Received: 14 December 2021 Revised: 27 January 2022 Accepted article published: 3 February 2022 Published online in Wiley Online Library: 24 February 2022

(wileyonlinelibrary.com) DOI 10.1002/pi.6378

A machine learning approach for the design of hyperbranched polymeric dispersing agents based on aliphatic polyesters for radiation-curable inks

Danny EP Vanpoucke,[a] Marie AF Delgove,[a] Jules Stouten,[a] Jurrie Noordijk,[a] Nils De Vos,[b] Kamiel Matthysen,[b] Geert GP Deroover,[b] Siamak Mehrkanoon[c] and Katrien V Bernaerts[a]*

Abstract

Polymeric dispersing agents were prepared from aliphatic polyesters consisting of δ-undecalactone (UDL) and β,δ-trimethyl-ε-caprolactones (TMCL) as biobased monomers, which were polymerized in bulk via organocatalysts. Graft copolymers were obtained by coupling of the polyesters to poly(ethylene imine) (PEI) in the bulk without using solvents. Various parameters that influence the performance of the dispersing agents in pigment-based UV-curable matrices were investigated: chemistry of the polyester (UDL or TMCL), polyester/PEI weight ratio, molecular weight of the polyesters and of PEI. The performance of the dispersing agents was modelled using machine learning in order to increase the efficiency of the dispersant design. The resulting models were presented as analytical models for the individual polyesters and the synthesis conditions for optimally performing dispersing agents were indicated as a preference for high-molecular-weight polyesters and a polyester-dependent maximum polyester/PEI weight ratio.

Supporting information may be found in the online version of this article.

Keywords: dispersant; polyester; poly(ethylene imine); structure–property relationships; machine learning

INTRODUCTION

Inks and coatings are typically formulations containing a colorant (e.g. pigment), a medium, a (polymeric) dispersing agent and some additives. Commercially available pigments are insoluble in the medium and are supplied as agglomerates. Especially for pigment-based inks, the application under consideration in this article, small particle sizes (<100–150 nm) are required to ensure good optical properties like gamut, gloss and covering power as well as good processability (no obstruction of the nozzles). In order to break the agglomerates into smaller (primary) particles, dispersion processes (e.g. milling) are required. A dispersion refers to a two-phase system consisting of small insoluble pigment particles homogeneously distributed in a liquid medium. In order to avoid re-agglomeration of the small pigment particles and to ensure long-term storage stability of the formulations (i.e. good colloidal stability), polymeric dispersants are added. Such dispersing agents typically contain a matrixophilic chain that should have good solubility in and compatibility with the ink medium to ensure steric stabilization, and a pigmentophilic chain that is anchored to the pigment, which is the dispersed phase (Fig. 1, left).[1,2] This type of dispersing agent has been reported in non-aqueous inks[3] as well as in aqueous coatings and inkjet inks.[4,5] In this work, radiation-curable monomers are the medium of the ink and they cure rapidly upon exposure to a radiation source. Compared to solventborne inks this avoids the release of environmentally unfriendly volatile organic compounds. Moreover, less

* Correspondence to: KV Bernaerts, Maastricht University, Aachen-Maastricht Institute for Biobased Materials (AMIBM), Brightlands Chemelot campus, Urmonderbaan 22, 6167 RD Geleen, The Netherlands. E-mail: katrien.bernaerts@maastrichtuniversity.nl

This article was published online on 24 February 2022. Subsequently, an error was found in supporting information, and the correction was published on 4 March 2022.

a Aachen-Maastricht Institute for Biobased Materials (AMIBM), Brightlands Chemelot campus, Maastricht University, Geleen, The Netherlands

b ChemStream, Edegem, Belgium

c Department of Data Science and Knowledge Engineering, Maastricht University, Maastricht, The Netherlands

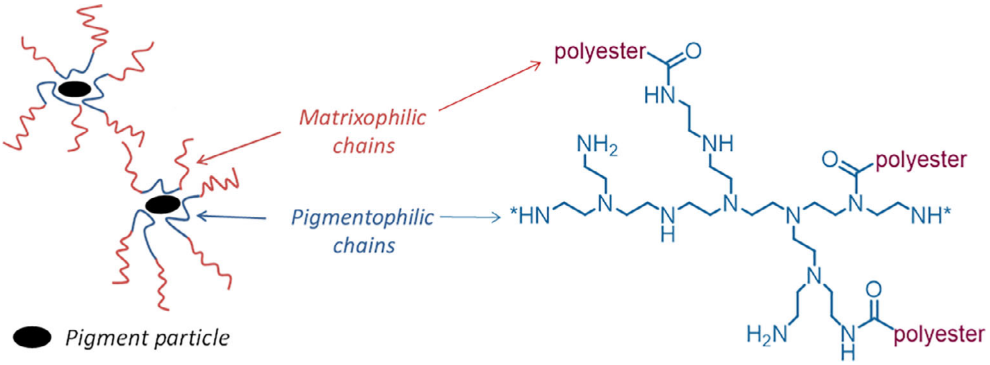

Figure 1. Schematic structure of hyperbranched dispersing agents.

energy is consumed compared to waterborne technology, which requires a lot of heat to evaporate the carrier after printing.

For non-aqueous pigmented inkjet inks, hyperbranched dispersing agents comprising poly(alkylene imine) chains as pigmentophilic chains and aliphatic polyesters as matrixophilic chains have been developed (Fig. 1, right).[3,6,7]

The polyesters used in state-of-the art systems are generally based on poly(ε-caprolactone). However, they suffer from a poor solubility in the ink medium and they tend to crystallize under use conditions. Copolymerization with other lactones like δ-valerolactone allows a reduction in the crystallinity and improvement in the solubility.[3,7,8] In the work reported here, we designed branched homopolyesters of δ-undecalactone (UDL) and β,δ-trimethyl-ε-caprolactones (TMCL), which avoids the need for copolymerization and results in fully amorphous polymers that are liquid at room temperature with T_g (polyTMCL) = −63 °C[9] and T_g (polyUDL) = −51 °C.[8] Moreover, extra attention was paid to sustainable synthesis of the polyesters and the hyperbranched dispersants. UDL is a biobased monomer that is synthesized from fatty acids, similarly to many δ-substituted valerolactones.[10] In the past, we developed a process for the enzymatic synthesis of TMCL monomers (a regio-isomeric mixture) using Baeyer–Villiger oxidation of 3,3,5-trimethylcyclohexanone.[11–13] The substrate can be prepared by the hydrogenation of isophorone which is obtained from the self-condensation of acetone, potentially sourced from renewables via the acetone–butanol–ethanol fermentation of lignocellulosic feedstocks. The enzymatic oxidation was shown to be more environmentally friendly compared to the chemical route in the case of the reaction performed at a laboratory scale with recycling of solvents and enzyme.[14] For the ring-opening polymerization (ROP) of lactones in bulk, organocatalysts[9] were used instead of metal-based catalysts. Lastly, dispersing agents of various grafting densities were prepared using a solvent-free process for the formation of an amide bond between carboxylic acid-functionalized matrixophilic polyester chains and the amines in the pigmentophilic poly(ethylene imine) (PEI) chains. The effect of structural variations in the dispersing agents on the dispersion quality of cyan pigments in radiation-curable matrices was studied and the quality of the dispersion was assessed based on the pigment particle size: the lower the particle size, the better the performance of the dispersion and thus of the final ink formulation.

Instead of using a traditional approach where the effect of systematic structural variations in the dispersant on the particle size was studied, we went a step further in the work reported here. A machine learning model was developed that predicts the structure–property relationships between the dispersant and the dispersion performance. The end goal is to become more efficient in the design of tailor-made dispersants with optimal application performance, while limiting laborious laboratory work to a minimum (in other words, a more sustainable process). Machine learning is revolutionizing modern materials design. In machine learning one typically starts from an existing experimental dataset of input features (e.g. structural variations of the dispersant) and response/target factors (quality of the dispersion measured via particle size). The dataset is split into training and testing datasets. The training dataset is the largest and is used to train the machine learning model, while the testing dataset is used to evaluate the quality of the developed model. Typically, the quality is expressed by comparison of the measured and predicted values, e.g. determined via the mean absolute error (MAE). The smaller the MAE, the better is the model. Where many of the successes of machine learning are rooted in very large datasets as a starting point to train the model,[15–17] the most common applications in academic and industrial materials design deal with small datasets of even less than 100 data points.[18–28]

Unfortunately, when using small datasets, the random splitting of the dataset introduces a significant chance factor, which transforms into an unacceptable variability in the predictive power of the model. Whereas in large datasets an extreme individual contribution of a single data point is tempered by the large number of other data points, this tempering effect is much reduced in small datasets. This makes the resulting model strongly dependent on the specific details of the data points used.[29]

Adapting machine learning methods for analyzing and making predictions based on sparse datasets is critical for advancing experimental research in complex material systems. In this context, we recently developed a machine learning framework aimed at such small datasets.[29,30] For small datasets, ensemble models were shown to outperform the individual model instances significantly, while at the same time giving rise to models that are more robust with regard to dataset sizes.

In a general ensemble model, the actual model is built as a (large) set of models each individually trained on an experimental dataset; this can be the same dataset or (non-)overlapping subsets of that dataset. The prediction of an ensemble model then becomes the (weighted) average of the predictions of the individual models. In our specific case, all models in the ensemble belong to the same family (e.g. polynomial regression models of order

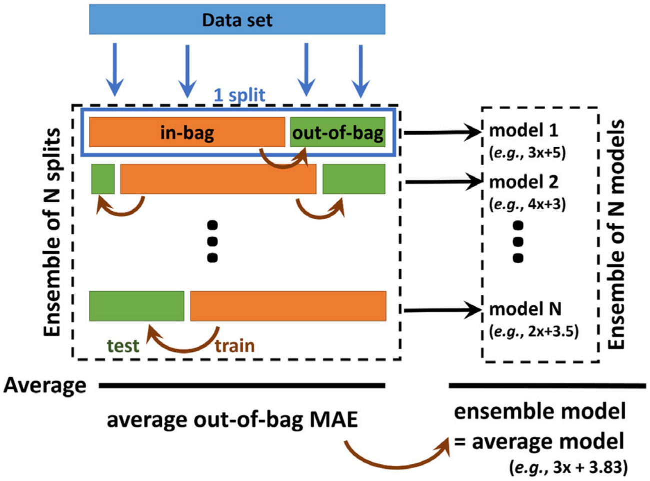

Figure 2. Schematic of the machine learning ensemble model approach.

2 such as $x^2 + bx + c$). Each of these models is trained on a different (random) splitting of our whole dataset, providing a different training and testing set for each individual model (Fig. 2). Within the context of ensemble models, training and testing sets are called '*in-bag*' and '*out-of-bag*' sets, respectively. Using the out-of-bag set, the quality of the associated individual model can be determined by calculating the MAE. In the case of small datasets, the different individual models show a strong variation, due to their sensitivity to individual data points, making them ill-suited for modeling purposes. The quality of the ensemble model, on the other hand, is estimated as the average of the out-of-bag MAE of the individual models, and was shown to be very robust in contrast to the MAE of the individual models.

Because all the individual models belong to the same model family, it is also possible to represent the ensemble model by a single individual model with the same properties as the ensemble, but with a significant reduction in computational cost for predictive purposes.

Using our previously developed ensemble approach for small datasets, a set of regression models was constructed for two dispersant datasets, namely PEI-*g*-polyTMCL and PEI-*g*-polyUDL. The quality of different models was compared. A total of 21 models were considered: linear regression; polynomial regression of orders 2 to 6; and polynomial regression models with LASSO regularization[†] of orders 1 to 15.[31, 32]

MATERIALS AND METHODS

Chemicals

Diphenyl phosphate (DPP; >99%, TCI), 1,5,7-triazabicyclo[4.4.0] dec-5-ene (TBD; 98%, Sigma-Aldrich), hexadecane (>99.5%, TCI), Pd/C (Sigma-Aldrich) and Celite® S (Sigma-Aldrich) were used as received. Benzyl alcohol (BnOH; 99%, Alfa Aesar) and UDL (>97%, Sigma-Aldrich) were distilled over CaH_2 prior to use. TMCL was synthesized chemically[8] or enzymatically[33] and distilled over CaH_2 under reduced pressure (1×10^{-3} mbar, 95–105 °C) prior to use. Solvents were supplied from Biosolve and used as received. Pigment blue 15:4 is Hostaperm™ Blue P-BFS, a CI Pigment Blue 15:4 from Clariant. PEI Epomin SP200 was received from Nippon Shokubai, PEI Lupasol® PR 8515 was obtained from BASF and PEI $M_w = 800$ Da was obtained from Sigma-Aldrich. Dipropylene glycol diacrylate (DPGDA) was supplied by Miwon under the tradename Miramer M222.

^{1}H NMR spectroscopy

NMR spectra were recorded with a Bruker Avance III HD Nanobay at 300 MHz for ^{1}H at ambient probe temperature in $CDCl_3$ with 16 scans.

Gel permeation chromatography

For the polyesters, gel permeation chromatography (GPC) was performed at 30 °C using a Waters GPC system equipped with a Waters 2414 refractive index detector. Tetrahydrofuran (THF) was used as eluent at a flow rate of 1 mL min^{-1}. Three linear columns were used (Styragel HR1, Styragel HR4 and Styragel HR5). Molecular masses were determined relative to polystyrene standards.

For the graft copolymers, the polymers were dissolved in 1,1,1,3,3,3-hexafluoroisopropanol (HFIP) with 0,019% NaTFA salt. Samples for GPC measurement were prepared by dissolving 5.0 mg of polymer in 1.5 mL of solvent. The solutions were filtered over a 0.2 μm PTFE syringe filter before injection. The GPC apparatus was calibrated with poly(methyl methacrylate) standards. Two PFG combination medium microcolumns with 7 μm particle size (4.6 × 250 mm, separation range 100–1 000 000 Da) and a precolumn PFG combination medium with 7 μm particle size (4.6 × 30 mm) with refractive index detector were used in order to determine molecular weight and dispersities.

Fourier transform infrared spectroscopy

Fourier transform infrared (FTIR) spectra were recorded using a Bruker Alpha FTIR equipped with a Zn–Se crystal for recording. All products were recorded as pure compounds.

Mass spectrometry

Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-ToF-MS) was conducted using a Bruker UltrafleXtreme spectrometer with a 355 nm Nd:YAG laser (2 kHz repetition pulse/Smartbeam-II™) and a grounded steel plate. *trans*-2-[3-(4-*tert*-Butylphenyl)-2-methyl-2-propenylidene]malononitrile (Sigma-Aldrich, >98%) was used as matrix (20 mg mL^{-1} in THF) and potassium trifluoroacetate (Sigma Aldrich, 98%) was used a cationization agent (10 mg mL^{-1} in THF). The polymers were dissolved in THF (10 mg mL^{-1}). Solutions of matrix, salt and polymer were mixed in volumetric ratios of 200:10:30, respectively. All mass spectra were collected in the reflector mode. Poly(ethylene glycol) standards with M_n equal to 5000, 10 000 and 15 000 g mol^{-1} were used for calibration. Data were processed using the FlexAnalysis (Bruker Daltonics) software package.

Measurement of particle size

Particle sizes of dispersions were measured using a PSS Nicomp 380, calibrated with a standard dispersion of 94 nm. Dispersions were diluted in EtOAc to parts per million concentrations and measurements made using dynamic light scattering.

Typical synthesis of polyUDL

In a vacuum-dried flask, BnOH (1.38 mL, 13.3 mmol, 1 eq) was added to UDL (81.0 mL, 425.9 mmol, 32 eq). The mixture was

[†] Regularization is an approach in which a penalty function is added to the regression objective and which generally leads to a reduction in model complexity (i.e. a reduction of model terms).

cooled at −10 °C using a cold bath of liquid nitrogen in ethylene glycol. The polymerization was started by adding TBD (1.85 g, 13.3 mmol, 1 eq) to the reaction mixture under nitrogen atmosphere. At the end of the reaction after 1 day, the reaction mixture was dissolved in cold chloroform at −20 °C (300 mL). The solution was washed with an acidified aqueous phase based on HCl until pH 2–3 (2 × 400 mL) and with water (1 × 400 mL). The solution was dried over magnesium sulfate and concentrated under vacuum. The polymer was dried under vacuum to obtain 84 g of polymer.

Typical synthesis of polyTMCL

In a vacuum-dried flask, BnOH (1.8 mL, 17.4 mmol, 1 eq) was added to a mixture of TMCL (55.0 mL, 323.2 mmol, 19 eq) and hexadecane as internal standard (typically 5% relative to the amount of TMCL). The mixture was heated to 60 °C in an oil bath. The polymerization was started by adding DPP (4.35 g, 17.4 mmol, 1 eq) to the reaction mixture under nitrogen atmosphere. The temperature was decreased to room temperature after 1 day. At the end of the reaction after 2 days, the reaction mixture was dissolved in cold chloroform at −20 °C (200 mL). The polymer was precipitated from cold methanol at −20 °C (1 L) and dried under vacuum to afford 33 g of polymer.

Typical hydrogenolysis procedure

An amount of 20 g of the polymer, in its benzyl-terminated form, was dissolved in 100 mL of ethyl acetate in a 250 mL flask. When dissolution was complete, Pd/C (1.0 g, 5 wt% with regard to the polymer) was added and the resulting slurry placed under nitrogen atmosphere. The flask was then placed under a hydrogen atmosphere by flushing the flask with hydrogen gas and the reaction is left to proceed for 3 days under a hydrogen atmosphere. After reaction, the slurry was filtered over a patch of Celite and the filtrate was evaporated resulting in the free acid-terminated polymer.

Typical coupling of polyester with PEI

Dispersant U4 in Table S1 was made by the addition of 0.5 g of PEI with normalized M_w 1.0 and 4 g polyUDL-COOH with normalized $M_{n,GPC,r} = 0.33$. After thorough mixing under a nitrogen atmosphere at 120 °C, the temperature was raised to 150 °C while maintaining a nitrogen flow over the viscous mixture. The reaction was allowed to proceed for 5 h, after which the mixture was cooled to room temperature.

Preparation of dispersions

Typically, PB15:4 and dispersing agent were added to the dispersion medium (DPGDA) in a ratio of 1:1:8 on weight basis. To this slurry, milling beads were added and the mixture was milled for 7 days at room temperature. The final dispersions were recovered by filtration to remove the milling beads.

Modeling and simulation

Modeling of the experimental data was performed using the machine learning framework previously developed by the authors.[29,30] In the current work, pasting-type ensembles of 1000 member instances with and 80/20 division between in-bag and out-of-bag data points were generated.[34] As base model types, we considered linear and polynomial regression (Figs S5 and S6), with and without LASSO regularization.[31,32] The regularization-strength hyperparameter of the LASSO was tuned for each individual model instance member of the ensemble using leave-one-out cross-validation (LOOCV)[35‡] performed on the in-bag dataset of the model instance.[36,37] The average of the out-of-bag evaluations was used as an estimator for the ensemble quality. The MAE was used as a quality measure to give equal weight to individual data points. MAE is the arithmetic average of the absolute errors, defined as

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

with y_i the prediction and x_i the experimental value.

The presented ensemble model surfaces were generated by predicting the particle size at each point of a 51×51 grid of $M_{n,GPC,r}$(polyester) and the polyester/PEI weight ratio. To obtain a two-dimensional surface, M_w(PEI) was kept constant. For the UDL dataset, surfaces of constant grafting density were plotted. The resulting surface was rendered using Gnuplot 5.2.[38]

RESULTS AND DISCUSSION

Preparation of dispersing agents

The goal of this study was to develop dispersing agents based on polyesters from alkyl-substituted lactones as matrixophilic chains and PEI as pigmentophilic chains (Fig. 3(c)). ROP of UDL and TMCL[9] using an alcohol as initiator was followed by hydrogenolysis in order to obtain acid-functionalized polyesters (Figs 3(a) and (b)). ^{1}H NMR spectroscopy confirms the disappearance of the benzyl end-group (Fig. S1).

While ^{1}H NMR did not allow the detection of the deprotected COOH signal, deprotection and COOH end-group presence were confirmed via MALDI-ToF-MS. This deprotection step was shown to maintain the structure of the polyester as demonstrated by the distribution of the polymer chains using MALDI-ToF-MS (Fig. S2). Additionally, the hydrogenolysis step did not significantly modify the molecular weight and dispersity of the polymers (Fig. S3). Matrixophilic chains based on polyUDL and polyTMCL of various molecular weights were prepared (Tables S1 and S2).

The dispersing agents were prepared by coupling the acid end-capped polyesters to PEI in bulk at 150 °C, resulting in graft copolymers PEI-*g*-polyester connected via amide bonds (Fig. 1(c)). Successful coupling of COOH-terminated polyesters was observed as evidenced from the appearance of peaks j and k next to amide bonds in ^{1}H NMR spectrum (Fig. 4(c)) and by the presence of amide bands in FTIR spectra at 1650 and 3325 cm^{-1} (broad) (Fig. 5(a)). Typical GPC curves show that a new high-molecular-weight peak appears upon coupling of PEI and the polyester, although still some unreacted PEI and/or polyester stays behind, as visible from the bimodal peak in Fig. 5(b).

[‡] LOOCV is a validation method appropriate for tuning hyperparameters on a small dataset. A model instance is trained on the entire dataset with one single sample removed. This is repeated for all possible configurations with one single sample removed (i.e. *N* times for a dataset of *N* samples). The performance of the model is then determined as the average of the errors obtained on the removed sample for each trained model instance (i.e. the average of *N* errors).

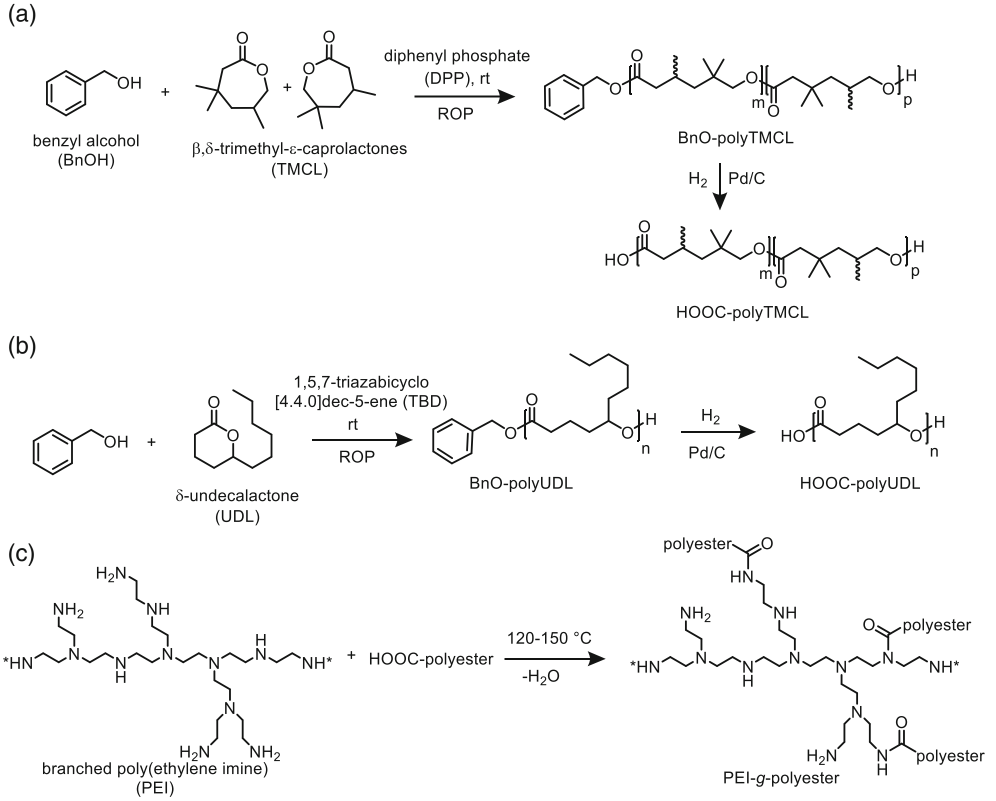

Figure 3. Synthesis of matrixophilic chains by ROP of lactones and hydrogenolysis for (a) polyTMCL and (b) polyUDL. (c) Synthesis of the PEI-*g*-polyester dispersant.

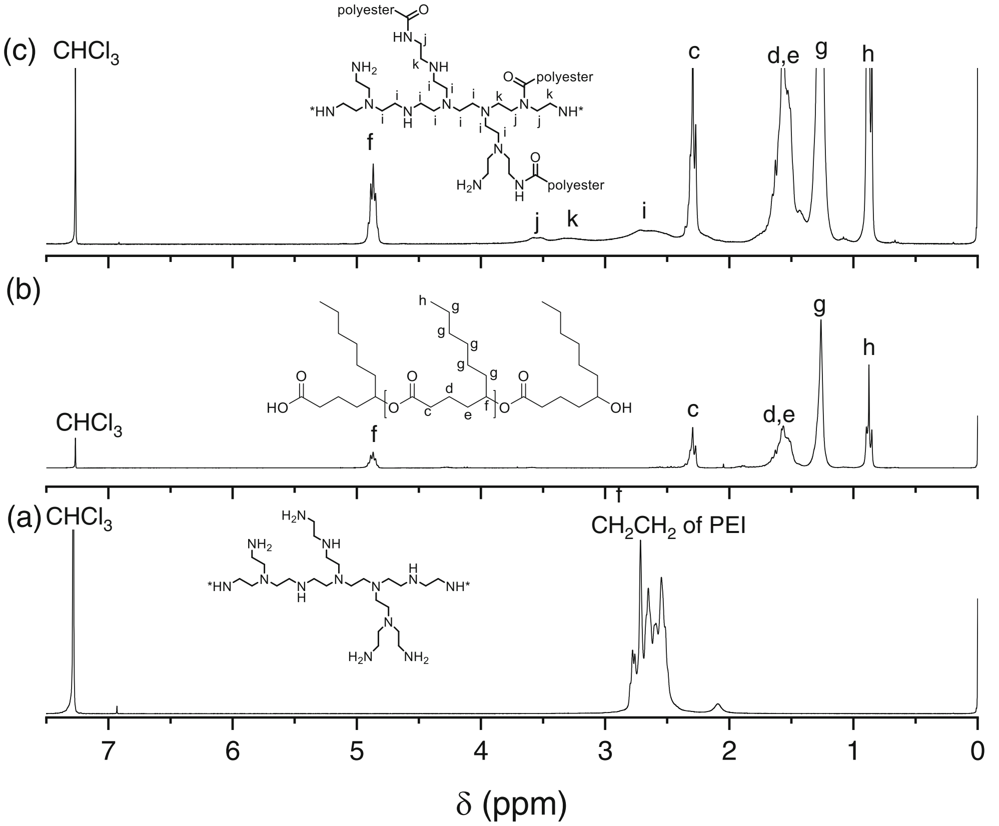

Figure 4. ^{1}H NMR spectra for (a) PEI, (b) polyUDL-COOH and (c) PEI-*g*-polyUDL.

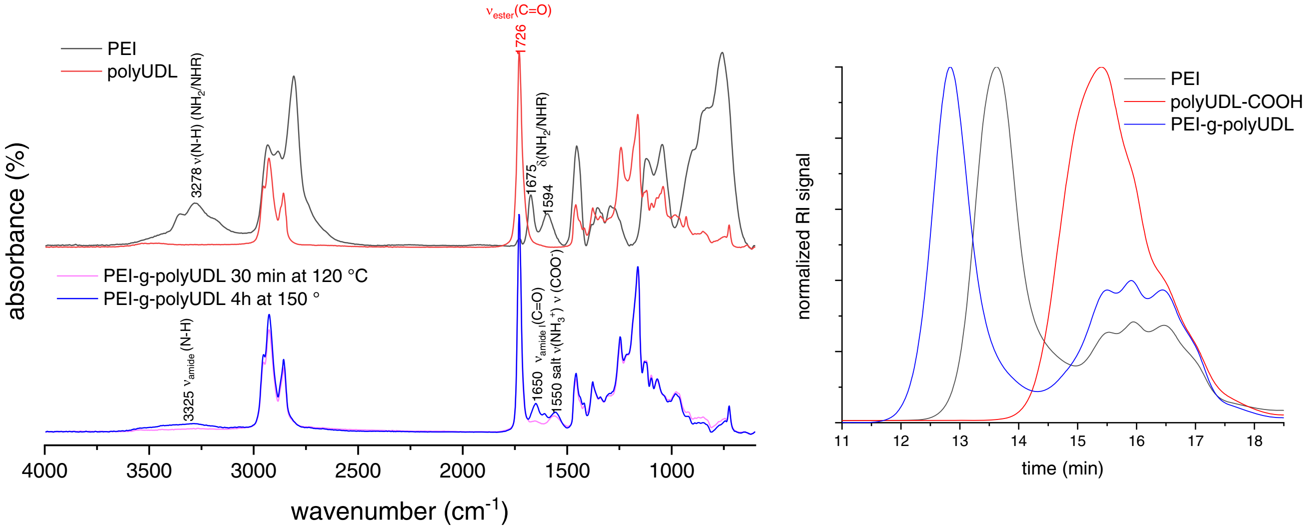

Figure 5. (a) FTIR spectra of a dispersing agent based on polyUDL-COOH coupled with PEI. The magenta line shows the formation of an amide salt after 30 min at 120 °C, and the blue line shows the result of coupling after 4 h at 150 °C. (b) GPC traces of PEI, polyUDL and PEI-*g*-polyUDL in HFIP + 0.019% NaTFA.

Machine learning model for evaluation of performance of dispersing agents

The performance of the dispersing agents for UV-curable inks was evaluated in dispersions made of a UV-curable matrix comprising DPGDA as UV-curable monomer and PB15:4 cyan as pigment. Those dispersions are in fact simplified ink formulations. The effect of different structural features in the dispersing agent was evaluated on its effectiveness in stabilizing pigments after the milling process was applied to make the dispersion: the chemistry of the polyester matrixophilic chains (polyUDL *versus* polyTMCL), the length of the polyester side chains ($M_{n,GPC,r}$), the length of the PEI pigmentophilic backbone and the polyester/PEI weight ratio as a measure for the amount of polyester side chains (Fig. 6).

The effect of systematic changes in the dispersant structure (called features in the context of machine learning) was quantified via particle size distribution (called target in the context of the machine learning) measurements of the dispersion with dynamic light scattering. Dispersing agents with good dispersing properties are expected to result in particle size distributions of low size, i.e. below 150 nm. On the contrary, dispersing agents with poor dispersing properties result in a higher particle size distribution, which indicates particle aggregate formation. Typical distributions are shown in Fig. S4.

To obtain a better understanding of the observed experimental results, we investigate the datasets using machine learning regression. The data are selected as discussed in the supporting information, resulting in a dataset of 25 and 16 data points for the TMCL- and UDL-based dispersing agents, respectively.

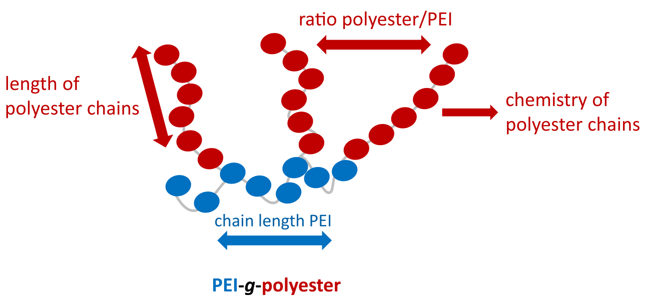

Figure 6. Schematic of different structural features varied in the dispersant design to tune the dispersion quality.

TMCL dataset

Preliminary regression modeling shows that an ensemble model based on simple linear regression – using the polyester/PEI weight ratio and the polyester and PEI molecular weights (TMCL1) – is rather hard to beat (Figs S5 and S6).[29] As polynomial models are prone to overfitting[§] small datasets (due to their large number of features), regularization of such models is required. Regularization is a method in which one adds a special penalty term to the polynomial model. This penalty term leads to the suppression (or even removal) of features (i.e. descriptors) which do not contribute significantly to the model. Of the various penalty functions available, we selected a penalty function with an L1 measure[**] known as LASSO regularization as it gives rise to a very strong regularization of models.[31,32] As our models are in essence ensemble models, counting the number of ensemble instances for which a feature is not removed provides a measure for the importance of this feature.[29] Analysis of the most important features in the regularized polynomial regression models indicates the presence of a strong functional relation with high polynomial order for the product of the polyester molecular weight and the polyester/PEI weight ratio. Further investigation of several feature combinations eventually leads us to select 11 relevant features for further study (see section 2.1 of the supporting information for a

[§] A model is overfitting when it has so much flexibility that it can perfectly predict the dataset on which it is trained, but fails badly on new and unknown data. For example, take four data points following a linear relation $y = 3x + 5 + \delta_{noise}$; a fourth-degree polynomial allows one to perfectly fit any four data points. However, any new point will in general be very poorly predicted. In contrast, any linear model will not perfectly model the four data points due to the noise term, but it will predict new data much more accurately than the fourth-degree polynomial.

[**] In an L1 measure, distances are defined as $\left\|\vec{x}\right\| = \sum_i |x_i|$. Euclidean distances are also known as L2 measures.

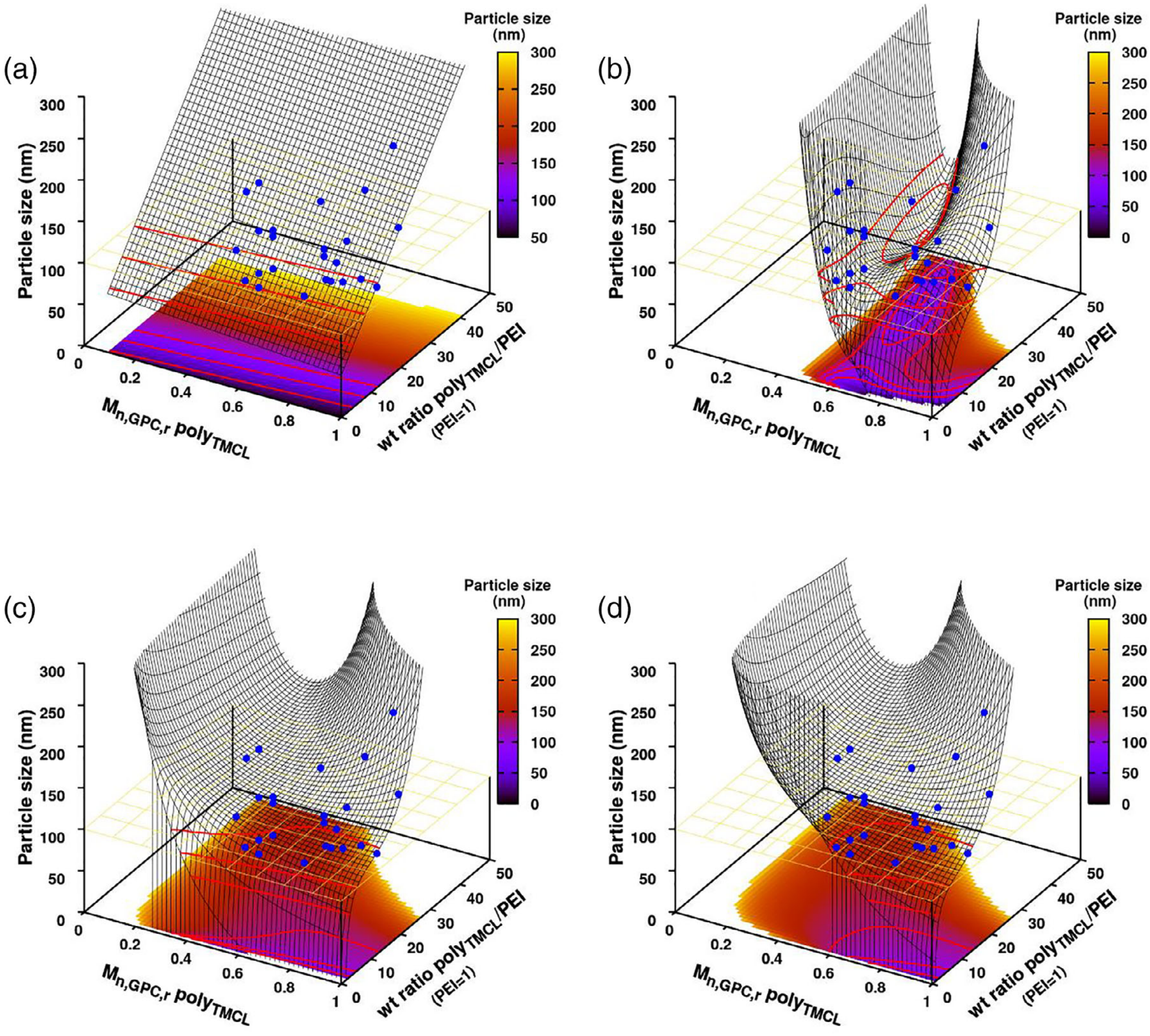

Figure 7. Various machine learning ensemble models, trained on the TMCL dataset: (a) TCML1 (animated gifs of these 3D graphs are provided as SI), (b) TCML2, (c) TCML3 and (d) TCML6. Contours are added at 75, 100 and 125 nm, and the plane of particle size 100 nm is indicated by the yellow grid.

Table 1. Properties of selected machine learning models

Model	Model type	No. features (selected >50% ensemble)	MAE_{oob} (nm)	CI_{low} (95%) (nm)[a]	CI_{high} (95%) (nm)[b]
TMCL					
TMCL1	Linear	3	45.4	44.5	46.3
TMCL2	Linear	11	55.8	53.5	58.5
TMCL3	LASSO O1	11 (4)	35.0	34.1	36.3
TMCL5	LASSO O1	5 (5)	34.3	33.5	35.1
TMCL6	LASSO O1	5 (5)	33.4	32.7	34.2
UDL					
UDL1	Linear	3	63.0	61.5	64.6
UDL2	LASSO O1	9 (8)	66.0	64.1	68.4
UDL3	LASSO O1	9 (7)	42.3	41.4	43.3
UDL4	LASSO O1	6 (5)	41.5	40.4	42.5

[a] CI_{low}, lower limit of the 95% confidence interval.
[b] CI_{high}, upper limit of the 95% confidence interval.

detailed discussion of how these features are selected). A simple linear (non-regularized) regression model with these 11 features is indicated as TMCL2 and is shown in Fig. 7(b). As is evident from Table 1, this model is outperformed by the original linear regression model (lower MAE for TMCL1), even though the graphical representation in Fig. 7 shows it to closely follow the experimental data. The TMCL2 model presents a very narrow canyon-like structure with some local valleys (Fig. 7(b)). The optimum dispersing agent is found for $M_{n,GPC,r}$(polyTMCL) > 0.6 and a polyester/PEI weight ratio of about 10. A second valley of interest is found for polyTMCL with $M_{n,GPC,r}$ around 0.7 and weight ratios in the range 20–45. This second valley is located outside of the range of the available experimental data making it hard to assess its quality. In this regard, it is important to remember that 11 features have been fitted to a dataset of only 25 data points, making the presence of the second valley likely to be an artifact due to overfitting.

When LASSO regularization is included, it becomes clear that not all features have equal importance (Table S3). The number of relevant features is significantly reduced, and the resulting model (TMCL3) outperforms the original linear regression model TMCL1 (Table 1). Removal of the least important features retains the model quality (TMCL5) while reducing model complexity. This final model is further stabilized by exponentiation of one specific feature (supporting information, section 2.2). The TMCL3 and TMCL6 models are shown in Figs 7(c) and (d). They present a much smoother surface, with a wide valley following the same general trend as the TMCL2 model. In this case however, the second valley has disappeared, and only the optimum region for $M_{n,GPC,r}$(polyTMCL) > 0.6 and lower polyester/PEI weight ratios (<15) remains. For higher weight ratios the smallest particle sizes are obtained with polyesters with lower molecular weight. Note, however, that with increasing weight ratios the valley minimum becomes narrower, and particle sizes rapidly increase for both too high and too low polyester molecular weights. Of the models presented, the TMCL6 model performs best in terms of numerical quality (the out-of-bag estimate of the MAE is the lowest), but also in terms of simplicity (only five features). Therefore, we propose the following analytical model for the TMCL-based dispersing agent:

$$\begin{aligned}\text{Particle size} = {} & 15.54\exp\left(\frac{AC-11.042}{6.441}\right)+161648.77\\ & \times\exp\left(\frac{9\times10^{-5}}{A}\right)-23\log_{10}B-218.28\frac{B}{C}\\ & +130.21\frac{1}{AC}-161555.28\end{aligned}$$

where A and B represent the normalized $M_{n,GPC,r}$(polyTMLC) and $M_{w,r}$(PEI) and C represents the dimensionless polyTMCL/PEI weight ratio.

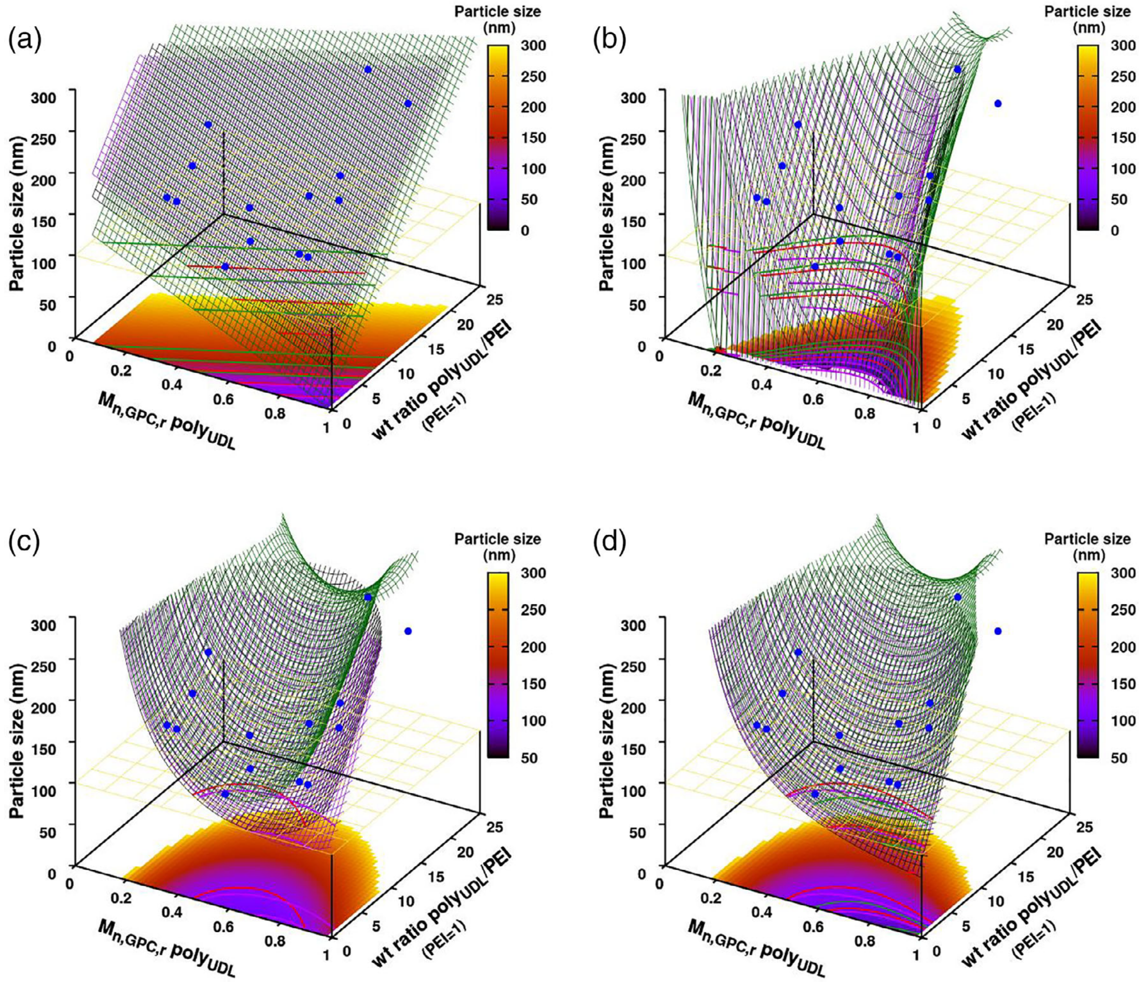

Figure 8. Various machine learning ensemble models, trained on the UDL dataset: (a) UDL1, (b) UDL2, (c) UDL3 and (d) UDL4. Contours are added at 75, 100 and 125 nm, and the plane of particle size 100 nm is indicated by the yellow grid. The model surfaces correspond to grafting densities of 2% (purple), 5% (black, with red contours) and 10% (green) (animated gifs of these 3D graphs are provided as SI).

UDL dataset

An important difference from the TMCL dataset is the fact that only two distinct features are available, as the molecular weight of PEI was identical in all experiments. With only two features, the quality of the linear regression ensemble model appears worse than what is found for the TMCL dataset. However, for both, the observed quality is in the range of the expected for datasets of 10–30 data points.[23,29]

During the preliminary regression modeling of the UDL dataset (supporting information, section 2.2.2) we found that a third simple feature could be included: the *grafting density*, which in terms of physical interpretation is related to the polyester/PEI weight ratio. The grafting density is based on the molar ratio between polyester and ethylene imine (see footnote to Table S1). This ratio is derived from $M_{n,NMR}$(polyester), which could not be determined with the highest accuracy possible because side reactions that have influence on the end-group functionality, e.g. water initiation, could not be excluded. It has been mentioned by Zhang and Ling[23] that in some cases low-quality data can even be used to improve the model. Although the terms polyester/PEI weight ratio and graft density are quite similar, they appear to be beneficial for different data points in the construction of polynomial models, hence both features are included in this work. At the level of a simple linear regression (ensemble) model (UDL1), the quality remains roughly unchanged: $MAE_{oob} \sim 65$ nm (Table 1). Investigation of the possible benefits by feature transformation via taking a logarithm or exponentiation (UDL2) provides additional hints for the construction of more complex features (supporting information, section 2.2.2). Based on these series of regularized polynomial regression models, nine complex features are constructed for the UDL3 model (Table 1). The resulting UDL3 model clearly outperforms the linear regression model. The resulting model, shown in Fig. 8(c), presents similarities to the TMCL model, but indicates it to be much harder for the UDL polyester to provide a satisfactory result. The shape of the model surface shows only little sensitivity to the grafting density. Within the UDL3 model, the optimum $M_{n,GPC,r}$(polyUDL) is above 0.5, with a polyester/PEI weight ratio of 5 or less.

The UDL3 model is further optimized through the removal of superfluous features (UDL4). This final model contains six features but presents the same model quality as UDL3 (Table 1). Although the general qualitative picture has not changed, the contribution of the grafting density is reduced, as is indicated by the shift of the contour lines delineating an optimum region. The qualitative picture is however retained. $M_{n,GPC,r}$(polyUDL) should be above 0.5 and the weight ratio below a factor 5.

The UDL4 model performs best in terms of numerical quality (the out-of-bag estimate of the MAE is the lowest), but also in terms of simplicity (only six features). We therefore propose the following analytical model for the UDL-based dispersing agent:

$$\begin{aligned}\text{Particle size} = {} & 4.62D + \frac{46.76}{A} + 2.43C\exp\left(\frac{A-0.59}{0.20}\right) \\ & + 2.67\frac{C^2}{D} - 0.67 \\ & \times \exp\left(2\left(\frac{C-10.15}{6.69} + \frac{D-8.19}{5.63}\right)\right) + \frac{51.86}{D} \\ & - 6.51\end{aligned}$$

where A represents $M_{n,GPC,r}$(polyUDL), C the dimensionless polyUDL/PEI weight ratio and D the grafting density (%).

Machine learning models versus *laboratory context*

As described in the above subsections, two machine learning models were developed based on the experimental datasets for the TMCL and UDL dispersing agents. The fitted best models are presented in Figs 7(d) and 8(d) for TMCL and UDL, respectively. For most of the experimental data points the fit is of good quality, with some exceptions which increase the overall MAE. This is to be expected for machine learning models trained on such extremely small datasets. The overall shape of the model surfaces is quite similar for both dispersing agents, as one would expect. For both dispersing agents we observe that the optimum behavior is expected for $M_{n,GPC,r} > 0.5$–0.6, indicating similar molecular weights for both TMCL and UDL. The maximum polyester/PEI weight ratios, on the other hand, show a somewhat different picture. Favorable results, i.e. small particle sizes, are obtained with a polyTMCL/PEI ratio < 15, while for polyUDL-based dispersants the weight ratio should be kept lower (polyUDL/PEI < 5). This means that there is less freedom in the chemical composition of PEI-*g*-polyUDL compared to PEI-*g*-polyTMCL.

CONCLUSIONS

PolyTMCL and polyUDL are efficient polyesters to be used as matrixophilic chains for the preparation of polymeric dispersing agents with PEI. Good stability of pigment dispersions was achieved for UV-curable matrices made of DPGDA with PB15:4 cyan as pigment, reaching low pigment particle sizes of about 100 nm. A machine learning regression model based on our previously developed ensemble approach for small datasets was developed to predict structure–property relationships between the dispersant and the dispersion quality. Normalized molecular weights of the polyester greater than 0.55 for polyTMCL and greater than 0.5 for polyUDL as well as a low polyester/PEI weight ratio (<15 for polyTMCL- and <5 for polyUDL-based dispersants) contributed to better pigment particle stabilization. The best model for PEI-*g*-polyTMCL dispersants gives MAE_{oob} of 33.4 nm, while for PEI-*g*-polyUDL, MAE_{oob} of 41.5 nm could be achieved.

ACKNOWLEDGEMENTS

MD, KB, ND, KM and GD received funding from the European Union (EU) project ROBOX (grant agreement no. 635734) under the EU's Horizon 2020 Programme Research and Innovation actions H2020-LEIT BIO-2014-1. The views and opinions expressed in this article are only those of the authors, and do not necessarily reflect those of the European Union Research Agency. The European Union is not liable for any use that may be made of the information contained herein. DV, JS and KB acknowledge the project D-NL-HIT carried out in the framework of INTERREG-Program Deutschland-Nederland, which is co-financed by the European Union, the MWIDE NRW, the Ministerie van Economische Zaken en Klimaat and the provinces of Limburg, Gelderland, Noord-Brabant and Overijssel.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

1 Schofield J, *Prog Org Coat* **45**:249–257 (2002).
2 Pirrung FOH, Quednau PH and Auschra C, *Chimia* **56**:170–176 (2002).

3 K. Bernaerts, I. Hoogmartens and G. Deroover, Non-aqueous pigmented inkjet inks. WO Patent 2008043677 (2008).
4 Schmitz J, Frommelius H, Pegelow U, Schulte H-G and Höfer R, *Prog Org Coat* **35**:191–196 (1999).
5 Spinelli H, *Adv Mater* **10**:1215–1218 (1998).
6 D. Thetford and J. D. Schofield, Dispersants. US Patent 5700395 (1997).
7 D. Thetford, J. D. Schofield and P. J. Sunderland, Dispersants. US Patent 6197877B1 (2001).
8 Delgove MAF, Luchies J, Wauters I, Deroover GGP, De Wildeman SMA and Bernaerts KV, *Polym Chem* **8**:4696–4706 (2017).
9 Delgove MAF, Wróblewska AA, Stouten J, van Slagmaat CAMR, Noordijk J, De Wildeman SMA *et al*., *Polym Chem* **11**:3573–3584 (2020).
10 A. L. Boog, A. L. Peters and R. Roos, Process for producing delta-lactones from 11-hydroxy fatty acids. US Patent 5215901A (1993).
11 Delgove MAF, Fürst MJLJ, Fraaije MW, Bernaerts KV and De Wildeman SMA, *ChemBioChem* **19**:354–360 (2018).
12 Delgove MAF, Elford MT, Bernaerts KV and De Wildeman SMA, *J Chem Technol Biotechnol* **93**:2131–2140 (2018).
13 Delgove MAF, Elford MT, Bernaerts KV and De Wildeman SMA, *Org Proc Res Develop* **22**:803–812 (2018).
14 Delgove MAF, Laurent A-B, Woodley JM, De Wildeman SMA, Bernaerts KV and van der Meer Y, *ChemSusChem* **12**:1349–1360 (2019).
15 Cendagorta JR, Tolpin J, Schneider E, Topper RQ and Tuckerman ME, *J Phys Chem B* **124**:3647–3660 (2020).
16 Curtarolo S, Setyawan W, Hart GLW, Jahnatek M, Chepulskii RV, Taylor RH *et al*., *Comput Mater Sci* **58**:218–226 (2012).
17 Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S *et al*., *APL Mater* **1**:011002 (2013).
18 Ghafari E, Bandarabadi M, Costa H and Júlio ENBS, *J Mater Civ Eng* **27**: 04015017 (2015).
19 Houben C, Peremezhney N, Zubov A, Kosek J and Lapkin AA, *Org Proc Res Develop* **19**:1049–1053 (2015).
20 Peremezhney N, Hines E, Lapkin A and Connaughton C, *Eng Optimizat* **46**:1593–1607 (2014).
21 Schweidtmann AM, Clayton AD, Holmes N, Bradford E, Bourne RA and Lapkin AA, *Chem Eng J* **352**:277–282 (2018).
22 Clayton AD, Schweidtmann AM, Clemens G, Manson JA, Taylor CJ, Niño CG *et al*., *Chem Eng J* **384**:123340 (2020).
23 Zhang Y and Ling C, *NPJ Comput Mater* **4**:25 (2018).
24 Rubens M, Vrijsen JH, Laun J and Junkers T, *Angew Chem Int Ed* **58**: 3183–3187 (2019).
25 Rubens M, Van Herck J and Junkers T, *ACS Macro Lett* **8**:1437–1441 (2019).
26 Coley CW, Thomas DA 3rd, Lummiss JAM, Jaworski JN, Breen CP, Schultz V *et al*., *Science* **365**:eaax1566 (2019).
27 Wang Z, Su Y, Shen W, Jin S, Clark JH, Ren J *et al*., *Green Chem* **21**:4555–4565 (2019).
28 Menon A, Gupta C, Perkins KM, DeCost BL, Budwal N, Rios RT *et al*., *Mol Syst Des Eng* **2**:263–273 (2017).
29 Vanpoucke DEP, van Knippenberg OSJ, Hermans K, Bernaerts KV and Mehrkanoon S, *J Appl Phys* **128**:054901 (2020).
30 D. E. P. Vanpoucke, AMADEUS v0.1. Available: https://github.com/DannyVanpoucke/Amadeus (2020).
31 Santosa F and Symes WW, *SIAM J Sci Stat Comput* **7**:1307–1330 (1986).
32 Tibshirani R, *J R Stat Soc B* **58**:267–288 (1996).
33 Delgove MAF, Valencia D, Solé J, Bernaerts KV, De Wildeman SMA, Guillén M *et al*., *Appl Catal Gen* **572**:134–141 (2019).
34 Breiman L, *Mach Learn* **36**:985–103 (1999).
35 Cartwright H, *Using Artificial Intelligence in Chemistry and Biology: A Practical Guide*. CRC Press, Boca Raton, FL (2008).
36 Lachenbruch PA and Mickey MR, *Dent Tech* **10**:1–11 (1968).
37 Molinaro AM, Simon R and Pfeiffer RM, *Bioinformatics* **21**:3301–3307 (2005).
38 T. Williams and C. Kelley, Gnuplot 5.2: an interactive plotting program. Available: http://gnuplot.sourceforge.net/ (2019).