Received: 30 November 2023 | Revised: 12 March 2024 | Accepted: 5 May 2024

DOI: 10.1002/aic.18494

AIChE JOURNAL

RESEARCH ARTICLE

Editor's Choice: Process Systems Engineering

Nonlinear manifold learning determines microgel size from Raman spectroscopy

Eleni D. Koronaki[1] | Luise F. Kaven[2] | Johannes M. M. Faust[2] | Ioannis G. Kevrekidis[3] | Alexander Mitsos[2,4,5]

[1]Faculté des Sciences, de la Technologie et de la Communication, Université de Luxembourg, Esch-sur-Alzette, Luxembourg

[2]Process Systems Engineering (AVT.SVT), RWTH Aachen University, Aachen, Germany

[3]Department of Chemical and Biomolecular Engineering and Department of Applied Mathematics and Statistics, Whiting School of Engineering, Johns Hopkins University, Baltimore, Maryland, USA

[4]JARA-CSD, Aachen, Germany

[5]Institute of Energy and Climate Research, Energy Systems Engineering (IEK-10), Forschungszentrum Jülich GmbH, Jülich, Germany

Correspondence
Alexander Mitsos, Process Systems Engineering (AVT.SVT), RWTH Aachen University, Forckenbeckstr. 51, 52074 Aachen, Germany.
Email: amitsos@alum.mit.edu

Funding information
Luxembourg National Research Fund (FNR), Grant/Award Number: 16758846; US Air Force Office of Scientific Research; Deutsche Forschungsgemeinschaft, Grant/Award Number: CRC 985; Horizon 2020 Framework Programme, Grant/Award Number: 890676 – DataProMat

Abstract

Polymer particle size constitutes a crucial characteristic of product quality in polymerization. Raman spectroscopy is an established and reliable process analytical technology for in-line concentration monitoring. Recent approaches and some theoretical considerations show a correlation between Raman signals and particle sizes but do not determine polymer size from Raman spectroscopic measurements accurately and reliably. With this in mind, we propose three alternative machine learning workflows to perform this task, all involving diffusion maps, a nonlinear manifold learning technique for dimensionality reduction: (i) directly from diffusion maps, (ii) alternating diffusion maps, and (iii) conformal autoencoder neural networks. We apply the workflows to a data set of Raman spectra with associated size measured via dynamic light scattering of 47 microgel (cross-linked polymer) samples in a diameter range of 208–483 nm. The conformal autoencoders substantially outperform state-of-the-art methods and results for the first time in a promising prediction of polymer size from Raman spectra.

KEYWORDS
diffusion maps, machine learning, nonlinear manifold, polymerization, Raman spectroscopy

1 | INTRODUCTION

Process analytical methods are crucial for optimizing process performance and product properties, especially in polymerization. In-line spectroscopic methods are advantageous, and, in particular, near-infrared and Raman spectroscopy are widely applied spectroscopic methods, see, for example, the reviews.[1–4] Evaluation methods for concentrations from (either online or offline) spectroscopic data are established and comprise regression models, such as univariate peak integration based on the Beer-Lambert law,[5] multivariate partial least squares (PLS), or artificial neural networks (ANNs),[6] and physically supported strategies such as multivariate curve resolution-alternating least squares[7] or indirect hard modeling (IHM).[8]

Size is a crucial product feature in several processes, for example, polymerization and crystallization. In contrast to concentrations, the size prediction from spectroscopic data remains a major challenge.

Herein, we consider cross-linked polymer networks called microgels, which are not considered particles.[9] Thus, we use the term *microgel size* or *polymer size* instead of the more common *particle size*. The fact that particles such as polymers influence spectroscopic measurements through light scattering is well established,[10] and experimental evidence of the correlation between Raman scattering and polymer size has been presented,[11] even for relatively large particles.[12] However, only a few approaches attempt to predict polymer sizes from Raman spectra.[11–16] These approaches rely on relatively small sets of data points and focus on training data-driven models. As most literature studies do not provide performance metrices or alternatively raw data publication, a comprehensive comparison of the quantitative performance metrices of studies from the literature is not feasible. However, qualitatively a lack of prediction accuracy of these approaches is observed compared to established methods such as dynamic light scattering (DLS). An overview of the state-of-the-art work on polymer size prediction from Raman spectra is presented in Table 1. Most of these methods are based on the linear methods PLS or principal component analysis (PCA), which reduce the predictors to a smaller set of uncorrelated components and perform least squares regression on these components instead of on the original data.

Each spectral measurement consists of many measured intensities, resulting in a large dimensionality of the input vector. However, the measured intensities are not independent and ideally depend on a small number of meaningful properties, for example, concentration and size. When the available data live in a low-dimensional, yet nonlinear manifold, linear methods often fail to capture the majority of the variance of the data, even with an increased number of principal components. Hence, we propose using nonlinear manifold learning approaches to achieve significant dimensionality reduction and, more importantly, to identify latent variables possessing specific desired properties. Dimensionality reduction replaces extensive data sets with a handful of latent variables. For the reduction, we employ diffusion maps (DMAPs),[17–19] to derive a parsimonious reduced description of the Raman spectra. Then, as one alternative, we predict the polymer size directly from the latent variables computed with the DMAPs algorithm. Moreover, we take additional steps regarding the characteristics of the latent space by applying two alternative machine learning methods to discover common latent variables between the spectra and the polymer sizes. In this sense, *common* describes a set of variables that has a one-to-one correlation to the desired observed quantity. The common variable between the spectra and the polymer size is used as a data-driven junction through which we can predict the polymer size given the spectrum of a new sample.

We propose three alternative machine learning methods (presented schematically in Figure 1): (i) direct prediction from DMAP

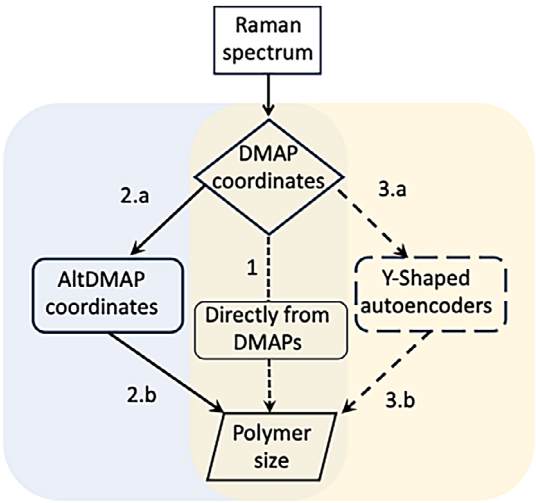


FIGURE 1 Schematic overview of the proposed machine learning approaches applied to low dimensional parameterization (provided by diffusion maps [DMAPs]) of measured Raman spectra: (i) in Approach 1 the polymer size is predicted directly from DMAPs, (ii) Approach 2 implements alternating diffusion maps (AltDMAPs) to find the data-driven common variable that is one-to-one with the polymer size (2.a), and in (2.b) predicts the polymer size from this common variable, and (iii) Approach 3 implements a Y-shaped conformal autoencoder, that identifies a "designer" latent space (3.a) from which it is possible to predict the polymer size (3.b).

TABLE 1 Overview of state-of-the-art approaches and our proposal (last row) to predict polymer sizes from Raman spectra.

References	Method	Polymer system	Size (nm)	Number of points	Spectrum (cm^{-1})
13	PLS	Styrene, butadiene, methyl methacrylate, acrylonitrile	80–200	47	100–4000
12	Focal depth	styrene	42–210	6	3200–3800
11	PLS	styrene, acrylic acid	55–150	23	400–4000
14	PLS	styrene, butyl acrylate, acrylic acid, methyl acrylate	20–200	N.S.	150–400, 150–1800
15	PLS	styrene	50–400	40	150–400, 0–4000
16	IHM+PLS	styrene	23–60	21	1020–2000
Nonlinear Manifold Learning		N-isopropylacryl-amide	208–483	47	100–3425

Note: The number of data points refers to the amount of samples measured via Raman spectroscopy.
Abbreviations: IHM, indirect hard modeling; PLS, partial least squares.

coordinates; (ii) an alternating diffusion maps (AltDMAPs) algorithm, initially introduced by Lederman et al.,[20] and later in the context of multimodal data fusion[21] und identification of "jointly smooth" functions on input and output manifolds;[22] and (iii) an ensemble of concurrently trained neural networks (NNs), named Y-shaped conformal autoencoders, introduced by Evangelou et al.[23] in the context of parameter non-identifiability.

Methods (ii) and (iii) are particularly appropriate for measurements that depend on various combinations of factors, the effect of which is not readily quantifiable. Herein, these methods are employed to identify the changes in the spectra explicitly attributed to the polymer sizes since the samples are not pretreated. As several factors beyond the size influence, for example, concentrations of monomer, inhibitor, surfactant, and other partaking species in the reaction, influence the spectra, the proposed methods act as a nonlinear filter that isolates (in a sense, disentangles) the spectral changes that are attributed to the differences in polymer size.

The common starting point for the machine learning approaches proposed in this work addresses the reduction of the high dimensionality of spectra, here achieved with DMAPs. The DMAPs algorithm is based upon (mathematical) diffusion processes on the data and facilitates discovering meaningful low-dimensional intrinsic geometric descriptions of data sets, even when the data is high-dimensional, nonlinear, and corrupted by (relatively small) noise. The algorithm has been successfully used for dimensionality reduction in applications relevant to reaction engineering in References 24 and 25, among others. We propose its use as an effective dimensionality reduction of full spectra, consisting of approximately 11,000 wavenumbers. This reduction enables efficient interpolation and regression since much fewer (typically < 10) variables are involved. Once the low-dimensional representation of the spectra is determined in an *offline step*, it is possible to translate between coordinates in the ambient (spectra) and the reduced space (DMAP coordinates). The mapping from high to low dimension is achieved with the Nyström extension,[26,27] whereas for the inverse, a particular implementation of Geometric Harmonics, called DoubleDMAPs, is selected.[28] Accurate reconstruction of the data set from the selected DMAP coordinates indicates that the latter is an adequate low-dimensional parameterization.

After the dimensionality reduction, three alternative machine learning workflows (i) to (iii) are compared. The first approach involves regression analysis to estimate the relationship between the latent variables (DMAP coordinates) and the polymer size (cf. Approach 1 denoted as "Directly from DMAPs" in Figure 1). For the implementation of AltDMAPs, a common variable between spectra (in their reduced representation) and polymer size is found here as an *offline* step (cf. Step 2.a in Figure 1). Then, the overall *online* computational workflow to predict the polymer size from a new spectrum starts by determining its low-dimensional DMAP coordinates with the Nyström extension. From that, the common variable, the AltDMAP coordinate, is predicted with an ANN or other regression methods, such as Extreme Gradient Boosting (XGBOOST) (cf. Step 2.a in Figure 1). Finally, the corresponding size is predicted from the AltDMAP coordinates with an ANN or with DoubleDMAPs (cf. Step 2.b in Figure 1).

In approach (iii), we exploit recent advances in conformal autoencoder NN techniques. The DMAP coordinates are used as inputs (and also outputs) to a Y-shaped autoencoder to disentangle the dependencies of the latent variables discovered by a traditional autoencoder architecture (cf. Step 3.a in Figure 1) and define the desired output, that is, the polymer size, as a function of a latent variable (cf. Step 3.b in Figure 1). Ultimately, given a new set of DMAPs, this "designer" NN predicts the corresponding polymer size (and reconstruct the DMAPs themselves).

We compare the proposed workflows with state-of-the-art techniques and show that it is essential to not only find *a generic* parsimonious low-dimensional parameterization of the data (here achieved with DMAPs) but to find the *appropriate one*, possessing a component that the polymer size can be written as a function of. We demonstrate that although the number of pairs of microgel size and spectral measurements is moderate (47, i.e., at least as high as in previous works, Table 1), the workflow is a promising direction in predicting polymer size in-line from Raman spectra.

The remainder of the article is structured as follows: First, the process of data collection is presented along with an overview of the state-of-the-art methods for polymer size prediction from spectra. Subsequently, the DMAPs and AltDMAPs methods are summarized, followed by a detailed description of the proposed workflow. The conformal autoencoder architecture is then presented, building on the successful dimensionality reduction achieved with DMAPs. Finally, results and conclusions are drawn from the proposed implementation.

2 | METHODS

The following sections include the description of the data set used in this contribution and the applied methods for size prediction: benchmark methods, and the proposed workflows, including the DMAPs approach, AltDMAPs workflow, and the conformal autoencoder.

2.1 | Data

We employ data from our samples taken from continuous synthesis of microgels.[29] The considered microgels here are based on N-isopropylacrylamide and cross-linked via *N*,*N*′-methylenebis(acrylamide). The reactor and measurement setup for the continuous synthesis are explained in our previous work.[30] Using the continuous flow reactor, microgels of different sizes are synthesized by changing the reactor temperature and flow rates and the initiator and surfactant concentration. As microgels are known to be of monodisperse size,[31-33] they represent an excellent system to study polymer size predictions from Raman spectroscopy. In our previous works,[29] we conducted in-line Raman spectra at reaction temperature at 60–80°C and DLS measurements at 50°C. In contrast, in this article, we acquire additional offline Raman measurements of the same samples but at 20°C and restricted conditions. The restrictions include measuring the

samples all within a small amount of time (over two experimentation days) and in glass vials filled to the exact same fluid level to ensure equal conditions for the acquisition of all spectra and to eliminate external influences on the measurements. Consequently, we also conduct further DLS measurements at 20°C.

The data consist of Raman spectra and DLS measurements of microgel samples. The samples are taken from the output of the continuous flow reactor, which runs at different experimental conditions for each sample. The samples are measured off-line without further treatment, for example, filtration or dialysis. In total, we use the data from 47 samples at different microgel sizes in the range of 208–483 nm in diameter. The determined polydispersity index of these microgels ranges well below 0.368, indicating monodisperse size distribution.

Microgels have a different size depending on a threshold in temperature: above approximately 32°C, they occur in a collapsed state; below the threshold temperature, they occur in a swollen state. Hence, the microgel sizes at 20°C are almost twice as big as at the reaction temperature. Each sample is measured three times via Raman spectroscopy. Raman spectra are taken with an acquisition time of 40 s using an RXN2 Raman Analyzer (Kaiser Optical Systems) with cosmic ray correction. DLS measurements of the samples diluted in ultrapure water are conducted via the Zetasizer Ultra (Malvern Panalytical) at 20°C with a scattering angle of 90°. Each DLS measurement is repeated four times, and the software ZS Xplorer analyzes the scattering intensities.

The Raman spectra comprise the Raman intensity measured over a range of wavelengths. The global range is between 100 to 3425 cm^{-1} correlating to 11,084 wavelengths per spectrum. Different spectra pretreatment methods can be applied to the spectral data. We compare using raw spectra and spectra with a linear fit or rubber band baseline correction in combination with standardization in the form of either Standard Normal Variate (SNV) or Min-Max normalization. The experimental data set is published open access[34] and comprises raw and pretreated Raman and evaluated DLS data.

We use the same data set for all subsequently described methods to predict microgel size from Raman spectra. Out of the 47 pairs of microgel size and Raman spectra, we take 40 for training and 7 for testing. The same split is applied to quantify the prediction performance of each considered method (state-of-the-art methods and our proposed workflow with nonlinear methods). We conduct the training for each prediction method with 10-fold cross-validation, which involves splitting the training set into 10 smaller subsets and using nine for training and one for testing. By repeating this process, using a different collection of subsets for training and validation each time, it is possible to define the best possible model hyper-parameters without sacrificing a lot of data. The number of hyper-parameters varies depending on the prediction method. Hence, the set of hyper-parameters for the individual method is described for each method separately in the following sections. The prediction performance of each method is evaluated based on commonly applied metrics: coefficients of determination (R^2), root mean squared error (RMSE), and mean absolute percentage error (MAPE) for training and testing.

The prediction accuracy is reflected in the %-error calculated as:

$$\%\text{-error} = 100 \cdot \frac{D_H^{\text{predicted}} - D_H^{\text{actual}}}{D_H^{\text{actual}}}, \quad (1)$$

where D_H is the microgel's hydrodynamic diameter. Based on the %-error the MAPE is calculated as the sum of the %-errors divided by the number of observations.

2.2 | Benchmark methods for polymer size prediction from Raman spectra

To benchmark our proposed method, we compare it against two state-of-the-art methods. These methods include the direct application of PLS to the spectral intensities and the application of PLS to fitted IHM parameters.

2.2.1 | PLS regression of spectral intensities

As conducted in the literature,[11,13,14,15] we apply a PLS model regression directly to the spectral data as first introduced by Ito et al.[13] We consider different spectral ranges for the calibration of our PLS models. These ranges include the global range and the so-called fingerprint (FP) region between 850 and 1800 cm^{-1}. Also, we apply pretreatment methods, combining two different types of baseline subtractions (linear fit and rubber band) and two normalization approaches (MinMax and SNV). Further, we normalize the data using the `zscore` function in Matlab. We analyze the results based on the metrics R^2, RMSE, and MAPE for calibration and validation. Based on the MSE for cross-validation, we chose the number of components (latent variables) for the PLS regression.

2.2.2 | Regression of hard model parameters

We conduct the regression of fitted hard model parameters for predicting microgel sizes, as we proposed in previous works.[16] First, an IHM[8] is developed. For that model, the spectral range of the FP is considered. The range between 1552 and 1560 cm^{-1} is excluded as it is attributed to an atmospheric oxygen signal. Besides the range restrictions, no further pretreatment is usually applied as per our findings.[16] However, we compare the PLS performance of the IHM parameters with and without pretreatment for a comprehensive comparison. The applied pretreatment for the comparison is MinMax or SNV normalization and linear fit or rubber band baseline subtraction. The IHM prediction model is calibrated using calibration measurements from our previous work.[35]

The model includes component hard models for the monomer, polymer, and water and a linear baseline. The hard model of each component consists of multiple characteristic peaks. The individual peaks are characterized by four parameters: position, intensity, shape

(fraction of the Gaussian part), and half-width at half maximum. The complete indirect hard model combines the component models with their distinctive peaks. The indirect hard model parameters are adjusted to suit the spectra of interest within the fitting process. The applied fitting mode constitutes a medium interaction method, where the weights of the components, the baseline, and the peak positions are varied. The weights of the components represent the magnitude of the individual component in the spectra. Each component (monomer, polymer, water) is accredited with one weight parameter during the model fitting process. The incorporated linear baseline is fitted with regard to its offset and slope. In this context, we restrict component shifts to avoid ambiguities due to overlapping spectral peak positions. The fitting mode follows our previous work.[16] The medium interaction fitting mode results in 49 modified parameter values (intercept and slope of the baseline, one weight for each of the three components, 23 monomer peak positions, 17 polymer peak positions, and four water peak positions) that serve as the input variables to the PLS regression model.

In addition, we compare the medium fitting mode with the fitting mode with high interaction. For the high interaction, in addition to the changes in the medium interaction, all peak parameters can be varied within the fitting process. Thus, the high interaction method results in 181 modified parameter values (intercept and slope of the baseline, one weight for each of the three components, 23 monomer peaks, 17 polymer peaks, and four water peaks, where each peak is characterized by the four parameters described previously).

The fitted IHM parameter values serve as input to the subsequent PLS regression. Again, we normalize the data using the `zscore` function in Matlab. We also analyze the results of the PLS regression based on the R^2, RMSE, and MAPE values for the hybrid modeling approach, combining IHM and PLS. Based on the MSE for cross-validation, we choose the number of components for the PLS regression.

2.3 | Diffusion maps

The following paragraphs highlight the functionality of DMAPs for dimensionality reduction. In addition, we establish a workflow for predicting polymer sizes directly from DMAP coordinates.

2.3.1 | DMAPs for dimensionality reduction

The DMAPs framework has been shown to facilitate the discovery of meaningful low-dimensional intrinsic geometric descriptions of data sets.[25,28,36] For a detailed description of the method, the interested reader is referred to the seminal papers[17,37] and also, among others .[19,24,25,28] Here, we use DMAPs to discover the dimensionality of the manifold that contains a collection of Raman spectra of microgel samples. Furthermore, DMAPs discover data-driven coordinates on (i.e., parameterizations of) the low-dimensional manifold where the data reside. These coordinates do not necessarily have a physical meaning, that is, they do not have to correlate to any physical quantity. The coordinates of the manifold are a few of the leading eigenvectors, ϕ_i, of a scaled affinity matrix, which contains the Euclidean distances between all the pairs of available data points.

Discovering which eigenvectors parameterize independent directions and do not span the same direction with different frequencies (harmonics) is essential. To distinguish independent eigenvectors, the local linear regression algorithm can be used as proposed by Dsilva et al.[38]

Notably, the proposed approach includes reverse mapping from the DMAP coordinates to the variables in ambient space, which allows for the translation between the high- and low-dimensional data description. To this end, Geometric Harmonics are proposed, introduced initially in Reference 17, as a scheme for extending functions defined on data $\mathbf{X}$, $f(\mathbf{X}) : \mathbf{X} \rightarrow \mathbf{R}$, for $x_{\text{new}} \notin \mathbf{X}$. Here, the DoubleDMAPs[28] algorithm, a particular implementation of Geometric Harmonics, is selected. DoubleDMAPs are suitable for low-dimensional data that can be parameterized by just a few nonharmonic eigenvectors. Generally, Geometric Harmonics construct an input-output mapping between the ambient coordinates $\mathbf{X}$ and a function of interest f defined on $\mathbf{X}$ by operating directly on the non-harmonic DMAP coordinates.

2.3.2 | Prediction directly from DMAPs

Establishing a parsimonious embedding of the spectra enables polymer size prediction directly from the DMAP coordinates. The direct prediction based on DMAP coordinates is achieved here with two different regression methods: (a) a neural network (NN) and (b) XGBOOST, using the DMAP coordinates as input and the polymer size as output. The schematic representation of the direct workflow is shown in Figure 2.

This direct approach is expected to perform well when the latent variables derived by DMAPs possess a one-to-one dependence on the desired observable, here the polymer size.

2.4 | Alternating DMAPs workflow

The following sections introduce to the AltDMAPs algorithm and describe the prediction workflow based on AltDMAPs for identifying common variables between the spectra and the polymer size.

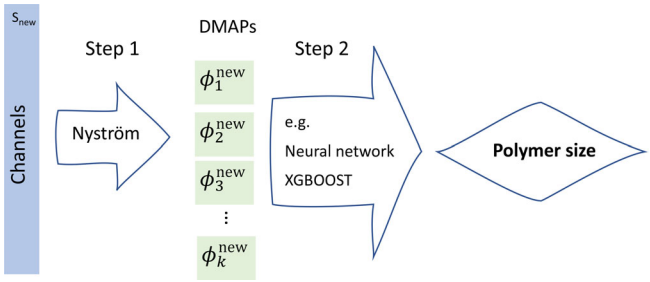


FIGURE 2 Schematic representation of the prediction strategy directly from the diffusion map (DMAP) coordinates that parsimoniously parameterize the spectra.

2.4.1 | Alternating DMAPs

AltDMAPs is a method, based on DMAPs, designed to parameterize the common variable between independent sets of observations, $\{s_i^{(1)}, s_i^{(2)}\}$. The method is introduced by Lederman et al.,[20] and the interested reader is referred to this paper and also to Reference 39 for a detailed presentation of the method. Here, only key points are presented for completeness.

At the core of the methodology lies the established DMAPs algorithm, which starts with a data set of N individual points (represented as d-dimensional real vectors, $x_1,\ldots,x_N$). A similarity measure between each pair of vectors (x_i,x_j), is computed using the standard Euclidean distance, based on which an affinity matrix is constructed.

A popular choice is the Gaussian kernel $w(i,j) = \exp\left[-\left(\frac{||x_i - x_j||}{\epsilon}\right)^2\right]$ where ϵ is a hyper-parameter that quantifies the kernel's bandwidth. To recover a parameterization insensitive to the sampling density, the normalization

$$\widetilde{\mathbf{W}} = \mathbf{P}^{-1}\mathbf{W}\mathbf{P}^{-1},$$

is performed, where $P_{ii} = \sum_{j=1}^{N} W_{ij}$ and W_{ij} the elements of the matrix $\mathbf{W}$. A second normalization applied on $\widetilde{\mathbf{W}}$,

$$\mathbf{K} = \mathbf{D}^{-1}\widetilde{\mathbf{W}}, \tag{2}$$

gives a $N \times N$ Markov matrix $\mathbf{K}$; here $\mathbf{D}$ is a diagonal matrix, collecting the row sums of the matrix $\widetilde{\mathbf{W}}$ eigenvectors ϕ_i.

Based to the DMAPs algorithm, the AltDMAPs algorithm defines two weighted kernel matrices, as defined in Equation (2): one that is based on the measurements from $s^{(1)}$ and the other based on the measurements from $s^{(2)}$, and constructs the alternating-diffusion operator as the product of the two normalized weight matrices:

$$\mathbf{K}^{\text{alt}} = \mathbf{K}^{(1)}\mathbf{K}^{(2)}.$$

It has been shown that the diffusion process defined by this operator is equivalent to one that would have been created from measurements of only the common variable.[20] This finding can be further explained in terms of alternating propagation steps using $K^{(1)}$ followed by $K^{(2)}$. Each propagation step can be considered a Markov chain that transitions to similar samples, as prescribed by the respective similarity matrix, first in the first sample set, followed by the second.

2.4.2 | Prediction workflow based on AltDMAPs

The proposed machine learning workflow consists of two *offline* learning steps, summarized in Figure 3, that take place once and create the tools for prediction, and three *online* application steps, shown schematically in Figure 4, that are implemented in line with the actual process for fast predictions.

The two offline steps are:

- Offline Step 1: Dimensionality reduction of the original collection of spectra, consisting of approximately 11,000 wavenumbers via

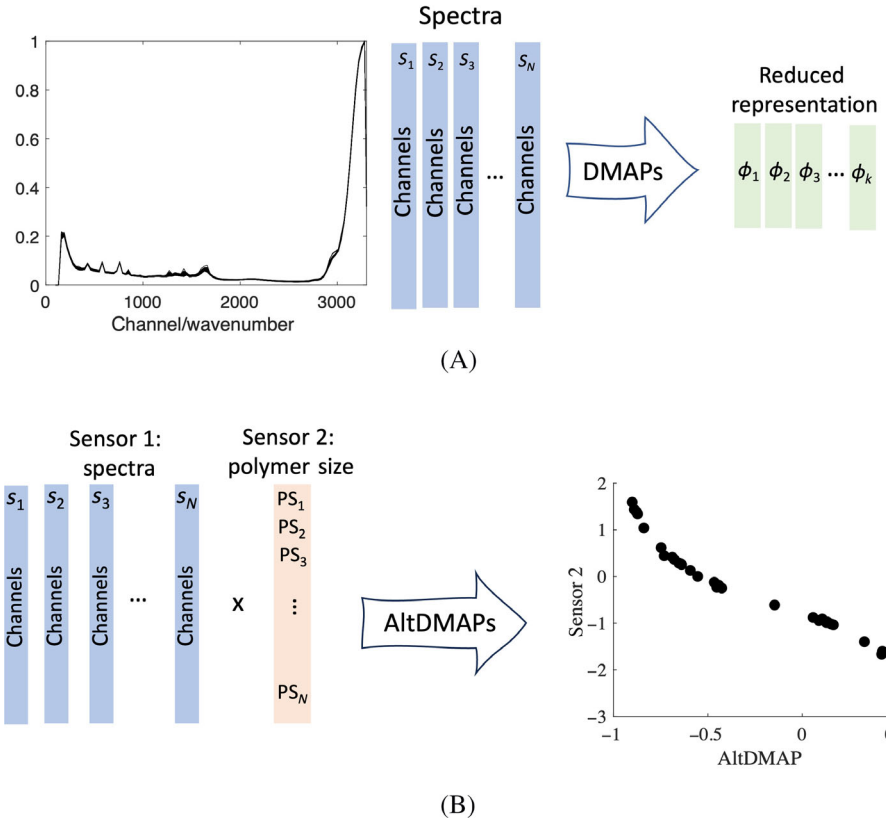


FIGURE 3 Schematic of the *offline* steps: (A) Step 1: Finding a reduced representation of the spectra with diffusion maps (DMAPs); (B) Step 2: Finding the common variable between spectra and polymer size with alternating diffusion maps (AltDMAPs).

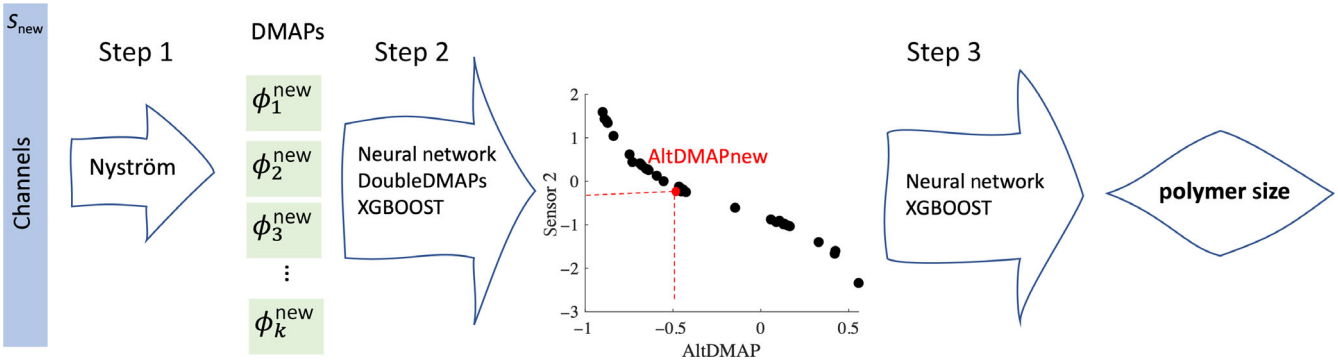


FIGURE 4 Schematic of the *online* application of size prediction: Given a measured (new) spectrum, the reduced description via diffusion maps (DMAPs) is determined (Step 1). Then, the common alternating diffusion map (AltDMAP) variable is inferred with neural networks (NNs), DoubleDMAPs, or XGBOOST methods (Step 2). Finally, the polymer size is predicted from the common variable, with NNs or XGBOOST (Step 3).

DMAPs. The goal is to reduce the number of variables to ideally < 10 and thus to re-state the high-dimensional data set in a low-dimensional coordinate system, parameterized by a small number of selected eigenvectors of the kernel matrix defined in Equation (2). The eigenvectors can be selected with the help of the local linear regression algorithm. In this article the eigenvectors are selected based on how accurately the high-dimensional variables are reconstructed, as quantified by an L^2-norm of the difference between predicted and actual spectra.

- Offline Step 2: Parameterizing the effect of polymer size on the spectra with the AltDMAPs algorithm. In this implementation of AltDMAPs, the spectra and the polymer size measurements are considered as the two independent sensor measurements ($s^{(1)}$ and $s^{(2)}$), respectively. The outcome of this algorithm represents the common variable(s) between the two modalities, that is, the polymer size.

Having established a reduced representation of the spectra and the common variable between the spectra and the size measurements, we proceed with the *online* prediction steps for a new spectrum. A schematic representation of the *online* workflow for size prediction from spectra is presented in Figure 4. To this end, three steps are proposed, designed to yield very fast (practically instantaneous) predictions, without further tuning of the model parameters, in an automated manner:

- Online Step 1: Compute the DMAP coordinates of a new spectrum using the Nyström extension.
- Online Step 2: Predict the common AltDMAPs variable(s) from the DMAP coordinates of the new spectrum:

$$\text{AltDMAP}_i = f_{\text{AltD}}(\text{DMAPs}_i), \tag{3}$$

i here is the ith new spectrum. The function f_{AltD} can be approximated by various methods, for example, a fully connected NN, DoubleDMAPS, or XGBOOST.

- Online Step 3: The polymer size, $D_{\text{H}}^{\text{predicted}}$ is predicted as a function of the AltDMAP coefficients as:

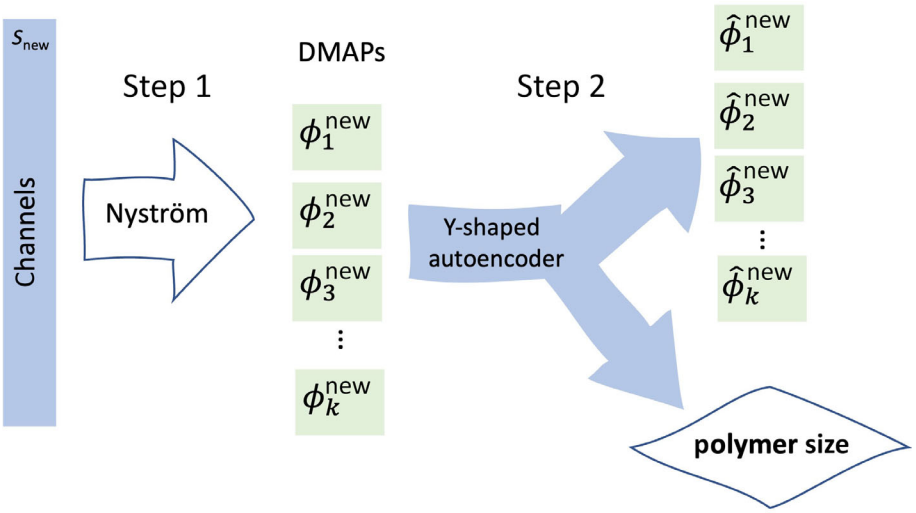
$$D_{\text{H,i}}^{\text{predicted}} = f_{\text{size}}(\text{AltDMAP}_i) \tag{4}$$

the function f_{size} can be approximated by a feed-forward NN or XGBOOST trained with AltDMAPs as input and polymer size as output.

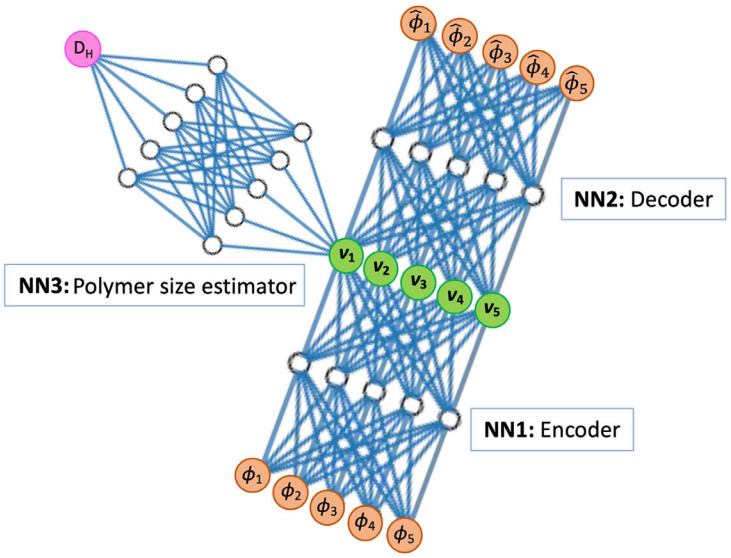
2.5 | Conformal autoencoders: Y-shaped architectures

Another alternative workflow to predict polymer sizes from Raman spectra involves an ensemble of concurrently trained NNs called Y-shaped conformal autoencoders. We use these Y-shaped conformal autoencoders to predict polymer sizes based on DMAP coordinates. The schematic representation of the workflow, including the Y-shaped autoencoder, is presented in Figure 5A. The Y-shaped autoencoder scheme, initially proposed in Reference 23, is summarized here as it was adjusted for the current application. More details on the implementation and training of this machine learning technology are presented in Reference 23. At the core of the scheme lies a regular autoencoder, that is, a NN where the inputs and outputs are the same, with the addition of an extra "sideways" NN component, as explained below. Overall, the Y-shaped scheme comprises three connected subnetworks (illustrated in Figure 5B):

- Encoder, NN1, which maps the DMAP coordinates, ϕ_i, to the autoencoder latent variables, ν_i:
  $(\phi_1,\phi_2,\phi_3,\phi_4,\phi_5) \mapsto (\nu_1,\nu_2,\nu_3,\nu_4,\nu_5)$
- Decoder, NN2, which can be thought of as the inverse transformation from the latent space of the autoencoder (ν_i) back to the DMAP coordinates ϕ_i:
  $(\nu_1,\nu_2,\nu_3,\nu_4,\nu_5) \mapsto (\hat{\phi}_1,\hat{\phi}_2,\hat{\phi}_3,\hat{\phi}_4,\hat{\phi}_5)$
- Polymer size estimator, NN3, which maps the right number of autoencoder latent variables, here one of them, ν_1, to the observed output data, here the polymer size:
  $(\nu_1) \mapsto D_{\text{H}}$



(A)



(B)

FIGURE 5 Schematic of the Y-shaped conformal autoencoder architecture: (A) Representation of the workflow; (B) Y-shaped autoencoder composition: NN1 is the encoder that maps the diffusion map (DMAP) coordinates to the latent variables of the autoencoder; NN2 is the inverse transformation, from the latent space back to the DMAP coordinates; NN3 maps one of the latent variables to the output of interest: polymer size.

The key feature is the loss function, consisting of several parts. Successful reconstruction of the input original parameters (the autoencoder part) constitutes the first part. Noteably, the reconstructed inputs, that is, the output of NN2, are not required further down in our analysis. Nevertheless, the accurate reconstruction of the inputs is important for the prediction step, because it ensures that the bottleneck variables, of which one is used for prediction, form a low-dimensional embedding of the data. This implies that the bottleneck variables are indeed a parsimonious representation of the original high-dimensional spectra. In theory, this autoencoder part of the NN architecture could also be used for dimensionality reduction of the original high-dimensional data by appropriately preselecting the size of the bottleneck layer. However, in the presented case study, the high dimension of the original data set (each spectrum contains approximately 11,000 wavenumbers) and the relatively limited number of data-points conclude the preferred strategy to first reduce the dimensionality with DMAPs and subsequently implement the NN with less variables.

In the following step, comes the ability of NN3, whose input is the single latent variable, here ν_1 is chosen, to reproduce the observed output, that is, the polymer size; this defines the polymer size as a function of single input, ν_1. To concurrently train the different NNs, an additional component to the loss function becomes necessary, which results from further imposing an orthogonality constraint on the conformal autoencoder's latent coordinates:

$$\langle d\nu_i, d\nu_j \rangle = 0, \ \forall i \neq j,$$

where $d\nu_i$ indicates the vector of partial derivatives of the latent coordinate ν_i with respect to of the input parameters ϕ_i and $\langle \cdot, \cdot \rangle$ indicates the inner product. This constraint is imposed using the automatic differentiation capabilities of the relevant code libraries and aims to

disentangle the combination of features that matters to the output from those combinations of features that do not affect it. To train the Y-shaped conformal autoencoder network, we follow a two-stage training procedure. First, we train the autoencoder networks (NN1 and NN2) to ensure sufficient reconstruction of the original input parameters. Next, we train NN3, which is restricted by the orthogonality constraint, to fit the polymer size data. Note that while training the network architecture, we continue to train the autoencoder weights to obtain optimal results.

3 | RESULTS

The results comprise the analysis of the size prediction from the benchmark methods, and the developed workflow, including predictions directly from DMAP coordinates via a NN or XGBOOST algorithm, implementation of AltDMAPs and of the Y-shaped autoencoder. The Raman spectra and size predictions from DLS measurements of microgel samples are available for transparency.[34] The codes implementing the different workflow steps can be found in the GitLab repository.[40]

3.1 | PLS regression of spectral intensities

The exhaustive evaluation of various combinations of pretreatment methods for Raman spectra as the basis for PLS regression is summarized in Table 2. Overall, the direct application of PLS regression to the spectral intensities results in poor prediction performances for any pretreatment method. The poor performance is indicated by the R^2 values significantly lower than 1 for the training and testing. Even R^2values below zero are encountered. Note that R^2 values below zero imply that the prediction would be more accurate using the mean value than the value predicted by the regression model. Comparing the performance of spectra in the FP and global spectral region yields that the prediction performance is independent of the spectral region used. For pretreatment involving MinMax normalization in combination with either linear fit or rubber band subtraction or solely linear fit or rubber band subtraction, the number of latent variables needed for the FP region is consistently higher than for the global region, although the global region comprises more prediction variables. In contrast, for spectra without pretreatment or with pretreatment involving SNV normalization, the number of latent variables needed for the FP region is lower than for the global region throughout.

In summary, the predictions using raw spectra with no pretreatment in the global region perform best in training and testing considering all performance metrices. Also, the prediction based on raw spectra in the FP region with no pretreatment shows a relatively good performance indicated by the second-best accumulated R^2 value (training and testing) of 1.520. Thus, the parity plots for these most promising configurations are shown in Figure 6 in comparison. Here, the gray circles represent the training data, and the red circles represent the test data. Over-fitting is precluded sufficiently, as the discrepancy between actual and predicted size is in the same range for the training and test data set. Furthermore, comparing the PLS results

TABLE 2 Performance of PLS regression on Raman spectra with different pretreatment methods.

		Training			Testing			Number of latent variables
Pretreatment	Spectral region	R^2	RMSE	MAPE	R^2	RMSE	MAPE	
Linear fit	FP	0.961	12.664	2.882	0.428	40.865	8.639	7
Linear fit	Global	0.284	54.449	13.079	0.573	35.288	8.419	2
MinMax, linear fit	FP	0.860	24.087	5.106	0.106	51.080	12.935	7
MinMax, linear fit	Global	0.416	49.157	11.369	0.696	29.795	7.389	3
MinMax, rubber band	FP	0.861	23.960	5.345	−0.206	59.322	15.700	6
MinMax, rubber band	Global	0.425	48.794	11.232	0.697	29.740	7.496	3
Raw	FP	0.790	29.463	6.750	0.730	28.038	7.887	5
Raw	Global	0.942	15.473	3.799	0.633	32.701	7.953	8
Rubber band	FP	0.993	5.445	1.359	−0.041	55.109	14.329	9
Rubber band	Global	0.509	45.073	10.800	0.662	31.401	7.614	3
SNV, linear fit	FP	0.181	58.212	14.004	0.616	33.455	8.286	1
SNV, linear fit	Global	0.937	16.184	3.944	0.329	44.230	9.777	6
SNV, rubber band	FP	0.175	58.431	13.992	0.620	33.300	8.336	1
SNV, rubber band	Global	0.937	16.173	3.942	0.342	43.821	9.708	6

Notes: The values for RMSE are given in nm and for MAPE in %. The intensity level of the gray shading visualizes the quality of the prediction model compared to the remaining models regarding each prediction performance metric: A lower gray shade corresponds to a better performance.
Abbreviations: FP, fingerprint; MAPE, mean absolute percentage error; PLS, partial least squares; RMSE, root mean squared error; SNV, standard normal variate.

based on the raw spectra in the FP region (Figure 6A) and the global region (Figure 6B) shows no significant improvement in the PLS performance caused by the spectral range. However, the limited size of the data set is probably a restrictive factor and with an extended data set the performance of the state-of-the-art methods might improve. Therefore, the findings regarding the application of the state-of-the-art methods to the employed data set can not be generalized.

3.2 | Regression of hard model parameters

In Table 3, we show the overview of different pretreatment methods in the IHM step. The fitted parameter values from the IHM are subsequently used in the PLS regression. We considered a high and medium fitting mode corresponding to more or less degrees of freedom for fitting the IHM evaluation model to the experimental spectra. Similarly to the direct regression on spectral intensities, we find that the overall performance of the predictions is unsatisfying. Again, we determine R^2 values significantly lower than 1 for the training and even below zero for testing in some cases. Overall, the medium fitting mode shows a poorer prediction performance than the high fitting mode. With respect to the latent variables, for pretreatment involving rubber band more latent variables are needed in high fitting mode than in medium fitting mode. In contrast to the expected outcome that the high fitting mode necessitates more latent variables, as we need to reduce a larger space of

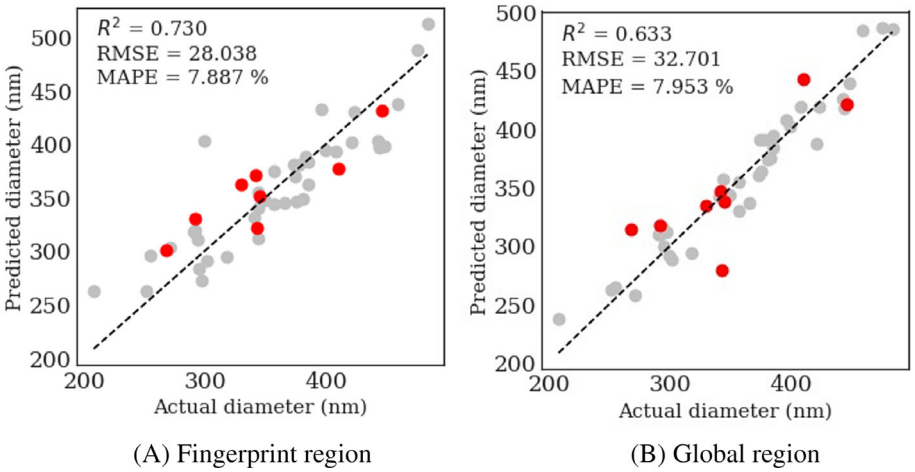


FIGURE 6 Parity plots of microgel size predictions via partial least squares (PLS) regression of raw spectra in (A) the fingerprint region and (B) the global region. Gray circles represent the training data, and red circles indicate the test data. (A) Fingerprint region; (B) Global region.

TABLE 3 Performance of PLS regression on IHM parameters from Raman spectra with different pretreatment methods.

		Training			Testing			Number of latent variables
Pretreatment	Fitting mode	R^2	RMSE	MAPE	R^2	RMSE	MAPE	
Linear fit	High	0.577	41.834	10.142	0.472	39.245	9.014	2
Linear fit	Medium	0.327	52.768	12.145	0.316	44.667	11.785	2
Raw	High	0.839	25.817	6.455	−0.257	60.565	14.040	6
Raw	Medium	0.327	52.780	12.145	0.329	44.237	11.637	2
Rubber band	High	0.646	38.255	8.739	0.335	44.043	10.582	6
Rubber band	Medium	0.538	43.745	9.917	−0.019	54.518	10.765	8
MinMax, linear fit	High	0.741	32.746	7.265	0.278	45.907	9.243	5
MinMax, linear fit	Medium	0.359	51.491	11.449	0.330	44.198	11.980	2
MinMax, rubber band	High	0.680	36.396	7.914	0.323	44.433	11.075	6
MinMax, rubber band	Medium	0.508	45.108	9.986	−0.153	57.985	11.675	8
SNV, linear fit	High	0.681	36.312	8.009	0.287	45.599	8.683	4
SNV, linear fit	Medium	0.366	51.220	11.413	0.283	45.723	12.380	2
SNV, rubber band	High	0.585	41.465	9.559	0.611	33.691	8.258	4
SNV, rubber band	Medium	0.511	44.967	10.015	−0.104	56.741	11.607	8

Notes: The values for RMSE are given in nm and for MAPE in %. The intensity level of the gray shading visualizes the quality of the prediction model compared to the remaining models regarding each prediction performance metric: A lower gray shade corresponds to a better performance.
Abbreviations: IHM, indirect hard modeling; MAPE, mean absolute percentage error; PLS, partial least squares; RMSE, root mean squared error; SNV, standard normal variate.

prediction variables, the remaining pretreatment methods not involving rubber band subtraction result in less latent variables for high than for medium fitting mode. Furthermore, no clear trend exists that one aspect of the pretreatment method performs better than the other.

In Figure 7, we present the results of the PLS regression based on IHM parameters from Raman spectra pretreated via SNV and rubber band baseline subtraction and high interaction during the fitting, as this configuration yields the relatively best performance according to Table 3. Also, we show the second best-performing configuration, namely regression based on spectra pretreated with MinMax normalization and a linear fit subtraction and fitted with high interaction. Similarly to the results from the direct application of PLS, the distribution of gray circles (training data) and the red circles test data) here shows that over-fitting is suppressed. In addition, the comparison of the PLS results based on the IHM parameter values from SNV plus linear fit pretreated spectra (Figure 7B) and spectra pretreated with MinMax normaliztion and a linear fit subtraction (Figure 7A) shows no notable improvement for the spectra pretreated via SNV in the PLS performance. In conclusion, the resulting prediction performance does not indicate reliable prediction accuracy. The reduction of the spectral range to IHM parameters constitutes a physically meaningful approach, but is shown to be insufficient for a limited size of data sets. For IHM in combination with PSL, the performance might improve with an extended data set and the findings can hence not be generalized but apply solely to the employed data sets.

3.3 | Prediction directly from DMAPs

Firstly, the DMAPs algorithm is implemented, and six DMAP coordinates, ϕ_1, ϕ_2, ϕ_3, ϕ_4, ϕ_5, and ϕ_6 are selected to parsimoniously represent the spectra that live in a high-dimensional ambient space (cf. Figure 2, Step 1). These are selected based on how accurately the original data can be reconstructed from those latent variables. Here the L^2 norm of the difference between the predicted and actual spectra, for the test set, is $3.34 \cdot 10^{-6}$. Subsequently, the polymer size is directly predicted from DMAPs coordinates (cf. Figure 2, Step 2).

To this end, a feed-forward NN is trained, having as inputs the DMAP coordinates that correspond to the training data and, as output, the polymer size. The hyper-parameters of the network here and in subsequent sections are fine-tuned using the *KerasTuner RandomSearch* hyper-parameter optimization framework. The results of direct

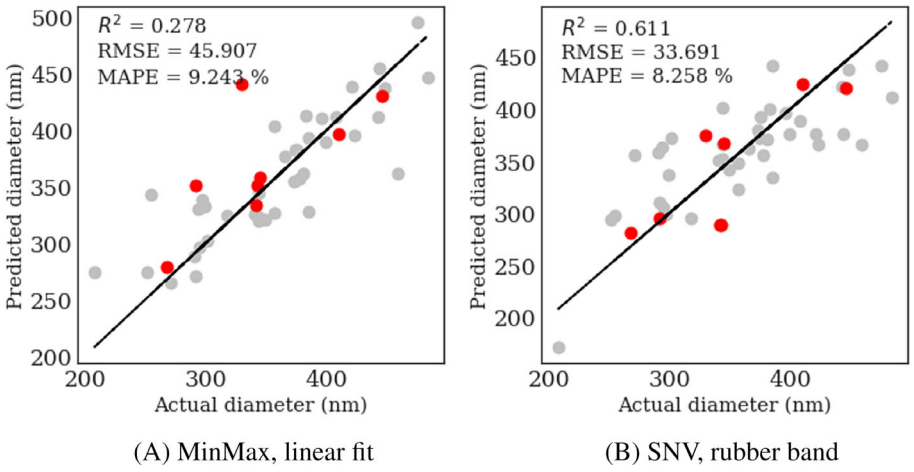


FIGURE 7 Parity plots of microgel size predictions via partial least squares (PLS) regression of indirect hard modeling (IHM) parameters from spectra fitted via high interaction. Gray circles represent the training data, and red circles indicate the test data. (A) MinMax, linear fit; (B) standard normal variate, rubber band.

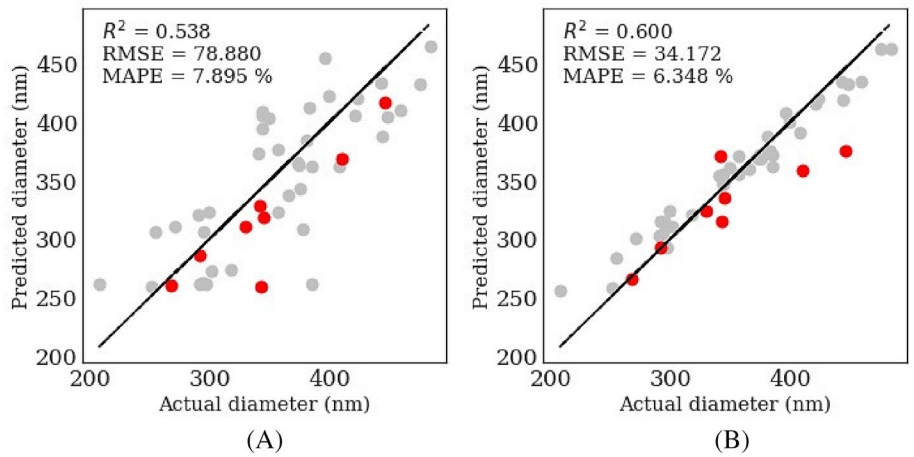


FIGURE 8 Parity plots of microgel size predictions from diffusion maps (DMAPs) (A) by a neural network and (B) using XGBOOST. Red points correspond to the test set, and gray points correspond to the training set values. The reported accuracy metrics (R^2, root mean squared error, and mean absolute percentage error) correspond to the test data.

predictions from DMAP coordinates are presented in Figure 8. It is possible to achieve regression, with MAPE = 7.895% and $R^2 = 0.538$, for the test set (cf. Figure 8A), and it is more challenging to identify the hyper-parameters that prevent over-fitting. Alternatively, the XGBOOST algorithm is used with randomized search on hyper-parameters of the XGBOOST estimator, using *RandomizedSearchCV* from the scikit-learn library in Python. The parameters of the estimator are optimized, here and in the implementations presented in subsequent sections by cross validated search over parameter settings. For efficiency, not all parameter values are tested, but rather a fixed number of parameter settings is sampled from the specified distributions. The performance of the method has similar accuracy to the NNs (cf. Figure 8B, MAPE = 6.348% and $R^2 = 0.600$).

3.4 | Prediction using alternating DMAPs

The AltDMAPs algorithm is implemented to find common variables, where the collection of spectra is considered as the s^1 (Sensor 1) measurements and the corresponding polymer sizes as s^2 (Sensor 2). Two significant common variables at AltDMAPs index 2 and 6 are found by implementing the local linear regression

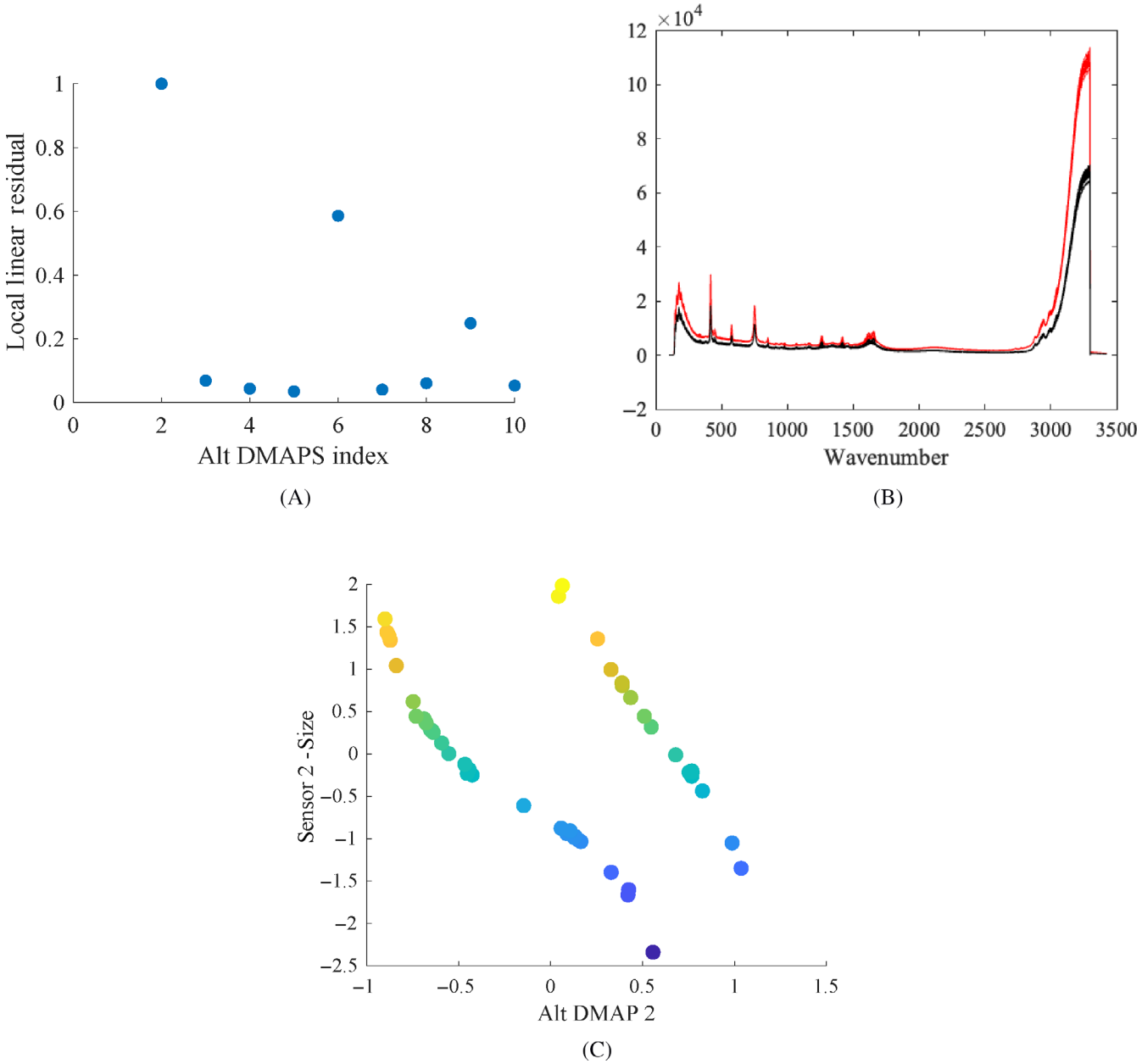


FIGURE 9 (A) Local linear regression indicates two significant common variables exist; (B) Plot of the raw spectra included in the data set; two clusters appear with respect to the spectral intensities, colored in red and black; (C) The (normalized) polymer size is plotted over the one common variable (AltDMAP) for both clusters in (B) illustrating the dependence of size on the latent variables (one-to-one within each individual cluster). To be able to predict the size from any spectrum, irrespective of the the cluster it belongs to, more than one AltDMAP is required.

algorithm, indicated by a high value of the local linear residual in Figure 9A.

The common variables parameterize the spectra variability attributed to the polymer size. It is worth noting that two clusters appear in the original data set of raw spectra concerning the spectral intensities (cf. Figure 9B). The common variable AltDMAP 2 is one-to-one with the size in each cluster separately, as visually demonstrated, in Figure 9C. Overall, for both clusters, the size is not a function of a single AltDMAP, and the AltDMAP variable with index 6 appears to be correlated to the polymer size, to some extent, since the local linear residual value (cf. Figure 9A) is nonnegligible.

Identifying a reduced description of the spectra and common variables between the spectra and the polymer size enables predicting the polymer size corresponding to a new spectrum by following the three *online* steps in the workflow (cf. Figure 4). As part of Online Step 1, the DMAP coordinates of a new set of spectra are determined by implementing the Nyström extension. The DMAP coordinates are accurately predicted by the Nyström extension, achieving a mean squared error, MSE$= 1.58 \times 10^{-7}$. Subsequently, in Online Step 2, the common variables, AltDMAPs, are computed as a function of the respective DMAP coordinates of the test set computed in the previous step. Two alternatives are implemented: Double DMAPs, with a prediction error of MSE$= 0.0248$, and the Extreme Gradient Boosting (XGBOOST) algorithm, which is slightly more accurate here, achieving MSE$= 0.0162$. In practice, including more AltDMAP coordinates for a new spectrum is necessary, here we include the first six AltDMAPs. The final step of the workflow involves predicting the polymer size given the common variables, AltDMAPs. The prediction results using a feed-forward NN and XGBOOST are shown in Figure 10. Here, the input is the AltDMAP coordinates, and the output is the polymer size. Both methods yield similar results regarding MAPE, RMSE, and R^2 for the test set.

The performance of the overall *online* workflow is shown in Figure 11. Specifically, four combinations of methods are implemented and reported: (i) the size is predicted by a NN from AltDMAPs predicted by Double DMAPs, leads to an $R^2 = 0.584$ and MAPE$= 8.101\%$ for the test set; (ii) the size is predicted by a NN from AltDMAPs predicted by XGBOOST, leads to an $R^2 = 0.330$ and MAPE$= 10.058\%$ for the test set; (iii) the size is predicted by the XGBOOST algorithm from AltDMAPs predicted by Double DMAPs, leads to an $R^2 = 0.345$ and MAPE$= 10.772\%$ for the test set; (iv) the size is predicted by the XGBOOST algorithm from AltDMAPs predicted by XGBOOST, leads to an $R^2 = 0.525$ and MAPE $= 8.481\%$ for the test set. As shown, the prediction of the polymer particle diameter from the *predicted* AltDMAPs is less accurate than the prediction of the *actual* AltDMAPs, which is attributed to the prediction error of the AltDMAPs from the DMAPs. Overall, case (i) is the best-performing one for the available data set for R^2 and MAPE. Considering the $\%$-error in size prediction, the combination of methods of case (i) is more advantageous regarding the maximum absolute values of the $\%$-error and error distribution.

Overall, the predictive accuracy directly from DMAPs is slightly better than when implementing the AltDMAP methodology, although they are characterized by the same order of magnitude. Given the size of the available data set and the difference between the performance metrics (R^2, MAPE), it is unclear whether this small difference is meaningful enough to persist given a different or larger data set.

3.5 | Y-shaped autoencoder

The conformal autoencoder architecture (Figure 5B) is trained with the same training set as the previous methods: The DMAP coordinates $(\phi_1,...,\phi_6)$ corresponding to the training set are the input to the encoder network NN1. These values are also the target values for the decoder network NN2. We set six latent variables in the bottleneck layer, and we require that the polymer size be defined as a function of ν_1, by concurrently training the neural network NN3.

The performance of this method is superior to both the AltDMAPs workflow and the direct prediction from DMAP coordinates with a R^2 value of 0.951 and MAPE value of 2.93$\%$ for the test set (cf. Figure 12). We believe that the reason for the enhanced performance lies in the fact that the Y-shaped architecture not only finds the latent variables of the data set, as do DMAPs, but with the second NN, specific properties are explicitly imposed on one of them: The

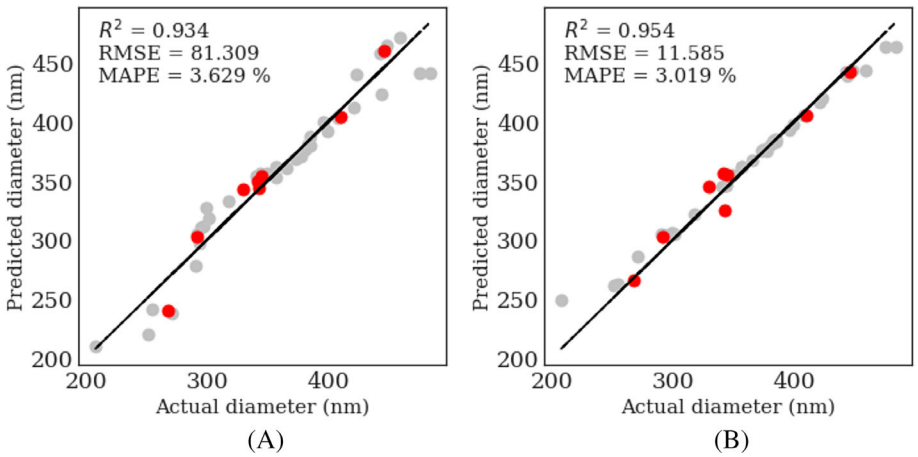


FIGURE 10 Parity plots of microgel size predictions from alternate diffusion maps (AltDMAPs), with (A) neural network, and (B) XGBOOST algorithm. The gray points correspond to the training set data, and the red points correspond to the test data. The reported accuracy metrics correspond to the test data.

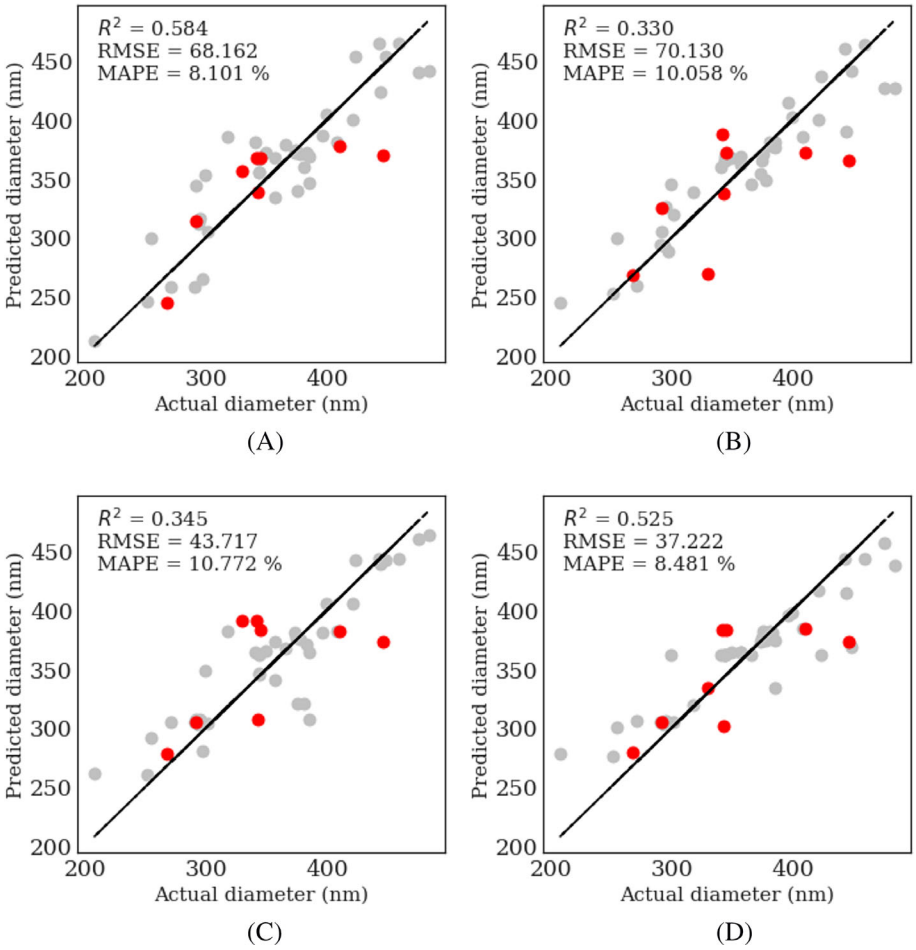


FIGURE 11 Parity plots of microgel diameter predictions via (A) a neural network (NN) from DoubleDMAPs-predicted alternate diffusion maps (AltDMAPs), (B) a NN from XGBOOST-predicted AltDMAPs, (C) XGBOOST from Double DMAPs-predicted AltDMAPs, and (D) XGBOOST from XGBOOST-predicted AltDMAPs. The gray points correspond to the training set data, and the red points correspond to the test data. The reported accuracy metrics (R^2, root mean squared error, and mean absolute percentage error) correspond to the test data.

polymer size must be a function of one latent variable orthogonal to all other latent variables. The latter property implies that only this one latent variable correlates to the particle diameter, which is a subtle but essential difference from the AltDMAPs approach (two common variables and even more required for accurate online prediction). This hypothesis merits further investigation, which lies beyond the scope of this article.

3.6 | Comparison of presented methods

In Table 4, we compare the results from the state-of-the-art to the proposed methods. We cluster the methods in three categories: state-of-the-art, alternating DMAPs, and prediction directly from DMAP coordinates. The results using the respective methods are described in detail in previous sections. Here, we present the best-forming configuration of pretreatment and spectral range for the direct PLS regression on Raman spectra and PLS regression on IHM parameters.

For Raman spectra studied here, the Y-shaped autoencoder outperforms the other considered methods of this work indicated by the only R^2 value considerably close to 1. Also, the low RMSE values of the Y-shaped autoencoder, being approximately a third of the values of the compared methods, indicate improved performance. Lastly, the MAPE value of the autoencoder-based method ranges at 2.930 %, which approximates the precision range of around 2 % expected from DLS ,[41] the established size measurement device. Thus, even though the evaluation is based on a data set with limited size, for the first time, a purely data-driven evaluation method based on Raman spectra advances the accuracy capability of the established size determination device. The next best-performing methods in the current study include the prediction directly from DMAP coordinates via the XGBOOST algorithm, the PLS regression from IHM parameters, and the direct PLS regression. Each method reaches a similar prediction performance of R^2 values of 0.600, 0.636, and 0.633, respectively. Also, their RMSE values are close to each other, ranging at 34.172, 34.840, and 32.700 nm. Finally, the MAPE value of the XGBOOST involving method is with 6.348 % slightly better than the PLS regression based on IHM parameters with 8.573 % and the direct PLS regression with 7.953 %.

All methods involving AltDMAPs only achieve an inferior prediction accuracy with R^2 values between 0.330 and 0.584. Within the AltDMAPs involving methods, there is no clear trend suggesting which combination of AltDMAPs prediction with size prediction enhances the performance.

On that note, it is worth looking further into the collection of methods that rely on manifold learning (DMAPs, AltDMAPs, and Y-shaped autoencoder), specifically on the number of latent variables required (cf. the last column of Table 4) and discuss their unique characteristics. The first observation is that DMAPs are able to meaningfully

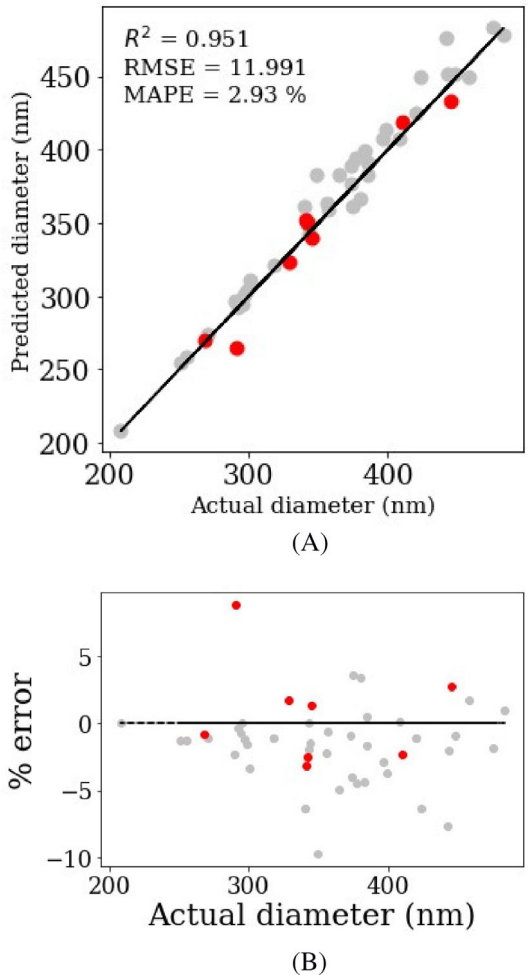


FIGURE 12 Prediction from Y-shaped autoencoder architecture: (A) Actual vs. predicted polymer size, and (B) % error for each one of the data points in the test set. Red points correspond to the test set, and gray points correspond to the training set values. The reported accuracy metrics (R^2, root mean squared error, and mean absolute percentage error) correspond to the test data.

embed the high-dimensional data using six coordinates. Even though a parsimonious embedding of the data is achieved, which is key in making machine learning tasks computationally tractable, it is not a quantitatively accurate predictor. Here, the ability of the AltDMAPs to identify the common variables between the measured observation (the size) and the low-dimensional description of the input (Raman spectra) by the DMAP coordinates is beneficial. Consequently, we further reduce the variables that are meaningful for the specific task of predicting polymer size from six to two latent variables, on the data.

It becomes clear that "designer" latent spaces, that is, ones with specific desired characteristics, become useful in disentangling the data features and creating latent variables that specifically map to the observable quantity of choice, here polymer size. This finding motivates the proposed conformal autoencoder architecture. The autoencoder network identifies a single latent variable, which not only maps to the size but is independent of other data features by design (imposed orthogonality condition in the loss function of the NN). This trait of the conformal autoencoder architecture enables our accurate prediction of polymer size here.

4 | CONCLUSIONS AND FUTURE WORK

This contribution considers the important open problem of predicting particle sizes online. We propose the prediction of monodisperse polymer sizes from Raman spectra via a data-driven approach. We apply the approach to our open-access data set of continuous microgel synthesis and demonstrate its capability. Further, we compare our approach against two state-of-the-art benchmark methods to underline the excellent prediction performance of our nonlinear approach.

All proposed nonlinear approaches rely on dimensionality reduction via DMAP coordinates. We find that the machine learning workflow combining the data reduction capability of the DMAPs algorithm with the recent advances of Y-shaped

TABLE 4 Testing performance of all considered prediction methods within this article.

Cluster	Method	R^2	RMSE	MAPE	Number of latent variables
State-of-the-art	Best configuration based on PLS regression directly to Raman spectra	0.633	32.701	7.953	8
	Best configuration based on PLS regression on IHM parameters	0.611	33.691	8.258	4
Alternating diffusion maps	Neural network based on DoubleDMAPs-predicted AltDMAPs	0.584	68.162	8.101	2 (+4)
	Neural network based on XGBOOST-predicted AltDMAPs	0.330	70.130	10.058	2 (+4)
	XGBOOST based on DoubleDMAPs-predicted AltDMAPs	0.345	43.717	10.772	2 (+4)
	XGBOOST based on XGBOOST-predicted AltDMAPs	0.525	37.222	8.481	2 (+4)
Prediction directly from DMAP coordinates	Neural network	0.538	78.880	7.895	6
	XGBOOST	0.600	34.172	6.348	6
	Y-shaped autoencoder	0.951	11.991	2.930	1

Note: The values for RMSE are given in nm and for MAPE in %.
Abbreviations: AltDMAP, alternate diffusion map; DMAP, diffusion map; IHM, indirect hard modeling; MAPE, mean absolute percentage error; PLS, partial least squares; RMSE, root mean squared error.

autoencoder outperforms all alternatively considered workflows significantly. The remaining proposed methods with altering configurations of involved algorithms accomplish prediction performances comparable to state-of-the-art linear methods. In contrast, the Y-shaped autoencoder approach enables drastically better prediction accuracy, similar to the established size measurement methods such as DLS. Therefore, we derive an algorithmic workflow for prediction of particle sizes from inline measurements that is competitive with offline analytical methods.

Predicting polymer sizes directly from inline Raman spectra taken from untreated samples, as labor-intensive DLS processing is circumvented and online reaction monitoring for closed-loop control is enabled. Furthermore, with the proposed method the spectra are not manipulated by spectral pretreatment, thus, no expert knowledge is necessary. In contrast to established machine learning workflows, the proposed algorithm exhibits high efficiency, as the algorithmic filtering enables a proficient prediction performance based on a data set of limited size. This aspect makes the proposed workflow especially relevant, as requiring experimental data is laborious. In addition, the workflow typically relies on less than ten coordinates in the reduced component space. Here, only the first six DMAP coordinates enable us to use the entire wavelength spectrum without exclusions, which would otherwise require problem-specific intuition.

Future works include the application of the proposed workflow to predict polymer concentrations and sizes simultaneously to highlight the application of our proposed readily available analysis tool. The simultaneous prediction allows a more comprehensive characterization via a single inline process analytical tool. Furthermore, subsequent investigations include the open challenge of predicting size distributions from inline Raman spectroscopy. In addition, future considerations involve extending the workflow to other systems beyond the presented application, which can involve crystallization processes, among others.

AUTHOR CONTRIBUTIONS

Eleni D. Koronaki: formal analysis (equal); investigation (lead); methodology (lead); software (lead); visualization (equal); writing – original draft (equal). **Luise F. Kaven:** data curation (lead); formal analysis (supporting); investigation (equal); methodology (equal); resources (lead); software (supporting); writing – original draft (equal); writing – review and editing (supporting). **Johannes M. M. Faust:** conceptualization (supporting); investigation (supporting); methodology (supporting); writing – review and editing (supporting). **Ioannis G. Kevrekidis:** conceptualization (supporting); investigation (supporting); methodology (supporting); resources (supporting); supervision (equal); writing – review and editing (equal). **Alexander Mitsos:** conceptualization (lead); funding acquisition (lead); investigation (supporting); methodology (supporting); resources (lead); supervision (equal); writing – original draft (equal); writing – review and editing (equal).

Eleni D. Koronaki: developed, implemented, and applied the DM, ADM, and Y-shaped autoencoder frameworks, developed the heuristic preprocessing, wrote initial draft; **Luise F. Kaven**: performed experimental design, preprocessed data, implemented and applied PLS and hybrid PLS+IHM method, wrote initial draft; **Johannes M. M. Faust**: performed preliminary numerical experiments with DM, guided analysis, edited manuscript; **Ioannis G. Kevrekidis**: guided the DM, ADM and Y-shaped autoencoder frameworks, edited manuscript; **Alexander Mitsos**: conceived the idea, initiated project, supervised Luise F. Kaven and Johannes M. M. Faust, wrote initial draft.

ACKNOWLEDGMENTS

This work was performed as a part of project B4 of the CRC 985 "Functional Microgels and Microgel Systems" funded by Deutsche Forschungsgemeinschaft (DFG). EDK was funded by the Luxembourg National Research Fund (FNR), grant reference 16758846. For the purpose of open access, the authors have applied a Creative Commons Attribution 4.0 International (CC BY 4.0) license to any Author Accepted Manuscript version arising from this submission. The work of YGK was partially supported by the US Air Force Office of Scientific Research. The authors thank Jörn Viell for scientific discussions and feedback on the manuscript and Andrij Pich for useful discussions on future tasks and impact of this work. Open Access funding enabled and organized by Projekt DEAL.

DATA AVAILABILITY STATEMENT

All data and code is provided online. We have provided an .xls with the data for each figure. An invention disclosure is submitted and a patent application is planed for the method. The code underlying this contribution is published in a GitLab repository .[40] The underlying data set is published via RWTH publications ,[34] including Raman spectra in untreated and pretreated form and the according DLS size predictions. Data from the article's figures are tabulated in Appendix S1.

ORCID

Alexander Mitsos https://orcid.org/0000-0003-0335-6566

REFERENCES

1. Chew W, Sharratt P. Trends in process analytical technology. *Anal Methods*. 2010;2(10):1412.
2. Beer T, Burggraeve A, Fonteyne M, Saerens L, Remon J, Vervaet C. Near infrared and Raman spectroscopy for the in-process monitoring of pharmaceutical production processes. *Int J Pharm*. 2011;417(1-2):32-47.
3. Pomerantsev A, Rodionova O. Process analytical technology: a critical view of the chemometricians. *J Chemometr*. 2012;26(6):299-310.
4. Simon L, Pataki H, Marosi G, et al. Assessment of recent process analytical technology (PAT) trends: a multiauthor review. *Organ Process Res Dev*. 2015;19(1):3-62.
5. Beer A. Bestimmung der absorption des rothen Lichts in farbigen Flüssigkeiten. *Annal Phys Chem*. 1852;162(5):78-88.
6. Marini F, Bucci R, Magrì A, Magrì A. Artificial neural networks in chemometrics: history, examples and perspectives. *Microchem J*. 2008;88(2):178-185.
7. Garrido M, Rius F, Larrechi M. Multivariate curve resolution-alternating least squares (MCR-ALS) applied to spectroscopic data from monitoring chemical reactions processes. *Anal Bioanal Chem*. 2008;390(8):2059-2066.
8. Alsmeyer F, Koß HJ, Marquardt W. Indirect spectral hard modeling for the analysis of reactive and interacting mixtures. *Appl Spectrosc*. 2004;58(8):975-985.

9. Keidel R, Ghavami A, Lugo D, et al. Time-resolved structural evolution during the collapse of responsive hydrogels: the microgel-to-particle transition. *Sci Adv*. 2018;4(4):eaao7086.
10. Bohren CF, Huffman DR. *Absorption and Scattering of Light by Small Particles*. Wiley; 1998.
11. Reis M, Araújo P, Sayer C, Giudici R. Evidences of correlation between polymer particle size and Raman scattering. *Polymer*. 2003; 44(20):6123-6128.
12. van den Brink M, Pepers M, van Herk A. Raman spectroscopy of polymer latexes. *J Raman Spectrosc*. 2002;33(4):264-272.
13. Ito K, Kato T, Ona T. Non-destructive method for the quantification of the average particle diameter of latex as water-based emulsions by near-infrared Fourier transform Raman spectroscopy. *J Raman Spectrosc*. 2002;33(6):466-470.
14. Houben C, Nurumbetov G, Haddleton D, Lapkin A. Feasibility of the simultaneous determination of monomer concentrations and particle size in emulsion polymerization using in situ Raman spectroscopy. *Ind Eng Chem Res*. 2015;54(51):12867-12876.
15. Ambrogi P, Colmán M, Giudici R. Miniemulsion polymerization monitoring using off-line Raman spectroscopy and in-line NIR spectroscopy. *Macromol React Eng*. 2017;11(4):1600013.
16. Meyer-Kirschner J, Mitsos A, Viell J. Polymer particle sizing from Raman spectra by regression of hard model parameters. *J Raman Spectrosc*. 2018;49(8):1402-1411.
17. Coifman R, Lafon S. Diffusion maps. *Appl Comput Harmonic Anal*. 2006;21(1):5-30.
18. Nadler B, Lafon S, Coifman R, Kevrekidis I. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Appl Comput Harmonic Anal*. 2006;21(1):113-127.
19. Coifman R, Kevrekidis I, Lafon S, Maggioni M, Nadler B. Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems. *Multiscale Model Simul*. 2008;7(2):842-864.
20. Lederman R, Talmon R. Learning the geometry of common latent variables using alternating-diffusion. *Appl Comput Harmonic Anal*. 2018; 44(3):509-536.
21. Katz O, Talmon R, Lo YL, Wu HT. Alternating diffusion maps for multimodal data fusion. *Inform Fusion*. 2019;45:346-360.
22. Dietrich F, Yair O, Mulayoff R, Talmon R, Kevrekidis I. Spectral discovery of jointly smooth features for multimodal data. *SIAM J Math Data Sci*. 2022;4(1):410-430.
23. Evangelou N, Wichrowski N, Kevrekidis G, et al. On the parameter combinations that matter and on those that do not: data-driven studies of parameter (non)identifiability. *PNAS Nexus*. 2022;1:154.
24. Chiavazzo E, Gear C, Dsilva C, Rabin N, Kevrekidis I. Reduced models in chemical kinetics via nonlinear data-mining. *Processes*. 2014;2(1): 112-140.
25. Koronaki E, Evangelou N, Psarellis Y, Boudouvis A, Kevrekidis I. From partial data to out-of-sample parameter and observation estimation with diffusion maps and geometric harmonics. *Comput Chem Eng*. 2023;178:108357.
26. Nyström E. *Über Die Praktische Auflösung von Linearen Integralgleichungen Mit Anwendungen Auf Randwertaufgaben der Potentialtheorie*. Akademische Buchhandlung; 1929.
27. Fowlkes C, Belongie S, Malik J. Efficient spatiotemporal grouping using the Nystrom method. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*. Vol 1. IEEE; 2001:I-II.
28. Evangelou N, Dietrich F, Chiavazzo E, Lehmberg D, Meila M, Kevrekidis I. Double diffusion maps and their latent harmonics for scientific computations in latent space. *arXiv preprint arXiv: 2204.12536*. 2022.
29. Kaven L, Schweidtmann A, Keil J, Israel J, Wolter N, Mitsos A. Data-driven product-process optimization of n-isopropylacrylamide microgel synthesis in flow. *arXiv preprint arXiv:2308.16724*. 2023.
30. Kaven L, Wolff H, Wille L, Wessling M, Mitsos A, Viell J. In-line monitoring of microgel synthesis: flow versus batch reactor. *Organ Process Res Dev*. 2021;25(9):2039-2051.
31. Kather M, Ritter F, Pich A. Surfactant-free synthesis of extremely small stimuli-responsive colloidal gels using a confined impinging jet reactor. *Chem Eng J*. 2018;344:375-379.
32. Wolff H, Kather M, Breisig H, Richtering W, Pich A, Wessling M. From batch to continuous precipitation polymerization of thermoresponsive microgels. *ACS Appl Mater Interfaces*. 2018;10(29):24799-24806.
33. Janssen F, Kather M, Ksiazkiewicz A, Pich A, Mitsos A. Synthesis of poly(N-vinylcaprolactam)-based microgels by precipitation polymerization: pseudo-bulk model for particle growth and size distribution. *ACS Omega*. 2019;4(9):13795-13807.
34. Kaven L, Mitsos A. Dataset to: nonlinear manifold learning determines microgel size from Raman spectroscopy. 2023. doi:10.18154/RWTH-2023-05604
35. Kaven L, Wolff H, Wille L, Wessling M, Mitsos A, Viell J. Dataset to: in-line monitoring of microgel synthesis: flow versus batch reactor. 2021. doi:10.18154/RWTH-2021-09666
36. Psarellis Y, Lee S, Bhattacharjee T, Datta S, Bello-Rivas J, Kevrekidis I. Data-driven discovery of chemotactic migration of bacteria via machine learning. *arXiv preprint arXiv:2208.11853*. 2022.
37. Coifman RR, Lafon S. Geometric harmonics: a novel tool for multiscale out-of-sample extension of empirical functions. *Appl Comput Harmonic Anal*. 2006;21(1):31-52.
38. Dsilva C, Talmon R, Coifman R, Kevrekidis I. Parsimonious representation of nonlinear dynamical systems through manifold learning: a chemotaxis case study. *Appl Comput Harmonic Anal*. 2018; 44(3):759-773.
39. Talmon R, Wu HT. Latent common manifold learning with alternating diffusion: analysis and applications. *Appl Comput Harmonic Anal*. 2019;47(3):848-892.
40. Koronaki E, Kaven L, Faust J, Kevrekidis I, Mitsos A. Code to: nonlinear manifold learning determines microgel size from Raman spectroscopy. 2023 https://gitlab.com/eleni.koronaki/mlforpolymersizeraman.git
41. MI Ltd. Zetasizer nano technical note MRK728-01 - the accuracy and precision expected from dynamic light scattering measurements. 2006 https://kdsi.ru/upload/iblock/357/be4ecfa4fcce215f870e4acb8eb229d6.pdf

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Koronaki ED, Kaven LF, Faust JMM, Kevrekidis IG, Mitsos A. Nonlinear manifold learning determines microgel size from Raman spectroscopy. *AIChE J*. 2024;70(10):e18494. doi:10.1002/aic.18494