J Mater Sci (2022) 57:13991–14002

Computation & theory

Modeling glass transition temperatures of epoxy systems: a machine learning study

Sven Meier[1], Rodrigo Q. Albuquerque[1,2], Martin Demleitner[1], and Holger Ruckdäschel[1,2,*]

[1] *Department of polymer engineering, University of Bayreuth, Universitätsstr. 30, Bayreuth 95447, Germany*
[2] *Neue Materialien Bayreuth GmbH, Gottlieb-Keim-Straße 60, Bayreuth 95448, Germany*

Received: 24 March 2022
Accepted: 24 May 2022
Published online:
22 July 2022

ABSTRACT

The use of machine learning (ML) models to screen new materials is becoming increasingly common as they accelerate material discovery and increase sustainability. In this work, the chemical structures of 16 epoxy resins and 19 curing agents were used to build an ML ensemble model to predict the glass transition (T_g) of 94 experimentally known thermosets. More than 1400 molecular descriptors were calculated for each molecule, of which 119 were chosen based on feature selection performed by principal component analysis. The quality of the trained model was evaluated using leave-one-out cross-validation, which yielded a mean absolute error of 16.15°C and an R^2 value of 0.86. The trained model was also used to predict T_g for 4 randomly selected resin/hardener combinations for which no experimental data were available. The same combinations were then prepared and measured in the laboratory to further validate the ML model. Excellent agreement was found between experimental and predicted T_g values. The current ML model was created using only theoretical features, but could be further improved by adding experimental or quantum mechanical properties of the individual molecules as well as experimental processing parameters. The results presented here contribute to improving sustainability and accelerating the discovery of novel materials with desired target properties.

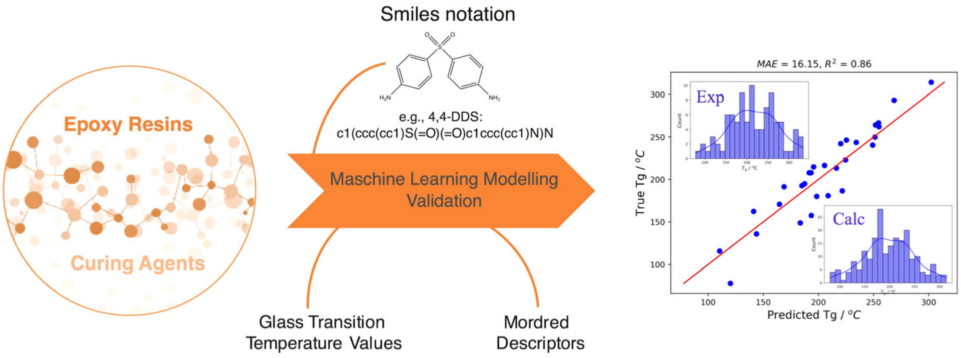
Handling Editor: Gregory Rutledge.

Address correspondence to E-mail: holger.ruckdaeschel@uni-bayreuth.de

https://doi.org/10.1007/s10853-022-07372-9

Springer

GRAPHICAL ABSTRACT



Introduction

In recent years, the demand of epoxy resins in industry has increased continuously because of their great potential in various sectors like adhesives, coatings, construction applications, electronic packaging materials, or composites. Due to their superior physical–mechanical properties, chemical and temperature resistance, dimensional stability, versatility, and processability, they are especially used for applications with increased performance requirements [1–4]. This especially holds true for glass or carbon fiber reinforced composites used in aerospace, automotive, and sports industries where their higher specific strength is superior to metallic materials such as aluminum or steel [5, 6]. The glass transition temperature (T_g) is considered a critical property for polymers because it is an important indicator of processing and performance properties such as stiffness, heat resistance, and adhesion. In addition, T_g defines and limits the application temperature of the material. Typically, T_g is measured using differential scanning calorimetry (DSC) or dynamic mechanical thermal analysis (DMTA) [7]. However, achieving a target value for T_g of a resin/hardener system is so far largely a trial-and-error process based on the experience of the researcher. As a result, the development of new thermoset systems and process optimization is often an inefficient, costly, and time-consuming process [8].

Due to the importance of T_g, there exist several modeling approaches in polymer research to predict it. Weyland et al. developed the basis of the group additive property method (GAP), which predicts the T_g of polymers from the sum of calibrated contributions with respect to typical monomer substructures [9]. This triggered a series of further works to improve the model's accuracy, which showed good predictive correlations, but were only applicable to polymers whose chemical structure groups had beforehand been experimentally investigated. Since the late 1980s various more general QSPR (quantitative structure–property relationships) models using descriptors for T_g prediction were developed [10–15]. However, all of the previous discussed prediction models were built on the basis of homopolymers, even though thermoset systems like epoxy-amine systems are of great importance for many high-performance applications.

Nevertheless, there are some approaches to predict the T_g of cross-linked thermoset systems, which, however, are limited to relatively few epoxy systems [16–18]. Bellenger et al. [16] presented a predictive model for the glass transition temperature of epoxy resin systems using an additivity law for copolymers and the contribution of cross-linked structures, using a total of 40 systems. Morrill et al. [18] developed a model that predicts the dependence of T_g on the stoichiometric ratio of co-monomers in amine-cured resins of bisphenol A diglycidyl ether.

The use of machine learning in materials science is a growing area which shows large potential in predicting material properties, such as the T_g of polymers. Data-driven approaches are a great way to optimize the thermoset development process in terms of efficiency, speed, and sustainability. In the field of homopolymers, there exist several promising attempts to predict the T_g using machine learning [19–23]. However, in the field of thermosets, there are few approaches that use machine learning applications for T_g prediction. These are trained either on very small datasets and/or predominantly on datasets obtained using molecular dynamics [24, 25]. Jin et al. [24] described an optimization method for multicomponent epoxy resin systems using an artificial neural network (ANN), where molecular dynamics simulation was used to obtain the input data for the ANN. In the same work, T_g was one of six predicted values of the model. In addition, only two different hardeners and resins based on 30 different data points were used, making it more difficult to use this model for screening more diverse resin/hardener combinations. Higuchi et al. [25] established a model which predicts the glass transition temperatures of linear homo-/heteropolymers and cross-linked epoxy resins. In addition, a consensus model is presented that links the two different kinds of models. The thermoset-specific model was trained with just 50 data points and six different epoxy resins. The data came exclusively from the literature where different methods (DSC, DMTA, TMA) and different specimen geometries were used. The performance of their consensus model (homo- and heteropolymers) was good (R^2 = 0.848), while the thermoset-only model gave poorer results (R^2 = 0.687). Yan et al. [26, 27] developed the state-of-the-art model currently available for predicting T_g of thermoset shape memory polymers (TSMPs), whose mean absolute percentage error for the test set was 13.91 %. The model is based on a variational autoencoder trained with drug molecules and adapted to TSMPs with transfer learning, combined with a weighted vector combination method and a convolutional neural network as regressor.

The use of atomistic molecular dynamics or multiscale simulations to add features to the dataset is another strategy to improve the performance of the model [28], but these theoretical methods are not trivial to implement for polymers, time-consuming, and their accuracy still needs to be improved, the latter possibly having a negative impact on the predictive ability of the model. On the other hand, using available experimental data from different works to extend and improve the dataset is also challenging, as the experimental techniques for preparing and measuring the T_g of materials are not standardized, as often found in the literature.

This work aims to predict the T_g for resin-hardener thermosets using an ML ensemble approach based solely on theoretical molecular descriptors that can be easily calculated for any molecular structure present in the thermoset. To this end, we will present a variety of systems studied in our group using, among others, standardized internal DMTA measurements of T_g to improve the generalizability of the final ML model. The molecular descriptors are calculated from the SMILES representation of the resins and hardeners, whose information is initially used to train different linear and nonlinear individual ML models, prior to combining them into one single ML ensemble model. Finally, the experimental validation of the ensemble model will be demonstrated by preparing and measuring new thermosets whose compositions and T_g were first suggested by the ML model itself.

Methods

The workflow for creating the machine learning model was inspired by the Crisp-DM model, the quasi-standard framework for data mining (DM) [29]. In total 6 steps were necessary to produce the results presented in this paper. The DM process is embedded in real experiments and measurements in the lab, both for data generation and for validation of the model, see Figure 1. The different steps of the workflow are chronologically described in detail in the next subsections.

Data generation/experiments

The experimental data were generated by DMTA carried out at a heating rate of 3 K/min in torsional mode at an elastic deformation of 0.1 % and an applied frequency of 1 Hz using a Rheometrics Scientific ARES RDA III (Germany). The specimens had a rectangular shape with 50x10x2 mm^3 according to standard ISO 6721-7. T_g was determined via DMTA measurements by the peak value of the loss factor (tan(δ)). For multiple measurements of the same material system, the average value for T_g was used.

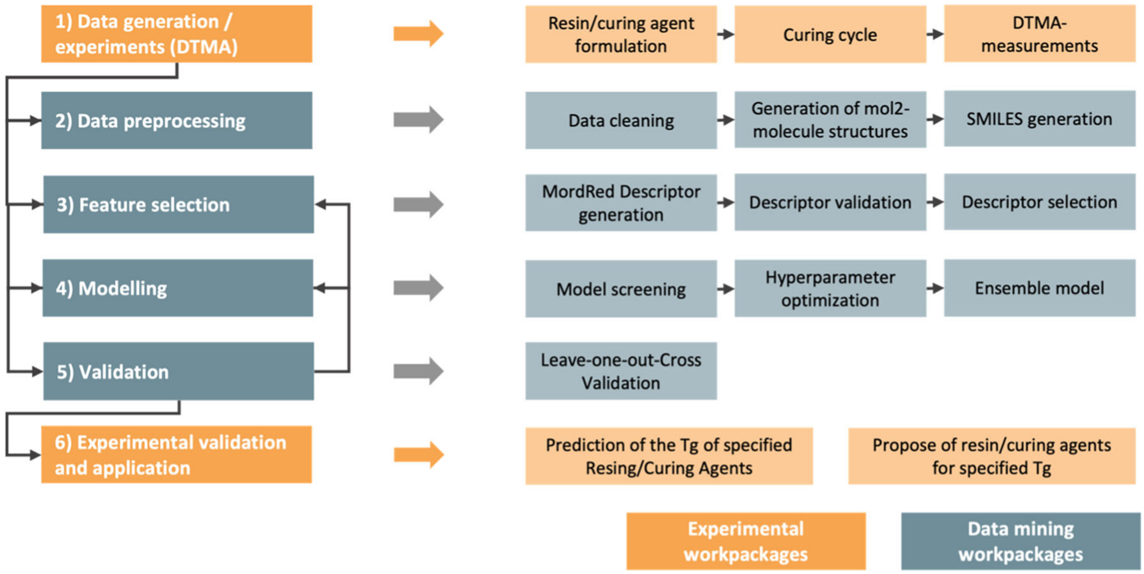

1) Data generation / experiments (DTMA)
Resin/curing agent formulation
Curing cycle
DTMA-measurements
2) Data preprocessing
Data cleaning
Generation of mol2-molecule structures
SMILES generation
3) Feature selection
MordRed Descriptor generation
Descriptor validation
Descriptor selection
4) Modelling
Model screening
Hyperparameter optimization
Ensemble model
5) Validation
Leave-one-out-Cross Validation
6) Experimental validation and application
Prediction of the Tg of specified Resing/Curing Agents
Propose of resin/curing agents for specified Tg
Experimental workpackages
Data mining workpackages

Figure 1 Workflow for creating the machine learning model and performing a practical application.

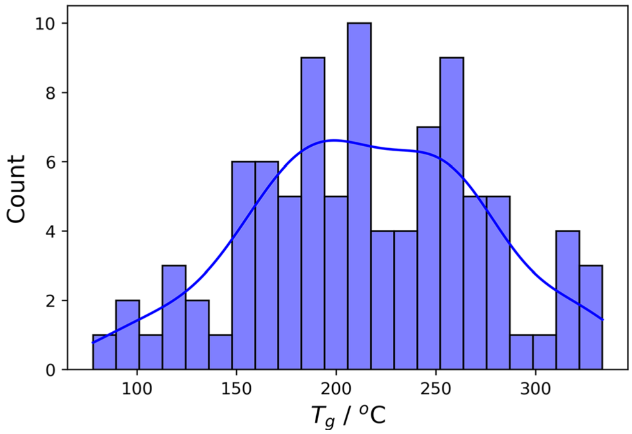


Figure 2 Distribution of experimental T_g values of the resin/hardener systems of the dataset; blue line represents the kernel density estimation.

The resins and hardeners combinations were mixed stoichiometrically and cured in the oven to the maximum degree of cure under optimized thermal conditions. Subsequently, multiple measurements were carried out in the DMTA for one material. The epoxy equivalent weight (EEW) of the epoxy resins from the dataset has a range from 101 g/eq to 187 g/eq, while the functionality ranges from 2 to 4.5. The amine equivalent weight (AHEW) of the curing agents has a range of 12.05 to 221.67, while the functionality is between 1 and 7. The technical data such as EEW were taken from the suppliers' technical data sheets. Figure 2 shows the distribution of T_g values in the dataset corresponding to 94 combinations of 16 resins and 19 hardeners. All resins and hardeners used in this work were those available in our laboratory and are shown in the SI. The lowest T_g value (77.91°C) is shown by the system DGEBF/D230, while the system Tactix742/4,4-DDS exhibited the highest T_g (333.51°C).

Data preparation and preprocessing

For each resin and hardener, the chemical structure of the molecule was drawn with the visualization program Avogadro and saved in the mol2 format, which is a plain text table format that represents a single or multiple chemical structures and contains atomic coordinates, chemical bonding information, and metadata of the molecule [30, 31]. The mol2 files were then converted into the corresponding SMILES (Simplified Molecular Input Line Entry System) notation [32]. The cheminformatics libraries Mordred [33] and RDKit [34] were used to convert each SMILES notation into the corresponding 2D descriptors, providing an initial set of 1613 descriptors per molecule. RDKit is a Python library for building, manipulating, analyzing, and automatically designing molecules [34]. MordRed is a descriptor generator available on Github, which can calculate more than 1800 descriptors [33]. A molecular descriptor transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment. Molecular descriptors are used to present molecular characteristics in cheminformatics and therefore are useful for feature engineering in data mining [33, 35]. The 1613 molecular descriptors available at the Mordred library for 2D structures were initially reduced to 1435 after removing non-numerical (e.g., Boolean) and missing/non-computed values, which were subsequently reduced to 119 descriptors per molecule (resin or hardener) after feature selection (see the next subsection). The molecular descriptors for the resin/hardener pair DEN431/D230 before and after feature selection are shown in the SI.

At the end of this stage, the dataset consisted of 94 samples, each containing 2 x 1435 features (or independent variables) corresponding to each resin/ hardener pair, and the target property (or dependent variable), T_g. 70 % of the dataset was used as training set, which was also used to optimize the hyperparameters of individual ML models by leave-one-out cross-validation to decrease overfitting. The final model was tested using 30 % of the dataset, as shown in the next sections. All features and target property were preprocessed to have unit variance and zero mean using the mean and variance of the training set.

Feature selection

In order to obtain the greatest possible information content for model training, feature selection was performed. Too many features lead to too much complexity of the dataset and may result in poorer performance. Therefore, feature selection was carried out via principal component analysis (PCA). PCA decomposes a multivariate dataset into a set of successive orthogonal components that explain a maximum of variance [36]. By using PCA, the descriptive variance of a feature can be determined over the dataset. By looking at the weight coefficient (also called loading) of each feature on the resulting uncorrelated linear combinations of each principal component, the amount of descriptive variance of the feature can be determined: The higher the weight of one feature, the greater its descriptive variance and thus the greater its information content. Features with little information content were filtered out of the dataset by setting a minimum threshold for the weight of all features. The evaluation of the model's performance for different thresholds allowed to determine the final descriptors to be used as features in the model, strongly reducing the total number of features from 1435 to 119 per molecule.

Modeling and validation

Modeling and evaluation do not take place separately in the workflow, but are performed iteratively (see Figure 1), as already shown in the previous subsection to find the best threshold for feature selection. To find the best ML model, the following three-step procedure was used:

1. Pre-screening of different linear and nonlinear ML models with unoptimized hyperparameters.
2. Hyperparameter optimization of the most accurate ML models found in a).
3. Creation of an optimized ensemble using the ML models found in b).

The leave-one-out validation method was used to improve statistics in all steps. Here, the learning algorithm is applied once for each data point, using all other instances as the training set and the selected data point as the test set with a single sample. The methodology is a special case of cross-validation where the number of folds is equal to the number of data points in the dataset [37] and is time-consuming for middle-to-large datasets, since one different ML model is trained for each sample. The mean absolute error (MAE), mean absolute percentage error ($MAPE$) and the coefficient of determination (R^2 score) were used as model evaluation metrics. MAE is given by:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i | \tag{1}$$

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. $MAPE$ is given by:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \frac{| y_i - \hat{y}_i |}{| y_i |} \cdot 100\%. \tag{2}$$

The coefficient of determination (R^2 score) is the quotient of the explained variation to the total variation in a regression model. It is often expressed as a percentage.

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{3}$$

Seven different regression models were prescreened, whereby their *MAE* and R^2 scores were compared via leave-one-out cross-validation. The ML methods are briefly summarized below and are described in more detail in refs [38, 39].

Linear Regression (LR). The LR fits a linear model with coefficients $w = (w_1, \ldots, w_p)$ to decrease the residual sum of squares between the observed data points, and the predicted data points by the linear approximation [40].

K-nearest-neighbors regression (KNN). The KNN classification methodology selects a set of K objects in the training set that are closest to the test object and appoints a label based on the most dominant class in that neighborhood. Neighbors-based regression can be used in cases where the data labels are continuous [41].

Random Forest Regression (RFR). Random Forest is an ensemble learning approach in which "weak learners" collaborate to form "strong learners" using a large collection of uncorrelated decision trees. However, instead of building a solution based on the results of a single deep tree, Random Forest aggregates the results of a set of shallow trees. RFR is built by growing trees depending on a random vector such that the tree predictor takes on numerical values as opposed to class labels [42].

Gradient Boosting Regression (GBR). GBR builds an additive model step wisely, allowing the optimization of arbitrary differentiable loss functions. At each stage, a regression tree is fitted to the negative gradient of the given loss function [43].

Kernel Ridge Regression (KRR). KRR merges ridge regression [44] with the kernel trick [45]. Thus, it learns a linear function in the space generated by the kernel and the data. This corresponds to a nonlinear function in the original space of the dataset.

Support Vector Regression (SVR). Support vector machines solve binary classification problems by formulating them as optimization problems to find the maximum edge that separates the hyperplane while correctly classifying as many training points as possible. The optimal hyperplane is represented with support vectors. SVR is achieved by formulating an optimization problem by defining a convex ϵ-insensitive loss function to be minimized and finding the flat tube that contains most of the training instances [46].

Lasso regression. Lasso regression is a linear model combined with the estimation of sparse coefficients via L1-regularization. This 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 [47].

The most promising models were followed up and had their hyperparameters optimized. In the last step, two or more single models were combined into an ensemble with the goal of minimizing *MAE*. For this purpose, the predictions of the individual ML models were combined via a weighted average whose weight coefficients were optimized to minimize *MAE*. Eq. 3 describes how to obtain the ensemble prediction:

$$\hat{y}_i = \frac{\sum_{j=1}^{m} a_j \cdot \hat{y}_{ij}}{\sum_{j=1}^{m} a_j} \tag{4}$$

where m is the number of individual ML models, a is the weight coefficient (or importance) of each model inside the ensemble, $\hat{y}_{ij}$ is the predicted target property of sample i carried out using model j, and $\hat{y}_i$ is the final ensemble prediction for sample i. The Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm was used to find the weighting factors of the individual ML models that minimize *MAE* of the ensemble predictions. All codes used in the workflow were written in Python (version 3) using Jupyter Notebook. The ML models were built using the Sklearn library [48].

Practical application

The optimized hyperparameters of the final ensemble model were then used to train a fresh model with the full dataset (training and test sets) and subsequently used to predict all 210 possible resin/hardener combinations for which no experimental data were available. Four of the new combinations were chosen to be prepared and measured in the laboratory to further validate the ML model. The first two formulations (Table 1) were chosen based on high and low target T_g values randomly selected. The remaining ones (Table 2) were randomly chosen from the pool of 210 possible combinations.

All four new systems were then prepared as follows. Resin and hardener were mixed stoichiometrically and degassed in a vacuum chamber below 50

Table 1 Model-proposed thermoset systems for targeted (high and low) T_g. The predicted T_g shown in the last column was calculated using the trained ensemble model. The supplier and trading names are given

Resin	Curing agent	Curing cycle	Target T_g
Hexion, Epikote 862[a]	Huntsman, Jeffamine D400[b]	1h@25°C 4h@60°C 1h@100°C	82°C
Huntsman, Tactix742[c]	Lonza, Lonzacure CAF[d]	1h@120°C 3h@180°C 2h@250°C	310°C

Chemical names: [a]DGEBF,[b]Poly(propylene glycol) bis(2-aminopropyl, ether), [c]Tris-(hydroxylphenyl)methane-based epoxy,[d]9,9-Bis(4-amino-3-chlorophenyl)fluorene

Table 2 Randomly selected formulations for evaluation of T_g predictions. The predicted T_g shown in the last column was calculated using the trained ensemble model. The supplier and trading names are given

Resin	Curing agent	Curing cycle	Target T_g
Huntsman Araldite MY 0610 CH[a]	Huntsmann Jeffamine D400[e]	1h@25°C 4h@60°C 1h@100°C	104°C
DIC HP7200H [b]	Huntsman, Aradur HY 906[c] + 1 wt% Evonik, Curezole 2E4MZ[d]	1h@120°C 3h@180°C 2h@250°C	259°C

Chemical names: [a]TGMAP,[b]Dicyclopentadiene-type epoxy, [c]NMA, [d]2E4MZ, [e]Poly(propylene glycol) bis(2-amino propyl, ether)

mbar. After degassing, the mixtures were poured into steel molds and curing was performed according to the specific curing cycles shown in Tables 1 and 2.

Results and discussion

This work shows that it is not only possible to obtain a predictive model for T_g of thermosets, but also to get some insight into the influence of molecular structures of resins and hardeners on the T_g of thermosets. Feature selection using PCA was carried out to reduce the total number of descriptors from 1435 to 119. The GBR and KRR models provided the best individual predictions, with MAE/R^2 values of 16.64/0.84 and 17.82/0.83, respectively, evaluated for the test set. The detailed evaluations of these models, together with their optimized hyperparameters, are shown in the Supporting Information.

Optimized ensemble model

The composition of the optimized ensemble model in terms of individual ML models and their corresponding model importances were: GBR (53 %) and KRR (47 %). The performance of the trained ensemble model evaluated using the test set is shown in Figure 3, where MAE = 16.15°C and R^2 = 0.86. The ensemble model was more accurate and had a larger coefficient of determination than the best individual ML model (GBR), which shows that the ensemble technique described here (Eq. 4) is an effective way to increase the models' performance. The results are

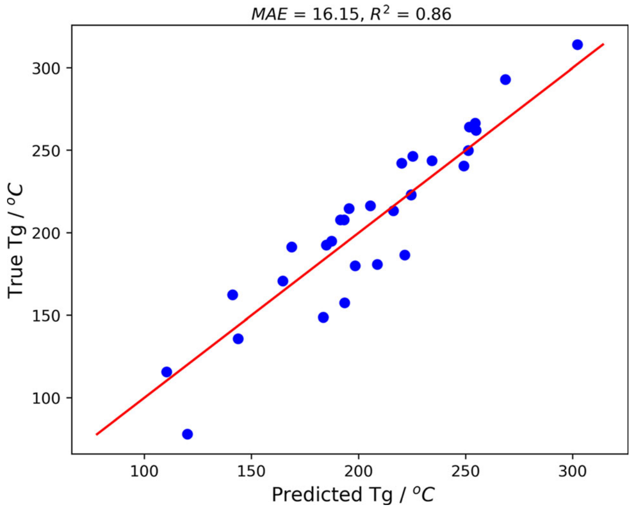


Figure 3 Comparison between experimental (True) and predicted T_g using the optimized ensemble model with composition 53 % GBR and 47 % KRR. The predictions were performed for the test (29 samples) using the model trained with the training set (65 samples). $MAPE$ = 9.38 % (accuracy = 90.62 %).

Table 3 Five most important molecular descriptors for T_g prediction according to the Lasso model using the whole dataset

Descriptor	Description	Molecule	Effect
MoeType SlogP_VSA	MOE-type descriptors using Crippen–Wildman atom classification system for atom-based calculation of octanol–water partition coefficient (log P) (hydrophobicity) [49]	Resin	+
MoeType SMR_VSA	MOE-type descriptors using Crippen–Wildman atom classification system for atom-based calculation of molar refractivity (sterics and polarizability) [49]	Resin	–
Autocorrelation ATS(0,'dv')	Broto–Moreau autocorrelation of a topological structure of order 0 weighted by valence electrons [50]	Resin	+
PathCount	Total number of paths of length m in the molecular graph [49]	Hardener	–
Autocorrelation ATS(0,'dv')	Broto–Moreau autocorrelation of order 0 weighted by valence electrons [50]	Hardener	–

promising, since the model's features were purely theoretical, solely based on the chemical structures of the molecules. The performance in terms of *MAPE* (9.38 %) was slightly better than relatively complex models published elsewhere [26, 27].

Final molecular descriptors

The most important descriptors affecting T_g were identified using the Lasso regression trained using the entire dataset, since it does by default feature selection [47], meaning that only the most important features will be associated with nonzero weight coefficients. In addition, the weight coefficients (c_i) of Lasso enable one to easily interpret the relation between the features and the target property (e.g., $property = c_1 \cdot feature_1 + c_2 \cdot feature_2 + \ldots + c_n \cdot feature_n + b$), which is not possible with other ML models like GBR or KRR. Table 3 shows the five descriptors that have the largest influence on the predicted T_g. The column "Molecule" describes whether the descriptor refers to the resin or hardener component of the thermoset. In addition, in the column "Effect," the characters (+, –) describe whether the descriptor has a positive or negative influence on T_g.

Descriptors directly related to aromaticity and polarity were also inside the pool of 119 descriptors used in the ML models, although their weight coefficients were not among the largest ones. The two most important descriptors were SlogP_VSA and SMR_VSA. The former is a metric of hydrophobicity and the latter accounts for molecular size and polarizability. The first two lines of Table 3 reveal that epoxy systems tend to exhibit higher T_g when the resin counterpart is more hydrophobic and smaller and less polarizable. These properties can be linked indirectly to aromatic structures, which generally possess higher T_g.

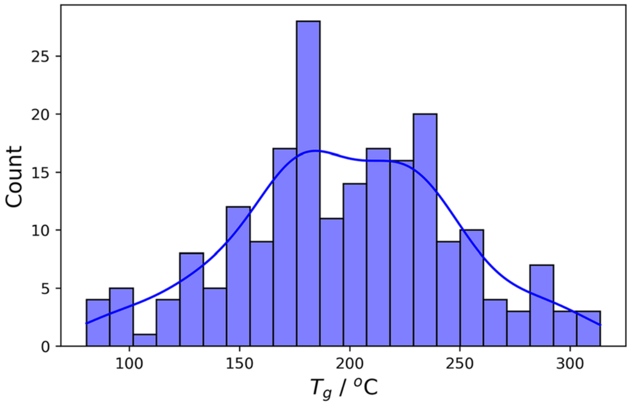


Figure 4 Distribution of the T_g values predicted by the optimized ensemble model for 210 new resin/hardener thermoset combinations.

The interpretation of different descriptors is not always simple or intuitive, and the fact that the final ML model considers a great number of them, shows that this task is even more challenging. There seems to exist a trade-off between interpretability and prediction capability: simpler models are very easy to interpret but do not necessarily deliver the best predictions.

New T_g predictions

Based on the dataset, which consists of 16 different resins and 19 different hardeners and experimental T_g values available for only 94 out of 304 total resin/

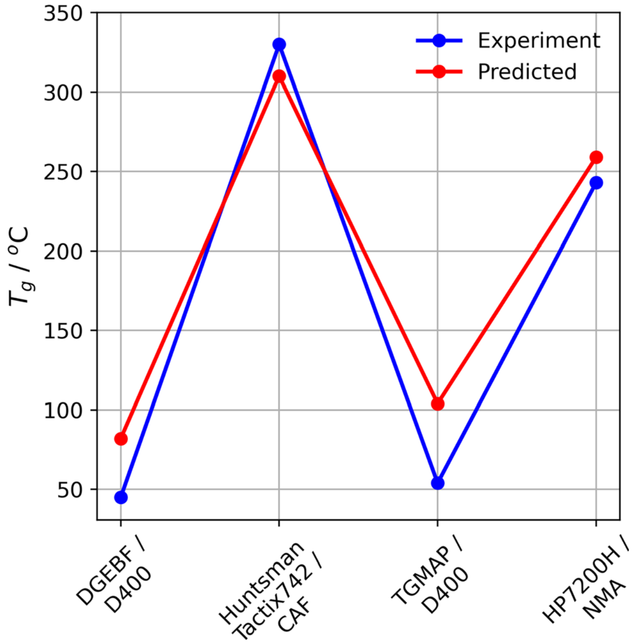


Figure 5 Comparison between experimental (True) and predicted T_g for the resin/hardener systems described in Tables 1 and 2.

hardener combinations, a total of 210 new resin/hardener combinations were predicted by the trained ML ensemble model. In the histogram of predicted T_g values, the distribution ranges from 81 to 318°C (Figure 4), which indicates that the model depicts the maximum variance of the available experimental glass transition temperatures.

Practical application

The comparison between the experimental and predicted T_g for the four selected resin/hardener systems described in Tables 1 and 2 is depicted in Figure 5.

The experimental trends were nicely reproduced by the predicted values. The predictions were much better for systems with middle to high T_g, as expected based on the greater availability of data points in that region to train the ML models, compared to regions of low T_g (see Fig. 2). In addition, the experimental T_g used to train the ML models were in the range 78-334°C (Figure 2), and therefore high and low predicted T_g values tend to accompany the same range. Using many experimental data points with T_g smaller than 78°C to train the ML models would have led to better predictions below 78°C.

Final considerations

Predicting the final T_g of a relatively complex polymeric material by simply using molecular descriptors of isolated resin/hardener molecules is a rough approximation assumed in this work. The first validation of this idea comes from the good model fittings shown in Fig. 3, which suggests that the property of individual molecules can influence their interaction and consequently the property exhibited by the final material. The second validation to support this approximation comes from Fig. 5, which shows that using molecular descriptors of individual molecules to predict the T_g of novel polymeric materials is feasible, especially in the case of middle-to-high T_g epoxy systems. Fig. 5 also shows that there is room for model improvement (vide infra). Most importantly, the results suggest that cross-linking is implicitly taken into account by the ML model via linear and nonlinear combinations of molecular descriptors of individual (isolated) molecules since T_g was predicted based on those descriptors. Some intuition on the approximation described here could be gained from a chemical point of view. For instance, if a molecular descriptor reveals that an (isolated) resin and its hardener counterpart have high aromaticities, one could expect to have a stabilizing interaction in the polymer between nearby molecular units, which can increase T_g, as in general observed experimentally. The assumption made here would have benefited from a larger number of measurements (as opposed to only 94 data points available) since this is also expected to improve the model accuracy.

Conclusions

In this work, it was shown that it is possible to predict the glass transition of thermoset systems based solely on the chemical structures of 16 resins and 19 hardeners using ML models. The best individual ML models combined via an ensemble approach and trained with 94 resin/hardener combinations were GBR and KRR, which contributed to 53 and 47 %, respectively, of the final predicted T_g value. The ML ensemble model performed quite well (MAE = 16°C and R^2 = 0.86). The analysis of the most important descriptors used in the Lasso regression has revealed that T_g of the thermoset is expected to increase for smaller, less polarizable, and more hydrophobic resins. The trained model was used to predict T_g for the 210 remaining resin/hardener combinations that had not been previously investigated experimentally. Four new resin/hardener combinations were randomly chosen from the 210 new predictions and experimentally prepared and characterized, giving T_g values in good agreement with the predictions: exp/theo [°C] = 45/82, 330/310, 54/104, 243/259. This strategy allows for greatly accelerating the development of new resin/hardener systems in a sustainable fashion, as one avoids trial-and-error procedures to obtain new thermosets exhibiting any specific T_g value. By describing the entire workflow, from data generation to ML modeling to the final practical application, this work demonstrates how machine learning can systematically unlock efficiencies in thermoset development and uncover new relationships. In conclusion, the results discussed help improve sustainability while accelerating the discovery of novel materials with desired target properties. Even though the rough approximation made here through which the features of isolated monomer units (here, hardener and resin) are used to predict the T_g of a much more complex, 3D polymeric material provided a good initial ML model, we are currently working on its improvement by adding quantum mechanical properties of the molecules to the current dataset to further increase the model accuracy.

Acknowledgements

We thank all colleagues and former co-workers for supporting us by providing DMTA measurements, which gave us the opportunity to generate our dataset. This publication was funded by the University of Bayreuth in the funding program Open Access Publishing.

Author's contributions

Sven Meier: Writing of the original draft, preparation of the dataset, discussion. Rodrigo Q. Albuquerque: ML modeling, paper correction, discussion. Martin Demleitner: Experimental setup and measurements, paper correction, discussion. Holger Ruckdäschel: Paper correction, discussion.

Funding

Open Access funding enabled and organized by Projekt DEAL.

References

[1] Mattar N, Langlois V, Renard E, Rademacker T, Hübner F, Demleitner M, Altstädt V, Ruckdäschel H, Rios de Anda A (2021) Fully bio-based epoxy-amine thermosets reinforced with recycled carbon fibers as a low carbon-footprint composite alternative. ACS Appl Polym Mater 3(1):426–435. https://doi.org/10.1021/acsapm.0c01187

[2] Memon H, Wei Y, Zhu C (2022) Recyclable and reformable epoxy resins based on dynamic covalent bonds-present, past, and future. Polym Test 105:107420. https://doi.org/10.1016/j.polymertesting.2021.107420

[3] Liu J, Sue H-J, Thompson ZJ, Bates FS, Dettloff M, Jacob G, Verghese N, Pham H (2009) Effect of crosslink density on fracture behavior of model epoxies containing block copolymer nanoparticles. Polymer 50(19):4683–4689. https://doi.org/10.1016/j.polymer.2009.05.006

[4] Guadagno L, Raimondo M, Vittoria V, Vertuccio L, Naddeo C, Russo S, De Vivo B, Lamberti P, Spinelli G, Tucci V (2014) Development of epoxy mixtures for application in aeronautics and aerospace. RSC Adv 4:15474–15488. https://doi.org/10.1039/C3RA48031C

[5] Ehrenstein, G.W.: Faserverbund-Kunststoffe. Werkstoffe - Verarbeitung - Eigenschaften. Carl Hanser Verlag GmbH & Co. KG, Munich (2006).

[6] Lengsfeld, H.,Altstädt, V.,Wolff-Fabris, F.,Krämer, J.:Composite Technologien. Carl Hanser Verlag GmbH & Co. KG,München(2014). https://doi.org/10.3139/9783446440807.http://www.hanser-elibrary.com/doi/book/10.3139/9783446440807

[7] Bard, S.,Demleitner, M.,Weber,R.,Zeiler, R.,Altstädt, V.: Effect of curing agent on the compressive behavior at elevated test temperature of carbon fiber-reinforced epoxy compositesPolymers **11**(6)(2019).https://doi.org/10.3390/polym11060943

[8] Demleitner, M.,Sanchez-Vazquez, S.A.,Raps,D.,Bakis,G.,Pflock, T., Chaloupka, A.,Schmölzer, S.,Altstädt, V.:Dielectric analysis monitoring of thermoset curing with ionic liquids: from modeling to the prediction in the resin transfer molding process. Polym Composite **40**(12)(2019).https://doi.org/10.1002/pc.25306

[9] Weyland HG, Hoftyzer PJ, Van Krevelen DW (1970) Prediction of the glass transition temperature of polymers. Polymer 11(2):79–87. https://doi.org/10.1016/0032-3861(70)90028-5

[10] Katritzky AR, Sild S, Lobanov V, Karelson M (1998) Quantitative structure-property relationship (qspr) correlation of glass transition temperatures of high molecular weight polymers. J Chem Inf Comput Sci 38(2):300–304. https://doi.org/10.1021/ci9700687

[11] Katritzky AR, Rachwal P, Law KW, Karelson M, Lobanov VS (1996) Prediction of polymer glass transition temperatures using a general quantitative structure-property relationship treatment. J Chem Inf Comput Sci 36(4):879–884. https://doi.org/10.1021/ci950156w

[12] Camelio P, Cypcar CC, Lazzeri V, Waegell B (1997) A novel approach toward the prediction of the glass transition temperature: application of the evm model, a designer qspr equation for the prediction of acrylate and methacrylate polymers. J Polym Sci, Part A: Polym Chem 35(13):2579–2590. https://doi.org/10.1002/(SICI)1099-0518(19970930)35:13<2579::AID-POLA5>3.0.CO;2-M

[13] Cypcar CC, Camelio P, Lazzeri V, Mathias LJ, Waegell B (1996) Prediction of the glass transition temperature of multicyclic and bulky substituted acrylate and methacrylate polymers using the energy, volume, mass (evm) qspr model. Macromolecules 29(27):8954–8959. https://doi.org/10.1021/ma961170s

[14] Lazzeri V (1996) Prediction of the glass transition temperature of multicyclic and bulky substituted acrylate and methacrylate polymers using the energy, volume, mass (evm) qspr model. Macromolecules 29(27):8954–8959. https://doi.org/10.1021/ma961170s

[15] Hopfinger AJ, Koehler MG, Pearlstein RA, Tripathy SK (1988) Molecular modeling of polymers.IV. estimation of glass transition temperatures. Polym Phys 26(10):2007–2028. https://doi.org/10.1002/polb.1988.090261001

[16] Bellenger V, Verdu J, Morel E (1987) Effect of structure on glass transition temperature of amine crosslinked epoxies. J Polym Sci, Part B: Polym Phys 25(6):1219–1234. https://doi.org/10.1002/polb.1987.090250604

[17] Lee G, Hartmann B (1983) Glass transition temperature predictions in some epoxy polymers. J Appl Polym Sci 28(2):823–830. https://doi.org/10.1002/app.1983.070280233

[18] Morrill JA, Jensen RE, Madison PH, Chabalowski CF (2004) Prediction of the formulation dependence of the glass transition temperatures of amine-epoxy copolymers using a qspr based on the am1 method. J Chem Inf Comput Sci 44(3):912–920. https://doi.org/10.1021/ci030290d

[19] Goswami S, Ghosh R, Neog A, Das B (2021) Deep learning based approach for prediction of glass transition temperature in polymers. Mater Today Proc 46(xxxx):5838–5843. https://doi.org/10.1016/j.matpr.2021.02.730

[20] Ma R, Liu Z, Zhang Q, Liu Z, Luo T (2019) Evaluating polymer representations via quantifying structure-property relationships. J Chem Inf Model 59(7):3110–3119. https://doi.org/10.1021/acs.jcim.9b00358

[21] Chen G, Tao L, Li Y (2021) Predicting polymers' glass transition temperature by a chemical language processing model. Polymers 13(11):1–14. https://doi.org/10.3390/polym13111898

[22] Tao,L.,Varshney, V.,Li, Y.: Benchmarking machine learning models for polymer informatics: an example of glass transition temperature. Journal Chem Inform Model (2021).https://doi.org/10.1021/acs.jcim.1c01031

[23] Karuth, A.,Alesadi, A.,Xia, W.,Rasulev,B.: Predicting glass transition of amorphous polymers by application of cheminformatics and molecular dynamics simulations. Polymer 218February,123495 (2020).https://doi.org/10.1016/j.polymer.2021.123495

[24] Jin, K.,Luo, H.,Wang, Z.,Wang, H.,Tao, J.: Composition optimization of a high-performance epoxy resin based on molecular dynamics and machine learning. Mater Design 194,108932 (2020).https://doi.org/10.1016/j.matdes.2020.108932

[25] Higuchi C, Horvath D, Marcou G, Yoshizawa K, Varnek A (2019) Prediction of the glass-transition temperatures of linear homo/heteropolymers and cross-linked epoxy resins.

ACS Appl Polym Mater 1(6):1430–1442. https://doi.org/10.1021/acsapm.9b00198

[26] Yan C, Feng X, Li G (2021) From drug molecules to thermoset shape memory polymers: a machine learning approach. ACS Appl Mater & Interface 13(50):60508–60521. https://doi.org/10.1021/acsami.1c20947

[27] Yan, C.,Feng, X.,Wick, C.,Peters, A.,Li, G.: Machine learning assisted discovery of new thermoset shape memory polymers based on a small training dataset. Polymer**214**(2021).https://doi.org/10.1016/j.polymer.2020.123351

[28] Gartner TE, Jayaraman A (2019) Modeling and simulations of polymers. a roadmap. Macromolecules 52(3):755–786. https://doi.org/10.1021/acs.macromol.8b01836

[29] Azevedo, A.,Santos, M.F.:Kdd, semma and crisp-dm: a parallel overview. IADIS European conference data mining, 182–185(2008)

[30] Hanwell, M.D.,mCurtis, D.E.,Lonie, D.C.,Vandermeersch, T., Zurek, E.,Hutchison, G.R.:Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. J Cheminform 41,17 (2012).https://doi.org/10.1186/1758-2946-4-17

[31] van de Waterbeemd H, Carter RE, Grassy G, Kubinyi H, Martin YC, Tute MS, Willett P (1997) Glossary of terms in computational drug design (iupac recommendations 1997). Pure Appl Chem 69(5):1137–1152. https://doi.org/10.1351/pac199769051137

[32] Anderson, E.,Veith, G.D.,Weininger, D.: SMILES: a line notation and computerized interpreter for chemical structures(1987). https://cfpub.epa.gov/si/si_public_record_report.cfm?dirEntryId=33186

[33] Moriwaki H, Tian YS, Kawashita N, Takagi T (2018) J Cheminform. Mordred: a molecular descriptor calculator 10(1):1–14. https://doi.org/10.1186/s13321-018-0258-y

[34] Landrum, G.:RDKit: Open-source cheminformatics software.https://www.rdkit.org/(2021)

[35] Wang R, Fu Y, Lai L (1997) A new method for calculating partition coefficients of organic compounds. Acta Physico - Chimica Sinica 13(1):615–621. https://doi.org/10.3866/pku.whxb19970101

[36] Archanah T, Sachin D (2015) Dimensionality reduction and classification through pca and lda. Int J Comput Appl 122(17):4–8. https://doi.org/10.5120/21790-5104

[37] Sammut, C.,Webb, G.I.:Leave-one-out cross-validation, pp. 600–601.Springer,Boston, MA2010. https://doi.org/10.1007/978-0-387-30164-8_469. https://doi.org/10.1007/978-0-387-30164-8_469

[38] Bishop CM (2006) Pattern recognition and machine learning (Series: Information Science and Statistics). Springer, London

[39] Murphy KP (2012) Machine learning: a probabilistic perspective (Series: Adaptive Computation and Machine Learning). MIT press, Cambridge

[40] Yan, X.,Su, X.G.: Linear regression analysis: theory and computing. World Scientific Publishing Co., Inc., USA 2009

[41] Altman NS (1992) An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat 46:175–185

[42] Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324

[43] Natekin A, Knoll A (2013) Gradient boosting machines, a tutorial. Front. Neurorobot. https://doi.org/10.3389/fnbot.2013.00021

[44] Vovk V (2013) Empirical inference, pp. 105–116. Springer, London. https://doi.org/10.1007/978-3-642-41136-6

[45] Hofmann T, Schölkopf B, Smola AJ (2008) Kernel methods in machine learning. Ann Stat 36(3):1171–1220. https://doi.org/10.1214/009053607000000677

[46] Smola A, Schölkopf B (2004) A tutorial on support vector regression. Stat Comput 14:199–222

[47] Tibshirani R (1996) Regression shrinkage and selection via the lasoselection via the lasso. J R Stat Soc Series B 58:267–288

[48] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830

[49] Wildman SA, Crippen GM (1999) Prediction of physicochemical parameters by atomic contributions. J Chem Inf Comput Sci 39(5):868–873. https://doi.org/10.1021/ci990307l

[50] Moreau G, Broto P (1980) The autocorrelation of a topological structure: a new molecular descriptor. Nouv J Chim 4(6):359–360

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.