Designing formulations of bio-based, multicomponent epoxy resin systems via machine learning

Rodrigo Q. Albuquerque, Florian Rothenhäusler, and Holger Ruckdäschel*

Petroleum-based epoxy resins are commonly used as a matrix in fiber-reinforced polymer composites. Bio-based epoxy resin systems could be a more environmentally friendly alternative to conventional epoxy resins. In this work, novel formulations of multicomponent, amino acid-based resin systems exhibiting high or low glass-transition temperatures (T_g) were designed via Bayesian optimization and active learning techniques. After only five high-T_g experiments, thermosets with T_g already higher than those of the individual components were obtained, pointing out the existence of synergistic effects among the amino acids used and confirming the efficiency of the theoretical design. Linear and nonlinear machine learning (ML) models successfully predicted T_g with a mean absolute error of 3.98°C and R^2 score of 0.91. A price reduction of up to 13.7% was achieved while maintaining the T_g of 130°C using an optimized formulation. The LASSO model provided information about the dependence of T_g on the number of active hydrogen atoms and aromaticity. This study highlights the importance of Bayesian optimization and ML to achieve a more sustainable development of epoxy resin materials.

Impact statement

This article shows how the sustainability of epoxy resin systems (ERSs) can be significantly improved by combining experimental and theoretical strategies. First, amino acids are used as curing agents in multicomponent formulations to produce bio-based ERSs. Second, the number of trial-and-error experiments required to obtain formulations with high or low glass-transition temperatures (T_g) is greatly reduced using machine learning (ML) strategies to design all experiments. Not only is it shown how T_g can be maximized in only five new theoretically designed formulations, but the economic advantages of the proposed approach are also discussed. The trends between T_g and the type of optimized biocomponents are discussed based on the unambiguous interpretation of the best-trained ML model. The results presented in this study pave the way for the theoretical design of more sustainable polymeric materials.

Introduction

Fiber-reinforced polymer composites are an important class of materials for lightweight structures due to their high weight specific modulus and strength.[1,2] Epoxy resins are commonly used as a matrix for fiber-reinforced polymer composites due to their low viscosity, good storability, and high glass-transition temperature (T_g).[3] Here, the T_g of the matrix is a crucial property as it determines the composite's maximum service temperature as well as the matrix' modulus and heat resistance.[4] However, epoxy resins and many of their curing agents, such as amines, anhydrides, and phenolic compounds, are harmful in case of skin contact or when ingested.[5–8] Furthermore, these compounds are derived via chemical processes from petroleum, which causes considerable CO_2 emissions. In addition, making materials more sustainable can help slow down climate change.

Sustainability during the design phase of thermoset formulations could be achieved by different means. First, petroleum-based components can be substituted or combined with bio-based components. For example, petroleum-based amine curing agents for epoxy resins can be substituted with amino acids, such as L-tryptophan.[9–14] Other amino acids were used in similar ways, as reported by Shibata et al.,[15] who investigated among other things, the thermo-mechanical and tensile properties of an epoxidized sorbitol polyglycidyl ether cured with L-cysteine, L-arginine, or L-lysine.

Rodrigo Q. Albuquerque, Department of Polymer Engineering, University of Bayreuth, Bayreuth, Germany; Neue Materialien Bayreuth GmbH, Bayreuth, Germany
Florian Rothenhäusler, Department of Polymer Engineering, University of Bayreuth, Bayreuth, Germany
Holger Ruckdäschel, Department of Polymer Engineering, University of Bayreuth, Bayreuth, Germany; Neue Materialien Bayreuth GmbH, Bayreuth, Germany; holger.ruckdaeschel@uni-bayreuth.de
*Corresponding author
doi:10.1557/s43577-023-00504-9

Rothenhäusler et al.[16,17] studied the glass-transition temperature, viscosity, and latency of a diglycidyl ether of bisphenol A (DGEBA) cured with L-arginine and its mechanical properties at different temperatures, as well as the mechanical properties of DGEBA cured with five other amino acids.[18] The mechanical performance of the resulting thermosets was slightly lower than that of their amine-cured counterparts.

There are a wide variety of amino acids with aliphatic, cyclic, or aromatic structures.[19] Aliphatic amino acids, such as L-arginine, L-citrulline, and L-glutamine, could possess numerous active hydrogen atoms. In contrast, L-tryptophan and L-tyrosine have only few active hydrogen atoms but incorporate large aromatic rings that are useful for achieving high T_g.[20] As amino acids have widely different structures, the combination of different amino acids as curing agents in one single material could be advantageous. Amino acids could react with one another via peptide reaction[21] (see **Figure 1**) to form a curing agent that possesses both numerous active hydrogen atoms as well as aromatic structures. Thus, there could be potential for synergistic effects when combining certain amino acids in distinct ratios.

Finding the optimal solution for one or more material properties when formulating new resin systems by trial and error is inefficient, time-consuming, and cost-intensive.[22] However, this can be overcome using machine learning (ML), which helps shorten material design phases.[20,23]

Pruksawan et al.[24] described a method for the optimization of epoxy-based adhesives with a small data set and four variables via active learning and Bayesian optimization. The tested thermosets consisted of one resin and one curing agent and the investigation was focused on optimizing the curing conditions and the epoxy amine ratio. In that work, the Bayesian optimization was performed after 47 experiments to find an adhesive joint strength ca. 13% higher than the largest property measured in the previous experiments.

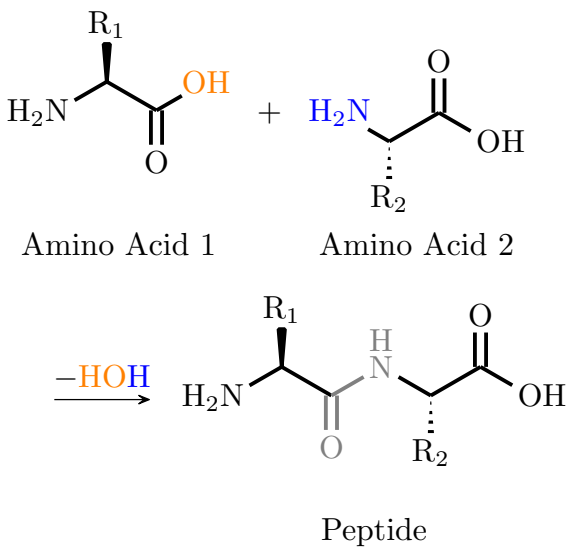

Figure 1. Peptide bond formation (gray) between two α-amino acids. The hydroxyl group of the carboxyl group of one amino acid reacts with one of the hydrogen atoms of the amino group of another amino acid.[21]

Similarly, Kang et al.[25] used an artificial neural network (ANN) for the prediction of lap shear strength and impact peel strength of epoxy adhesives. They analyzed the influence of the thermoset composition (weight ratios of resin, filler, curing agent, and flexibilizer) on the resulting mechanical properties. With 50 datapoints for lap shear strength and impact peel strength each, the ANNs did not show a high performance, with R^2 of 0.642 and 0.588, respectively.

Another ML approach described in the literature to predict T_g of one-component epoxy resin systems based on the chemical structure of the molecular units of the mixture has been recently published by Ruckdäschel et al.[20] After generating ca. 1800 molecular descriptors, feature selection was used to get the most important ones, from where an ML ensemble model was trained to predict T_g, giving R^2 and mean absolute error of 0.86 and 16.15°C, respectively, for the test set.

Ramprasad et al.[26] have reported a virtual experiment using different active learning (AL) strategies to screen 736 different polymers to find high-T_g ones. In this investigation, more than 100 local and global descriptors related to the chemical structure and morphology of different polymers were used as (fingerprint) features and T_g as the target property to train a ML model. The model showed an R^2 score of 0.66 for the comparison of experimental and predicted T_g using a data set of 42 samples.

To the best of our knowledge, the ML-based prediction and optimization of mechanical properties of bio-based multicomponent thermosetting systems were not yet addressed in the literature. Therefore, the aim of this investigation is to optimize and predict the glass-transition temperature of DGEBA cured with a mixture of seven amino acids, whose reaction is accelerated by a substituted urea. The goal is to check whether ML models can help find the maximum and minimum T_g in the nine-component system and predict T_g for randomly chosen mixtures with as few experiments as possible. Despite starting with only a couple of experiments, a very efficient Bayesian optimization can still be performed to achieve novel thermosets with optimized properties. Economical considerations come naturally from the great diversity of designed formulations exhibiting similar T_g, as will be shown in the results.

Materials and methods

Materials

D.E.R. 331 with an epoxide equivalent weight of 187 g mol^{-1} was purchased from Blue Cube Assets GmbH & Co. KG, Olin Epoxy (Stade, Germany). L-Arginine (purity 98.9%), L-citrulline, γ-aminobutyric acid (GABA) (purity 100%), L-glutamine (purity 100%), L-proline (purity 100%), L-tryptophan (purity 100%), and L-tyrosine were purchased from Buxtrade GmbH (Buxtehude, Germany). The reaction

Table I. Molecular weight (M_w), number of active hydrogen atoms (f), and resulting amine equivalent weight (AEW) of the amino acids used as curing agents.

Amino Acid	M_w (g mol^{-1})	f	AEW (g mol^{-1})
L-arginine	174.2	7	24.89
L-citrulline	175.2	6	29.20
GABA	103.1	3	34.37
L-glutamine	146.2	5	29.24
L-proline	115.1	2	57.55
L-tryptophan	204.2	4	51.05
L-tyrosine	181.2	3	60.40

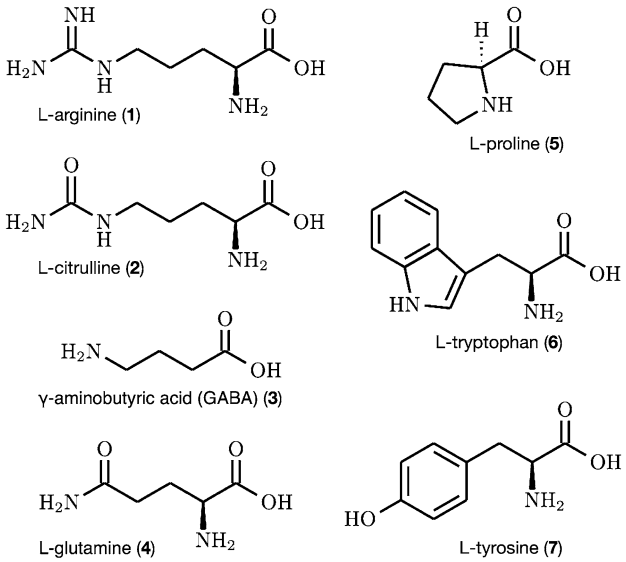

Figure 2. Molecular structures of the amino acids used as curing agents.

between epoxy resin and amino acids is accelerated by DYHARD UR400, a substituted urea, which was bought from Alzchem Group AG (Trostberg, Germany). The curing agents' molecular weight (M_w), number of active hydrogen atoms (f), and resulting amine equivalent weight (AEW) are shown in **Table I**.

Resin formulation

For each amino acid, one epoxy amino acid masterbatch was prepared via three-roll milling of the resin amino acid mixture (**Figure 2**). The preparation follows the procedure already applied and described in the literature.[16] All seven masterbatches are prepared so that the stoichiometric ratio R of epoxy groups to active hydrogen atoms is equal to 1. For each experiment, the masterbatches are weighed according to the ratios P_n of the respective experiment. Here, the ratios correspond to the percentage of epoxy groups that react with the hydrogen atoms of the amino acid of the respective masterbatch. Thus, the sum of the ratios P_n is equal to one (see Equations 1 and 2).

$$\mathbf{P} = \left[P_1\ P_2\ P_3\ P_4\ P_5\ P_6\ P_7\right], \quad (1)$$

$$\sum_{n=1}^{7} P_n = 1. \quad (2)$$

For example, in a formulation with the ratios

$$\mathbf{P} = \left[0.5\ 0.4\ 0.1\ 0\ 0\ 0\ 0\right],$$

half of the epoxy groups ideally would react with L-arginine, 40% with L-citrulline, and 10% with GABA. After weighing in the corresponding weight ratios of the masterbatches, one weight percentage of the accelerator (DYHARD UR400) was added before mixing in a centrifuge speed mixer by Hauschild Engineering (Hamm, Germany) at 3000 min^{-1} for 120 s. The mixture was degassed for 60 min at 10 mbar to ensure the elimination of entrapped air prior to curing.

Curing cycle and sample preparation

The amino acid epoxy mixture was poured into aluminum molds that were preheated at 90°C. Afterward, the material systems were cured for 2 h at 120°C and 2 h at 170°C and cooled down to room temperature over 4 h in a Memmert ULE 400 convection oven from Memmert GmbH + Co. KG (Schwabach, Germany). Dynamic mechanical analysis (DMA) specimens were prepared from the cured plates according to standard ISO 6721-7 with a Mutronic DIADISC5200 diamond plate saw from MUTRONIC Präzisionsgerätebau GmbH & Co. KG (Rieden am Forggensee, Germany).

Thermal characterization

Glass-transition temperature T_g was determined via DMA on specimens with dimensions 50 mm × 10 mm × 2 mm using a Rheometrics Scientific ARES RDA III from TA Instruments Inc. (New Castle, Del., USA). Here, a shear strain amplitude of 0.1% with a frequency of 1 Hz was applied during heating with a rate of 3 K min^{-1}. For this investigation, the temperature of the peak value of the loss factor tan δ was chosen as T_g. Two specimens were tested per formulation and their average T_g taken as target property for the ML modeling.

Design of experiments

Initially, the ratios P_1 to P_7 of the amino acids used in five different formulations were randomly generated and the corresponding materials were subsequently prepared. After each material preparation, T_g was measured. The initial data set had five samples (or formulations), each described by seven features (P_1–P_7) and one target property (T_g). The Bayesian optimization was then performed twice using Gaussian process regression (GPR), where new samples were queried using 10^6 virtual experiments to suggest the next two formulations to be prepared: one formulation to maximize T_g and another one to minimize T_g. In real applications, one would either maximize or minimize T_g, but both situations were investigated here as a proof of concept.

The two new suggested formulations were used to prepare the corresponding materials and their T_g was measured.

The data set was then updated with these new datapoints and this procedure continued, being intercalated with AL (see the "Active learning" section). After achieving 29 samples, six new samples were added to the data set to improve the final model. These last samples were selected from the virtual experiments using kernel ridge regression (KRR), which screened samples that could exhibit T_g in the range of 80–100°C because the current data set was composed mostly of samples with higher T_g (>100°C).

Models

The ML models were built using the Scikit-learn library.[27]

Different models were screened using all 35 samples of the current data set. Default hyperparameters were used in each model (see the Supplementary information), unless stated otherwise. The hyperparameter optimization is tricky for such a small data set and would involve splitting it into a training set, a validation set, and a test set. This is expected to generate models with very high variance, as shown in the "Results and discussion" section.

The models were evaluated using *k*-fold cross-validation (CV), which was repeated for all values of *k* in the range of 2–10, from where the best *k* parameter was obtained for each model.

The mean absolute error (MAE) and the coefficient of determination (R^2 score) were used as model evaluation metrics. MAE is given by

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i |, \qquad 3$$

where n is the number of samples and y_i and $\hat{y}_i$ are the true and predicted target property, respectively, for sample i. R^2 is the quotient of the explained variance to the total variance in a regression model.

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \qquad 4$$

Overfitting was evaluated by training ML models using the training set and performing predictions on both the training and test sets, from where their prediction errors (MAE) were compared. The investigated models are briefly summarized below and are also described in more detail in Reference 23.

Gaussian process regression (GPR). GPR uses a multivariate Gaussian fitted on the data set to perform predictions on new data. One usually adopts for the GPR a mean of zero and the covariance matrix given by a kernel function.[28] Predictions are also described by a Gaussian distribution, from where one readily gets the corresponding mean and the standard deviation, automatically giving uncertainty values for the predictions.

Kernel ridge regression (KRR). This method uses an L2 regularization term and the so-called kernel trick to make predictions. Regularization means that larger weight coefficients in the linear combinations of features are penalized more than smaller ones. With L2 regularization, the penalty is proportional to the square of those coefficients and the latter tend to become small, but not necessarily zero.[29,30]

K-nearest neighbors (KNNs). The predictions are based on the similarity between datapoints, which is often calculated using the simple Euclidean distance between them. This method is very efficient and is considered nonparametric because no real training is required, only the distances between datapoints are computed.[31]

Gradient boosting regression (GBR). GBR builds an additive model stepwisely, allowing the optimization of arbitrary differentiable loss functions. At each stage, a regression tree is fitted to the negative gradient of the current loss function. This method uses the gradient descent technique to add new estimators (in this case, regression trees) one at a time to create an optimized ensemble model.[32]

Support vector regression (SVR). The general idea of SVR is to find the best hypertube (defined by the weighting coefficients and the bias) passing through most of the samples in the data set, where the maximum acceptable deviation from the target property is given by the positive parameter ε: most of the samples are therefore inside a multidimensional ε-tube (also called an ε-insensitive tube).[33]

Least squares (LS). This method finds a linear combination of the features that minimizes the sum of squares of the errors between the true and the predicted target property. By default, the LS model has no regularization term and is one of the simplest models to build. The weighting coefficients of the linear combination are found by minimizing the loss function, which is the mean squared error.[34]

LASSO. This is basically an LS model with an L1 regularization term that penalizes large weighting coefficients via a term that is linear on the weighting coefficients themselves. The L1 regularization is particularly useful in the context of feature selection, as it tends to favor solutions with fewer nonzero coefficients, effectively reducing the number of features upon which the given solution depends.[35]

Random forests (RFs). This model averages the predictions of many uncorrelated decision trees, each of which considering different (randomly generated) subsets of the features and samples. Each decision tree consists of a sequence of simple rules, each based on a single feature. After all uncorrelated trees are grown, the predicted target property of any sample is calculated by simply averaging the predictions for that sample using all trees.[36]

Active learning

Active learning (AL) is an excellent tool to choose the next sample to increase the size of the current data set aiming at enhancing the predictive capability of the ML model in use. The simplest way to perform AL is by training different models with randomly chosen subsets (bootstraps) of the original data set and using these models to predict the target properties of the same sample. The best sample (out of

the pool of virtual experiments) that is chosen to be added to the data set is the one for which the average prediction exhibited the largest standard deviation or uncertainty. This technique is called uncertainty sampling.[37] In other words, if the ML models are not certain about the prediction for a given sample, this means that adding this sample to the data set would help the model to describe new situations that it was not able to describe before. All AL steps carried out in this work were done using the KRR model with default hyperparameters and 10 bootstraps with size of 70% of the data set.

Bayesian optimization

A GPR model was used as the regressor in the Bayesian optimization approach. The Matérn kernel, which is a generalization of the radial basis function kernel, was used as the kernel for the GPR model. After training the GPR model using the training set, predictions were performed on all virtual experiments (each virtual experiment was a formulation [i.e., a 7D normalized vector]) and the mean and standard deviation of the predictions were used to build an acquisition function (here, the maximum expected improvement) from where the next sample was selected. This procedure gave samples with potentially high T_g. To find samples with low T_g, $-T_g$ was used as the property to be maximized.

Figure 3 shows how new experiments were suggested via Bayesian optimization. In the case of AL-suggested experiments, the next virtual sample chosen is simply the one with highest uncertainty (SD).

Experimental evaluation

The first experimental evaluation was performed by continuously training successive models with increasingly large data sets in the following way. The predicted T_g for sample 6 was obtained from a fresh ML model trained with all previous samples (1–5) and compared with the experimental T_g for that sample. Then, samples 1–6 were used to train another fresh ML model to predict T_g for sample 7, which was then compared with the T_g measured for that sample and so on, until T_g of sample 35 was predicted using a model trained with samples 1–34. This was done with all investigated models described in the "Models" section.

The second experimental evaluation was performed by training a KRR model using the first 29 samples to screen the virtual experiments to find six new samples exhibiting T_g in the range of 80–100°C. The predicted and experimental T_g for these samples were then compared.

The last experimental evaluation was performed by randomly choosing five experiments from the pool of virtual ones and predicting T_g for all of them using the investigated models.

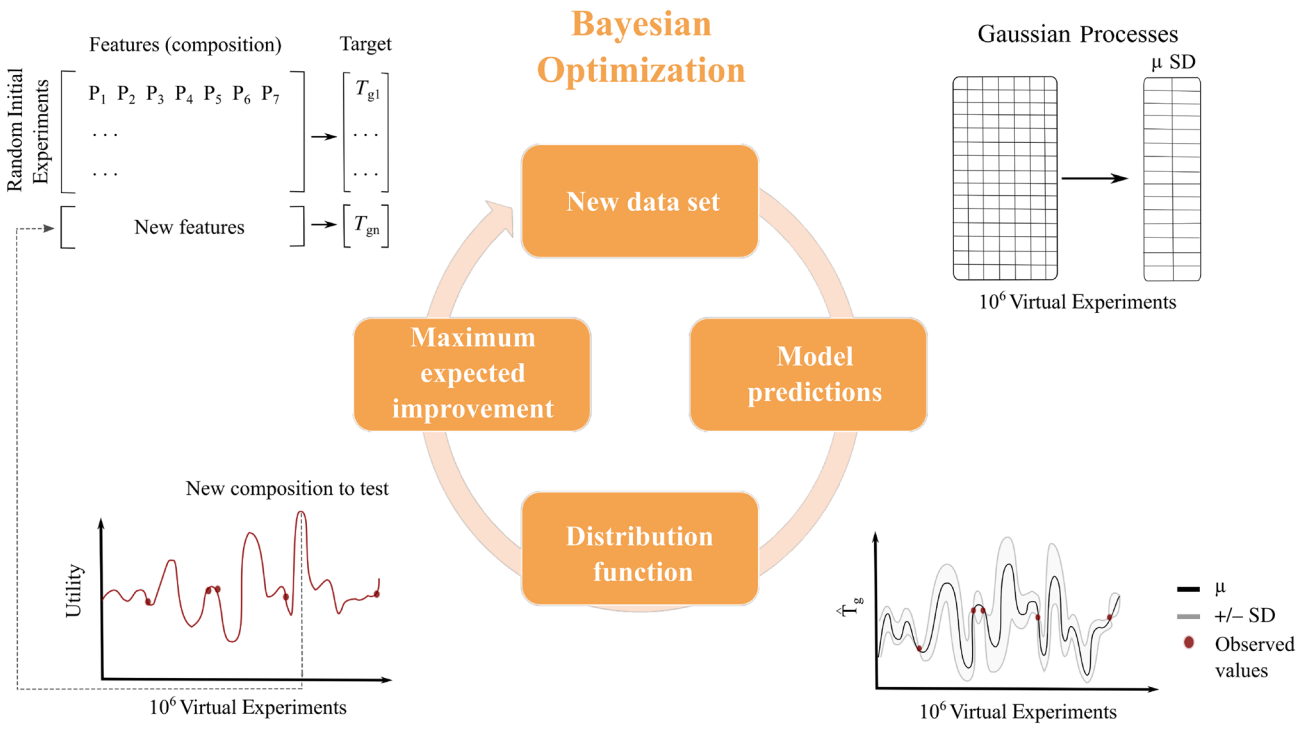

Figure 3. Workflow used in the Bayesian optimizations. Starting with an initial, random data set (top, left), a Gaussian process regression model is trained and used to perform predictions (μ) and uncertainties (SD) for a large virtual data set (top, right), which is graphically shown below (bottom, right). μ and SD are combined to obtain the utility for all virtual samples (bottom, left). The virtual sample with the highest utility is experimentally measured and added to the initial data set.

These experimental validations are discussed in the "Results and discussion" section.

Results and discussion

Designed experiments

The strategy used to obtain extreme T_g formulations that could be concomitantly used to create/train a good ML predictive model with as few experiments as possible is shown in **Figure 4**. Region I ("Rdn") contains five different experimental formulations (samples 1 to 5), which were randomly selected from the pool of 10^6 virtual formulations, also called virtual experiments. These initial samples were used to train two GPR models to perform predictions of T_g on all the virtual experiments, from where the corresponding mean and variance were used as the basis for the Bayesian optimization procedure (region II, "BO"), which finally suggested formulations leading to high (triangles) or low (squares) T_g. Note that for the prediction of formulations with maximum T_g, the Bayesian optimization consisted of exploitation steps only (T_g increased monotonically), whereas in the case of predictions related to low T_g, exploitation and exploration steps were observed (i.e., the model has also suggested formulations not having extreme [low] T_g whenever this has led to a substantial enhancement of the model's predictive capability), which happens after exploring/visiting new regions of the seven-dimensional configurational space of the formulations. In region II, the Bayesian optimization was already able to maximize T_g with only few experiments. The high T_g of sample 13 (131°C) seemed to be a (local) maximum. Due to the very large number of possible combinations between seven different amino acids and to the fact that the true function that defines T_g is not known, it cannot be confirmed that this is a global maximum. No convergence of low T_g was found in region II (squares) after carrying out 14 experiments in total. To explore different regions of the multidimensional formulation space and to check other possible local maxima, as well as to achieve lower T_g, three new experiments suggested via AL (region III) were performed. As expected, after experiments 15–17, a new round of Bayesian optimization (region IV) was then able to find a much lower T_g via an exploitation step (sample 19, square), while no T_g higher than that of experiment 13 was suggested (sample 18, triangle) as a result of an exploration step.

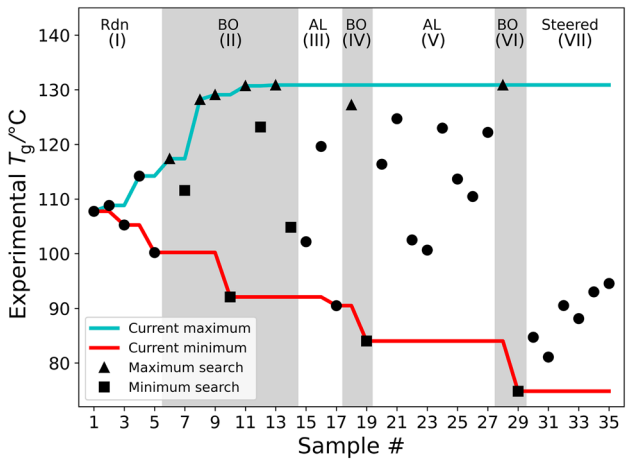

Figure 4. Measured T_g of formulations suggested via random drawings (Rdn), Bayesian optimization (BO), active learning (AL), and steered drawings, which are separated into regions I to VII. The steered drawings suggested formulations that could possess T_g in the range of 80–100°C.

Once the strategy of using AL steps before the Bayesian optimization has proven to be efficient, a new AL round was again performed (region V), this time suggesting eight new experiments. After that, a last Bayesian optimization was done (region VI), which found two formulations (samples 28 and 29) exhibiting high and low T_g, respectively.

At this stage, the data set consisted of 29 formulations and their corresponding T_g values. Although these few experiments have already met one of the goals of this investigation that was to find new formulations having high and low T_g, training a ML model with so few datapoints would not give a very accurate model or allow to perform a good model evaluation. Because most of the datapoints concentrated at higher T_g (>100°C), a steered-based procedure was performed (region VII), where a KRR model trained with the whole data set was used to suggest six new formulations out of the virtual experiments having T_g in the specific range of 80–100°C. Indeed, the measured T_g of the new formulations were in the theoretically expected range (see the datapoints in region VII), which can already be seen as a first experimental validation of the ML model, as also discussed in the next sections.

Reactions and synergistic effects

The current maximum and minimum T_g measured for all suggested formulations are depicted in Figure 4 as cyan and red lines, respectively. Starting from five random points having T_g in the relatively small range of 100–115°C (region I), the Bayesian optimization-based design of experiments was able to detect formulations with T_g, which cover a much wider range (76–131°C). DMA of the individual masterbatches reveals that the highest and lowest T_g were 129.48 (L-citrulline) and 80.91°C (GABA), respectively (see **Table II**). Because this range (81–129) is smaller than the one found for the new suggested formulations (76–131), synergistic interactions between the different amino acids seem to take place when they are mixed, causing T_g to be higher than those of the individual amino acids.

The number of theoretically possible reactions among seven different amino acids is extremely high (see **Figure 5**) and discussing them is not in the scope of this investigation. These reactions are, however, responsible for the expansion of the range of T_g beyond the T_g's of the individual components. When only considering the amine-epoxy reaction (blue), the peptide reaction (orange), and the esterification of hydroxyl groups with carboxyl groups (green), there are already more than 70 possible reactions. However, this does not take into account that the reaction of carboxyl groups with different

Table II. T_g measured[a] for masterbatches containing only one amino acid.

Ratio Label	Amino Acid	T_g ± STD in °C
P_1	L-arginine	112.78 ± 0.18
P_2	L-citrulline	129.48 ± 0.41
P_3	GABA	80.91 ± 0.16
P_4	L-glutamine	113.84 ± 0.18
P_5	L-proline	–
P_6	L-tryptophan	124.94 ± 0.11
P_7	L-tyrosine	119.36 ± 0.60

The stoichiometric ratio R of active H atoms to epoxy functional groups is 1.

[a]Experiments done in duplicates.

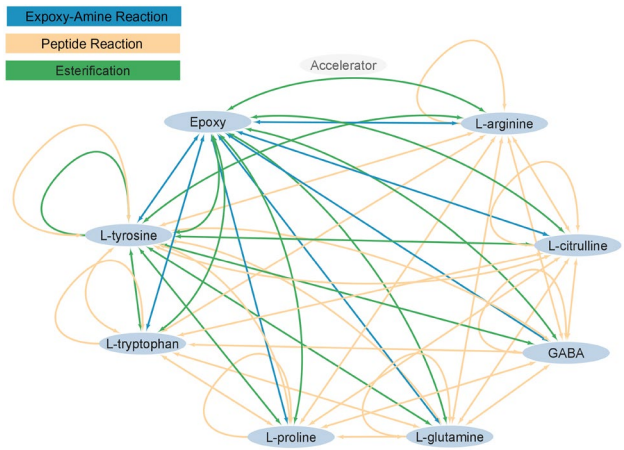

Figure 5. Overview about the theoretically possible reactions between individual components of the investigated material.

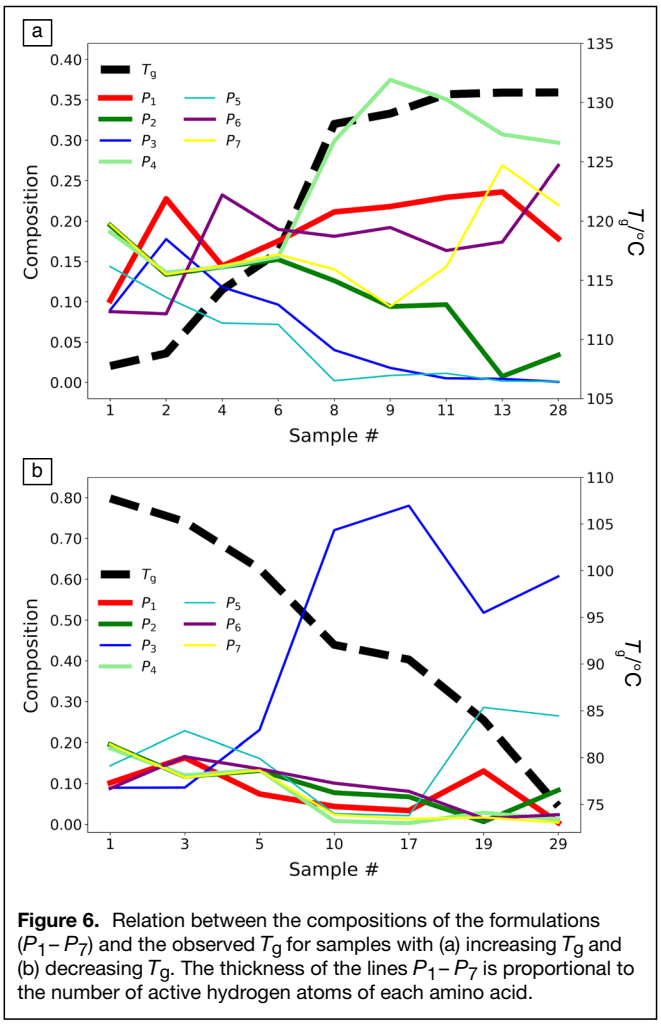

Figure 6. Relation between the compositions of the formulations ($P_1 - P_7$) and the observed T_g for samples with (a) increasing T_g and (b) decreasing T_g. The thickness of the lines $P_1 - P_7$ is proportional to the number of active hydrogen atoms of each amino acid.

amino groups of an amino acid leads to a completely different product and that the number of ways that seven different amino acids can be sequentially combined in the same peptide is exceedingly large. This shows that the material, although limited in the number of its components, results in a highly complex, heterogeneous network, once cured.

The trends observed in the composition and T_g of the designed and prepared formulations are shown in **Figure 6**, where the ratios P_1–P_7 refer to the amino acids as shown in Table II. Note that the selected samples shown in the x-axis of Figure 6 exhibit T_g (thick-dashed lines) that either continuously increase (Figure 6a) or continuously decrease (Figure 6b). Before starting a more general discussion on the overall trends of the composition of the formulations, the composition of sample 13, which has a very high T_g (130.84°C), is examined. As shown in Figure 6a, this sample is mostly composed of L-glutamine ($P_4 = 0.31$), L-tyrosine ($P_7 = 0.27$), L-arginine ($P_1 = 0.24$), and L-tryptophan ($P_6 = 0.17$), whose T_g's lie in the range of 112–124.94°C. This means that the T_g of the mixture is about 6°C higher than the highest T_g (L-tryptophan) and about 18°C higher than the lowest T_g (L-arginine) of its main components, which again points out to the existence of synergistic effects by combining different amino acids.

Taking into account the selected samples shown in Figure 6, the largest positive and negative ratio variations for those formulations were +0.181 (P_6) and −0.161 (P_2) for samples with increasing T_g and +0.517 (P_3) and −0.192 (P_7) for samples with decreasing T_g. This anticipates that formulations with higher amounts of L-tryptophan and lower amounts of L-citrulline tend to exhibit high T_g (Figure 6a) and vice versa. Similarly, formulations tend to exhibit low T_g (Figure 6b) if they have higher amounts of GABA and lower amounts of L-tyrosine. In fact, this is in part expected because L-tryptophan and GABA give rise to individual materials with one of the highest T_g (124.94°C) and the lowest T_g (80.91°C), respectively, as shown in Table II. Interestingly, the amino acid with the highest individual T_g (L-citrulline) does not seem to help getting high T_g materials. In fact, the formulation with the highest T_g (sample 28, T_g = 130.86°C) has only a very small amount of L-citrulline ($P_2 = 0.034$).

The thickness of the lines P_1–P_7 in Figure 6 is proportional to the respective numbers of active hydrogen atoms ($\equiv f$) in

the structure of each amino acid. For the low T_g case, this indicates that smaller f values are associated with low T_g. For the high T_g case, the influence of f is less clear at first sight. In addition, there is an optimum ratio of aliphatic amino acids, which have a high f, to aromatic amino acids, which do have only few active hydrogen atoms. One hypothesis is that the aliphatic amino acids (L-arginine and L-glutamine) react via peptide bond formation with the aromatic ones (L-tryptophan and L-tyrosine) thereby forming an aromatic curing compound with a high number of active hydrogen atoms. Even though the hypothesis discussed above is based on a relatively simple reaction between aliphatic and aromatic amino acids, there are indeed numerous ways for this reaction to occur. A more thorough analysis of the influence of the composition and functionality on the final T_g of the prepared materials, performed via the LASSO model and using all samples, is discussed in the "Model interpretation" section.

Economical considerations

The optimization of the material properties via the design of thermoset formulations is of key interest for polymer engineers. Economical aspects also play a strong role in finding the best formulation. This is particularly easy to take into account when formulations exhibit similar target properties, where cheaper formulations are clearly given priority. **Table III** shows the prices of curing agents of some high T_g samples (price CA), as well as the prices of the corresponding curable epoxy resin amino acid mixtures (price M), which includes the price of DGEBA.

Although the T_g of the listed samples is almost the same, their prices vary considerably. For instance, choosing sample 11 instead of only L-citrulline changes the price M from 9.74 to 8.40 Euro kg^{-1}, which represents a drop of 13.7% in cost. Note that the T_g of sample 11 is even slightly higher than that of the material with only L-citrulline as the curing agent. When the same comparison is performed using the price CA, this drop is even more pronounced (33.6%). This economical aspect becomes crucial whenever a material is produced at an industrial scale.

Table III. Prices of curing agents (CA) and curable epoxy resin–curing agent mixtures (M), in Euro kg^{-1}, and corresponding measured T_g in °C.

CA	T_g	Price CA	Price M
L-citrulline	129.48	46.49	9.74
Sample 8	128.19	32.69	8.73
Sample 9	129.07	31.78	8.46
Sample 11	130.70	30.88	8.40
Sample 13	130.84	30.25	8.67
Sample 28	130.86	34.31	9.47

The resin used was DGEBA (price in June 2022 = 4 Euro kg^{-1}).

Model evaluations

Different models were initially tested with default hyperparameters (see the Supplementary information) and their performances evaluated by k-fold CV, as shown in **Figure 7**. The statistics of the evaluation was improved by splitting the 35-sample data set (into k folds) 200 times, each generating different training/test sets, from where the error bars in Figure 7 were obtained. In addition, the parameter k used in the k-fold CV evaluation was also optimized for each model–the optimal k value corresponding to the lowest MAE error found in each case is shown in parentheses below the model's name in Figure 7. The models had somehow similar performances for the test set (MAE = 2.6–5.4°C), where the nonparametric GPR, together with SVR exhibited the lowest MAE and highest R^2 values. Even the largest MAE value obtained for the test set, calculated for the RF model (5.4°C), was about three times lower than the error obtained for the prediction of T_g of epoxy systems and evaluated on the test set (16.2°C), as reported in our previous work.[20]

The comparison between the model performances for predictions on the training and test sets reveals more pronounced differences among the models (compare the red and pink bars in Figure 7). When a model performs well for the training set and performs much worse for the test set, it is said to overfit the data, while similar performances for both training and test sets indicate that overfitting is minimized. The models GPR, GBR, and KNN have exhibited strong overfitting because they have an error of zero for the training set and errors in the range of 2.6–5.1°C for the test set. The LASSO model, on the other

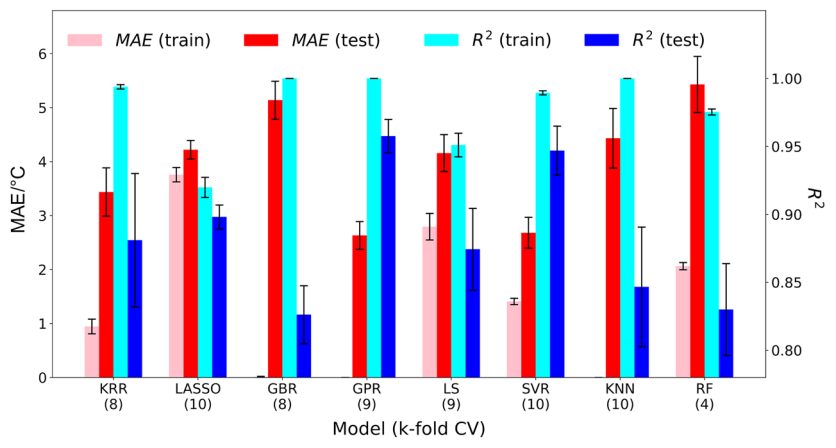

Figure 7. k-fold cross-validation (CV) evaluation of the different models investigated. Default hyperparameters were used (see the Supplementary information). The performances refer to predictions on the training and test sets averaged more than 200 random k-fold splittings. The number in parentheses is the best k value found for each model. Error bars give ±1 standard deviation.

hand, has shown very similar errors for the training and test sets, which indicates that overfitting mostly does not take place. All other models have shown nonnegligible overfittings. For instance, in the case of the KRR model, the MAE error for the test set is more than three times larger than that for the training set (see Figure 7).

Further decreasing overfitting can be done, for instance, by fully optimizing the hyperparameters, especially those related to regularization. However, this is tricky because of a very small number of samples that need to be further divided into (training + validation) and test sets. Taking the KRR model as an example, the optimized hyperparameters have shown a strong dependence on the training/validation set used. Performing 10 different hyperparameter optimizations on this model (see the Supplementary information), each one with a different set of 22 randomly chosen samples for the training/validation set and the remaining 13 samples for the test set gave MAE = 3.348°C (±1.292) and $R^2 = 0.849$ (±0.155) for the predictions on the different test sets. The model performance on the corresponding training sets gave MAE = 1.490°C (±0.490) and $R^2 = 0.985$ (±0.007), which means that KRR is still overfitting, although a bit less, as the test error is roughly two times the training error instead of more than three times, as previously discussed. Most importantly, the variance of the model becomes very large, as concluded from the standard deviation obtained for the MAE error (1.292°C) as compared to that shown for the non-optimized model (0.448°C).

It is worth to point out that further improving the model performance and also decreasing overfitting can be more easily done by simply increasing the size of the data set far beyond 35 samples, which is out of the scope of this work because, among others, this is not a sustainable solution for the efficient design of experiments.

Based on the model performances achieved on the test set (MAE and R^2), as well as on the overfitting considerations, the LASSO model was chosen to be discussed here in more detail (see the Supplementary information for more details on the other models). This model is also very important to help interpret the relation between the composition of the formulations and the observed T_g (see the "Model interpretation" section). **Figure 8**a shows the comparison between the experimental (blue line) and predicted (red line) T_g for samples 6–35. Each predicted T_g was calculated using a fresh LASSO model trained with all previous samples, which means that samples 1–5 were used to train a model to predict T_g for sample 6, then samples 1–6 were used to train another model to predict T_g for sample 7, and so on. The MAE error for each prediction is shown in Figure 8b (bars), where the red line is a moving average of MAE over periods of five samples.

Figure 8c shows the evaluation of the LASSO model using a k-fold CV (best $k = 10$) with all 35 samples. The meaning of $k = 10$ is that nine parts or folds of the data set are used to train a fresh model, which then performs predictions on the tenth, left-out fold–see the "Materials and methods" section for more details on the CV procedure). An average error of about 4°C and a reasonably good R^2 parameter were obtained. The performance of the LASSO model, trained with all 35 samples, was then evaluated on the randomly selected experimental validation set (Figure 8d), which gave a small MAE error (<5°C). The experimental validation is also discussed further on.

Model interpretation

In order to interpret the model in terms of the relation between T_g and the composition P_n used for the formulations, the LASSO model was used. This model does feature selection because of the L1 regularization term in its loss function, from where the weight coefficients of some of the features can completely vanish, facilitating the interpretation of the results.

The LASSO model trained with all 35 samples gave the following relation between the predicted T_g ($= \hat{y}$) and the composition ($P_1 - P_7$) of the amino acid formulation:

$$\hat{y} = -0.47P_3 - 0.52P_5 + 0.02P_6 + 0.05P_7, \quad (5)$$

where the weight coefficients for P_1, P_2, and P_4 were zero. Equation 5 reveals that increasing the fractions of the amino acids GABA (P_3) and L-proline (P_5) in a formulation strongly decreases T_g, whereas L-tryptophan (P_6) and L-tyrosine (P_7) exhibit a much weaker but positive influence on T_g. This trend is better understood by checking the measured T_g of epoxy systems having only individual amino acids (Table II), where L-tryptophan and L-tyrosine have relatively high T_g's (125ºC and 119°C, respectively) when compared with GABA (81°C). This means that one needs to use high T_g components in the formulations to increase the final T_g of the cured thermoset and vice versa, as intuitively expected if one ignores possible reactions between amino acids during curing (vide infra for a counter proof). By examining the number f of active hydrogen atoms in each amino acid (see Table I) and taking into account Equation 5, it becomes clear that aliphatic amino acids with large f values as in the case of L-arginine, L-citrulline, and L-glutamine ($f \geqslant 5$) do not influence positively or negatively T_g of the multicomponent material. Aliphatic or cyclic amino acids with a small number of active hydrogen atoms ($f \leqslant 3$) seem to decrease the thermosets' T_g. On the contrary, the fraction of aromatic amino acids (P_6 and P_7) positively influences T_g (see the positive/negative signs in Equation 5). The positive influence aromatic structures have on T_g has also been observed in other materials.[20]

Interestingly, the amino acid L-citrulline (P_2), which has the highest T_g (129°C) for the pure epoxy has shown no influence on the T_g of the investigated seven-component formulations, according to Equation 5. As explained, there are many possible complex interactions involving all seven amino acids present in the epoxy material and this is the reason why counting exclusively on intuition is not always the best way to design new experiments with multidimensional parameters. Instead, using Bayesian optimization to select the best formulations seems to be a better approach (see Figure 4). Ideally, however, the combination of intuition and modeling for such tasks is preferred, especially

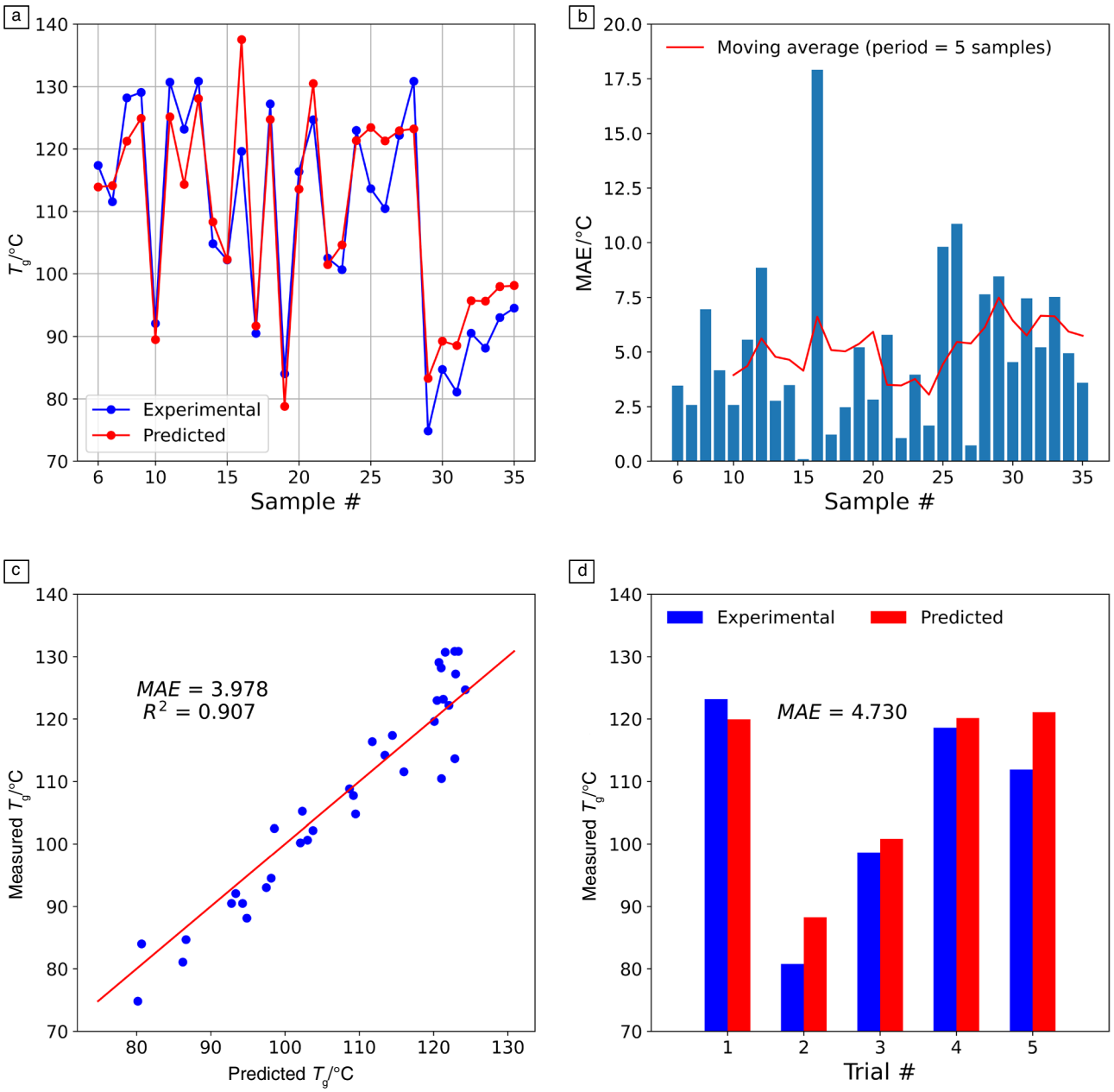

Figure 8. (a) Comparison between predicted and experimental T_g. The prediction for sample *i* was performed by training a fresh model using samples 1 to *i*–1. (b) Error for the predictions shown in (a) (bars) and their moving average (red line). (c) Tenfold cross-validation evaluation of the performance of the model including all 35 samples. (d) Comparison between the predicted and experimental T_g for the experimental validation set of randomly selected formulations. The predictions shown in all subplots were performed using a LASSO model with default, non-optimized hyperparameters (α = 0.1).

if the number of parameters gets large (typically, >20), when the Bayesian optimization then becomes much less efficient.

Although it was not possible to cure a sample with only L-proline in the formulation due to the pronounced porosity observed during curing, Equation 5 suggests that this material would have a considerably low T_g.

To a first approximation, some linearity between features and target can be assumed because the LASSO model was indeed consistent with some experimental observations and previous experimental findings. However, the nonlinear KRR model performed only slightly worse than LASSO and could also have led to the conclusion that a nonlinear relationship between features and target is not unlikely, although this relationship cannot be interpreted directly, as was the case with LASSO. A thorough investigation of linearity between features and target to get as close to ground truth as possible needs to be done after hyperparameter

optimization to obtain more reliable results and should especially be done with a much larger data set (>>35 samples). However, a much larger data set is in conflict with the main goal of this manuscript, which is, among others, to propose a sustainable solution for the development of new biomaterials with as few experiments as possible, as we have shown here.

Experimental validation

The steered drawing of six new formulations from the virtual experiments exhibiting T_g in the desired range of 80–100°C (see Figure 4, region VII) was a first experimental validation of the model trained with only 29 samples. The second experimental validation was discussed in the frame of Figure 8a, where every new measured T_g was compared with the T_g predicted from a fresh model trained using all previous measurements. The final experimental validation was performed by selecting five random formulations from the pool of virtual experiments and comparing the measured T_g of the newly prepared samples with the predicted T_g calculated using the model trained with all 35 previous samples, as already shown in Figure 8d, which gave a MAE error of 4.730°C. The use of other models (see **Table IV** and the Supplementary information) gave similar errors. Note that the randomicities inherent to the GBR and RF models were taken into account by averaging the predictions of 200 different runs, from where standard deviations were calculated, as shown in Table IV. According to our experience and that of our academic and industrial partners, being able to predict T_g for any new formulation with an absolute error smaller than about 10°C already enables one to design new materials for different applications in a reliable fashion. In another investigation,[20] the MAE error for the prediction of T_g for an experimental validation set of novel epoxy resin systems was about 31°C, which is considerably worse than the current model.

Final considerations

Although the linear LASSO model has provided an equation that agreed with the experimental findings and the same equation has been used to predict T_g for five new experiments randomly selected from a pool of 10^6 virtual experiments, yielding a very small error (4.73°C), small nonlinearities in the data set cannot be excluded because of the reduced size of our data set. Further increasing the size of the data set to address this issue in more detail is beyond the scope of this paper, as mentioned earlier. We refer the reader to the work of Sofos et al.,[38,39] who have recently discussed how to extract physically meaningful equations from larger data sets using numerical and analytical ML approaches applied to other systems, from where model linearities can be better discussed.

Table IV. MAE error for the prediction of T_g for the experimental validation set calculated using the models previously evaluated in Figure 7.

Model	MAE in °C (±STD)
KRR	5.485
LASSO	4.730
GBR	5.700 (±0.040)
GPR	6.618
LS	6.339
SVR	7.408
KNN	9.505
RF	5.940 (±0.334)

Conclusion

It was shown that bio-based epoxy resin systems with tailored T_g can be efficiently designed with a minimum number of experiments via Bayesian optimization and AL techniques. The highest/lowest T_g measured for the designed formulations was higher/lower than those of the individual components of the formulations, which pointed out the synergistic effects when combining different amino acids as curing agents. The efficiency of the presented method is highlighted by the convergence of the high-T_g formulations after five iterations of Bayesian optimization. In this paper, sustainability was achieved by the use of bio-based curing agents and the implementation of Bayesian optimization during the material design phase, leading to shorter material development phases and epoxy resin systems with optimized properties.

In view of the exploration/exploitation steps during the theoretical design of experiments, very diverse formulations with extreme T_g were found, from where economical aspects could be easily considered. For the examples discussed, it was shown that the price reduction of the thermoset and the curing agent was 13.7% and 33.6%, respectively. Consequently, Bayesian optimization could help to save significant costs when producing thermosets at industrial scale.

Based on the weakest tendency to overfit and on the highest accuracy toward the experimental validation set, the best model found in this investigation was the LASSO. This feature selection-based model also provided an easy interpretation of the influence of the chemical structure (aromaticity and number of active hydrogen atoms, f) on the final T_g for the corresponding thermoset. Amino acids with very high f values ($\geqslant 5$) did not seem to influence T_g positively or negatively. For the low-f amino acids ($f \leqslant 3$), those with aromatic moieties had a positive impact on T_g, in agreement with literature reports.

The findings discussed in this work pave the way toward more sustainable solutions to efficiently design epoxy resin system exhibiting desired properties. Future works may discuss the Bayesian optimization approach developed here to design optimal formulations of tailored thermosets by optimizing different target properties simultaneously.

Acknowledgments

The authors want to thank S. Taumann and A. Himsel for their support during the experiments.

Author contributions

R.Q.A. contributed to ML modeling, paper writing, correction, and discussion; F.R. contributed to writing of the original draft, correction, experimental measurements, and discussion; H.R. contributed to paper correction and discussion.

Funding

Open access funding enabled and organized by Projekt DEAL. Parts of this work were funded by the "Bayerischen Staatsministerium für Wissenschaft und Kunst" (Grant No. F.2-M7426.10.2.1/4/16, Germany). 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 No. 20E1907A)."

Conflict of interest

The authors declare no conflicts of interest or other disclosures.

Supplementary information

Details on all (ML) predictive models discussed in this work are available as a PDF file. The online version contains supplementary material available at https://doi.org/10.1557/s43577-023-00504-9.

References

1. M. Ashby, *Materials Selection in Mechanical Design*, 1st edn. (Butterworth-Heinemann, Oxford, 1992)
2. G.W. Ehrenstein, *Faserverbund-Kunststoffe* (Carl Hanser Verlag, Munich, 2006)
3. H. Lengsfeld, F. Wolff-Fabris, J. Krämer, J. Lacalle, V. Altstädt, *Faserverbundwerkstoffe—Prepregs und ihre Verarbeitung* (Hanser Publishers, Munich, 2014). https://www.hanser-elibrary.com/doi/book/10.3139/9783446440807
4. F. Henning, E. Moeller, *Handbuch Leichtbau: Methoden, Werkstoffe, Fertigung* (Carl Hanser Verlag, Munich, 2020)
5. L. Bourne, F. Milner, K. Alberman, *Am. J. Ind. Med.* **16**(2), 81 (1959)
6. H. Greim, D. Bury, H.-J. Klimisch, M. Oeben-Negele, K. Ziegler-Skylakakis, *Chemosphere* **36**(2), 271 (1998)
7. K. Venables, *Br. J. Ind. Med.* **46**, 222 (1989)
8. W. Anku, M. Mamo, P. Govender, *Phenolic Compounds in Water: Sources, Reactivity, Toxicity and Treatment Methods* (IntechOpen, London, 2017)
9. Y. Li, F. Xiao, C.P. Wong, *J. Polym. Sci. A Polym. Chem.* **45**(2), 181 (2007)
10. A. Motahari, A. Omrani, A. Rostami, *Comput. Theor. Chem.* **977**, 168 (2011)
11. A. Motahari, A.A. Rostami, A. Omrani, M. Ehsani, *J. Macromol. Sci. Part B Phys.* **54**(5), 517 (2015)
12. L. Mazzocchetti, S. Merighi, T. Benelli, L. Giorgini, "Evaluation of Tryptophan—Late Curing Agent Systems as Hardener for Epoxy Resin," *AIP Conf. Proc.* **1981**(1) (2018), No. 020170
13. P. Gnanasekar, N. Yan, *Polym. Degrad. Stab.* **163**, 110 (2019)
14. S. Merighi, L. Mazzocchetti, T. Benelli, L. Giorgini, *Processes* **9**, 42 (2021)
15. M. Shibata, J. Fujigasaki, M. Enjoji, A. Shibita, N. Teramoto, S. Ifuku, *Eur. Polym. J.* **98**, 216 (2018)
16. F. Rothenhäusler, H. Ruckdäschel, *Polymers* **1**(1), 4331 (2022). https://doi.org/10.3390/polym14204331
17. F. Rothenhäusler, H. Ruckdäschel, *Polymers* **14**(21), 4696 (2022)
18. F. Rothenhäusler, H. Ruckdäschel, *Polymers* **15**(2), 385 (2023). https://doi.org/10.3390/polym15020385
19. D.L. Nelson, M.M. Cox, *Lehninger Principles of Biochemistry* (Macmillan Learning, New York, 2021)
20. S. Meier, R.Q. Albuquerque, M. Demleitner, H. Ruckdäschel, *J. Mater. Sci.* **57**(29), 13991 (2022). https://doi.org/10.1007/s10853-022-07372-9
21. P. Larsen, *6—Physical and Chemical Properties of Amino Acids* (Academic Press, London, 1980)
22. M. Demleitner, S.A. Sanchez-Vazquez, D. Raps, G. Bakis, T. Pflock, A. Chaloupka, S. Schmölzer, V. Altstädt, *Polym. Compos.* **40**(12), 4500 (2019). https://doi.org/10.1002/pc.25306
23. R.Q. Albuquerque, C. Brütting, T. Standau, H. Ruckdäschel, *e-Polymers* **22**, 318 (2022). https://doi.org/10.1515/epoly-2022-0031
24. S. Pruksawan, G. Lambard, S. Samitsu, K. Sodeyama, M. Naito, *Sci. Technol. Adv. Mater.* **20**(1), 1010 (2019)
25. H. Kang, J.H. Lee, Y. Choe, S.G. Lee, *Nanomaterials* (Basel) **11**(4), 872 (2021)
26. C. Kim, A. Chandrasekaran, A. Jha, R. Ramprasad, *MRS Commun.* **9**(3), 860 (2019). https://doi.org/10.1557/mrc.2019.78
27. 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, E. Duchesnay, *J. Mach. Learn. Res.* **12**, 2825 (2011)
28. C.E. Rasmussen, C.K.I. Williams, *Gaussian Processes for Machine Learning* (MIT Press, New York, 2006)
29. V. Vovk, *Empirical Inference* (Springer, Cham, 2013)
30. T. Hofmann, B. Schölkopf, A.J. Smola, *Ann. Stat.* **36**(3), 1171 (2008)
31. N.S. Altman, *Am. Stat.* **46**, 175 (1992)
32. A. Natekin, A. Knoll, *Front. Neurorobot.* **7**, 21 (2013)
33. A. Smola, B. Schölkopf, *Stat. Comput.* **14**, 199 (2004)
34. X. Yan, X.G. Su, *Linear Regression Analysis: Theory and Computing* (World Scientific Publishing, Singapore, 2009)
35. R. Tibshirani, *J. R. Stat. Soc. Ser. B Methodol.* **58**(1), 267 (1996)
36. L. Breiman, *Mach. Learn.* **45**(1), 5 (2001). https://doi.org/10.1023/A:1010933404324
37. C.C. Aggarwal, *Data Mining: The Textbook*, 1st edn. (Springer, Cham, 2015)
38. T.E. Karakasidis, F. Sofos, C. Tsonos, *Fluids* **7**(10), 321 (2022)
39. F. Sofos, A. Charakopoulos, K. Papastamatiou, T.E. Karakasidis, *Phys. Fluids* **34**(6), 062004 (2022). https://doi.org/10.1063/5.0096669