Received: 11 January 2024 | Revised: 28 February 2024 | Accepted: 3 March 2024

DOI: 10.1002/app.55422

RESEARCH ARTICLE

Journal of Applied Polymer Science WILEY

Strategies for the fast optimization of the glass transition temperature of sustainable epoxy resin systems via machine learning

Florian Rothenhäusler | **Holger Ruckdaeschel**

Department of Polymer Engineering, University of Bayreuth, Bayreuth, Germany

Correspondence
Holger Ruckdaeschel, Department of Polymer Engineering, University of Bayreuth, Universitätsstraße 30, 95444 Bayreuth, Germany.
Email: holger.ruckdaeschel@uni-bayreuth.de

Funding information
German Federal Ministry for Economic Affairs and Climate Action (BMWK), Grant/Award Number: #20E1907A

Abstract

Aligned with the prevailing sustainability paradigm, the imperative adoption of bio-based substitutes for constituents within petroleum-derived epoxy resin becomes evident. Blending bio-based and petroleum-based epoxy resins and curing agents, establishes a synergistic compromise addressing both sustainability imperatives and the mechanical efficacy of thermosets. The conventional approach to discovering optimal compositions for multi-component mixtures under specific boundary conditions includes empirical trial and error and is seen as a protracted and inefficient endeavor. Conversely, leveraging machine learning might afford a streamlined and confident resolution to this challenge. This investigation elucidates the requisite strategies for maximizing the efficiency of material property optimization through the application of Bayesian optimization and active learning. Illustratively, the study demonstrates the proficient optimization of the glass transition temperature within a four-component epoxy resin system. This optimization is conducted across varying ranges of bio-content and cost considerations. The study underscores the utility of machine learning in achieving this task with notable efficiency. The efficacy of least squares, kernel ridge regression, Gaussian process regression, and artificial neural networks, is meticulously evaluated through comprehensive seven-fold cross-validation and validated against experimental data.

KEYWORDS

Bayesian optimization, differential scanning calorimetry, glass transition temperature, machine learning, sustainability, thermosets

1 | INTRODUCTION

Epoxy resins play a pivotal role in various industrial applications due to their remarkable mechanical properties and thermal stability.[1,2] As the demand for advanced materials continues to rise, there is an increasing urgency to explore sustainable alternatives to traditional epoxy resins.[3] The imperative to find greener alternatives is driven not only by environmental concerns but also by the necessity to reduce reliance on petroleum-based resources in the manufacturing process.[4]

In the realm of bio-based epoxy resins and curing agents, chemically modified vegetable oils have risen in importance as commercially viable components.[5,6] However, the long carbon chains in vegetable oils contribute to low glass transition temperatures (T_g) in resulting thermosets compared to their petroleum-based counterparts, due to limited steric hindrance.[7] To address this limitation, researchers have investigated alternative compounds such as cardanol, extracted from cashew nut shell liquid, which incorporates a phenyl group, offering enhanced steric hindrance.[8] Despite these efforts, substituting conventional epoxy resins with epoxidized vegetable oils (EVO) typically results in thermosets with lower performance.[9] Consequently, only a fraction of petroleum-based components is usually replaced with EVO to strike a balance between mechanical properties, cost-effectiveness, and environmental impact reduction.[10]

Nevertheless, integrating bio-based components into epoxy resins represents a promising avenue toward sustainability. Balancing the bio-based content in epoxy resin formulations with petroleum-based constituents presents a critical challenge. This challenge arises from the need to maintain mechanical performance while adhering to specific bio-content and price constraints. Achieving this delicate equilibrium demands a multidimensional approach, integrating sustainable practices with cost-effectiveness and performance metrics. However, this pursuit is intricate, requiring meticulous exploration of various compositions to optimize critical material properties.[11]

In tackling this intricate optimization task, machine learning and Bayesian optimization (BO) emerge as powerful tools.[12–14] The complexity of the composition optimization problem, coupled with the myriad factors influencing T_g, necessitates sophisticated techniques for efficient exploration of the compositional space. T_g is a key property strongly correlated with the mechanical and thermal performance of epoxy resins, making its optimization crucial for ensuring the viability of sustainable alternatives.[15] BO, with its ability to model complex, nonlinear relationships and make informed decisions with limited experimental data, becomes instrumental in navigating the vast design space of epoxy resin formulations.[16] While many studies have employed machine learning to achieve desired results, there have been fewer efforts focused on creating viable strategies for the optimization of resin system formulations, emphasizing the need for a comprehensive framework for the optimization process itself.

This paper aims to contribute to this growing body of knowledge by investigating strategies for the fast optimization of the T_g in epoxy resin systems. The focus will be on a mixture of petroleum-based and bio-based components, exploring different minimum bio-contents within specified price ranges. Subsequently, the data will be utilized to model the epoxy resin system using least squares (LS), kernel ridge regression (KRR), Gaussian process regression (GPR), and artificial neural networks (ANN). The models' predictions will be validated through a seven-fold cross-validation (CV) and experimental validation, culminating in a robust understanding of the intricate relationship between composition and T_g. This research paves the way for the development of sustainable and economically viable epoxy resin formulations in the future.

2 | EXPERIMENTAL

2.1 | Key strategies

Prior to delving into the specifics of the experimental procedures, it is imperative to explain the foundational strategies implemented in this study. As previously underscored, the principal objective is to elucidate the framework essential for the fast optimization of resin systems, employing the T_g as an exemplar. Such an ambitious goal necessitates an interdisciplinary approach, seamlessly integrating the domains of materials science and data science.

Firstly, the composition of the epoxy resin system demands meticulous consideration. An optimal and pragmatic approach involves the selection of one bio-based and one petroleum-based component for both the epoxy resin and the curing agent. The decision-making process is contingent upon factors such as cost, bio-content of the bio-based constituents, as well as the viscosity and reactivity of each component.[15] Moreover, the molecular structure of the bio-based components should ideally lead to a heightened cross-link density and the formation of rigid network segments.

The number of dimensions of the feature vector is crucial.[17] Analogous to the exclusion of superfluous components from the resin system, diminishing the number of feature variables expedites the optimization procedure. Components that remain constant, such as a fixed percentage of an accelerator, need not be incorporated into the feature space, given their unchanging nature during optimization. Alternatively, the normalization of specific features with respect to one another serves to reduce the feature space. Furthermore, normalized and standardized properties prove advantageous for algorithms employing gradient descent methodologies.[18]

The ensuing consideration pertains to the experimental protocol itself, wherein the determination of T_g can be accomplished through various methods, including

TABLE 1 EEW, AHEW, bio-contents and prices of the epoxy resins and curing agents comprising the four-component system.

Label	Name	EEW	AHEW	Bio-content (%)	Price (Euro/kg)
R_1	DGEBA	187 g mol^{-1}	–	0	3
R_2	NC-514	425 g mol^{-1}	–	65	8
CA_1	IPDA	–	42.58 g mol^{-1}	0	8
CA_2	Mergamid L450	–	90 g mol^{-1}	73	5.5

differential scanning calorimetry (DSC), dynamic mechanical analysis (DMA), and thermal mechanical analysis (TMA).[19–21] However, both DMA and TMA necessitate cured specimens, implying additional steps encompassing the curing of mixtures in molds and subsequent cutting into specific dimensions. In stark contrast, the utilization of DSC mitigates this process intricacy by eliminating the need for molding and sawing, thereby facilitating the measurement of up to 12 data points per day, with each formulation tested in duplicate. These measures collectively serve to drastically reduce the optimization iteration cycle time, ensuring expeditious results.

The ultimate consideration mandates expertise from both polymer materials science and data science, specifically in the judicious selection of the kernel function and acquisition function. The kernel function, elucidating the covariance matrix, assumes a pivotal role in determining the mean and standard deviation calculated by GPR of the virtual experiments. While the knowledge of the precise equation underlying the data set is not obligatory during BO, making informed assumptions about anticipated trends when varying features remains advantageous.[22] The rationale behind the advantageous selection of a specific kernel function in the present study is explicated in the experimental design (see Section 2.5.1).

It is important to note that these considerations are universally applicable, extending beyond the optimization of T_g to encompass diverse research inquiries, including the optimization of polymer processing parameters and other material properties. The crux lies in a continual evaluation of whether all avenues for enhancing workflow efficiency have been exhaustively explored, thereby potentially uncovering untapped opportunities for reducing research time and costs.

2.2 | Materials

Diglycidyl ether of bisphenol A (DGEBA) resin with an epoxide equivalent weight (EEW) of 187 g mol^{-1} was obtained from Blue Cube Assets GmbH & Co. KG, Olin Epoxy (Stade, Germany). The cardanol-based epoxy resin NC-514 (EEW = 425 g mol^{-1}) was provided by Cardolite Corporation (Pennsylvania, USA). Isophorone diamine (IPDA) with an active hydrogen equivalent weight (AHEW) of 42.58 g mol^{-1} was bought as Aradur®42 BD from Huntsman Corporation (Texas, USA). The vegetable oil based amine curing agent Mergamid L450 with an AHEW of 90 g mol^{-1} was provided by HOBUM Oleochemicals GmbH (Hamburg, Germany). The selected bio-based epoxy resin and curing agent have a similar reactivity as their petroleum-based counterparts, thereby avoiding a delayed curing of the bio-based components and formation of inhomogeneities. The epoxy resins' EEW, the curing agents' AHEW, the bio-contents and prices of all components are shown in Table 1.

2.3 | Resin formulation

The components constituting the resin system are employed in their as-received state. Usually, the weight percentages of all components of a system are used as features for the optimization. Typically, the sum of all components in such systems equals 100% or 1, leading to high-dimensional feature spaces that pose challenges in terms of comprehension and visualization. To mitigate the complexity associated with the high-dimensional feature spaces, a normalization strategy is implemented. Specifically, the number of epoxy groups from the petroleum-based epoxy resin $\left(n_{\text{epoxy,petro}}\right)$ is normalized to the total number of epoxy groups in the resin system $\left(n_{\text{epoxy,petro}} + n_{\text{epoxy,bio}}\right)$. This normalization yields a molar ratio X_R ranging from 0 to 1, where 0 signifies that no epoxy groups originate from the petroleum-based resin, and 1 indicates that all epoxy groups are derived from the petroleum-based resin (see Equation (1)).

$$X_R = \frac{n_{\text{epoxy,petro}}}{n_{\text{epoxy,petro}} + n_{\text{epoxy,bio}}}. \quad (1)$$

A parallel approach is adopted for the curing agents, normalizing the number of active hydrogen atoms from the petroleum-based curing agent $\left(n_{\text{H,petro}}\right)$ to the total number of active hydrogen atoms in the curing agents $\left(n_{\text{H,petro}} + n_{\text{H,bio}}\right)$, resulting in a molar ratio X_{CA} (see Equation (2)).

$$X_{CA} = \frac{n_{H,petro}}{n_{H,petro} + n_{H,bio}}. \quad (2)$$

By transitioning from the conventional optimization of weight percentages to the normalized approach, the dimensionality of the feature vector to be optimized is effectively reduced from four to two, encapsulated by the vector (see Equation (3)):

$$\mathbf{X} = [X_R \; X_{CA}]. \quad (3)$$

For simplicity, the stoichiometric ratio (R) between the active hydrogen atoms of the curing agents and the epoxy groups of the resins is maintained at a fixed value of $R = 1$. Weight ratios for each component are then computed, taking into account their EEW and AHEW (see Table 1). Subsequently, the mixtures are homogenized using a centrifuge speed mixer from Hauschild Engineering (Hamm, Germany) operating at 3000 min^{-1} for 60 s.

2.4 | Differential scanning calorimetry

The analysis of the T_g for the cured resin formulations was conducted using a Mettler Toledo DSC 1 instrument (Columbus, Ohio, USA). The resin mixtures underwent curing in a dynamic DSC measurement spanning from 25°C to 200°C, with a heating rate of 10 K min^{-1}. The upper temperature limit ensures complete curing of the resin system, while maintaining a sufficiently low temperature to prevent thermal degradation of the aliphatic, bio-based components. Following curing, the specimens were cooled below their T_g with a cooling rate of -20 K min^{-1}. To ascertain T_g, the specimens were subsequently heated to 200°C. The determination of T_g for machine learning (ML) models utilized the inflection point of the DSC thermograms during the second heating cycle. The nitrogen flow rate was maintained at 50 mL min^{-1}, and the sample mass was controlled at 10 ± 2.5 mg. Each curing cycle involved testing two specimens for robust analysis.

2.5 | Modeling

2.5.1 | Experimental design

Procedure

The initial optimization series aims to assess the efficacy of various kernels and acquisition functions in optimizing the T_g of the resin system without imposing a minimum bio-content constraint. Three random data points, featuring bio-contents between 35% and 57.5%, serve as the starting point for this series. Kernels (radial basis function (RBF), Matérn, and dot product (DP)) are individually paired with two acquisition functions: upper confidence bound (UCB) and maximum expected improvement (EI) (see Figure 1).

The second optimization series extends the task by seeking the highest T_g for mixtures with a minimum 5% bio-content, employing the DP kernel in conjunction with both acquisition functions (UCB and EI). The DP UCB and DP EI approaches utilize the initial three random data points along with the data points proposed by themselves in the first optimization series.

This process is replicated in the third optimization series, with a minimum bio-content requirement increased to 10%. Given the higher number of mixtures feasible with close to 10% bio-content, and the likelihood of data points proposed by BO being clustered in the feature space with low bio-content, AL is introduced to enhance the data set and predictive capabilities of the final models. The results of AL and BO are included in the prediction of the next AL and BO experiments.

In the industrial context, product price is a critical factor. Therefore, the final round of virtual experiments focuses on mixtures with a minimum bio-content of 25% and a maximum price of 5.5Euro/kg. The data points obtained from all kernel functions and acquisition functions, including AL, serve as input ($n = 29$). An additional five random data points are introduced to diversify the dataset, resulting in a total of 35 data points for modeling and experimental validation. Subsequently, five

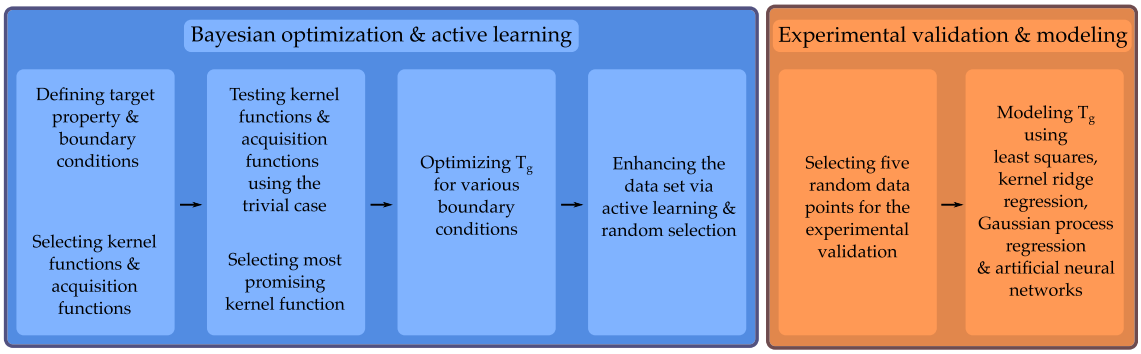

FIGURE 1 Procedure of this study. [Color figure can be viewed at wileyonlinelibrary.com]

unrelated random data points are selected from the virtual experiments as a validation set. Least squares, kernel ridge regression, Gaussian process regression, and artificial neural networks are then employed to predict the T_g of the validation set.

Virtual experiments

To explore the independent variations of X_R and X_{CA} between 0 and 1, a grid approach with a step size of 0.01 was adopted for the virtual experiments. This step size balances the need for small expected differences between neighboring points and the desire to induce notable changes in T_g. Consequently, a grid of 10,201 (101 by 101) virtual experiments was established (see Figure 2).

Boundary conditions

From a materials science standpoint, it is clear that the aromatic or cyclic structures of DGEBA and IPDA result in stiff network segments. In contrast, the aliphatic structures of the bio-based components, lead to network segments with a high mobility, which decreases the T_g. Furthermore, the lower equivalent weights of the petroleum-based components compared to that of their bio-based counterparts result in an increased cross-link density. Both factors, network segment stiffness and cross-link density correlate well with the T_g. In conclusion, it is to be expected that the mixture with $X_R = X_{CA} = 1$ has the highest T_g and that the addition of bio-based components decreases T_g in a non-linear way. Given the lower temperature limit of the DSC for measuring T_g and the inverse relationship between T_g and bio-content, a boundary condition was imposed to exclude mixtures with a bio-content higher than 57.5%. This resulted in the elimination of 1978 virtual experiments with a foreseeable very low T_g, leaving 8223 virtual experiments in the data set. Subsequent optimization series focusing on a minimum bio-content of 5% and 10% further reduced the number of virtual experiments by an additional 71 and 195, respectively. The series optimizing for a minimum bio-content of 25% and a maximum price of 5.5Euro/kg considered 1325 virtual experiments.

2.5.2 | Bayesian optimization

Gaussian processes serve as surrogate models to approximate the T_g of the resin mixtures.[23] These GP models predict both the mean and standard deviation of virtual experiments through the application of a kernel function. The choice of acquisition function determines the utility for all virtual experiments. Subsequently, the mixture with the highest utility is recommended as the next sample to be measured and incorporated into the data set. The BO process is realized through the implementation of the BayesianOptimizer class within ModAL.[24]

Kernel functions

Several kernel functions were employed in this study:

1. The RBF kernel with a length scale of 0.2 and length scale boundaries ranging from 10^{-12} to 10^{15} was utilized. It is multiplied with a constant kernel having a constant value of 10^3 and constant boundaries from 10^{-3} to 10^4.
2. The Matérn kernel with a length scale of 0.2, length scale boundaries from 10^{-12} to 10^{15}, and ν set to 1.5 was employed. Similar to the RBF kernel, it is multiplied with a constant kernel having a constant value of 10^3 and constant boundaries from 10^{-3} to 10^4.
3. One DP kernel with a sigma value of 1.0 and sigma boundaries ranging from 10^{-4} to 10^2 is squared and multiplied by another DP kernel with the same parameters raised to the power of four. This product is further multiplied by a constant kernel with a constant value of 0.1 and constant boundaries from 10^{-4} to 10^2. This configuration is designed to mimic a higher-order polynomial

Acquisition functions

Two acquisition functions were employed:

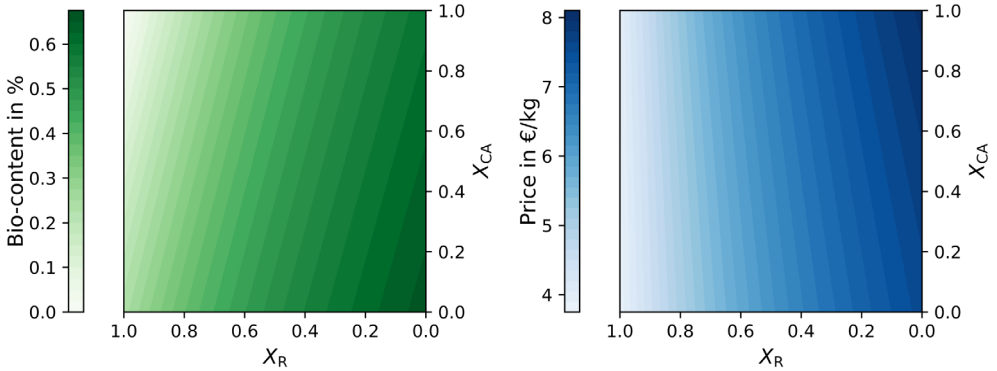

FIGURE 2 Left: bio-content of all the possible mixtures in the resin system. Right: price of the mixtures. [Color figure can be viewed at wileyonlinelibrary.com]

1. The UCB is defined as follows

$$\mathrm{UCB}(\mathbf{X}) = \mu(\mathbf{X}) + \beta\sigma(\mathbf{X}), \tag{4}$$

where $\mu(\mathbf{X})$ is the predicted mean of the GP for the virtual experiments $\mathbf{X}$, β is a hyperparameter controlling the trade-off between exploitation and exploration (set to 1 in this study), and $\sigma(\mathbf{X})$ represents the predictive standard deviation of the GP model.

2. The EI is defined as follows

$$\mathrm{EI}(\mathbf{X}) = \begin{cases} (\mu(\mathbf{X}) - f(\mathbf{X}^+) - \xi)\cdot\Phi(Z) + \sigma(\mathbf{X})\cdot\phi(Z) & \text{if } \sigma(\mathbf{X}) > 0 \\ 0 & \text{if } \sigma(\mathbf{X}) = 0 \end{cases}, \tag{5}$$

where

$$Z = \begin{cases} \dfrac{\mu(\mathbf{X}) - f(\mathbf{X}^+) - \xi}{\sigma(\mathbf{X})} & \text{if } \sigma(\mathbf{X}) > 0 \\ 0 & \text{if } \sigma(\mathbf{X}) = 0 \end{cases}, \tag{6}$$

where $\mu(\mathbf{X})$ is the predictive mean of the GP model, $f(\mathbf{X}^+)$ is the best observed function value so far, ξ is the exploration-exploitation trade-off (set to 0), $\Phi(Z)$ is the cumulative distribution function of the standard normal distribution evaluated at Z, $\phi(Z)$ is the probability distribution function of the standard normal distribution evaluated at Z and $\sigma(\mathbf{X})$ is the predictive standard deviation of the GP model at input $\mathbf{X}$.

2.5.3 | Active learning

Active Learning (AL) was integrated using the ActiveLearner class from ModAL, employing GPR as the estimator.[24] In each iteration cycle, the virtual experiment with the highest standard deviation was identified and selected for inclusion in the next iteration. This approach aimed to prioritize data points with higher uncertainty, contributing to the enhancement of the model's predictive capabilities.

2.5.4 | Models

Least squares

LS modeling was performed using scikit-learn 1.3.2.[25] The T_g was represented by a second-order polynomial incorporating terms such as 1, X_R, X_{CA}, X_R^2, $X_R X_{CA}$ and X_{CA}^2.

Kernel ridge regression

KRR was implemented using scikit-learn 1.3.2.[25] The T_g was modeled using a polynomial kernel, and α was set to 0.1, determined through grid search and CV.

Gaussian process regression

GPR was implemented using the BayesianOptimizer module from ModAL and two kernel functions were employed:

1. The first RBF kernel had a length scale of 0.1 and length scale boundaries ranging from 10^{-5} to 10^5. It was multiplied with a constant kernel having a constant value of 10^2 and constant boundaries from 10^{-3} to 10^4. Additionally, another RBF kernel with a length scale of 0.2 and similar boundaries was used, multiplied by a constant kernel with a constant value of 10^3 and constant boundaries from 10^{-3} to 10^4.
2. The second set of kernels included one DP kernel with a sigma value of 1.0 and sigma boundaries ranging from 10^{-6} to 10^6, squared, and multiplied by a constant kernel with a constant value of 10 and constant boundaries from 10^{-6} to 10^6. Another DP kernel with a sigma value of 2.0 and similar boundaries was used, squared, and multiplied by a constant kernel with a constant value of 10 and constant boundaries from 10^{-6} to 10^6.

Artificial neural network

The ANN was implemented using PyTorch 2.1. The normalized properties X_R and X_{CA} served as inputs. The ANN comprised five hidden layers, each with 256 neurons and using the ReLU activation function. The output of the ANN was the predicted T_g of the resin mixtures. The ANN was trained on the training set via the Adam algorithm using the mean squared error (MSE) as the loss function, with a learning rate of 0.001 until the MSE fell below 0.025. For the evaluation of the validation set, the model was then trained on the 35 data points with a learning rate of 0.0005 until MSE was smaller than 0.001.

Cross validation

For a more accurate estimate of each model's accuracy, seven-fold CV was performed 100 times. The coefficient of determination for the T_g of the training set $\left(R^2_{\text{train}}\right)$, along with the corresponding mean average error $(\text{MAE}_{\text{train}})$, were determined. Similarly, the models were used to predict the T_g of the test set, and R^2_{test} and MAE_{test} were calculated. Finally, the models trained on the initial 35 data points were used to predict the T_g of the validation set, determining R^2_{val} and MAE_{val}.

3 | RESULTS AND DISCUSSION

3.1 | Performance of different kernels and acquisition functions

In the initial optimization series, the objective is to optimize T_g without imposing a minimum bio-content (0%). Considering the aliphatic structures of the bio-based components, it is reasonable to expect a non-linear decrease in T_g concerning decreasing X_R and X_{CA}. Thus, all kernels and acquisition functions are tasked with finding the mixture with the maximum T_g which has $X_R = X_{CA} = 1$.

The top row of Figure 3 illustrates BO using the RBF kernel. Both acquisition functions, UCB and EI, require three BO rounds to find the fully petroleum-based mixture with $X_R = X_{CA} = 1$ and the maximum T_g. The RBF kernel initially proposes [1 0] and continues its search along the $[1\,X_{CA}]$ line. Interestingly, the data points proposed by both RBF UCB and RBF EI are identical.

The maximum T_g is about 145°C which is in range of literature values for T_g (144°C–153°C).[26,27] This as well as the absence of any exothermal heat flow during the second heating show that the resin mixtures are sufficiently cured. Note that the contour plot in Figure 3 is fitted using a second-order polynomial using all 40 points

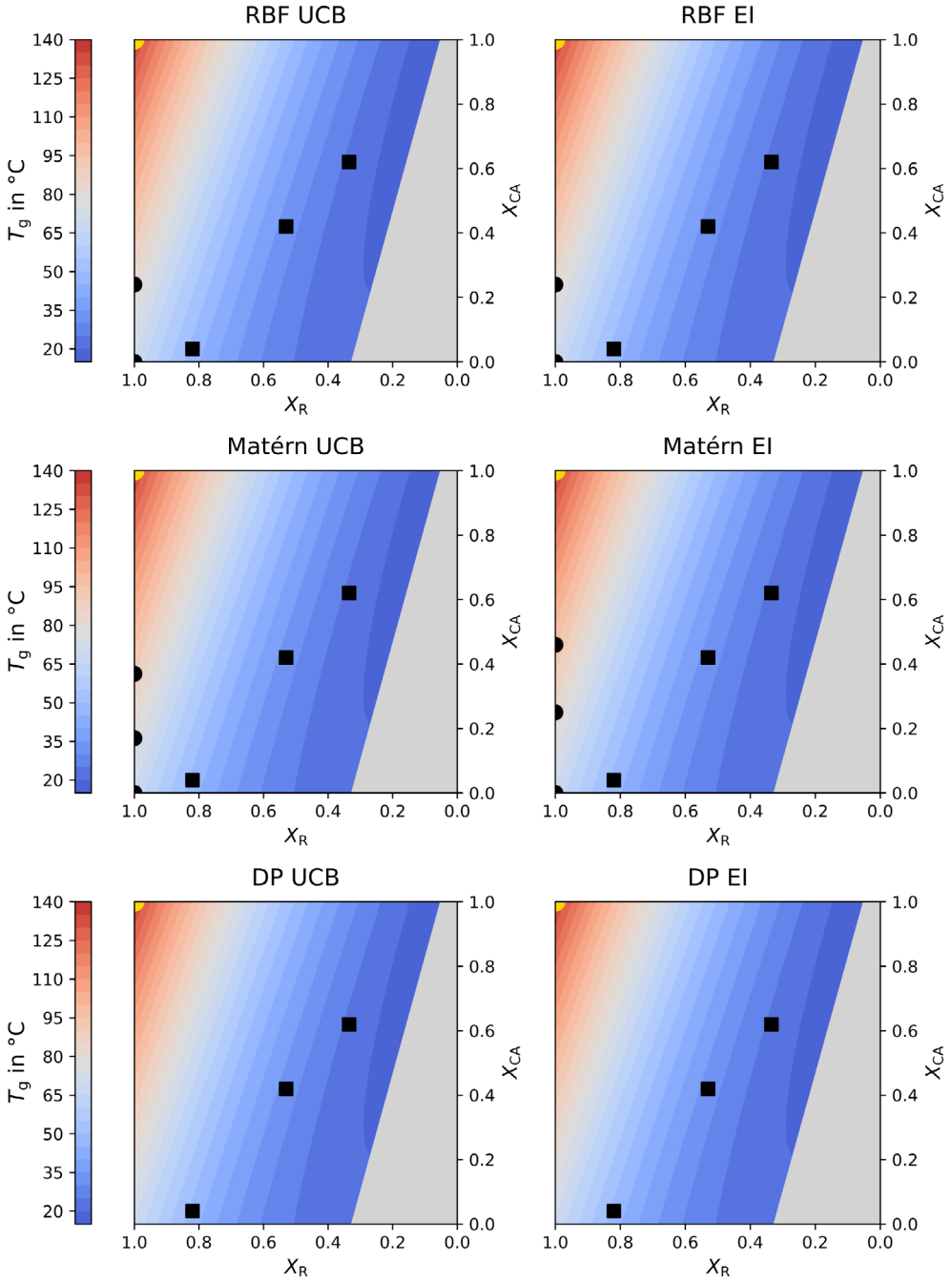

FIGURE 3 BO of T_g without a minimum bio-content using RBF (top), Matérn (center) and DP (bottom) kernel functions, in combination with UCB (left) and EI (right). Random data points are marked as squares (■) while data points proposed by the BO are marked as circles (•). The champions are marked as gold circles (●). [Color figure can be viewed at wileyonlinelibrary.com]

of the data set. However, this was only available after the different optimization runs were finished and the modeling done. It is only included for visual guidance.

The center row of Figure 3 displays BO using the Matérn kernel, which takes four rounds to find the optimum composition ([1 1]). Similar to the RBF kernel, the search initiates at [1 0], proceeding along the $[1\,X_{CA}]$ line.

In contrast, the DP kernel in the bottom row requires only one BO round to discover the maximum T_g. This underscores the importance of selecting kernel functions aligned with the assumed trends in the data based on materials science. As already mentioned, it was assumed that the decrease in T_g can be modeled with an n^{th}-order polynomial and that T_g decreases monotonously for an increasing bio-content. This assumption is confirmed by the modeling and experimental validation in Section 3.3. The efficient optimization is facilitated by reducing the feature space dimensions through normalization and informed consideration of the epoxy resin system's materials science.

Remarkably, all proposed mixtures have the form of $[X_R\,1]$ or $[1\,X_{CA}]$, avoiding the simultaneous presence of both bio-based components in the resin system. Rather, the bio-content is maximized in either the epoxy resin or the curing agent while it is minimized in the other. The same trend will be shown for the champions of the optimization series with minimum bio-contents of 5% and 10%, respectively.

The materials science explanation lies in the fact that the T_g depends on the network segment stiffness and cross-link density. Epoxy resins may only react with amine curing agents and vice versa, thereby linking resins to only curing agents and curing agents to only epoxy resins. Introducing one bio-based component weakens the network around that segment, limiting its influence because the bio-based network segment is connected to a petroleum-based component. On the other hand, introducing bio-based components into both the epoxy resin and curing agent allows the formation of connected bio-based network segments, decreasing local cross-link density and T_g. These findings emphasize the significance of informed choices in the optimization process, incorporating both materials science and data science for efficient and meaningful results.

3.2 | Optimizing for different boundary conditions

The second optimization series focuses on a resin system with a maximum T_g and a minimum bio-content of 5%. Continuing with the DP kernel function, as it proved efficient in the first series, the data points selected by each optimization approach in the initial series are utilized as inputs for the second series. The contour plot in Figure 4 is fitted using a second-order polynomial based on all 40 data points, offering visual guidance.

In this series, DP UCB, being more deterministic, identifies the maximum T_g in the first BO round, while DP EI, with a more exploratory approach, scans the extremes of the feature space and finds it in the second BO round. After five rounds of BO, the optimization is halted, as no additional high T_g mixtures are proposed. The final prediction, using all 40 data points, confirms that no mixture in the virtual experiments surpasses the identified maximum T_g.

High T_g mixtures in this dataset are predominantly close to 5% bio-content, supporting the assumption that the addition of aliphatic, bio-based components decreases T_g. The proposed mixtures continue to follow the pattern of $[X_R\,1]$ or $[1\,X_{CA}]$. The maximum T_g is approximately 128°C, achieved by [0.95 1], followed closely by [0.96 0.98] with a T_g of 126°C. Notably, the antagonism of interconnected bio-based network segments is evident, with [0.95 1] having a higher bio-content (5.7%) yet exhibiting a higher T_g than [0.96 0.98] (5.2% bio-content).

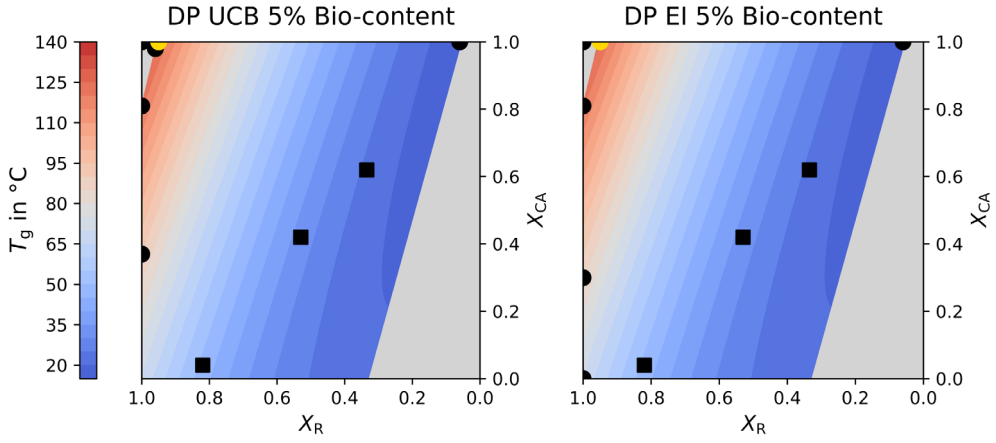

FIGURE 4 BO of T_g with a minimum bio-content of 5% using DP kernel functions, in combination with UCB (left) and EI (right). Random data points are marked as squares (■) while data points proposed by the BO are marked as circles (•). The champions are marked as gold circles (●). [Color figure can be viewed at wileyonlinelibrary.com]

This emphasizes the strategy of maximizing bio-content in the epoxy resin while minimizing it in the curing agent, maintaining an overall bio-content close to the minimum threshold of 5%.

The third optimization series follows a similar approach to the second one. However, due to the larger number of virtual experiments with a bio-content close to 10%, there are more possible champions, making it more challenging. To address this, AL is combined with BO in each optimization cycle, with one mixture proposed by BO and one by AL. Both mixtures are prepared and tested, with their T_g values used as inputs for BO and AL, respectively. The datasets of DP UCB and DP EI are treated separately. The contour plot in Figure 5 is fitted using a second-order polynomial based on all 40 data points, offering visual guidance.

DP UCB identifies the champion of the dataset with a T_g of 114.7°C at [0.91 0.99] (bio-content 10.2%). The mixture [0.90 1] (bio-content 10.9%) has a slightly lower T_g (113°C), approximately one standard deviation lower than that of [0.91 0.99]. The strategy of maximizing bio-content in one part of the resin system while minimizing it in the other, maintaining bio-content close to the minimum threshold of 10%, remains consistent.

In contrast, the EI acquisition function does not identify the same champion as DP UCB. Instead, it proposes [1 0.6] with a T_g of 101°C. Concluding from the results of the second and third optimization series, investing bio-content in the bio-based resin component instead of investing it in the curing agent is more effective and decreases T_g less severely. An examination of the chemical structure of the bio-based components reveals that the phenolic moiety in the bio-based epoxy resin NC-514 introduces steric hindrance, leading to stiffer network segments compared to the aliphatic bio-based curing agent Mergamid L450.

AL scans the feature space for virtual experiments with a high standard deviation resulting from the GPR. Typically, data points proposed by AL are far away in the feature space from already investigated points, focusing on the extremes of the feature space. This proves optimal for enhancing the dataset and exploring the feature space. After five rounds of BO, the optimization of this series is halted. The final prediction, using all 40 data points, confirms that no mixture in the virtual experiments surpasses the identified maximum T_g (Figure 5).

The last optimization series demonstrates how the investigated data points can efficiently design resin mixtures, considering realistic limitations on bio-content and price. For instance, the minimum bio-content is set to 25%, while the maximum price of the mixtures is set to 5.5Euro/kg. To expedite the process, the previously investigated data points of all kernel functions and acquisition functions are used as input ($n = 29$).

Figure 6 shows that DP UCB proposes the mixture [0.76 0.9] as the champion, which is closest to the [1 1] data point. This mixture has a T_g of 81.15°C and a bio-content of 25%. Notably, the GPR model predicts a T_g of 80.2°C for the mixture [0.76 0.9] which is about only 1°C off of its true T_g, showing the predictive capabilities of GPR models. Once again, the same trends regarding the composition of the mixture apply as for the optimization of mixtures with 5% and 10% bio-content. The bio-content is not evenly split between the epoxy resin and the curing agent; rather, it is minimized in the curing agent and maximized in the epoxy resin. Simultaneously, the total bio-content remains close to the minimum of 25% (see Figure 6).

After adding the [0.76 0.9] mixture to the dataset, BO proposes it as the mixture with the highest T_g. Consequently, BO cannot find a mixture with a higher T_g under the given criteria, indicating that the highest T_g in the dataset has already been identified.

3.3 | Modeling & experimental validation

The dataset, comprising 35 points, is modeled using different regression techniques, including LS, KRR, GPR

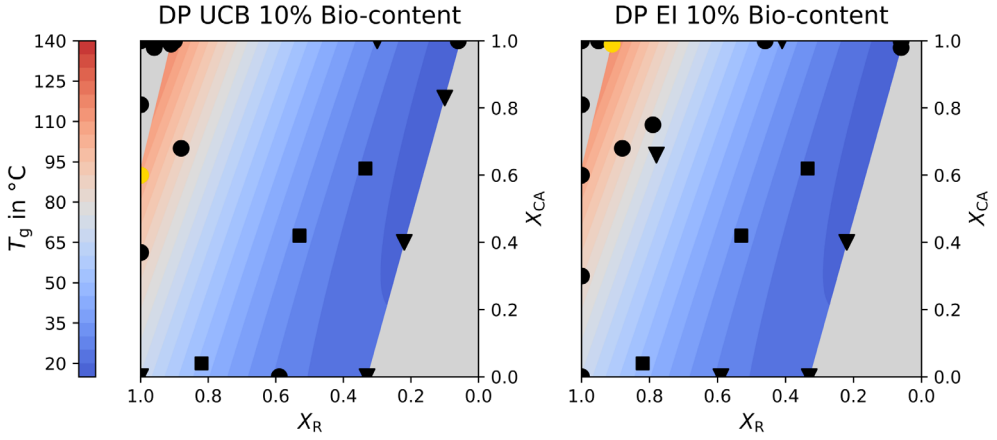

FIGURE 5 BO of T_g with a minimum bio-content of 10% using DP kernel functions in combination with UCB (left) and EI (right). Random data points are marked as squares (■), data points proposed by BO and AL are marked as circles (•) and (▼), respectively. The champions are marked as gold circles (●). [Color figure can be viewed at wileyonlinelibrary.com]

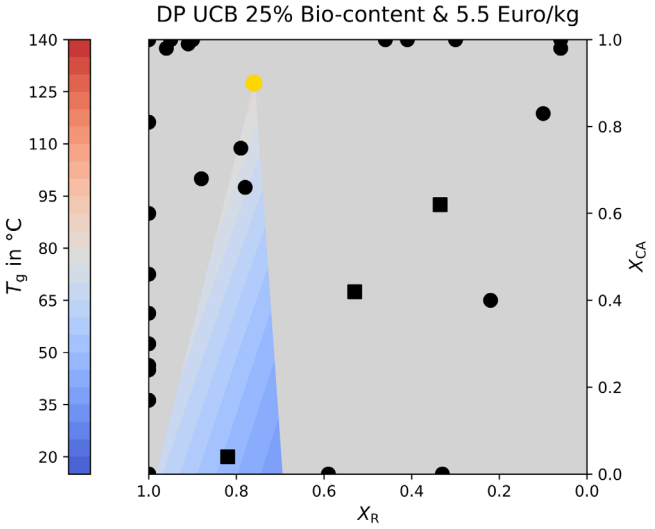

FIGURE 6 BO of T_g with a minimum bio-content of 25% and maximum price of 5.5 Euro/kg using DP UCB. Random data points are marked as squares (■), data points resulting from BO and AL are marked as circles (•). The champion is marked as gold circle (●). [Color figure can be viewed at wileyonlinelibrary.com]

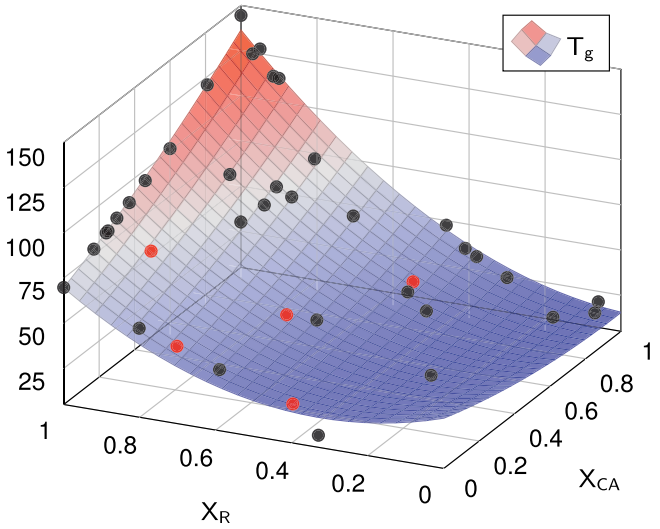

FIGURE 7 The 40 data points of the entire dataset from which 35 (•) were determined via random selection, BO or AL, while the remaining five data points belong to the validation set (●). [Color figure can be viewed at wileyonlinelibrary.com]

with RBF and DP kernels, and ANN. The models' predictive capabilities are evaluated based on their coefficient of determination (R^2) and mean average error (MAE) on the training, testing, and validation sets. Objective assessments of model accuracy involve seven-fold CV for each model, repeated 100 times. Figure 7 illustrates the location of the validation set in the feature space along with the dataset obtained from previous optimization series. While the validation set's points are randomly selected in the feature space, some points are relatively close to existing dataset points. However, selecting five points out of 8223 that are entirely distant from the initial 35 is improbable. Despite the apparent proximity, the gradients between the validation set points and the nearest dataset points may be substantial. The T_g of the investigated mixtures spans from 11.7°C ([0.33 0]) to 144.5°C ([1 1]).

Figure 8 shows the R^2 and MAE of the train, test and validation sets after 100 rounds of seven-fold CV using LS, KRR, GPR with RBF and DP kernels, and ANN. This breakdown provides a structured analysis of each model's performance, allowing for a clear comparison and discussion of strengths and weaknesses. The LS model exhibits outstanding performance on the training and validation sets, as evidenced by high R^2_{train} and R^2_{val} of 0.989 each. However, a noticeable drop in R^2_{test} (0.960) raises concerns about potential overfitting or a lack of generalization. Similarly, MAE_{train} is only about 2.64°C, while MAE_{test} is roughly 3.68°C. Despite the potential overfitting indicated by the test set, the relatively low MAE values, particularly on the validation set (1.39°C), suggest good predictive accuracy. Overall, the R^2 and MAE demonstrate a strong fit to the data, indicating that the T_g of mixtures not included in the dataset can be reliably predicted by the second-order polynomial. It's noteworthy that increasing the order of the polynomial from two to three or four improved the metrics for the training set but led to worse performance on the test and validation sets. This clear indication of overfitting underscores the justification for choosing a second-order polynomial.

The KRR model demonstrates high R^2 values on all sets, indicating a good fit to the data. The reduction in R^2 from R^2_{train} (0.986) to R^2_{test} (0.967) is smaller than with the LS model, suggesting that regularization has mitigated the risk of overfitting. The MAE values for the KRR model $\left(MAE_{val} = 2.68^{\circ}C\right)$ are higher compared to the LS model, indicating slightly lower predictive accuracy. However, the KRR model generalizes better to unseen data compared to the LS model, as evidenced by the smaller drop in R^2 between training and test sets.

The GPR RBF model exhibits a perfect R^2_{train} and MAE_{train}, indicating overfitting. GPR is highly flexible and can fit complex patterns in the training data. However, there is a noticeable drop in performance on the test and validation sets, suggesting challenges in generalizing to new data. The relatively higher MAE on the test and validation sets compared to LS suggests some limitations. The reason is likely that the bell shape of the RBF kernel function does not align with the expected trend (second-order polynomial) of the function underlying the data.

The GPR DP model exhibits high R^2 values across all sets, indicating a good fit to the data. Similar to the RBF kernel, this model performs well on the training set but

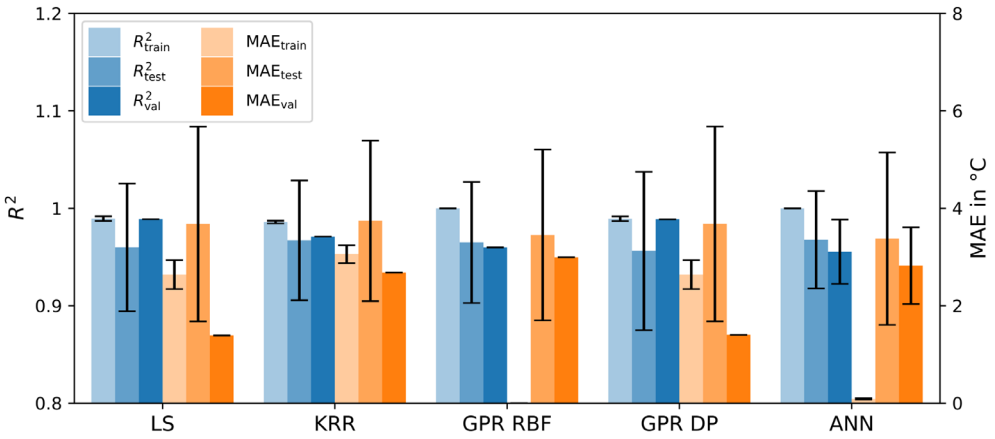

FIGURE 8 R^2 and MAE of the train, test and evaluation sets after 100 rounds of seven-fold CV using LS, KRR, GPR with RBF and DP kernels, and ANN. [Color figure can be viewed at wileyonlinelibrary.com]

faces a drop in performance on the test set $\left(R^2_{test} = 0.956\right)$. The slight drop in R^2 on the test set indicates some potential overfitting, although it's not as pronounced as in other models. The MAE values are comparable to the LS model $\left(MAE_{val} = 1.4^{\circ}C\right)$, suggesting good predictive accuracy.

The ANN shows exceptional performance on the training set with a near-perfect R^2_{train} of 1. ANN, especially with extensive layers, can capture intricate patterns in the training data. However, there is a drop in performance on the test and validation sets, indicating potential challenges in generalization $\left(R^2_{val} = 0.956\right)$. The model demonstrates generally low MAE values $\left(MAE_{val} = 2.82^{\circ}C\right)$, indicating good predictive accuracy, but there is an increase on the test and validation sets compared to the training set.

All models show strong performance on the training set, suggesting they can capture the training data well. However, there are signs of overfitting in some models, as evidenced by the decrease in performance on the test and validation sets. The LS model and GPR DP model have relatively lower MAE_{val} values, suggesting better predictive accuracy. Still, model selection remains context-dependent, considering trade-offs between overfitting and predictive accuracy.

The nature of the dataset, including its size and complexity, influences the model performance. The observed overfitting tendencies in some models may be attributed, in part, to the restricted data size, impacting the generalizability of the models. More complex models, like ANN, have a higher capacity to fit the training data but are prone to overfitting. To address potential overfitting, further exploration of techniques such as regularization or increasing the dataset size could be beneficial. Additionally, fine-tuning hyperparameters for models showing promise but exhibiting overfitting tendencies might enhance overall performance in future studies.

4 | CONCLUSION AND OUTLOOK

In summary, this investigation exemplifies an effective methodology for optimizing resin mixture formulations within the constraints of bio-content and cost considerations. Key strategies, such as utilizing DSC for characterization, a parsimonious selection of components, and the application of normalized features, contribute to the success of the optimization process. The incorporation of specific kernel functions aligned with principles from materials science further bolster efficiency. The optimization of a four-component epoxy resin system is successfully achieved with a modest investment of 5 days in measurement time and a mere 30 data points. Notably, the optimization series reveals a trend where champions sought to maximize bio-content in either the resin or curing agent while adhering to the minimum bio-content criteria. This phenomenon is explicated by the interconnection of compliant network segments formed by the bio-based components, exerting an influence on the T_g. This observation prompts inquiry into the necessity of multiple bio-based components in both epoxy resins and curing agents. While such complexity does not confer an advantage in maximizing T_g, it may prove beneficial for optimizing other mechanical or thermal properties. Modeling the dataset using various regression techniques and ANN provides valuable insights, with LS and GPR DP demonstrating robust performance with the limited dataset. However, continuous improvement is possible with the addition of more data points and fine-tuning of hyperparameters. The study concludes by highlighting the potential application of proposed strategies to optimize other key properties or consider additional boundary conditions, such as CO_2 footprint or ecological impact. Future studies may focus on aspects not extensively addressed in this investigation, paving the way for further advancements in resin mixture optimization.

AUTHOR CONTRIBUTIONS

Florian Rothenhäusler: Conceptualization (lead); data curation (lead); formal analysis (lead); investigation (lead); methodology (lead); software (lead); validation (lead); visualization (lead); writing – original draft (lead). **Holger Ruckdaeschel:** Project administration (lead); supervision (lead); writing – review and editing (lead).

ACKNOWLEDGMENTS

On the occasion of the 261th anniversary of the publication of "An Essay towards solving a Problem in the Doctrine of Chances," the authors would like to thank Thomas Bayes for his ground-breaking work in the field of statistics which is the foundation for the Bayesian optimization, one of the most efficient tools for solving optimization problems today. We would like to thank Rodrigo Queiroz de Albuquerque for his invaluable contributions and his relentless effort to teach and support the Department of Polymer Engineering in the field of machine learning. Open Access funding enabled and organized by Projekt DEAL.

FUNDING INFORMATION

Parts of the research documented in this manuscript have been funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) within the research project "EcoPrepregs—Grundlagenforschung zur Klärung der Struktur-Eigenschaftsbeziehungen von Epoxidharzen und Fasern aus nachwachsenden Rohstoffen zur Anwendung in der Sekundärstruktur von Flugzeugen" (grant #20E1907A, Germany).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Florian Rothenhäusler https://orcid.org/0000-0001-9948-3310

Holger Ruckdaeschel https://orcid.org/0000-0001-5985-2628

REFERENCES

[1] G. W. Ehrenstein, *Faserverbund-Kunststoffe*, Carl Hanser Verlag GmbH & Co. KG, Munich **2006**.

[2] H. Lengsfeld, V. Altstädt, F. Wolff-Fabris, J. Krämer, *Composite Technologien*, Carl Hanser Verlag GmbH & Co. KG, München **2014**. https://doi.org/10.3139/9783446440807

[3] S. Kumar, S. K. Samal, S. Mohanty, S. K. Nayak, *Polym.-Plast. Technol. Eng.* **2018**, *57*, 133.

[4] D. Welsby, J. Price, S. Pye, P. Ekins, *Nature* **2021**, *597*, 230.

[5] R. P. Wool, in *Bio-Based Polymers and Composites* (Eds: R. P. Wool, X. S. Sun), Academic Press, Burlington **2005**, p. 202. https://www.sciencedirect.com/science/article/pii/B9780127639529500083

[6] C. F. Frias, A. C. Serra, A. Ramalho, J. F. J. Coelho, A. C. Fonseca, *Ind. Crops Prod.* **2017**, *109*, 434.

[7] F. C. Fernandes, K. Kirwan, D. Lehane, S. R. Coles, *Eur. Polym. J.* **2017**, *89*, 449.

[8] A. Mora, D. Mélanie, G. David, S. Caillol, *Eur. J. Lipid Sci. Technol.* **2019**, *06*, 121.

[9] E. Kinaci, E. Can, J. J. L. Scala, G. R. Palmese, *Polymer* **2020**, *12,* 1956.

[10] Cardolite Corporation, Epoxy resins, diluents and modifiers. **2023** https://www.cardolite.com/products/epoxy-modifiers/ (assessed: August 2023).

[11] R. Q. Albuquerque, F. Rothenhäusler, H. Ruckdäschel, *MRS Bull.* **2024**, *49*, 59.

[12] S. Pruksawan, G. Lambard, S. Samitsu, K. Sodeyama, M. Naito, *Sci. Technol. Adv. Mater.* **2019**, *20*, 1010.

[13] J. R. Deneault, J. Chang, J. Myung, D. Hooper, A. Armstrong, M. Pitt, B. Maruyama, *MRS Bull.* **2021**, *46*, 566.

[14] K. Park, Y. Kim, M. Kim, C. Song, J. Park, S. Ryu, *Compos. Sci. Technol.* **2022**, *220*, 109254. https://www.sciencedirect.com/science/article/pii/S0266353821006102

[15] R. Q. Albuquerque, F. Rothenhäusler, P. Gröbel, H. Ruckdäschel, *ACS Appl. Eng. Mater.* **2023**, *1*, 11.

[16] X. Xu, W. Zhao, L. Wang, J. Lin, L. Du, *Chem. Sci.* **2023**, *14*, 10203.

[17] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, N. de Freitas, *Proc. IEEE* **2016**, *104*, 148.

[18] James G, Witten D, Hastie T, Tibshirani R, Taylor J. *An Introduction to Statistical Learning: With Applications in Python,* Springer, Berlin **2023**.

[19] M. J. Richardson, N. G. Savill, *Polymer* **1975**, *16*, 753.

[20] R. P. Chartoff, P. T. Weissman, A. Sircar, *ASTM Spec. Tech. Publ.* **1994**, *1249*, 88.

[21] S. E. Keinath, R. F. Boyer, *J. Appl. Polym. Sci.* **1981**, *26*, 2077.

[22] D. Duvenaud, Automatic model construction with Gaussian processes. PhD thesis. **2014**.

[23] C. E. Rasmussen, C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, USA **2006**.

[24] T. Danka, P. Horvath, modAL: A modular active learning framework for Python. https://github.com/cosmic-cortex/modAL, https://arxiv.org/abs/1805.00979

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, *J. Mach. Learn. Res.* **2011**, *12*, 2825.

[26] N. Thuc Boi Huyen, A. Maazouz, *Polym. Int.* **2004**, *05*, 591.

[27] E. Mounif, V. Bellenger, P. Mazabraud, F. Nony, A. Tcharkhtchi, *J. Appl. Polym. Sci.* **2010**, *116*, 969.

How to cite this article: F. Rothenhäusler, H. Ruckdaeschel, *J. Appl. Polym. Sci.* **2024**, *141*(21), e55422. https://doi.org/10.1002/app.55422