs significant cost and/or time constraints. The primary benefit of PREP is the potential to optimize the use of experimental resources by reducing the number of required iterations, thereby minimizing material and operational costs. The method is theoretically applicable to datasets with any number of dimensions, as its nearest-neighbor selection is based on the number of latent variables plus two. However, further experimental validation in higher-dimensional spaces will be necessary to fully assess its scalability. In particular, the PREP approach effectively tracked the TDS with fewer iterations, effectively identifying relevant samples within fewer cycles. Moreover, it provides a valuable mechanism for discerning whether all potential DS (PDS) candidates truly belong to the actual DS by preparing and analyzing both lowest-score and highest-score PREP parameter samples, enabling improved estimates of how close the process is to identifying actual design space members. Tested across several highly nonlinear datasets, our method outperformed two widely used competing approaches, achieving target results across each dataset/target evaluated in fewer iterations and, for harder-to-achieve targets, typically half or less the number of iterations the other methods require. These findings underscore the PREP method's potential as a reliable and efficient tool for real-world applications with complex underlying data structures.

Acknowledgement

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.

During the preparation of this work, the authors used ChatGPT to enhance the readability of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

3.6 References

1. Tomba, E., M. Barolo, and S. García-Muñoz, *General framework for latent variable model inversion for the design and manufacturing of new products.* Industrial & engineering chemistry research, 2012. **51**(39): p. 12886-12900.
2. ICH, I., *Quality Implementation Working Group Points to Consider (R2).* ICH-Endorsed Guide for ICH Q, 2011. **8**.
3. Tomba, E., et al., *Latent variable modeling to assist the implementation of Quality-by-Design paradigms in pharmaceutical development and manufacturing: A review.* International journal of pharmaceutics, 2013. **457**(1): p. 283-297.
4. Stockdale, G.W. and A. Cheng, *Finding design space and a reliable operating region using a multivariate Bayesian approach with experimental design.* Quality Technology & Quantitative Management, 2009. **6**(4): p. 391-408.
5. N. Politis, S., et al., *Design of experiments (DoE) in pharmaceutical development.* Drug development and industrial pharmacy, 2017. **43**(6): p. 889-901.
6. Destro, F. and M. Barolo, *A review on the modernization of pharmaceutical development and manufacturing–Trends, perspectives, and the role of mathematical modeling.* International Journal of Pharmaceutics, 2022. **620**: p. 121715.
7. Zhang, L. and S. Garcia-Munoz, *A comparison of different methods to estimate prediction uncertainty using Partial Least Squares (PLS): a practitioner's perspective.* Chemometrics and intelligent laboratory systems, 2009. **97**(2): p. 152-158.
8. Bano, G., et al., *Uncertainty back-propagation in PLS model inversion for design space determination in pharmaceutical product development.* Computers & Chemical Engineering, 2017. **101**: p. 110-124.
9. Faber, K. and B.R. Kowalski, *Prediction error in least squares regression: Further critique on the deviation used in The Unscrambler.* Chemometrics and Intelligent Laboratory Systems, 1996. **34**(2): p. 283-292.
10. Van Huffel, S. and J. Vandewalle, *The partial total least squares algorithm.* Journal of computational and applied mathematics, 1988. **21**(3): p. 333-341.
11. Phatak, A., P. Reilly, and A. Penlidis, *An approach to interval estimation in partial least squares regression.* Analytica chimica acta, 1993. **277**(2): p. 495-501.
12. Denham, M.C., *Prediction intervals in partial least squares.* Journal of Chemometrics: A Journal of the Chemometrics Society, 1997. **11**(1): p. 39-52.
13. Serneels, S., P. Lemberge, and P.J. Van Espen, *Calculation of PLS prediction intervals using efficient recursive relations for the Jacobian matrix.* Journal of Chemometrics: A Journal of the Chemometrics Society, 2004. **18**(2): p. 76-80.
14. Helland, I.S., *Partial least squares regression and statistical models.* Scandinavian journal of statistics, 1990: p. 97-114.
15. Faber, N.K.M., *Uncertainty estimation for multivariate regression coefficients.* Chemometrics and intelligent laboratory systems, 2002. **64**(2): p. 169-179.
16. Efron, B. and R.J. Tibshirani, *An introduction to the bootstrap*. 1994: Chapman and Hall/CRC.
17. Laky, D., et al., *An optimization-based framework to define the probabilistic design space of pharmaceutical processes with model uncertainty.* Processes, 2019. **7**(2): p. 96.
18. Hua, S., G. Qu, and S.S. Bhattacharyya. *Exploring the probabilistic design space of multimedia systems*. in *14th IEEE International Workshop on Rapid Systems Prototyping, 2003. Proceedings.* 2003. IEEE.
19. Bano, G., et al., *Probabilistic Design space determination in pharmaceutical product development: A Bayesian/latent variable approach.* AIChE Journal, 2018. **64**(7): p. 2438-2449.
20. Kusumo, K.P., et al., *Bayesian approach to probabilistic design space characterization: A nested sampling strategy.* Industrial & Engineering Chemistry Research, 2019. **59**(6): p. 2396-2408.

21. Jaeckle, C.M. and J.F. MacGregor, *Industrial applications of product design through the inversion of latent variable models.* Chemometrics and intelligent laboratory systems, 2000. **50**(2): p. 199-210.
22. Facco, P., et al., *Bracketing the design space within the knowledge space in pharmaceutical product development.* Industrial & Engineering Chemistry Research, 2015. **54**(18): p. 5128-5138.
23. Paris, A., C. Duchesne, and É. Poulin, *Establishing multivariate specification regions for incoming raw materials using projection to latent structure models: comparison between direct mapping and model inversion.* Frontiers in Analytical Science, 2021: p. 7.
24. García-Muñoz, S., et al., *Optimization of batch operating policies. Part I. Handling multiple solutions#.* Industrial & engineering chemistry research, 2006. **45**(23): p. 7856-7866.
25. MacGregor, J.F., K. Muteki, and T. Ueda, *On the rapid development of new products through empirical modeling with diverse data-bases*, in *Computer Aided Chemical Engineering*. 2006, Elsevier. p. 701-706.
26. Bano, G., et al., *A novel and systematic approach to identify the design space of pharmaceutical processes.* Computers & Chemical Engineering, 2018. **115**: p. 309-322.
27. Palací-López, D., et al., *New tools for the design and manufacturing of new products based on latent variable model inversion.* Chemometrics and Intelligent Laboratory Systems, 2019. **194**: p. 103848.
28. Palací-López, D., et al., *Improved formulation of the latent variable model inversion–based optimization problem for quality by design applications.* Journal of Chemometrics, 2020. **34**(6): p. e3230.
29. Borràs-Ferrís, J., et al., *Defining multivariate raw material specifications in industry 4.0.* Chemometrics and Intelligent Laboratory Systems, 2022. **225**: p. 104563.
30. Wold, H., *Systems analysis by partial least squares.* 1983.
31. Geladi, P. and B.R. Kowalski, *Partial least-squares regression: a tutorial.* Analytica chimica acta, 1986. **185**: p. 1-17.
32. Tayebi, S.S., 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 Applied Polymer Materials, 2022. **4**(12): p. 9160-9175.
33. Zhao, F., et al., *Novel formulations of flexibility index and design centering for design space definition.* Computers & Chemical Engineering, 2022. **166**: p. 107969.
34. Zhao, F., et al., *Design space description through adaptive sampling and symbolic computation.* AIChE Journal, 2022. **68**(5): p. e17604.
35. Sachio, S., et al., *Computer-aided design space identification for screening of protein A affinity chromatography resins.* Journal of Chromatography A, 2024. **1722**: p. 464890.
36. Ochoa, M.P., et al., *Novel flexibility index formulations for the selection of the operating range within a design space.* Computers & Chemical Engineering, 2021. **149**: p. 107284.
37. Geremia, M., F. Bezzo, and M.G. Ierapetritou, *Design space determination of pharmaceutical processes: Effects of control strategies and uncertainty.* European Journal of Pharmaceutics and Biopharmaceutics, 2024. **194**: p. 159-169.
38. Ding, C. and M. Ierapetritou, *A novel framework of surrogate-based feasibility analysis for establishing design space of twin-column continuous chromatography.* International Journal of Pharmaceutics, 2021. **609**: p. 121161.

3.7 Supporting Information

In addition to the dataset discussed in the main manuscript, three additional highly nonlinear and complex datasets were generated to further evaluate the robustness of the PREP method under diverse conditions. Below, we provide the equations, target values, and summarized results for each dataset. These datasets were selected to introduce varying mathematical complexities and nonlinearities, challenging the methods to perform effectively under a range of conditions. While the outcomes align closely with those presented in the main text, only key findings are discussed here for context, along with accompanying plots.

3.7.1 Third Simulated Dataset

This dataset introduces nested power-law terms with variable-dependent exponents (α=4 and β=0.2), creating a highly nonlinear relationship between X and Y. The complexity is further amplified by the weighted combinations of the input variables in the power-law terms, making this dataset particularly challenging.

$$Y_1 = (0.4X_1 + 0.5X_2)^{\alpha X_3{}^{\beta}} + (0.3X_1 + 0.7X_3)^{\alpha X_2{}^{\beta}} + (2.7X_2 + 3.3X_3)^{\alpha X_1{}^{\beta}} \quad eq.\ s1$$

$$Y_2 = (0.2X_1 + 0.8X_2)^{\beta X_3{}^{\alpha}} + (0.6X_1 + 0.4X_3)^{\beta X_2{}^{\alpha}} + (3.2X_2 + 2.8X_3)^{\beta X_1{}^{\alpha}}$$

• *Third Simulated Dataset Results Summary*

PREP outperformed both Tomba and Original MI methods in all targets, requiring an average of 3 to 3.3 iterations across targets, compared to 11–16 for Original MI and 5–6.3 for Tomba. PREP also demonstrated greater robustness, as evidenced by its lowest variability in iteration counts across different sample sets.

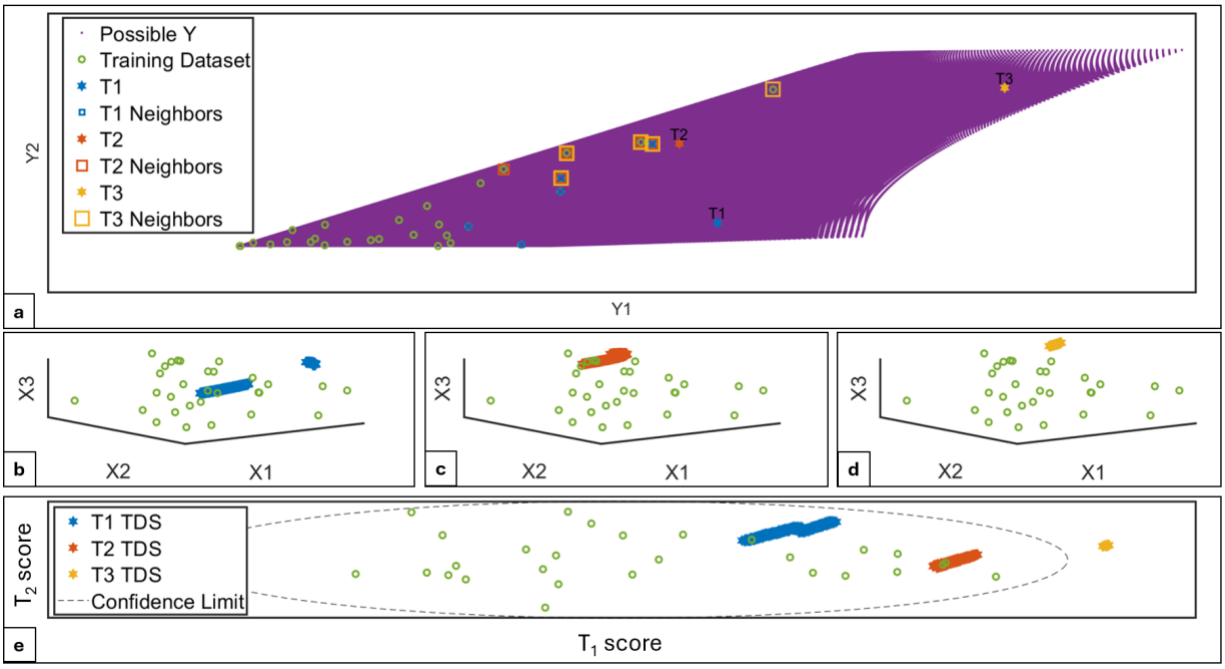


Figure S1. Distributions of the third simulated dataset and targets: (a) Selected targets 1 (T1), 2 (T2), and 3 (T3) along with corresponding nearest neighbors for each target; (b-d) TDS for T1 (b), T2 (c), and T3 (d) in the original space; (e) Projection of each TDS from the second row into the latent space.

Table S1. R^2 values from PLS conducted on the entire dataset compared to using the 5 nearest neighbors corresponding to each target for the third simulated dataset.

PLS Num Components	**Entire Dataset**		**T1 Neighbors**		**T2 Neighbors**		**T3 Neighbors**	
	R^2X (%)	R^2Y (%)	R^2X (%)	R^2Y (%)	R^2X (%)	R^2Y (%)	R^2X (%)	R^2Y (%)
1	42.1	74.1	37.9	75.7	64.0	77.1	53.7	77.2
2	71.7	78.4	97.1	83.2	95.3	86.3	82.7	92.4
3	100.0	78.4	100.0	84.7	100.0	100.0	100.0	99.9

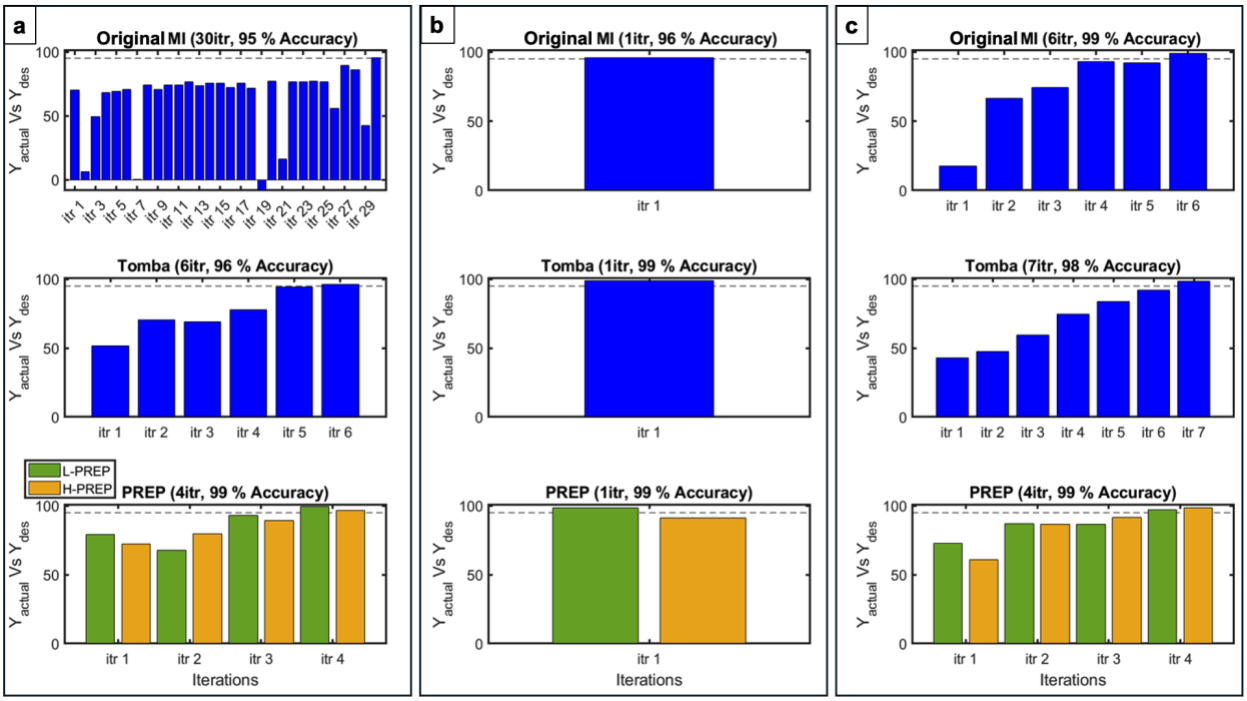


Figure S2. Performance of different methods in reaching (a) T1, (b) T2, and (c) T3 using the third simulated dataset. The first row displays the results from the Original Model Inversion, the second row displays the results from the Tomba method, and the third row illustrates the results from the proposed PREP method.

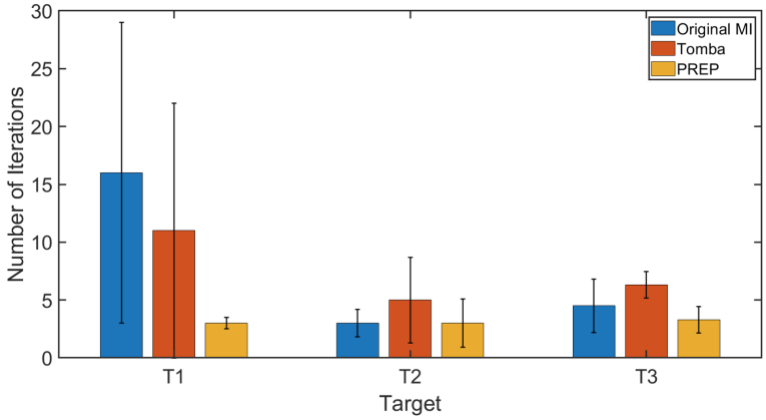


Figure S3. Comparison of the performance of different methods (original model inversion, Tomba, and PREP) in achieving the same targets shown in Figure S1 using various calibration datasets for the third simulated dataset

3.7.2 Fourth Simulated Dataset

This dataset incorporates a mix of trigonometric, exponential, and power-law functions, increasing complexity by introducing periodicity, rapid oscillations, and nonlinearity. The interdependence between input variables adds another layer of difficulty.

$$Y_1 = (X_2 \,.X_1) * \cos\left(\frac{\pi}{6}X_1\right) + \ (X_2 \,.X_1) * \exp\left(\frac{\pi}{6}X_3\right) + \cos\left(\frac{\pi}{2}X_2\right) \quad eq.s2$$

$$Y_2 = + \left({X_3}^{X_2}\right) * \cos\left(\frac{\pi}{2}X_2\right) + \left({X_1}^{X_3}\right) * \cos(\pi X_2) + \left({X_3}^{X_2}\right) * \cos(\pi X_1)$$

$$+ \sin\left(\frac{\pi}{3}X_1\right) + \ (X_2 \,.X_1) * \cos(2\pi X_3) \ + \ X_3$$

• *Fourth Simulated Dataset Results Summary*

PREP consistently required fewer iterations to reach the target, with averages ranging from 5 to 11 iterations across the three targets compared to 12–30 for Original MI and 12–25 for Tomba. Its robustness was also again evident from the lower variability in iteration counts observed with different initial datasets.

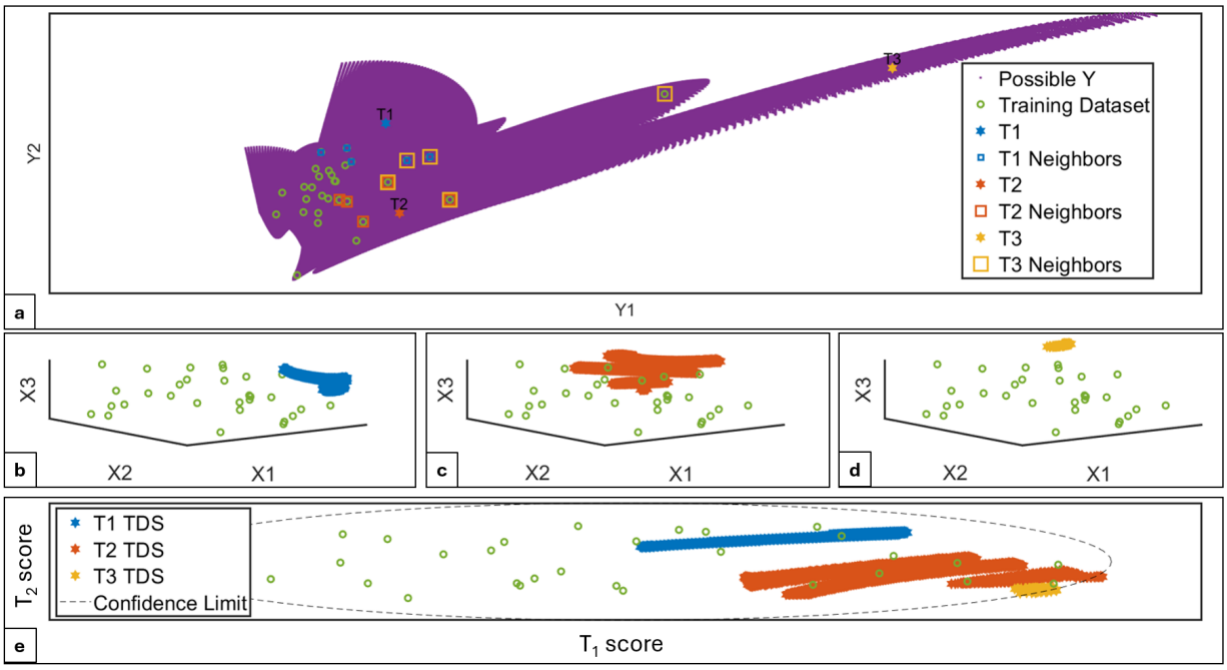


Figure S4. Distributions of the fourth simulated dataset and targets: (a) Selected targets 1 (T1), 2 (T2), and 3 (T3) along with corresponding nearest neighbors for each target; (b-d) TDS for T1 (b), T2 (c), and T3 (d) in the original space; (e) Projection of each TDS from the second row into the latent space.

Table S2. R^2 values from PLS conducted on the entire dataset compared to using the 5 nearest neighbors corresponding to each target for the fourth simulated dataset.

PLS Num Components	**Entire Dataset**		**T1 Neighbors**		**T2 Neighbors**		**T3 Neighbors**	
	R^2X (%)	R^2Y (%)	R^2X (%)	R^2Y (%)	R^2X (%)	R^2Y (%)	R^2X (%)	R^2Y (%)
1	37.3	26.2	79.7	53.2	44.5	46.1	52.5	62.6
2	69.4	32.8	99.5	62.2	72.9	92.3	68.3	85.4
3	100.0	35.3	100.0	64.1	100.0	95.3	100.0	86.1

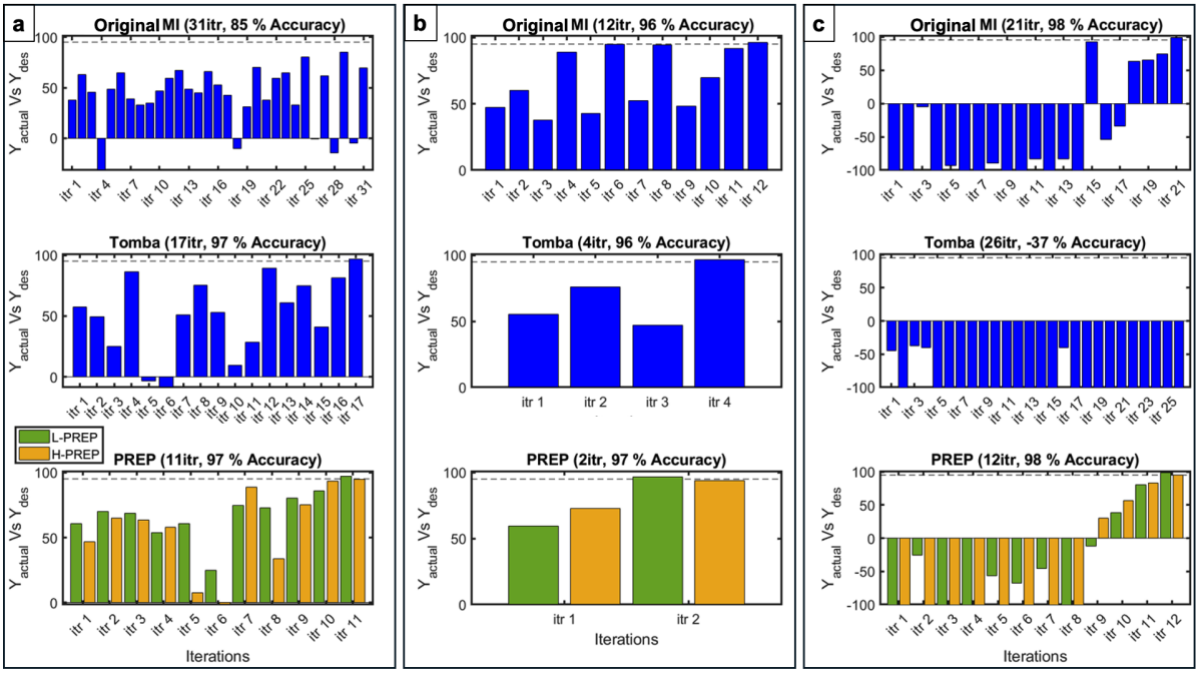


Figure S5. Performance of different methods in reaching (a) T1, (b) T2, and (c) T3 using the fourth simulated dataset. The first row displays the results from the Original Model Inversion, the second row displays the results from the Tomba method, and the third row illustrates the results from the proposed PREP method.

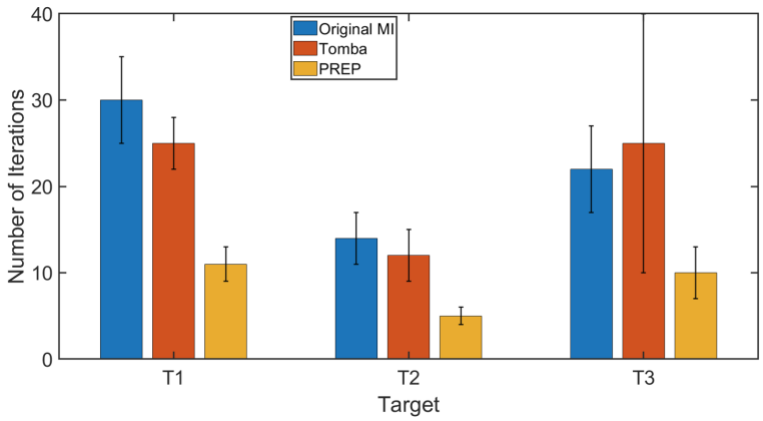


Figure S6. Comparison of the performance of different methods (original model inversion, Tomba, and PREP) in achieving the same targets shown in Figure S4 using various calibration datasets for the fourth simulated dataset

3.7.3 Fifth Simulated Dataset

This dataset combines sinusoidal, logarithmic, and square-root terms, creating moderate nonlinearity while maintaining a simpler structure compared to the other datasets. It serves as a contrasting example to evaluate the method's performance in less complex scenarios.

$$Y_1 = \sin(\pi X_1) + \log(X_2 + 1) + \sqrt{X_3} \quad eq.s3$$

$$Y_2 = (X_1^2 + X_2)^{0.3} + \exp\left(\frac{X_3}{2}\right)$$

• *Fifth Simulated Dataset Results Summary*

PREP again demonstrated superior performance, achieving targets in an average of 2 to 7 iterations compared to 5–20 for Original MI and 5–10 for Tomba. Its robustness was particularly notable, with the lowest standard deviations in iteration counts.

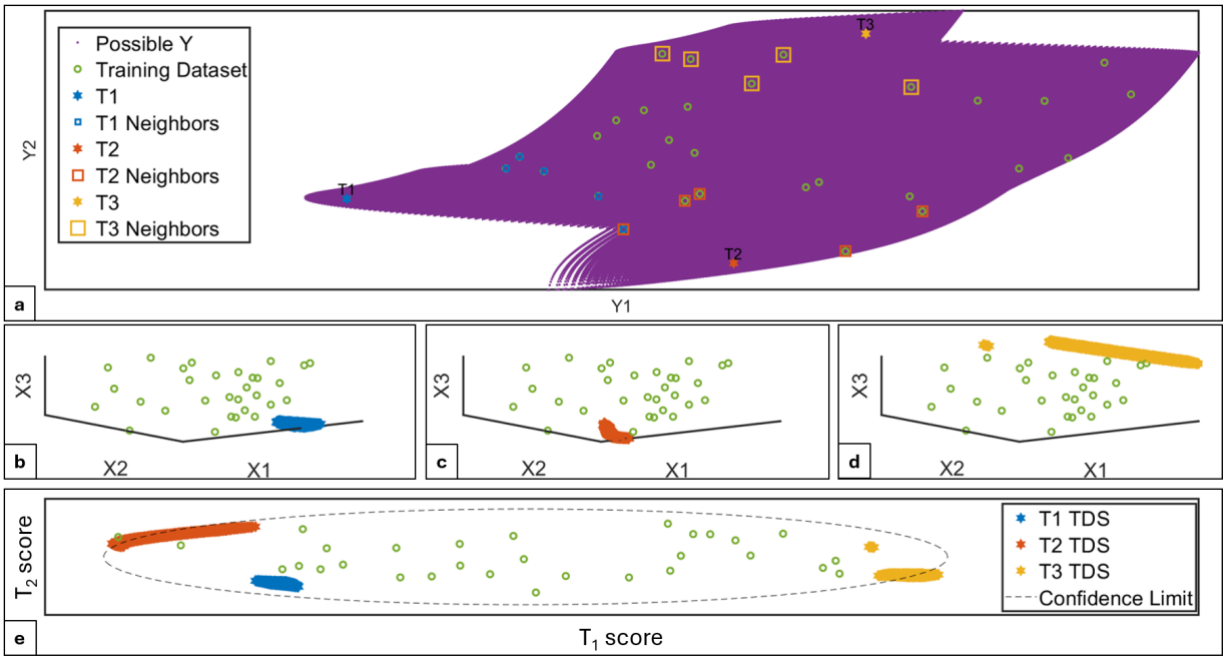


Figure S7. Distributions of the fifth simulated dataset and targets: (a) Selected targets 1 (T1), 2 (T2), and 3 (T3) along with corresponding nearest neighbors for each target; (b-d) TDS for T1 (b), T2 (c), and T3 (d) in the original space; (e) Projection of each TDS from the second row into the latent space.

Table S3. R^2 values from PLS conducted on the entire dataset compared to using the 5 nearest neighbors corresponding to each target for the fifth simulated dataset.

PLS Num Components	**Entire Dataset**		**T1 Neighbors**		**T2 Neighbors**		**T3 Neighbors**	
	R^2X (%)	R^2Y (%)	R^2X (%)	R^2Y (%)	R^2X (%)	R^2Y (%)	R^2X (%)	R^2Y (%)
1	8.3	50.5	65.2	85.7	53.8	42.5	45.3	57.5
2	67.0	82.3	96.6	91.5	94.2	76.4	93.9	72.8
3	100.0	83.3	100.0	93.4	100.0	81.3	100.0	88.2

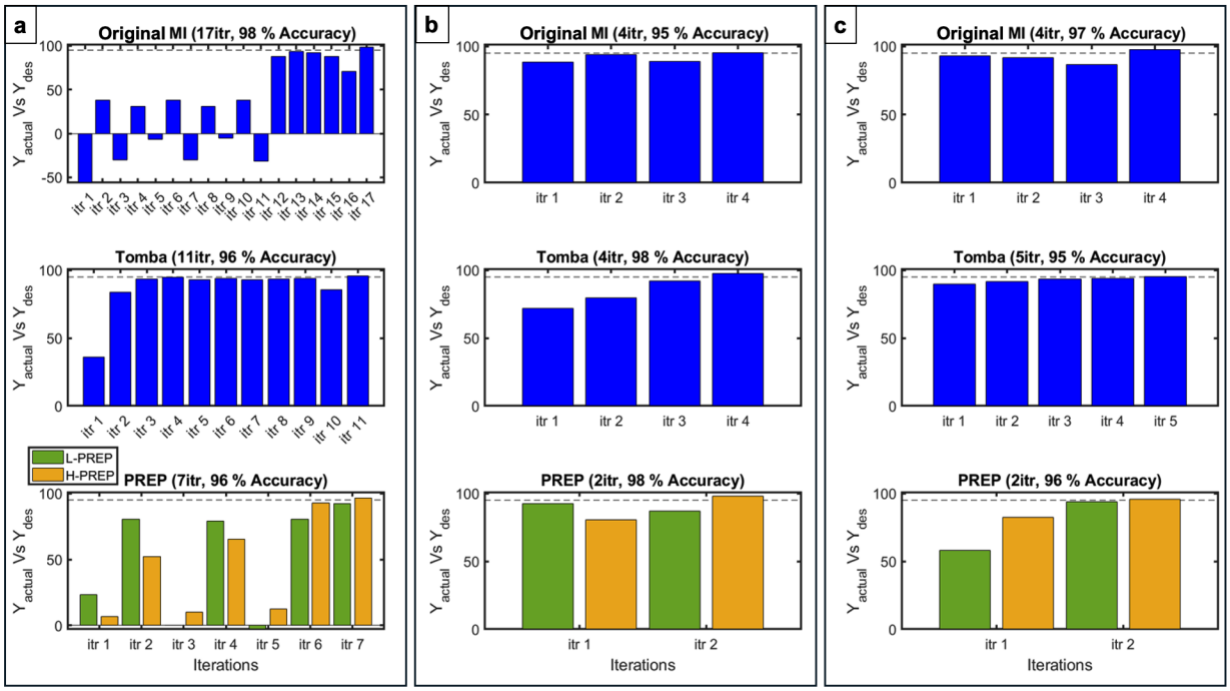


Figure S8. Performance of different methods in reaching (a) T1, (b) T2, and (c) T3 using the fifth simulated dataset. The first row displays the results from the Original Model Inversion, the second row displays the results from the Tomba method, and the third row illustrates the results from the proposed PREP method.

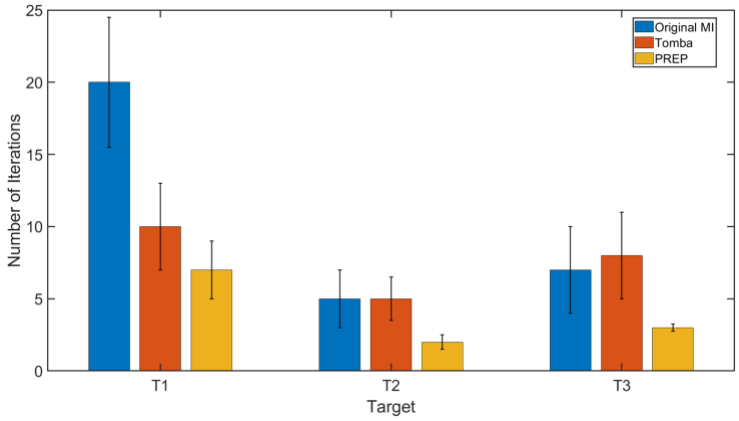


Figure S9. Comparison of the performance of different methods (original model inversion, Tomba, and PREP) in achieving the same targets shown in Figure S7 using various calibration datasets for the fifth simulated dataset

These results emphasize the robustness and efficiency of the PREP method across datasets with diverse levels of complexity and nonlinearity, making it a reliable tool for challenging design problems.

Chapter 3:

4 Data-Driven Optimization of Nanoparticle Size Using the Prediction Reliability Enhancing Parameter (PREP)

The contents of this chapter have been published in Nanoscale Journal.

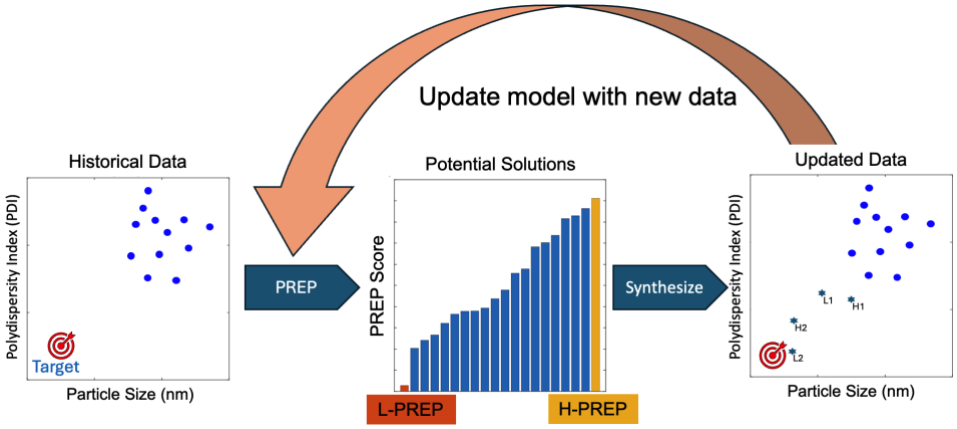
Seyed Saeid Tayebi, Nate Dowdall, Todd Hoare*, and Prashant Mhaskar*

Department of Chemical Engineering, McMaster University, 1280 Main St. W., Hamilton, Ontario, Canada L8S 4L7

Authorship Contribution

Seyed Saeid Tayebi: Conceptualized the idea, developed the methodology, wrote the code, tested the models, experimentally synthesized the samples for the microgel case study, wrote the manuscript draft, and addressed comments during the revision process. Nate Dowdall: Conducted the experimental work for the polyelectrolyte case study, contributed to the manuscript writing, particularly in the section describing the experimental setup and results for the polyelectrolyte case study. Todd Hoare: Contributed to the conceptualization of the experimental approach, provided supervision, reviewed and contributed to the writing of the manuscript, and managed project administration and funding. Prashant Mhaskar: Contributed to the conceptualization of the modeling approach, provided supervision, reviewed and contributed to the writing of the manuscript, and managed project administration and funding.

Abstract



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 s 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.

***Keywords*:** nanoparticles, microgels, self-assembled nanoparticles, particle size, data-driven modeling, PREP method

4.1 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 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 [28-34]. 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, 35] 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, 35-37].

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 [38-41]. 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 [42-49]. 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 [50]. 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 [39, 51-53]. 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.

4.2 Preliminaries

4.2.1 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 Figure 4-1; for detailed mathematical formulations, and data-blocking configurations, the reader is referred to our prior manuscript [50].

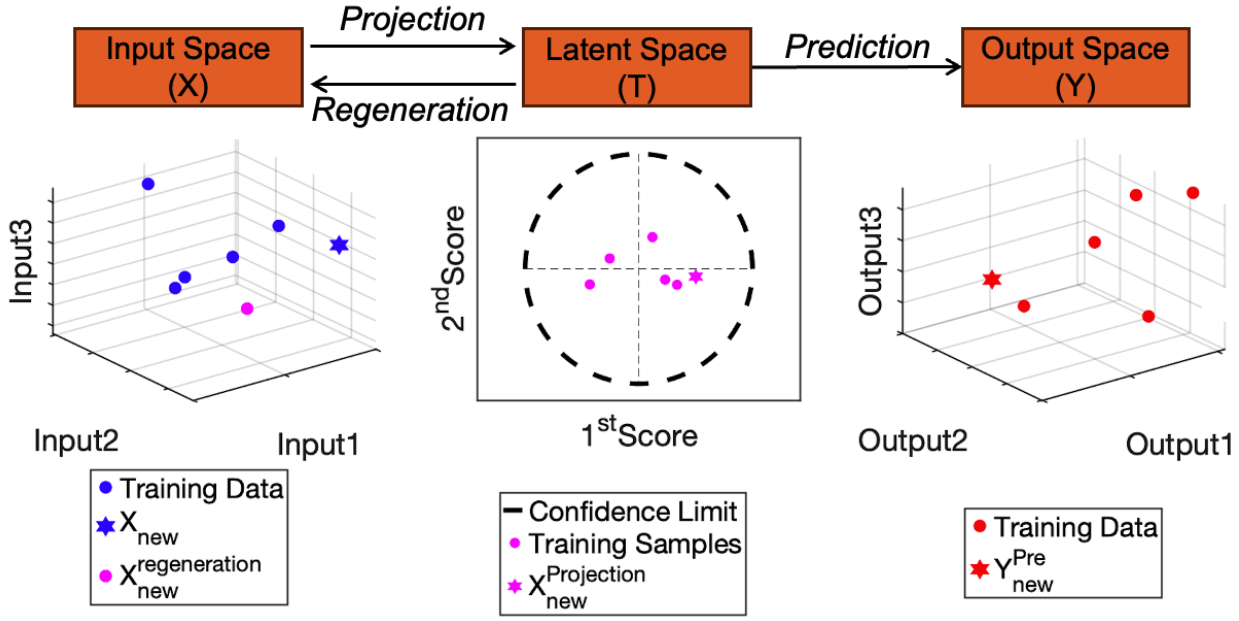


Figure 4-1: General latent variable modeling framework.

4.2.2 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:

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.

Figure 4-1 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).

4.3 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 Equation 1[50], in which the values of $c_i \in (0,1]$ and $p_i \in (0,1]$ are determined specifically for each dataset through an optimization algorithm.

$$\mathrm{PREP}=c_1\,\mathrm{hotelingT2}_{\mathrm{pls}}^{p1}+c_2\,\mathrm{SPE}_{\mathrm{x,pls}}^{p2}+c_3\,\mathrm{hotelingT2}_{\mathrm{pca}}^{p3}+c_4\,\mathrm{SPE}_{\mathrm{pca}}^{p4}+c_5\,\mathrm{h}_{\mathrm{pls}}^{p5}+c_6\,\mathrm{h}_{\mathrm{pca}}^{p6} \qquad \text{(eq.1)}$$

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 Equation 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. Figure 4-2 illustrates the general scheme of the method, with further details available in the original paper [50].

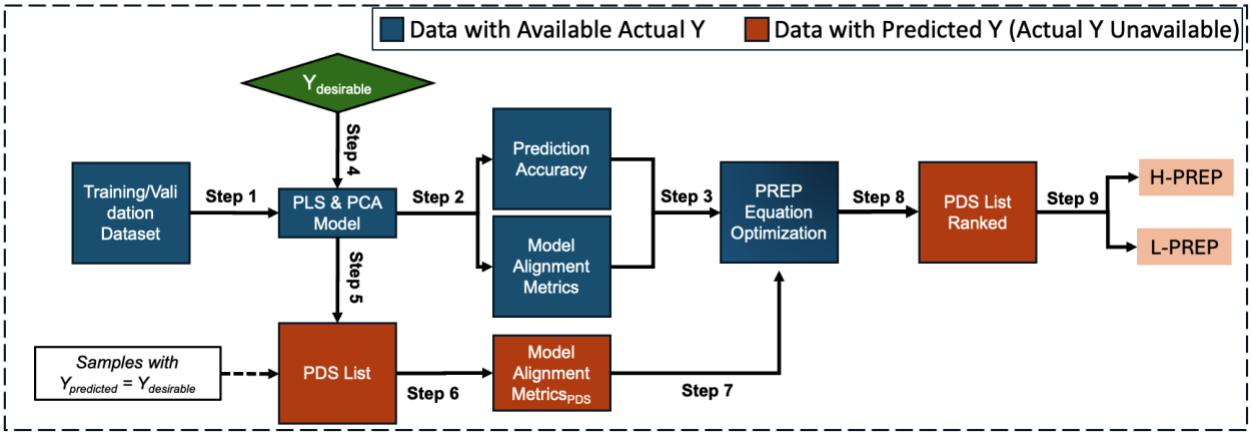


Figure 4-2: Schematic illustration of the proposed PREP method. The green box represents the desired target output set. Blue boxes indicate the training and validation data in which actual Y values are known and used for optimizing the PREP equation. Orange boxes depict the dataset of potential candidates, for which only X values are available. Candidates selected through the PREP method are prioritized for experimental testing.

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 [50]. 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).

4.4 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.

4.4.1 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-sensitive 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 [39]. 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 [54]) 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 4-1. 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 [55]. 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.

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

Microgel Synthesis

The initial dataset used in this study is summarized in Table 4-1. 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.

Table 4-1: 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* (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***	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***	1.6	**160**	**342**	**0**	50	**954**
13	1.6	**173**	**45**	**42**	50	**190**
14	1.6	**244**	**176**	**24**	50	**332**
*15***	1.6	**160**	**228**	**57**	50	**300**

**Sizes correspond to the intensity-averaged effective diameter measured at pH=7.4 and 37°C*

***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*

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 ($Y_{predicted}$) closely match the desired targets ($Y_{desirable}$) 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 Equation 2.

$$\min_{x^{new}} \left\{ w_1 \left(\hat{y}^{new}\text{-}y^{des}\right)\Gamma\left(\hat{y}^{new}\text{-}y^{des}\right)^T + w_2 \text{ Hoteling } T^2_{pls} + w_3 \text{ } SPE_{x^{new}} \right\} \quad \text{eq.2}$$

s.t.

$$\hat{y}^{new} = \tau Q^T$$

$$\hat{x}^{new} = \tau P^T$$

$$\tau = x^{new} W^*$$

where Γ is a [L*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 $w_1 \left(\hat{y}^{new} - y^{des}\right)$ in which w_i 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 [56] the eigenvalue-less-than-one rule [57], 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 Equation 2 was applied, resulting in the recipe outlined in Table

4-2 (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 4-2) 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 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. Figure 4-3 depicts all available datapoints and five nearest neighbors to the target in both the input (a) and output (b) spaces.

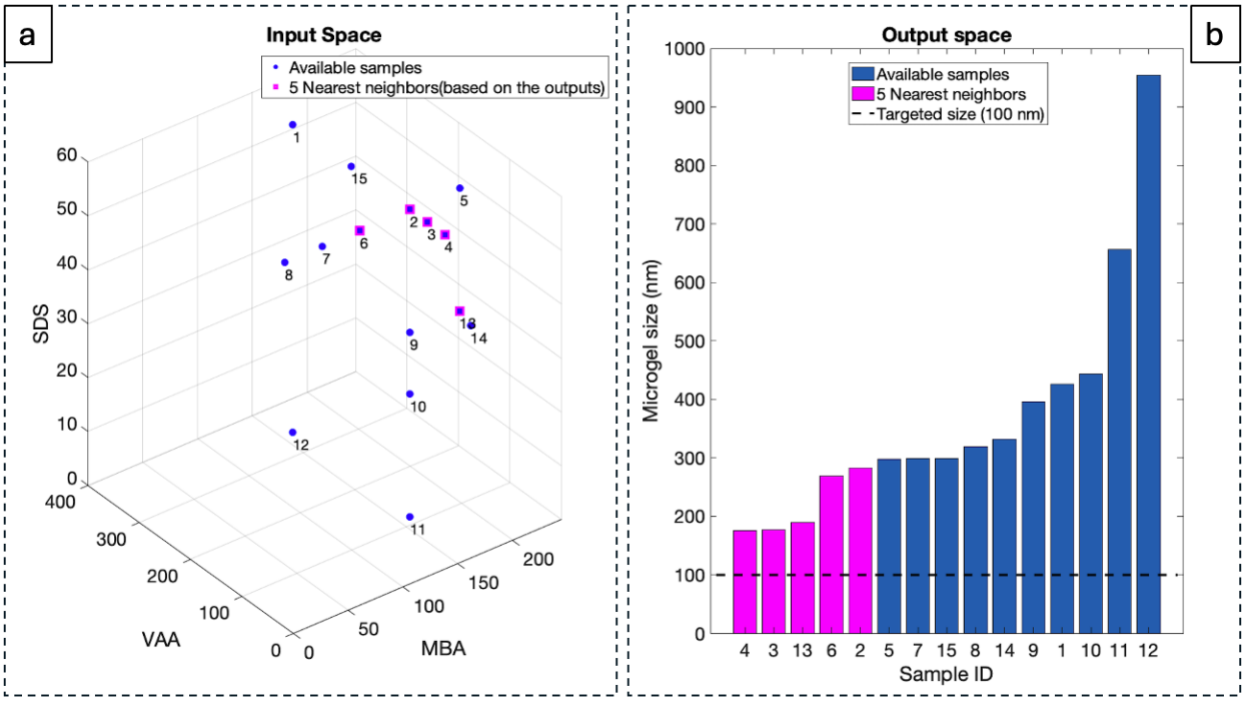


Figure 4-3: Visualization of all available datapoints alongside the five nearest neighbors to the target in both the input (a) and output (b) spaces derived from the pre-existing dataset (Table 4-1).

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 Figure 4-4(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 Figure 4-4(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 Equation 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 Figure 4-4 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 Table 4-2, 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).

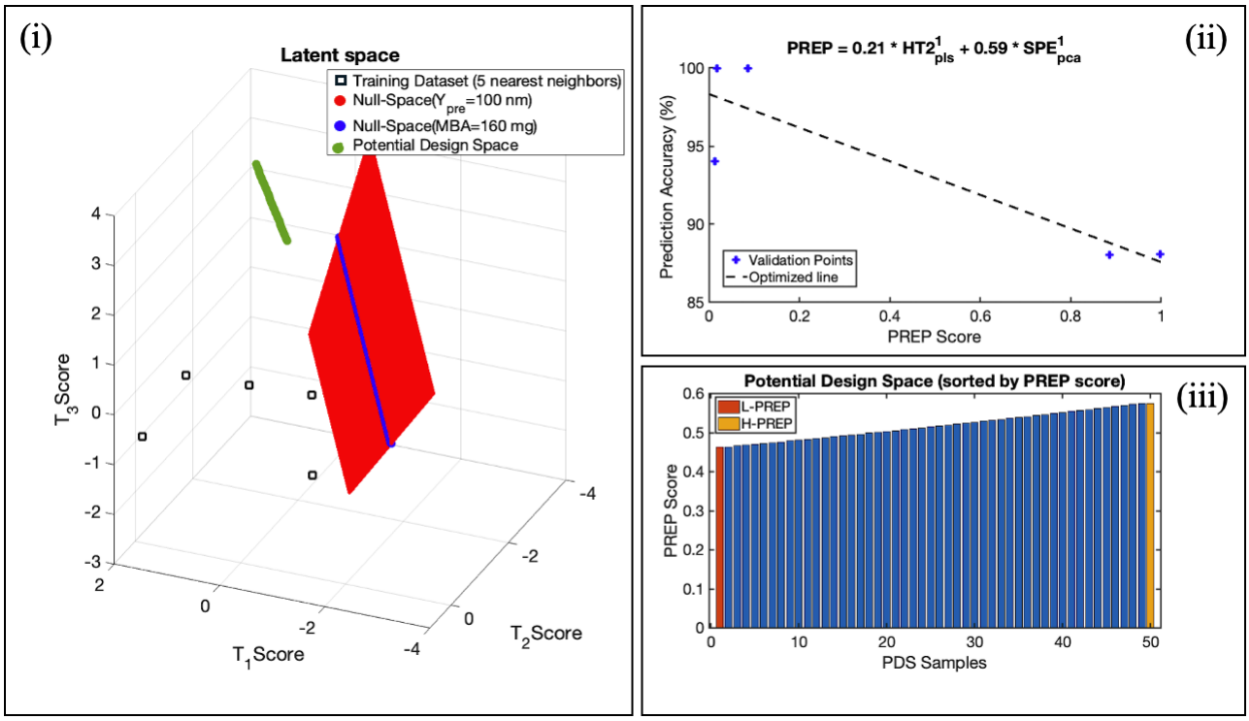


Figure 4-4: Results from iteration 1 of the PREP implementation on microgel optimization. Sub-panel (i) represents the visualization of the Potential Design Space (PDS) in the latent space, (ii) shows the outcome of the PREP equation optimization demonstrating Results from iteration 1 of the PREP implementation 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.

The updated latent space based on the revised dataset are shown in Figure 4-5(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 Figure 4-5(iii) subsequently synthesized. As shown in Table 4-2, the L-PREP sample demonstrates exceptional

proximity to the target particle size, achieving a size of 104 nm. Correspondingly, as shown in Figure 4-5 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.

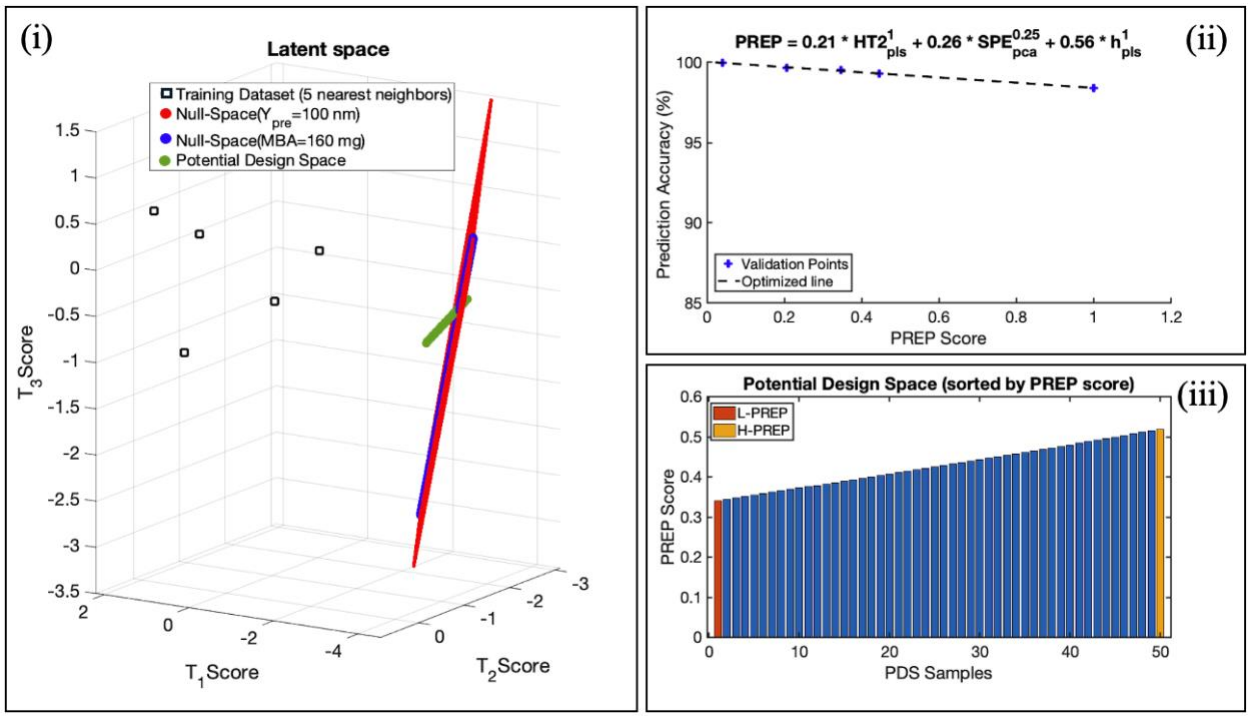


Figure 4-5: Results from iteration 2 of the PREP implementation on microgel optimization. Sub-panel (i) represents the visualization of the Potential Design Space (PDS) in the latent space, (ii) shows the outcome of the PREP equation optimization demonstrating the alignment of validation data points 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.

Table 4-2 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 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.

Table 4-2: 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**	

4.4.2 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 [58-60]. 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 [61, 62]) or as an additive with tunable release based on the ionic

interactions between the charged drug and its counterion polymer [63, 64]. 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 [58, 65], 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 [66]) 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.

• *Experimental Details*

Materials

Sulfated yeast beta glucan (glucan sulfate, GS) from *S. cerevisiae* was prepared as described by Williams et al. [67] (M_n = 13.5 kDa, Đ = 5.5, sulfur degree of substitution = 0.33, charge density = 1.54 ± 0.06 μeq/mg). Cationic dextran (Dex-GTAC) was prepared via functionalization with

glycidyltrimethylammonium chloride in the presence of NaOH according to previous methods [68, 69] (M_n = 3.7 kDa, Đ = 1.05, nitrogen degree of substitution = 0.50, charge density = 2.09 ± 0.1 μeq/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 4-3. GS, Dex-GTAC, and DOX were dissolved in PBS prepared at the ionic strength identified in Table 4-3, 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.

Table 4-3: 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]*	*Pre-Assembly GS Conc. [mg/mL]*	*Pre-Assembly Dex-GTAC Conc. [mg/mL]*	*Pre-Assembly DOX Conc. [mg/mL]*	*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**

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 4-3 was used as the model output. Figure 4-6 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 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 [70-72]. 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. Figure 4-6 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).

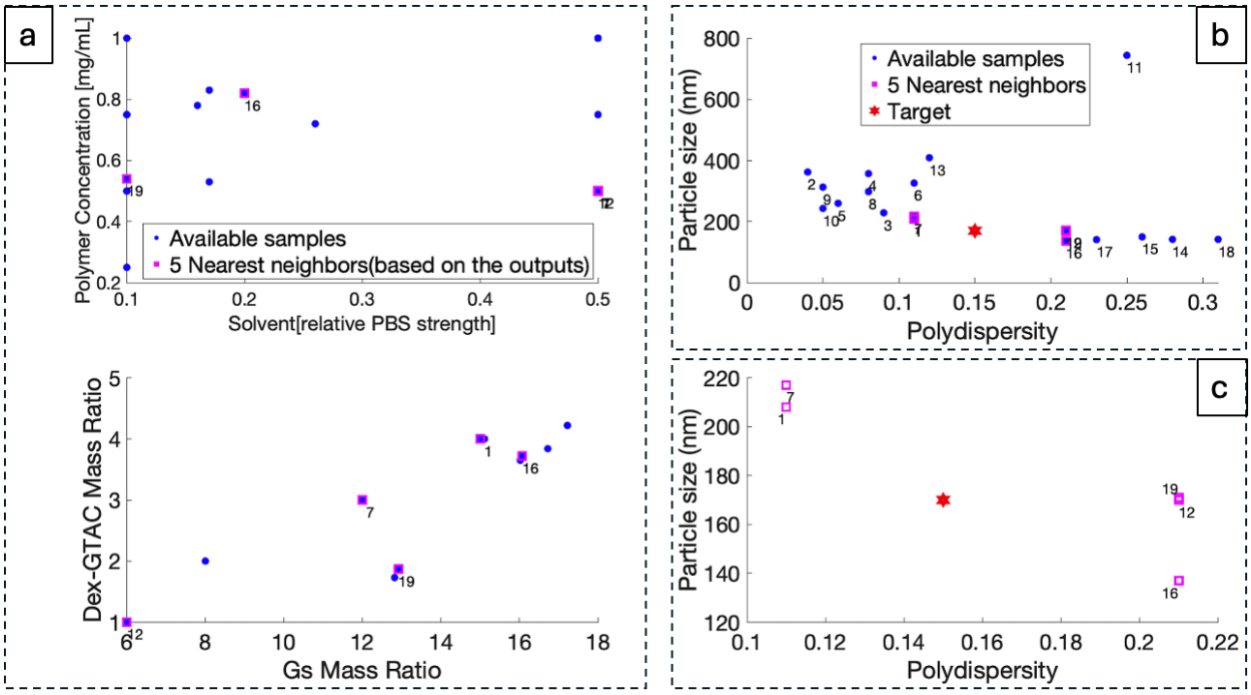


Figure 4-6: Visualization of all available data points along with the five nearest neighbors to the target in the input space (a) and output spaces showing all samples (b) and only the nearest neighbors (c) as derived from the pre-existing dataset summarized in Table 4-3.

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 Figure 4-7. 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 (Figure 4-7(a)), with two of the validation data points yielding prediction accuracy values as low as 60%. However, in the second iteration (Figure 4-7(b)), model accuracy improved substantially, with the lowest prediction accuracy among the validation data points showing a prediction accuracy of 85%. Table 4-4 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 (Figure 4-8(a) and 8(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 4-4) that precisely matched the model's targeted dispersity value. Note that by the fourth iteration (Figure 4-8(b)) even the least accurate validation sample achieved a prediction accuracy above 93%, showing the relevance of the PREP method to improve model outputs in minimal iterations. It is important to note that conducting the PREP algorithm over another two iterations (Table 4-4) 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 (Figures 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 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.

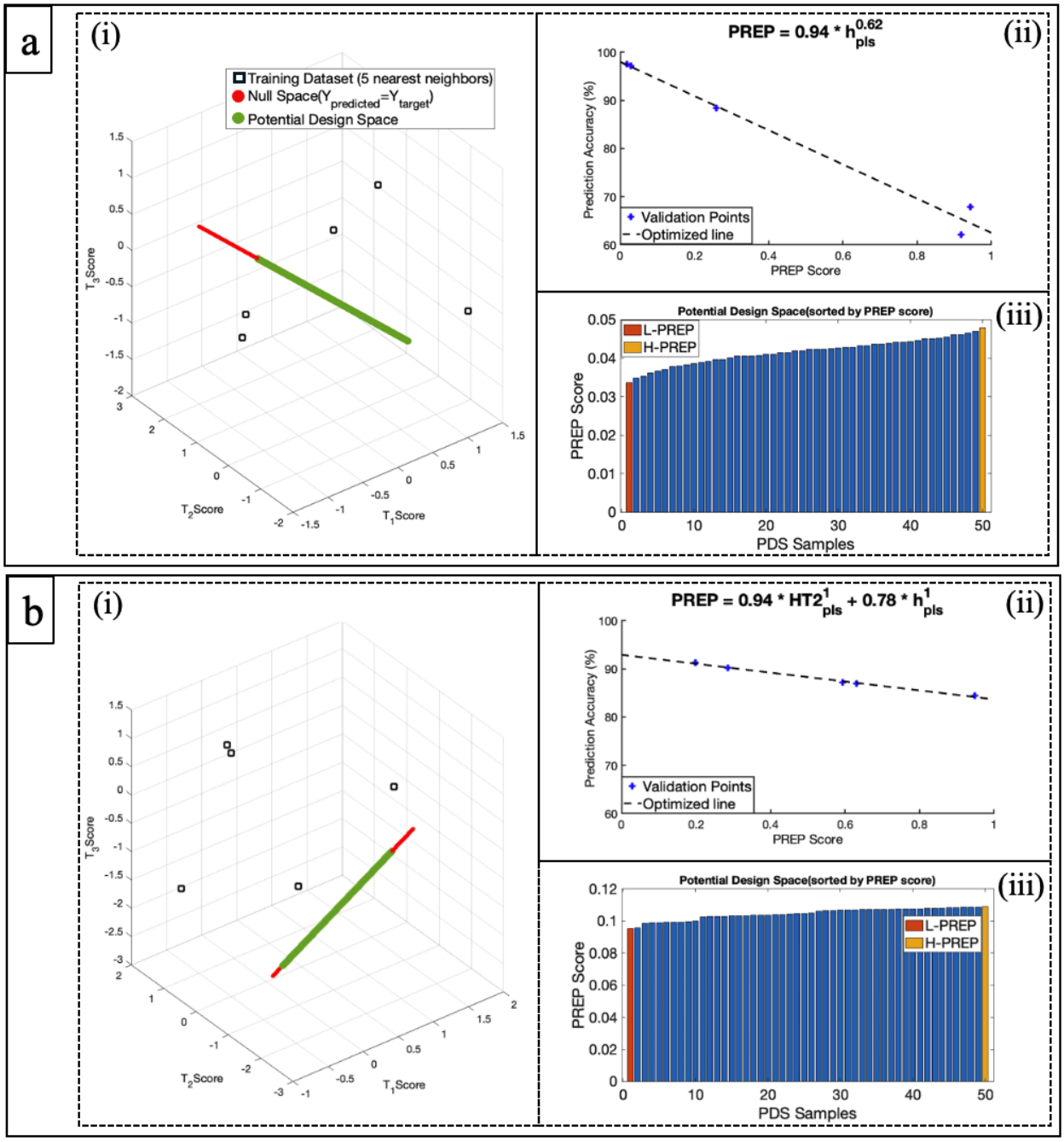


Figure 4-7: Results from iteration 1 (a) and iteration 2 (b) of the PREP implementation on PEC optimization. In each sub-panel, (i) represents the visualization of the Potential Design Space (PDS) in the latent space, (ii) shows the outcome of the PREP equation Results from iteration 1 (a) and iteration 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.

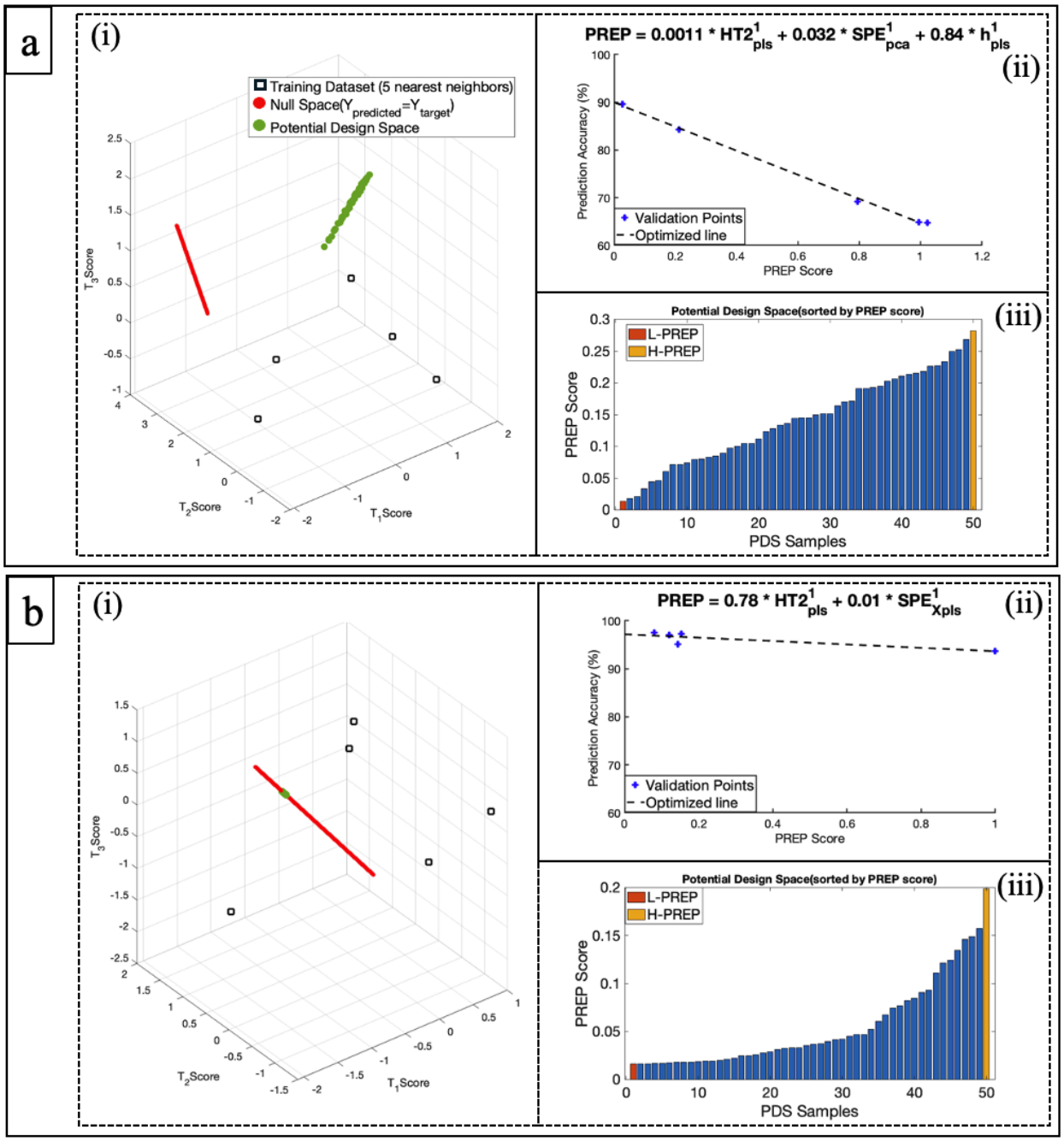


Figure 4-8: Results from iteration 3 (a) and iteration 4 (b) of the PREP implementation on PEC optimization. In each sub-panel, (i) represents the visualization of the Potential Design Space (PDS) in the latent space, (ii) shows the outcome of the PREP equation optimization demonstrating the alignment 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.

Table 4-4: 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]*	*Pre-Assembly GS Conc. [mg/mL]*	*Pre-Assembly Dex-GTAC Conc. [mg/mL]*	*Pre-Assembly DOX Conc. [mg/mL]*	*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***	**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***	**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**

** *Best performing samples*

Figure 4-9 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 [73, 74]) 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.

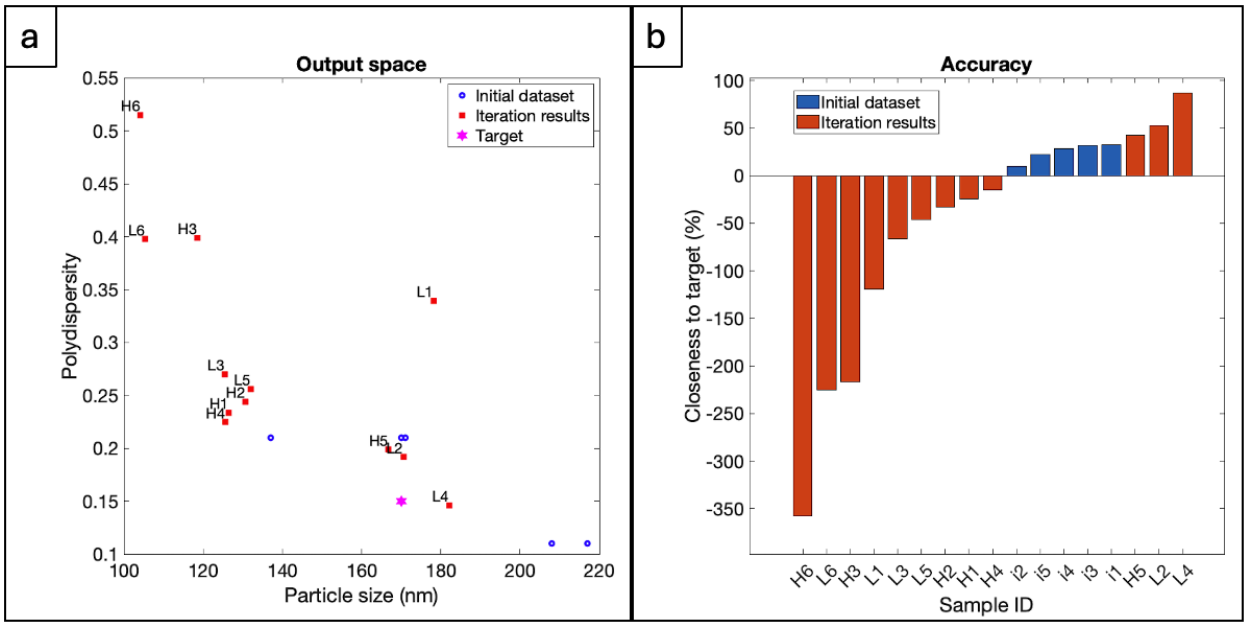


Figure 4-9: Assessment of iteration results relative to the target particle size and polydispersity expressed relative to (a) the actual output space and (b) the proximity of each datapoint to the target size.

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.

4.5 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 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 properties 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.

4.6 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.

Acknowledgments

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.

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.

Supporting Information

- PREP optimization results for the second case study (PEC) including outcomes from Iterations 5 and 6.

4.7 References

1. Hickey, J.W., et al., *Control of polymeric nanoparticle size to improve therapeutic delivery.* Journal of Controlled Release, 2015. **219**: p. 536-547.
2. Sun, N., T. Wang, and S. Zhang, *Radionuclide-labelled nanoparticles for cancer combination therapy: a review.* Journal of Nanobiotechnology, 2024. **22**: p. 728.
3. Gavas, S., S. Quazi, and T.M. Karpiński, *Nanoparticles for cancer therapy: current progress and challenges.* Nanoscale research letters, 2021. **16**(1): p. 173.
4. Kim, K.-R., et al., *Sentinel lymph node imaging by a fluorescently labeled DNA tetrahedron.* Biomaterials, 2013. **34**(21): p. 5226-5235.
5. Proulx, S.T., 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): p. 5128-5137.
6. Chou, L.Y., K. Zagorovsky, and W.C. Chan, *DNA assembly of nanoparticle superstructures for controlled biological delivery and elimination.* Nature nanotechnology, 2014. **9**(2): p. 148-155.
7. Patravale, V.B., A.A. Date, and A.B. Jindal, *Nanomedicines for the Prevention and Treatment of Infectious Diseases*. Vol. 56. 2023: Springer.
8. Sharifi, E., et al., *Nanostructures for prevention, diagnosis, and treatment of viral respiratory infections: from influenza virus to SARS-CoV-2 variants.* Journal of Nanobiotechnology, 2023. **21**(1): p. 199.
9. Adjei, I.M., C. Peetla, and V. Labhasetwar, *Heterogeneity in nanoparticles influences biodistribution and targeting.* Nanomedicine, 2014. **9**(2): p. 267-278.
10. Yoo, J.-W., N. Doshi, and S. Mitragotri, *Adaptive micro and nanoparticles: temporal control over carrier properties to facilitate drug delivery.* Advanced drug delivery reviews, 2011. **63**(14-15): p. 1247-1256.
11. Wu, L., J. Zhang, and W. Watanabe, *Physical and chemical stability of drug nanoparticles.* Advanced drug delivery reviews, 2011. **63**(6): p. 456-469.
12. Wong, C., et al., *Multistage nanoparticle delivery system for deep penetration into tumor tissue.* Proceedings of the National Academy of Sciences, 2011. **108**(6): p. 2426-2431.

13. Win, K.Y. 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): p. 2713-2722.
14. Williford, J.-M., et al., *Shape control in engineering of polymeric nanoparticles for therapeutic delivery.* Biomaterials science, 2015. **3**(7): p. 894-907.
15. Choi, S.H., 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): p. 1864-1870.
16. Albanese, A., et al., *Tumour-on-a-chip provides an optical window into nanoparticle tissue transport.* Nature communications, 2013. **4**(1): p. 2718.
17. Wang, X., et al., *Size control synthesis of melanin-like polydopamine nanoparticles by tuning radicals.* Polymer Chemistry, 2019. **10**(30): p. 4194-4200.
18. Song, Z., et al., *Size control of copper nanodrugs through emulsion atom transfer radical polymerization.* Polymer Chemistry, 2024. **15**(17): p. 1777-1785.
19. Mendes, A.C., et al., *Self-assembly in nature: using the principles of nature to create complex nanobiomaterials.* Wiley interdisciplinary reviews: nanomedicine and nanobiotechnology, 2013. **5**(6): p. 582-612.
20. Grzelczak, M., et al., *Directed self-assembly of nanoparticles.* ACS nano, 2010. **4**(7): p. 3591-3605.
21. Valencia, P.M., et al., *Microfluidic platform for combinatorial synthesis and optimization of targeted nanoparticles for cancer therapy.* ACS nano, 2013. **7**(12): p. 10671-10680.
22. Liu, D., et al., *A versatile and robust microfluidic platform toward high throughput synthesis of homogeneous nanoparticles with tunable properties.* Adv. Mater, 2015. **27**(14): p. 2298-2304.
23. Karnik, R., et al., *Microfluidic platform for controlled synthesis of polymeric nanoparticles.* Nano letters, 2008. **8**(9): p. 2906-2912.
24. Saswade, R., et al. *Data-driven Model Predictive Control of Nanoparticle Production in Modular Reactors*. in *2024 European Control Conference (ECC)*. 2024. IEEE.
25. Koronaki, E.D., et al., *Nonlinear manifold learning determines microgel size from Raman spectroscopy.* AIChE Journal, 2024. **70**(10): p. e18494.
26. Dong, S., et al., *Gaussian processes modeling for the prediction of polymeric nanoparticle formulation design to enhance encapsulation efficiency and therapeutic efficacy.* Drug Delivery and Translational Research, 2024: p. 1-17.
27. Sahai, N., M. Gogoi, and N. Ahmad, *Mathematical modeling and simulations for developing nanoparticle-based cancer drug delivery systems: a review.* Current Pathobiology Reports, 2021. **9**: p. 1-8.
28. Keßler, S., 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**: p. 9-18.
29. Keßler, S., F. Schmid, and K. Drese, *Modeling size controlled nanoparticle precipitation with the co-solvency method by spinodal decomposition.* Soft Matter, 2016. **12**(34): p. 7231-7240.
30. Stepanyan, R., et al., *Controlled nanoparticle formation by diffusion limited coalescence.* Physical review letters, 2012. **109**(13): p. 138301.
31. López-Domínguez, P., 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): p. 4467.
32. Lahiq, A.A. and S.M. Alshahrani, *State-of-the-art review on various mathematical approaches towards solving population balanced equations in pharmaceutical crystallization process.* Arabian Journal of Chemistry, 2023. **16**(8): p. 104929.
33. Dogra, P., et al., *Mathematical modeling in cancer nanomedicine: a review.* Biomedical microdevices, 2019. **21**: p. 1-23.

34. Sheibat-Othman, N., et al., *Is modeling the PSD in emulsion polymerization a finished problem? An overview.* Macromolecular Reaction Engineering, 2017. **11**(5): p. 1600059.
35. Mikayilov, E., et al., *Role of computational modeling in the design and development of nanotechnology-based Drug Delivery systems.* Chemical and Biochemical Engineering Quarterly, 2024. **38**(2): p. 97-110.
36. Al Najjar, T., N.K. Allam, and E.N. El Sawy, *Anionic/nonionic surfactants for controlled synthesis of highly concentrated sub-50 nm polystyrene spheres.* Nanoscale Advances, 2021. **3**(19): p. 5626-5635.
37. Ahmed, M.A., J. Erdőssy, and V. Horváth, *The role of the initiator system in the synthesis of acidic multifunctional nanoparticles designed for molecular imprinting of proteins.* Periodica Polytechnica Chemical Engineering, 2021. **65**(1): p. 28-41.
38. Palací-López, D., et al., *Improved formulation of the latent variable model inversion–based optimization problem for quality by design applications.* Journal of Chemometrics, 2020. **34**(6): p. e3230.
39. Tayebi, S.S., 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 Applied Polymer Materials, 2022. **4**(12): p. 9160-9175.
40. Yacoub, F. and J.F. MacGregor, *Product optimization and control in the latent variable space of nonlinear PLS models.* Chemometrics and intelligent laboratory systems, 2004. **70**(1): p. 63-74.
41. Zhang, L. and S. Garcia-Munoz, *A comparison of different methods to estimate prediction uncertainty using Partial Least Squares (PLS): a practitioner's perspective.* Chemometrics and intelligent laboratory systems, 2009. **97**(2): p. 152-158.
42. Denham, M.C., *Prediction intervals in partial least squares.* Journal of Chemometrics: A Journal of the Chemometrics Society, 1997. **11**(1): p. 39-52.
43. Serneels, S., P. Lemberge, and P.J. Van Espen, *Calculation of PLS prediction intervals using efficient recursive relations for the Jacobian matrix.* Journal of Chemometrics: A Journal of the Chemometrics Society, 2004. **18**(2): p. 76-80.
44. Faber, N.K.M., *Uncertainty estimation for multivariate regression coefficients.* Chemometrics and intelligent laboratory systems, 2002. **64**(2): p. 169-179.
45. Helland, I.S., *Partial least squares regression and statistical models.* Scandinavian journal of statistics, 1990: p. 97-114.
46. Faber, K. and B.R. Kowalski, *Prediction error in least squares regression: Further critique on the deviation used in The Unscrambler.* Chemometrics and Intelligent Laboratory Systems, 1996. **34**(2): p. 283-292.
47. Van Huffel, S. and J. Vandewalle, *The partial total least squares algorithm.* Journal of computational and applied mathematics, 1988. **21**(3): p. 333-341.
48. Phatak, A., P. Reilly, and A. Penlidis, *An approach to interval estimation in partial least squares regression.* Analytica chimica acta, 1993. **277**(2): p. 495-501.
49. Efron, B. and R.J. Tibshirani, *An introduction to the bootstrap*. 1994: Chapman and Hall/CRC.
50. Tayebi, S.S., T. Hoare, and P. Mhaskar, *Fast-Tracking Design Space Identification with the Prediction Reliability Enhancing Parameter (PREP).* Computers & Chemical Engineering, 2025: p. 109159.
51. Petrusic, S., et al., *Properties and drug release profile of poly (N-isopropylacrylamide) microgels functionalized with maleic anhydride and alginate.* Journal of Materials Science, 2013. **48**: p. 7935-7948.
52. Campora, S., et al., *Functionalized poly (N-isopropylacrylamide)-based microgels in tumor targeting and drug delivery.* Gels, 2021. **7**(4): p. 203.
53. Das, A., et al., *Poly (N-isopropylacrylamide) and its copolymers: a review on recent advances in the areas of sensing and biosensing.* Advanced Functional Materials, 2024. **34**(37): p. 2402432.
54. Hoare, T. and R. Pelton, *Highly pH and temperature responsive microgels functionalized with vinylacetic acid.* Macromolecules, 2004. **37**(7): p. 2544-2550.

55. Mok, Z.H., *The effect of particle size on drug bioavailability in various parts of the body.* Pharmaceutical Science Advances, 2024. **2**: p. 100031.
56. Osten, D.W., *Selection of optimal regression models via cross-validation.* Journal of Chemometrics, 1988. **2**(1): p. 39-48.
57. Cliff, N., *The eigenvalues-greater-than-one rule and the reliability of components.* Psychological bulletin, 1988. **103**(2): p. 276.
58. Meka, V.S., et al., *A comprehensive review on polyelectrolyte complexes.* Drug discovery today, 2017. **22**(11): p. 1697-1706.
59. Sarika, P. and N.R. James, *Polyelectrolyte complex nanoparticles from cationised gelatin and sodium alginate for curcumin delivery.* Carbohydrate polymers, 2016. **148**: p. 354-361.
60. Birch, N.P. and J.D. Schiffman, *Characterization of self-assembled polyelectrolyte complex nanoparticles formed from chitosan and pectin.* Langmuir, 2014. **30**(12): p. 3441-3447.
61. Liu, W., et al., *An investigation on the physicochemical properties of chitosan/DNA polyelectrolyte complexes.* Biomaterials, 2005. **26**(15): p. 2705-2711.
62. Oupický, D., et al., *DNA delivery systems based on complexes of DNA with synthetic polycations and their copolymers.* Journal of Controlled Release, 2000. **65**(1-2): p. 149-171.
63. Zhang, L., et al., *Preparation of polyelectrolyte complex nanoparticles of chitosan and poly (2-acry1amido-2-methylpropanesulfonic acid) for doxorubicin release.* Materials Science and Engineering: C, 2016. **58**: p. 724-729.
64. Luo, Y., et al., *Preparation, characterization and drug release behavior of polyion complex micelles.* International journal of pharmaceutics, 2009. **374**(1-2): p. 139-144.
65. Schatz, C., et al., *Formation and properties of positively charged colloids based on polyelectrolyte complexes of biopolymers.* Langmuir, 2004. **20**(18): p. 7766-7778.
66. Yuan, H., et al., *Effect of the modifications on the physicochemical and biological properties of β-glucan—A critical review.* Molecules, 2019. **25**(1): p. 57.
67. Williams, D.L., et al., *Development of a water-soluble, sulfated (1→ 3)-β-D-glucan biological response modifier derived from Saccharomyces cerevisiae.* Carbohydrate Research, 1992. **235**: p. 247-257.
68. Thomas, J.J., M. Rekha, and C.P. Sharma, *Dextran–glycidyltrimethylammonium chloride conjugate/DNA nanoplex: A potential non-viral and haemocompatible gene delivery system.* International journal of pharmaceutics, 2010. **389**(1-2): p. 195-206.
69. Bendoraitiene, J., et al., *Peculiarities of starch cationization with glycidyltrimethylammonium chloride.* Starch-Stärke, 2006. **58**(12): p. 623-631.
70. Mohan, T., et al., *Highly protein repellent and antiadhesive polysaccharide biomaterial coating for urinary catheter applications.* ACS Biomaterials Science & Engineering, 2019. **5**(11): p. 5825-5832.
71. Mocchiutti, P., et al., *Cationic and anionic polyelectrolyte complexes of xylan and chitosan. Interaction with lignocellulosic surfaces.* Carbohydrate polymers, 2016. **150**: p. 89-98.
72. Kodiyan, A., et al., *Surface modification with alginate-derived polymers for stable, protein-repellent, long-circulating gold nanoparticles.* Acs Nano, 2012. **6**(6): p. 4796-4805.
73. Olson, R.D., et al., *Doxorubicin cardiotoxicity may be caused by its metabolite, doxorubicinol.* Proceedings of the National Academy of Sciences, 1988. **85**(10): p. 3585-3589.
74. Ashikawa, K., et al., *Evidence that activation of nuclear factor-κB is essential for the cytotoxic effects of doxorubicin and its analogues.* Biochemical pharmacology, 2004. **67**(2): p. 353-364.

4.8 Supporting information of last chapter

The following figure provides the PREP optimization results for Iterations 5 and 6 of the second case study (PEC formulation), which are referenced in the main text but have been included here for completeness and to support the discussion of convergence behavior in later iterations.

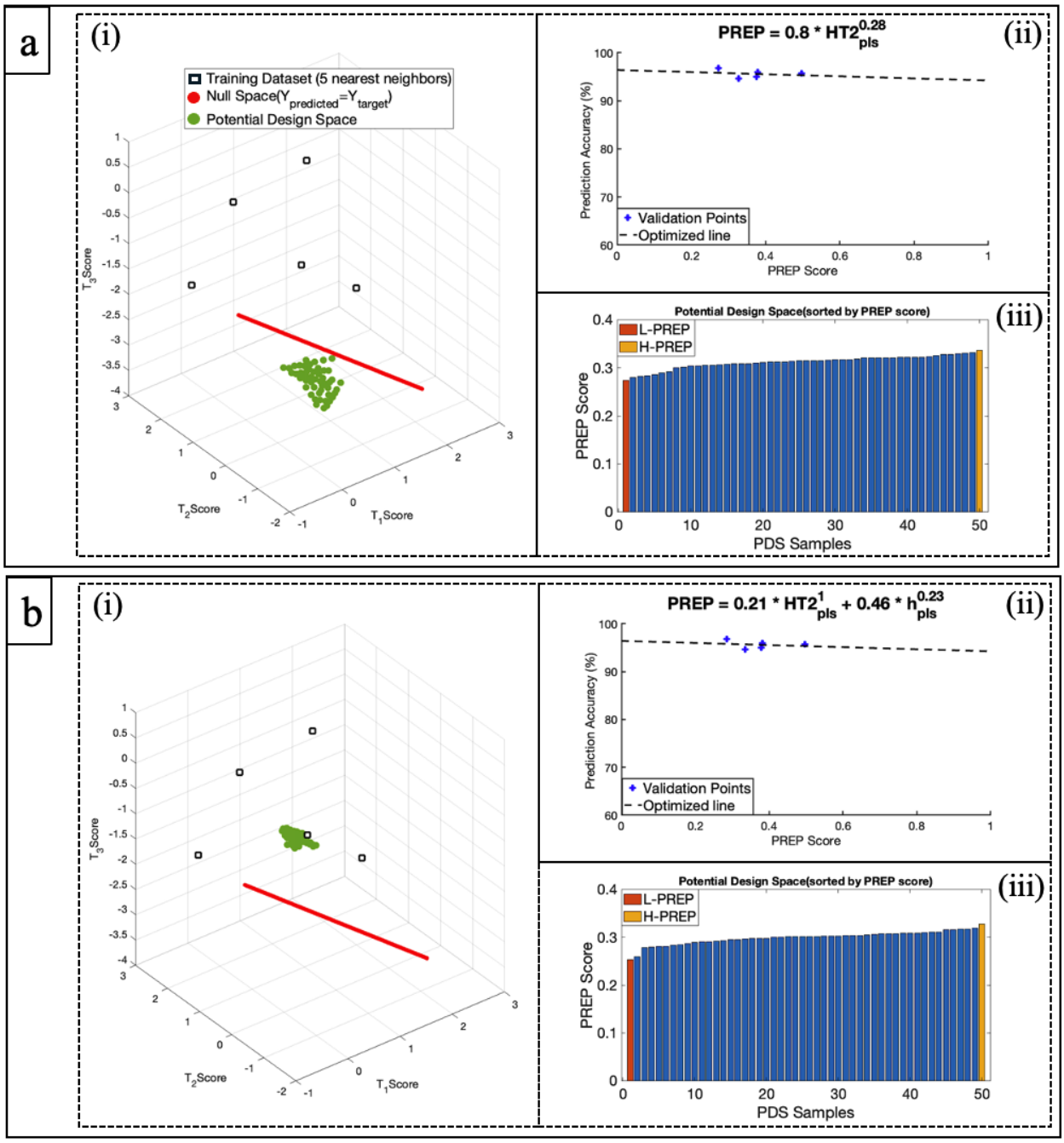


Figure S1. Results from iteration 5 (a) and iteration 6 (b) of the PREP implementation on PEC optimization. In each sub-panel, (i) represents the visualization of the Potential Design Space (PDS) in the latent space, (ii) shows the outcome of the PREP equation optimization demonstrating the alignment 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.

Chapter 5

5 Conclusions and Recommendations for Future Works

The main focus of this thesis was to address the challenges associated with model development over limited-sample datasets that exhibit high complexity and intricate relationships among variables. Solving these challenges has direct applications in formulation optimization, product design, and any batch dataset scenario for which historical data is available and the goal is either to understand how the system works or achieve a targeted final outcome. Specifically, this thesis aimed to enhance model prediction accuracy for small datasets and introduced a novel metric that consolidates multiple evaluation parameters into a single, unified score. This score helps in decision-making and guides dataset expansion towards identifying the optimal solution. In this chapter, we will summarize the key contributions of the thesis, highlighting the methodologies and experimental validations, and propose potential future research directions.

5.1 Conclusions

In Chapter 2, a novel approach was developed by coupling Partial Least Squares (PLS) modeling with data clustering to enhance the consistency of the LVM calibration dataset and, in turn, improve model predictions. The clustering of samples based on their similarities in input variables (X), output variables (Y), or a combination of both was explored. The results showed that this clustering approach offered advantages over traditional non-clustered PLS models. When applied to a real experimental case—predicting the properties of dual-responsive microgels—two new experimental samples were synthesized whose formulation recipes fell within the reliable prediction zone (green region) defined by the clustered model. The predicted swelling profiles for these samples were in strong agreement with the measured experimental outcomes, providing clear evidence that the model was both predictive and reliable for practical design scenarios. This clustering-based approach thus holds potential to replace the time-

consuming trial-and-error methods currently used in product design, offering a more efficient pathway to achieving specific target properties.

In Chapter 3, a new methodology called the Prediction Reliability Enhancing Parameter (PREP) was introduced to improve the identification of samples that yield more reliable predictions compared to alternative candidates by expediting the process of identifying the Design Space (DS). By applying PREP iteratively, we were able to achieve much faster convergence toward the optimal solution compared to conventional methods based on simulated and highly non-linear data sets. The efficiency of this approach proved especially beneficial in scenarios in which quick DS identification is essential and/or sample preparation is costly and time-consuming. A key advantage of PREP lies in its ability to optimize the allocation of experimental resources by reducing the number of required iterations, thus minimizing material and operational expenses. This method is versatile and can be applied to datasets of any dimensionality.

Finally in Chapter 4, the PREP method was effectively employed to optimize the synthesis of nanoparticles with precise control over particle size and size distribution. The approach was tested in two distinct case studies, each involving different nanoparticle types and fabrication techniques: the precipitation polymerization of dual-responsive microgels and the physical self-assembly of polyelectrolyte complex nanoparticles. In both cases, PREP successfully achieved the desired particle size and polydispersity characteristics within just two iterations, delivering highly accurate results under challenging design constraints that led to traditional methods failing to achieve the target size. The successful application of PREP in these nanoparticle design scenarios demonstrates its strong potential for addressing a wide range of complex optimization problems in materials design, particularly in high-dimensional spaces.

5.2 Future Works

The findings of this research open several avenues for future investigation.

- One promising direction for future research is the integration of fuzzy clustering with PLS models. In some cases, general rules may apply within specific clusters while other rules might hold true across all clusters. Fuzzy clustering allows for the relative inclusion of information from multiple clusters, offering potential advantages over traditional methods that disregard such relationships. There are two possible approaches for implementing fuzzy clustering within the PLS framework. The simpler approach involves developing PLS models based on the already introduced cluster-based PLS method in Chapter 2 and using fuzzy clustering to determine the degree of membership of a new observation. The model prediction for the new data point would then be weighted according to its membership in each cluster. Alternatively, a more complex approach could involve developing PLS models for each cluster using all available data, weighted by their respective membership values. This approach allows each cluster's model to incorporate broader information, enhancing its expertise within its specific domain and potentially improving overall prediction accuracy.
- Another promising avenue for future work involves integrating propensity models with conventional PLS approaches, such as Bayesian PLS, to account for model parameter uncertainty. Incorporating this uncertainty into the existing PREP equation could further enhance decision-making by providing a more comprehensive evaluation of prediction reliability. This approach would require the development of multiple PREP equations, each corresponding to different levels of model parameter variance. By applying these varied PREP equations to the members of the potential design space, it would be possible to identify the samples with both lower PREP scores and more robust uncertainty estimates, thereby enabling more informed and reliable choices for experimental exploration.

- An additional avenue for future research involves the incorporation of nonlinear and kernel-based techniques, such as Kernel PLS and Kernel PCA, into the PREP methodology. These techniques are particularly well-suited for systems characterized by higher complexity or higher-dimensional datasets. Traditional PLS, as a linear regression-based model, often struggles to capture inherent nonlinear patterns present in the data. By integrating nonlinear methods at the core of the PREP framework, it may become possible to reveal and model these complex relationships more effectively. This extension would expand the applicability of the PREP method, enabling its use in a broader range of systems in which nonlinearities play a crucial role. Furthermore, this approach could improve the accuracy and robustness of predictions in high-dimensional settings, potentially opening new avenues for the application of PREP in diverse industries such as materials science, biotechnology, and beyond.
- The last promising avenue for future work is the extension of the PREP methodology to incorporate a three-block data structure that includes not only the traditional blocks of input variables (X) and response variables (Y) but also a third block representing the system's state, such as initial conditions. This structure is particularly relevant in many fields such as personalized medicine in which the state of the system (e.g. a patient's initial condition) plays a critical role in determining the optimal intervention. For example, in pharmaceutical applications, patient-specific data (S) can be used to determine the most appropriate drug dosage (X) needed to achieve the desired therapeutic outcome (Y). By adapting the PREP method to handle this three-block structure, PREP could then evaluate not only how the system responds to changes in input variables but also how the initial state conditions impact the outcome. Such an advancement would expand the utility of PREP in personalized healthcare, enabling more accurate and tailored drug dosage recommendations that could improve patient treatment plans and enhance overall therapeutic efficacy.