Materials & Design 267 (2026) 116265

Contents lists available at ScienceDirect

Materials & Design

journal homepage: www.elsevier.com/locate/matdes

Glass transition temperature prediction in lignin polyurethanes using machine learning on small experimental dataset☆

Silviu Florin Acaru [a], Marc Comí [b], Panagiotis Falireas [b], Danny E.P. Vanpoucke [c,d], Richard Vendamme [b], Katrien V. Bernaerts [a,*]

[a] Sustainable Polymer Synthesis Group, Aachen-Maastricht Institute for Biobased Materials (AMIBM), Faculty of Science and Engineering, Maastricht University, Brightlands Chemelot Campus, Urmonderbaan 22, 6167 RD Geleen, Netherlands (the)
[b] Sustainable Polymer Technologies Team, Flemish Institute for Technological Research (VITO N.V.), Mol, Belgium
[c] Hasselt University, Institute for Materials Research (IUMAT), Quantum & Artificial inTelligence design Of Materials (QuATOMs), Martelarenlaan 42, B-3500 Hasselt, Belgium
[d] imec, IUMAT, Wetenschapspark 1, B-3590 Diepenbeek, Belgium

ARTICLE INFO

Keywords:
Lignin
Polyurethane
Machine learning
Small dataset

ABSTRACT

Lignin-based polyurethanes (PUs) offer a compelling route towards sustainable material development, yet the challenge of designing chemical formulations with targeted properties, such as glass transition temperature (T_g), remains unresolved. In this work, we present a systematic approach to explore key structural parameters – such as lignin content, co-polyol chain length, isocyanate functionality, and mixing ratios – across 136 unique formulations, creating a diverse dataset of lignin-based PUs. By harnessing this small dataset, we develop a machine learning (ML) stacking ensemble model capable of reliably predicting T_g, with a mean absolute error of 13.41 °C on the validation set, surpassing the performance of all individual models. Additionally, we enhance model interpretability by integrating advanced mapping techniques and employ an adaptive grid search algorithm to explore extrapolative scenarios. Our workflow, paired with a user-friendly interface, enables rapid discovery and optimization of formulations with desired properties. This study not only deepens the understanding of structure–property relationships in lignin-PUs but also provides a scalable ML-driven tool for designing sustainable materials with precision, highlighting the transformative potential of artificial intelligence in green chemistry and materials innovation.

1. Introduction

The global polyurethane (PU) industry is experiencing growing pressure to develop environmentally friendly and biobased alternatives in response to ecological concerns. This shift is driven by the increasing demand for high-performance materials in sectors like construction, automotive, furniture, and footwear, where sustainability has become a key priority . In this regard, lignin, a naturally abundant biopolymer, emerges as a particularly promising feedstock, owing to its multifunctional polyol structure containing both aliphatic and phenolic hydroxyl groups, which makes it intrinsically suitable for polyurethane synthesis. Although lignin is also compatible with other polymer families such as polyethers, polyacetals, and polyesters, polyurethanes remain the most widely employed across these industrial sectors, spanning thin films, coatings, foams, elastomers, and composites, positioning lignin-based PUs as a highly relevant platform for sustainable materials development [1].

Despite this promise, the creation of new materials has traditionally been a laborious process, heavily reliant on extensive laboratory experimentation. While crucial for material innovation, this conventional approach often proves inefficient, lengthy, and expensive. A prime example is the challenge of regulating thermomechanical properties in PUs, which necessitates multiple experiments to adjust ingredient proportions, catalysts, and production methods, potentially hindering the rapid commercialization of new materials [2–8].

Recent progress in data-centric methodologies, however, offers a promising avenue for expediting PU discovery. The field of materials science has seen a surge in the application of artificial intelligence (AI)

☆ This article is part of a special issue entitled: 'AI Materials' published in Materials & Design.

* Corresponding author.
E-mail address: katrien.bernaerts@maastrichtuniversity.nl (K.V. Bernaerts).

https://doi.org/10.1016/j.matdes.2026.116265
Received 12 December 2025; Received in revised form 15 May 2026; Accepted 19 May 2026
Available online 19 May 2026

and machine learning (ML) techniques. Researchers are now employing sophisticated algorithms to forecast polymer structures and properties with remarkable precision, thereby reducing the need for costly physical experiments [9,10]. These models achieve this by automating the prediction of critical performance metrics such as the glass transition property (T_g), even when working with sparse datasets [11–14].

Despite these advancements, not all ML methods can perform effectively with small experimental datasets. One of the primary limitations of traditional ML algorithms is their tendency to overfit when data is limited, meaning the model captures noise rather than underlying patterns [15]. This issue becomes particularly significant when developing advanced materials like PUs, where small deviations in formulation can lead to large variations in performance metrics. Furthermore, certain models struggle with capturing non-linear relationships that are often present in experimental datasets, requiring more sophisticated approaches to make accurate predictions from limited data [16].

To overcome these limitations, ensemble methods have emerged as powerful tools. These ML techniques combine multiple models, known as base learners, to improve prediction accuracy and robustness. Random Forests (RF), for instance, construct multiple decision trees and aggregate their predictions. Tao *et al.,* evaluated 79 different ML models to predict the T_g of polymers, using various polymer representations and feature engineering approaches. RF models were among the top performers, particularly when using a polymer's repeat unit as the structural representation [17]. Ensemble models using polynomial base models have also been shown to improve prediction accuracy, moreover, in contrast to RF-based models, the final predictive model can be reformulated as a single model instance significantly improving computational efficiency at production level [18]. Such models were successfully used for designing hyperbranched polymers as dispersing agents in inks [19]. Gradient Boosting Regression (GBR), another powerful ensemble method, has also proven effective in materials informatics. Unlike RF, which builds trees in parallel, GBR constructs trees sequentially, with each new tree correcting the errors of the previous ones. Albuquerque *et al.* used GBR to analyze various formulations of biobased epoxy resins [20]. However, the limited dataset in that study (35 samples) resulted in significant overfitting, highlighting the need for more robust approaches to handle small data scenarios. In contrast, the LASSO model demonstrated strong generalization performance.

In addressing the challenges, this study proposes an innovative approach that combines domain expertise with ensemble ML techniques for optimizing sustainable, lignin-based PUs. Traditional methods such as cross-validation and hyperparameter tuning are effective for refining individual models, but ensemble approaches—like stacking and bagging—offer the advantage of improving overall performance by leveraging multiple algorithms. Stacking, for instance, uses a meta-model to combine the predictions of various base models, which has been shown to significantly enhance predictive accuracy. Esmaeili et al. applied this technique to improve the prediction of polymer solubility parameters by integrating Support Vector Machines (SVMs), Neural Networks (NNs), and RFs, with a linear regressor serving as the meta-model [21]. Similarly, bagging, which trains models on random subsets of the data to reduce variance, has proven effective for managing heterogeneous datasets. The utility of bagging was demonstrated for improving the prediction of polymer mechanical properties and the robustness of polymer degradation models, respectively [22,23].

This study builds on these advancements by combining stacking and bagging techniques with domain-specific knowledge to enhance the predictive accuracy and formulation optimization of lignin-based PUs. This material system is particularly amenable to a data-driven approach, as the isocyanate-hydroxyl reaction underlying PU synthesis proceeds efficiently under mild conditions, enabling the rapid preparation of large numbers of formulation variants—a prerequisite for generating the experimental datasets required by ML workflows. A range of models – linear (*e.g.*, Lasso, ElasticNet), kernel-based (*e.g.*, SVMs), and tree-based (*e.g.*, RF, GBR) – are employed to capture both linear and non-linear relationships within the data. Linear models, in particular, have shown consistent performance in polymer research [24,25]. By integrating the strengths of these diverse models, the proposed approach seeks to establish more reliable data driven structure–property relationships and improve the prediction accuracy of sustainable PU formulations, particularly when working with small experimental datasets. To this end, the T_g, measured by differential scanning calorimetry (DSC), was selected as the target property: it is a fundamental thermal parameter directly linked to PU network structure—specifically segmental mobility, crosslink density, and free volume—and is routinely used to evaluate performance across all PU applications. Moreover, DSC autosampler operation allows high-throughput T_g determination with minimal sample quantities, facilitating the systematic data collection necessary for robust model training. Building on this foundation, a user-friendly interface that incorporates the model as its core is introduced, enabling chemists to visualize recipe formulations based on desired properties. This interface simplifies the interaction with the model, removing the need for chemists to have in-depth knowledge of ML techniques and allowing them to focus on leveraging its capabilities with ease.

The central research question guiding this study is: *Can ML models, in combination with domain expertise, effectively predict and optimize the formulations of sustainable, lignin-based PUs, especially in the context of limited data*?

2. Materials and methods

2.1. Lignin PU experiments

2.1.1. Materials

All commercially available solvents, reagents, and chemicals were used as received without further purification unless otherwise stated. Polytetrahydrofurans ($PTHF_s$) with $M_n = 250$, 650 and 1000 g/mol and corresponding OH values of 8.68, 3.20 and 2.10 mmol/g, were purchased from Merck Life Sciences. Stannous octoate ($Sn(Oct)_2$) was purchased from Alfa Aesar, while methyl ethyl ketone (MEK) and anhydrous tetrahydrofuran (THF) were purchased from VWR Chemicals. Hexamethylene diisocyanate (HDI) and NCO trimer Desmodur® ultra N 3600 (further mentioned as "HDI trimer" (HDIt), $f_{NCO} \sim 3.5$ isocyanate groups/molecule) were purchased from Covestro.

For this study, Kraft lignin, specifically Indulin AT (KL) from Ingevity (North Charleston, SC, USA), was selected as the most widely available industrial lignin, obtained as a by-product of the paper industry. To reduce the intrinsic heterogeneity of raw Kraft lignin, we applied a simple solvent extraction step using an organic solvent, yielding the soluble extracted Kraft lignin (EKL) fraction. This upgrading technique is known to improve batch-to-batch reproducibility by providing lignin fractions with more uniform molecular weight distribution and hydroxyl group content, two parameters that strongly influence PU network formation. By using this lignin source and extraction method, we ensure that the variability observed in T_g originates from the controlled formulation parameters (lignin wt%, PTHF molecular weight, NCO/OH ratio, isocyanate functionality, etc.) rather than from uncontrolled lignin heterogeneity.

2.1.2. Construction of the dataset

A total of 180 lignin-based PU films were prepared with different structural characteristics by changing relevant parameters in the formulation. Four main structural parameters during the PU fabrication were modified: (1) the amount of lignin in the polyol mixture, (2) the length of the co-polyol polytetrahydrofuran (PTHF), (3) the type and functionality of the isocyanate and (4) the mixing ratio of isocyanate to polyol. As observed in Fig. 1, lignin-based PUs were fabricated to present different thermal behavior by exchanging the lignin content in the polyol mixture (Parameter 1), the molar mass of the co-polyol PTHF

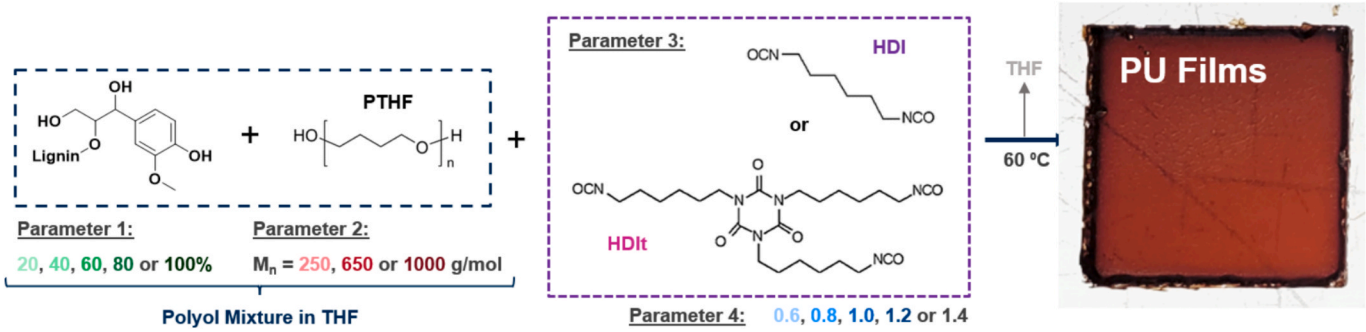

Fig. 1. Scheme representing the 4 structural parameters modified during the preparation to obtain the 150 PUs.

(Parameter 2), the functionality of the isocyanate used for the synthesis (Parameter 3) and the mixing ratio (Ratio = $[NCO_{(HDI\ or\ HDIt)}]/[OH_{(lignin\ +\ PTHF)}]$ (Parameter 4). Regarding the latter, the use of either a difunctional isocyanate HDI or HDIt created two main groups of specimens whereas in each group five subgroups were created by gradually increasing the mixing ratio from 0.6 to 1.4. Each subgroup synthesized with a particular ratio was further split into 3 subgroups by using PTHF of molar mass 250, 600 and 1000 g/mol where the impact of the polyol length was examined. Finally, each subgroup formulated with a certain isocyanate, mixing ratio, and PTHF was further divided in 5 final subgroups by changing the lignin content in the polyol mix. In all cases, the same batch of extracted kraft lignin (EKL) was used.

2.1.3. Synthesis procedure

EKL: in a three neck 2 L round bottom flask equipped with mechanical stirring, 800 g of technical MEK was added to 200 g of Kraft lignin powder. The mixture was stirred (200 rpm) at room temperature (22 °C) for 16 h. After, a basic filtration under gravity was used to remove the insoluble fraction. The filtrate containing the EKL in MEK solution was dried under vacuum and in the oven at 45 °C for 4 h. EKL is characterized by an OH value of 6.42 mmol/g and $M_n = 1590$ g/mol and a dispersity of $Đ = 1.9$ (as determined from GPC in THF calibrated with PS standards).

PU fabrication: the series of PU films were prepared at various EKL to PTHF mass ratios (polyol mixture) using either HDI or HDIt as isocyanate comonomer, according to Eq. (1).

$$EKL(\%, w/w) = \frac{w_{EKL}}{w_{EKL} + w_{PTHF}} \times 100 \quad (1)$$

The [NCO]/[OH] ratio was equal to 0.6, 0.8, 1.0, 1.2 or 1.2 for every formulation and was calculated according to the isocyanate groups from either the HDI or HDIt and the total hydroxyl molar content from EKL (both phenolic and aliphatic groups) and PTHF based on Eq. (2):

$$Ratio = \frac{[NCO]}{[OH]} = \frac{w_{NCO} \times [NCO]_{NCO}}{w_{PTHF} \times [OH]_{PTHF} + w_{EKL} \times [OH]_{EKL}} \quad (2)$$

Where w_{NCO}, w_{PTHF}, w_{EKL} represent the masses (g) of HDI or HDIt, PTHF and EKL, respectively; [NCO] is the molar content of isocyanate groups in HDI or HDIt, in mmol/g; $[OH]_{PTHF}$ and $[OH]_{EKL}$ are the molar content of total hydroxyl groups in the co-polyol and EKL in mmol/g, respectively. The experimental procedure carried out for the synthesis of EKL PU films is described below.

In a vial, catalyst $(Sn(Oct)_2$, 0.8 mol % with respect to the total $[OH]_{PTHF}$ and $[OH]_{EKL}$ (polyol mixture) content in EKL and co-polyol PTHF were dissolved in dry tetrahydrofuran (THF) (8 mL for 2 g of total formulation) and stirred at room temperature until complete dissolution. Afterwards, the isocyanate was added to the solution and the temperature was raised to 60 °C. The reaction time was dependent of the formulation parameters and ranged from 1-30 min. The solution was poured into Teflon molds (70 × 35 × 2 mm), the solvent was evaporated slowly overnight at room temperature and the curing was further performed in an oven for 6 h at 105 °C. Finally, the film was peeled off from the Teflon mold, was vacuum-dried at 40 °C for 18 h to evaporate any residual THF and was stored in a desiccator before testing. The thickness of the obtained films was about 0.5 mm.

2.1.4. Characterization techniques

Determination of Swelling Ratio: swelling tests were carried out with round disc-shaped specimens with an average mass of 600 mg using THF to perform the solubility tests. The specimens were immersed in the solvent for 24 h at 25 °C. Subsequently, the swollen samples were weighed then dried overnight under vacuum at 80 °C to obtain the dry weight. The swelling ratio was calculated using Eq. (3).

$$Swellingratio(\%) = \frac{m_{swollen} - m_{dry}}{m_{dry}} \times 100 \quad (3)$$

Differential Scanning Calorimetry (DSC): DSC was performed using a TA Instruments Discovery DSC 250 equipped with a refrigerated cooling system that allows cooling to −90 °C. The samples were measured in T_{zero} pans with perforated T_{zero} hermetic lids to allow a nitrogen atmosphere around the sample. DSC thermograms were recorded with a heating rate of 10 °C min^{-1}. Only experimental data obtained from the second heating step were reported. A representative DSC thermogram is shown in Fig. S1 of the Supporting Information.

2.2. Machine learning

2.2.1. Hardware and software

All calculations were performed on a HP laptop with a 13th Gen Intel® Core™ i7-1365U processor running at 1.80 GHz. Python v. 3.10 was used as programming language.

2.2.2. Dataset preparation and preprocessing

The experimental raw data (dataset.csv in Supporting Information), initially compiled in excel spreadsheet format, was converted to a CSV file and imported into a Python environment for preprocessing. The preprocessing consisted of several steps: categorical feature encoding using LabelEncoder to transform nominal variables into numerical representations, dropping instances with missing values and feature scaling using RobustScaler to normalize the dataset comprising both numeric and categorical features. We chose RobustScaler for its resilience to outliers and suitability for heterogeneous data types.

The preprocessed dataset comprised nine input features and one target variable (T_g). The first eight features represented experimental parameters used in the synthesis process. The ninth feature (swelling ratio) and the target variable (T_g) were both characterization values obtained from the synthesized product (see Table 1). The raw dataset contained instances with no values in the target (*e.g.,* "NaN" for T_g) as well as instances with 0 wt% lignin content. We removed these rows, which reduced the dataset from 180 instances to 136.

Table 1
Dataset features for machine learning model; * mandatory features used in the wrapper method for feature selection.

Amount of lignin	Amount of Co-polyol by weight percent	Co-polyol type	Amount of isocyanate by weight percent	Amount of isocyanate based on ratio	Isocyante type	Mixing molar ratio [NCO]/[OH]	Catalyst amount	Swelling ratio	Glass transition temperature
Lignin (wt %)*	Co-polyol (wt%)	Co-polyol type (PTHF)*	Isocyanate (wt%)	Isocyanate (mmol NCO)	Isocyanate type	Ratio*	Tin(II) octoate (wt %)	Swelling ratio (%)	T_g (°C)

2.2.3. Feature selection

We applied two methods to identify the features most critical for developing the ML model.

The first method was Principal Component Analysis (PCA). PCA is a statistical technique used in feature selection through reduction of the number of variables. It works by transforming the original features of a dataset into a new set of uncorrelated variables called principal components (PCs). These components are linear combinations of the original features and are ordered so that the first few capture the most significant variance in the data. In order to determine the most important features, we aimed to identify those that explain more than 90% of the total variance. This threshold provides an effective balance between retaining the majority of the data's variability and reducing the dataset's dimensionality [26,27]. We computed the loading for each feature,

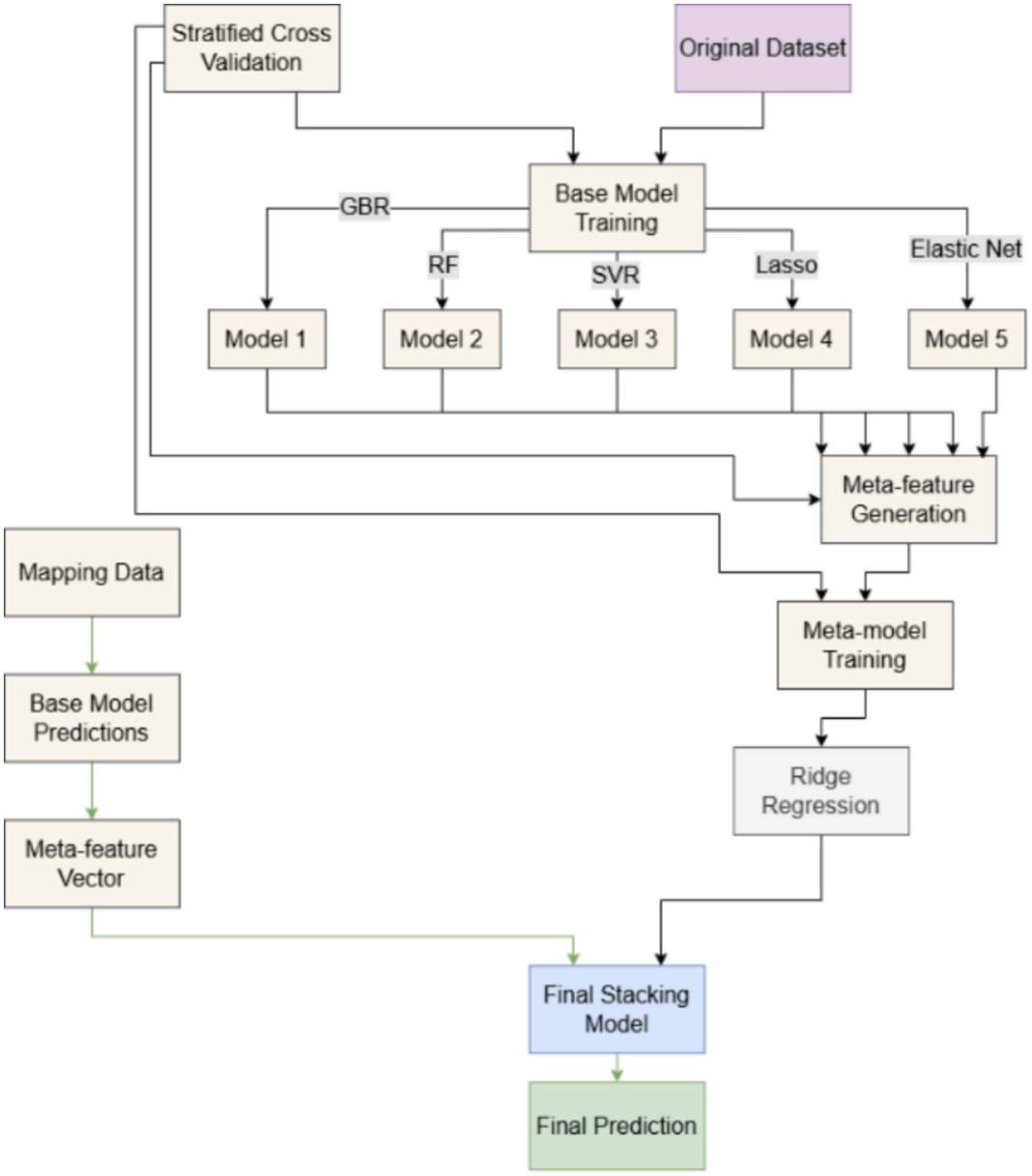

Fig. 2. Process description for developing the stacked ensemble model and fitting of unseen data (*e.g.*, the mapping data). A stratified fixed splitting strategy was used to divide the dataset into training (approximately 104 samples), validation (16 samples), and test (16 samples) sets, ensuring representative target variable distribution across all partitions.

which represents the contribution of that feature in forming each PC. Higher loading indicates a stronger influence on the corresponding component.

We ranked the features based on two key metrics: the average absolute loading and the maximum absolute loading. The average absolute loading is the mean of the absolute values of a feature's loading across the selected PCs. This reflects its overall importance across all considered components. The maximum absolute loading, on the other hand, identifies the highest absolute value of a feature's loading across the PCs. This highlights its strongest contribution to any single component.

The second approach employed a wrapper method to identify the most effective combination of features for predictive performance. This method systematically evaluated different feature subsets in conjunction with model training and optimization. We defined two feature lists: "*mandatory*" features always included in the model and "*optional*" features explored in every possible combination (see Table 1 for feature designation). The wrapper method generated all subsets of optional features and combined them with the mandatory features, creating diverse feature combinations. For each combination, we trained the model using the ensemble method detailed in Fig. 2 to predict the T_g. These results are presented in section 3.2. By iterating through all possible features and model combinations and evaluating their performance based on predefined metrics, the wrapper method identified the optimal combination that yielded the best predictive accuracy. This approach ensured that feature selection was directly tied to the model's performance. Notably, during the early stages of the study, we observed low performing models when using only the four input variables (*e.g.*, the parameters shown in Fig. 1). This prompted us to integrate additional features, which included the swelling ratio output as an input feature. The results presented in Supporting Information section 3.2 demonstrate how datasets without swelling ratio return a higher Mean Absolute Error (MAE) than when swelling ratio is used in certain ML architectures we experimented with (*e.g.*, out-of-fold cross validation).

After completing the PCA and wrapper analyses, we assessed the results and we compared the outcomes against domain expertise and a correlation analysis of features. The selection of the dataset features was ultimately based on the model performance with the lowest MAE values. This integration guided our selection of the optimal feature set for the final ML model, balancing predictive power with interpretability for T_g prediction. Additional results of the PCA are available in Supporting Information section 3.1.

2.2.4. Ensemble learning regression

The ensemble method employed in this study combined multiple base models with a meta-model to improve predictive performance. The implementation scheme is shown in Fig. 2 and it is commonly referred to as stacking ensemble.

There are two other methods such as bagging and boosting and they differ in the way they combine the base models and the objectives they aim to achieve [28]. Bagging, short for bootstrap aggregating, aims to reduce variance by training multiple models on bootstrapped samples of the original dataset and averaging their predictions. Boosting, on the other hand, focuses on reducing bias by sequentially training models that give more weight to misclassified instances from previous iterations. While bagging models are typically independent, boosting creates a strong dependency between successive models in the ensemble. The key distinguishing feature of stacking is its use of a meta-learner, which learns to combine the predictions of diverse base models in an optimal way, potentially capturing complex interactions between these models and leading to superior overall performance [29]. However, within the stacking ensemble we implemented, the tree-based models on their own can be considered as bagging methods.

Furthermore, the process description, which graphically outlines the model development and highlights feature importance and model strengths, is divided into three subsections. The first section focused on the tuning and performance assessment of the base models. The base models were trained and validated using the dataset derived from the analysis carried out under the feature selection section 2.2.3. This dataset contained the predictive features and the target variable T_g.

To evaluate the model's performance, a stratified fixed splitting approach was used. Rather than employing cross-validation splits, a stratified splitting technique was implemented to ensure that the validation and test sets were representative of the entire target variable distribution, which was important given the relatively small sizes of these sets. The stratified splitting process involved three sequential steps: (1) sorting all samples by the target variable (T_g) in ascending order, (2) applying systematic sampling to select 16 validation samples evenly distributed across the sorted range, and (3) applying systematic sampling again to the remaining samples to select 16 samples, with all remaining samples (104) allocated to the training set. This systematic approach ensured that low, medium, and high T_g values were proportionally represented in all three data partitions, thereby preventing the bias that could arise from random sampling with small validation and test sets (see Fig. S4 in Supporting Information for stratified statistics). A consistent random seed was used to guarantee reproducibility and enable fair comparison across different feature combinations and model configurations.

Then, for each fold, we trained a variety of base models, (*e.g.*, GBR, RF, Support Vector Regressor (SVR), Lasso, and ElasticNet algorithms). Each algorithm underwent hyperparameter optimization using GridSearchCV with predefined parameter grids (see Table S3 in Supporting Information). To leverage the fixed validation set for hyperparameter selection while maintaining proper data separation, the training and validation sets were temporarily combined for GridSearchCV, with a custom cross-validation split that designated training samples as the training fold and the validation samples as the validation fold. This ensured that hyperparameter selection was based on performance of the fixed validation set. After identifying the optimal hyperparameters, each base model was retrained exclusively on the training set to prevent data leakage. For the tree-based models (GBR and RF), the number of base estimators was also tuned, testing values starting from 1, 10, 50, 100, and then incrementing by 100 up to 1000, to determine the optimal number of estimators. The remaining hyperparameters (*e.g.*, learning rate, max depth, regularization strength) were tuned via GridSearchCV. The trained base models then generated predictions on all three data partitions (training, validation, and test sets), creating the meta features for the second layer of the ensemble. All five base models were used to construct the ensemble. Therefore, the *meta*-feature matrix had rows corresponding to the number of data points (*i.e.,* 104) and columns corresponding to the number of base models (*i.e.,* 5), while the validation and test sets each produced (16×5) meta-feature matrices.

The second section focused on the performance assessment of the meta-learner and the ensemble as a whole. The ridge regression algorithm was used as the meta-model. The meta-model was trained on the training set meta-features and evaluated on both the validation and test set meta-features. Unlike the base models, the meta-model did not undergo hyperparameter tuning; instead, it was trained with default Ridge regression parameters.

In both instances, the base models and the meta-model were monitored with regards to the metrics defined in section 2.2.5. Metrics were computed separately for the training, validation, and test sets, providing comprehensive insight into model fitting, hyperparameter selection quality, and generalization performance respectively All metrics were calculated on the original scale of the target variable (°C) by applying inverse transformation using the fitted scaler, ensuring interpretability of the results.

The final model was identified based on the test set performance, specifically minimizing test MAE while monitoring validation MAE to assess consistency as well as other metrics as described in 2.2.5. This was used to determine new formulations by means of mapping unseen data or the use of an adaptive grid search method described in section 2.3 and section 2.4.

2.2.5. Model evaluation metrics

The evaluation of the ML models was based on three key metrics: R^2, mean square error (MSE) and MAE.

R^2 is the coefficient of determination, and it indicates the proportion of variance in the dependent variable predictable from the independent variable(s). R^2 ranges from 0 to 1, where 1 indicates a perfect fit.

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

MSE measures the average squared difference between estimated and actual values. It is always non-negative, and values closer to zero are better. MSE gives a higher weight to larger errors, as the differences are squared before they are averaged.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (5)$$

MAE measures the average magnitude of errors in a set of predictions, without considering their direction. It is the average over the test sample of the absolute differences between prediction and actual observation. MAE is less sensitive to outliers compared to MSE.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (6)$$

where $y_i$is the actual value, $\hat{y}_i$is the predicted value, $\bar{y}_i$is the me an of actual value, and $n$is the number of observations.

2.3. Mapping for model behavior

After developing the ML model, we sought to understand the model's behavior and the relative importance of different features in determining the output. We did this by means of mapping, through a systematic exploration of the entire feature space defined by the input parameters. We established discrete ranges for each feature (see Table 2). Using these predefined ranges, then we generated all possible combinations of feature values. For each combination, the input data was scaled using a pre-trained scaler, and then fed into the best base models developed above. The predictions from these best base models (described as meta-features) were then input into the meta-model, which produced the final prediction (Fig. 2). Through mapping the model performance, we acquired knowledge on the synthesis formulation of over 4 million sample combinations.

2.4. Predicting new synthesis parameters for formulations using an adaptive grid search

The analysis of our model's performance through mapping enabled us to develop new synthesis formulations effectively. However, we encountered significant challenges in the mapping process related to input feature increments. Using larger increments led to gaps in coverage of the desired target T_g values, while using smaller increments (as detailed in Table 2) generated over 4 million unique synthesis combinations. This vast number of combinations exceeded the practical limits of our computing hardware, namely a 13th Gen Intel® Core™ i7-1365U processor running at 1.80 GHz. More specifically, the visualization library (*i.e.*, plotly) failed to render the parallel coordinates plot, our chosen method for visualizing the relationships between synthesis parameters.

Furthermore, our initial mapping strategy's parameter-input approach, while effective at demonstrating feature-T_g relationships, lacked direct output control. This limitation necessitated developing an alternative strategy that could balance fine-grained target output resolution with reasonable computational demands.

To address these challenges, we implemented an adaptive grid search algorithm that optimizes input parameters for T_g prediction using our ensemble ML model. The algorithm operates on the seven input parameters initialized as a coarse grid (Table 3). For each target T_g, it systematically evaluates combinations of these parameters, using both base and meta-models for T_g prediction.

The algorithm works through an iterative refinement process. Initially, for each target T_g, it systematically evaluates all parameter combinations within the current grid using both base and meta-models for prediction. The absolute difference between each predicted T_g and the target T_g is then calculated to identify the parameter combination that yields the smallest difference, representing the best match. Following this identification, the algorithm narrows the search range around these optimal parameters while maintaining the grid density. This entire process is repeated for three iterations, each time working within a more focused parameter space to progressively improve prediction accuracy.

We applied this method across target T_g values from −17 °C to 100 °C in 2.39 °C increments. This systematic approach allowed us to identify optimal synthesis parameters for each target T_g temperature while avoiding the computational limitations of our initial exhaustive mapping strategy. The final output provided a comprehensive set of synthesis formulations.

This process generated a dataset of optimal parameter combinations and their corresponding predicted T_g values for each target. By setting a wider range for T_g than the data points available in the training dataset, we also explored the extrapolation capacities of the ML model. We plotted the input determination dataset using parallel coordinates. This plot is advantageous primarily because it can display multiple variables in a single 2D graphic. It also allows visualization for easy identification of relationships between different parameters and the T_g and can ensure the adoption of the ML model by experimental chemists.

3. Results and discussion

3.1. Dataset overview

3.1.1. Correlation of features

The analysis of correlations with T_g reveals intriguing relationships among various features (Fig. 3).

Lignin (wt%) exhibits the strongest positive correlation with T_g, indicating its significant influence on this property. This observation aligns with established chemical reasoning from the literature. Lignin, a complex polymer composed of phenolic compounds, is known to affect the crosslinking density of lignin-PU materials. The presence of aromatic

Table 2
Mapping parameters to explore the T_g ¬ prediction capabilities.

Feature	Interval	Representation
Lignin (wt%)	Discrete integers ranging from 0 to 70 (step of 1)	0, 1, 2, …, 70
Co-polyol type (PTHF)	Discrete numeric values representing molecular weights	250, 650, 1000
Ratio	17 equidistant points in the range from 0.6 to 1.4 (step of 0.05)	0.6, 0.65, 0.7, 0.75…1.4
Co-polyol (wt%)	5 evenly spaced values between 0 and 66	0, 16.5, 33, 49.5, 66
Isocyanate (wt%	Discrete values ranging from 0 to 20 (step of 0.4)	0, 0.4, 0.8 …, 19.6, 20

Table 3
Grid features and their parameters.

Features	Values
Lignin (wt%)	0, 33.33, 66.67, 100
Co-polyol type (PTHF)	250, 650, 1000
Ratio	0.6, 0.8, 1.0, 1.2, 1.4
Co-polyol (wt%)	(Calculated as 100 – Lignin (wt%)): 0, 33.33, 66.67, 100
Isocyanate(wt%)	0, 5, 10, 15, 20

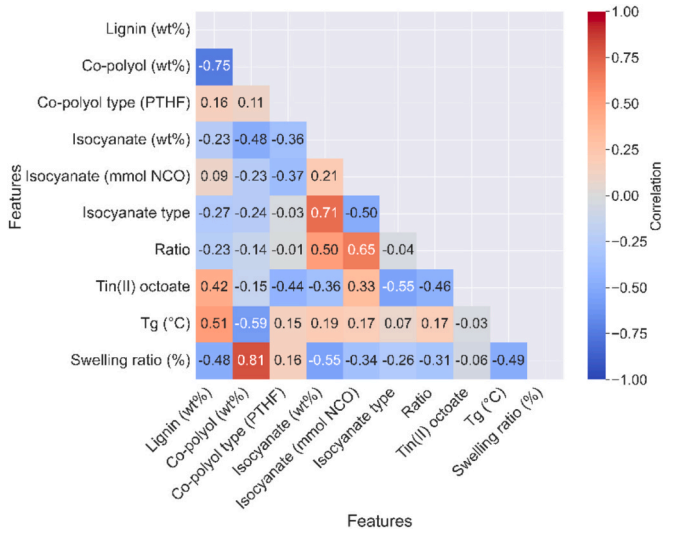

Fig. 3. Heat map showing the correlation coefficients between the features within the dataset, 1 indicates a perfect positive correlation, −1 indicates a perfect negative correlation and 0 indicates no correlation.

groups and π-π interactions can enhance the rigidity and thermal stability of the material. As the lignin content increases, the network of crosslinks becomes denser, restricting molecular mobility and leading to a higher T_g [30–32]. Conversely, the *Co-polyol (wt%)* demonstrates the most substantial negative correlation with T_g. This observation aligns with previous knowledge, where an increase in the co-polyol as flexible soft segment leads to an increase in free volume and a decrease in T_g [33]. Isocyanate-related features *Isocyanate (wt%)* and *Isocyanate (mmol NCO)* show similar positive correlations, followed closely by the *Ratio*. As the isocyanate increases, the crosslinking density within the polymer network increases thereby restricting segmental motion and elevating T_g [34]. These correlations can also be logically explained by the inherent relationships between these features within the dataset, providing valuable insights into the interconnected nature of the variables under study.

Examining correlations with swelling ratios as experimentally accessible proxy for network density uncovers a different pattern. The co-polyol feature *Co-polyol (wt%)* displays the strongest positive correlation, followed by its type *Co-polyol Type (PTHF)*. The strong positive correlation between *Co-polyol (wt%)* and *Swelling ratio* suggests that as the co-polyol content increases, the swelling ratio also increases. As the co-polyol content increases, the amount of multifunctional lignin decreases. By consequence, the lignin-PU network becomes less rigid and more flexible, reducing crosslink density, increasing free volume, and improving THF (solvent for swelling studies) affinity, all of which contribute to a higher swelling ratio. *Isocyanate (wt%)* shows the most pronounced negative relationship with the swelling degree. As the isocyanate content increases, the lignin-PU network becomes more crosslinked, less flexible, and less permeable to THF, leading to a lower swelling degree [35].

3.1.2. Clustering of data

In Fig. 4, we used density contours and scatter plots to deepen the understanding of the experimental dataset.

From Fig. 4 (A), which used HDIt within its synthesis formulation, the distinct bimodal clustering of data points shows that low lignin content (10–20 wt%) correlates with low T_g and increased swelling ratio. Higher lignin content (30–40 wt%), highlighted by green and blue dots, generally shows increased T_g with lower swelling ratios, though some data points appear dispersed in the lower T_g regions, particularly for lignin 30 wt%. Higher lignin content introduces more aromatic units, which increase T_g, while also enhancing crosslink density, thereby reducing the swelling ratio. A more in-depth discussion of the swelling ratio and feature influence is provided in Supporting Information section 2.

Fig. 4 (B), analyzing *Co-polyol type (PTHF)* and *Ratio*, demonstrates that higher molecular weights (M_n 1000 g/mol) for the co-polyol tend to cluster in the higher T_g region, particularly when combined with higher lignin contents (as observed in conjunction with Fig. 4 (A)). The intermediate M_n (650 g/mol) shows a bimodal distribution, with points present in both high and low T_g clusters. The $M_n = 250$ g/mol co-polyol data points show a distribution pattern that correlates strongly with lignin concentrations. Notably, samples with lower Ratios (0.6–0.8) tend to appear as outliers with higher swelling ratios, what can be explained by the low crosslink efficiency due to isocyanate deficiency.

Fig. 4 (C) shows the HDI dataset, where the bimodal distribution remains evident but with different clustering characteristics than HDIt. The higher lignin content (45 wt%) samples maintain the trend seen with HDIt: higher T_g with lower swelling ratio when the polymers get more crosslinked. Notable outliers include samples showing isolated clusters around 400–500% swelling ratio and low lignin content, giving rise to loosely crosslinked networks with high swelling. In Fig. 4 (D), the distribution of co-polyol molecular weight types maintains similar patterns to those observed in Fig. 4 (B) of the HDIt dataset, preserving the bimodal distribution but with distinct clustering characteristics.

This data analysis reveals that the dataset exhibits clear bimodal distributions with strong negative correlations (−0.56 and −0.48), where the composition of compounds significantly impacts the properties of the PUs. These results reinforce the challenge we are addressing: identifying optimal combinations through purely experimental methods would be an extremely time-consuming and resource-intensive process.

3.2. Feature selection

PCA revealed that four components were necessary to explain 95% of the variance in the data, and three would reach 87% of the variance (see Fig. S3, Supporting Information). We selected four since this exceeds the pre-set acceptance criteria. To interpret the feature importance, we examined both the average and maximum absolute loadings (see Fig. 5). The average absolute loading indicates a feature's overall importance across all PCs, while the maximum absolute loading highlights its peak contribution to any single PC.

Ratio exhibited the highest average absolute loading of 0.35, followed by Tin(II) octoate at 0.32 and Lignin (wt%) and *Isocyanate (wt%)* at 0.30. The remaining features showed lower average loadings, with Co-polyol type (PTHF) recording the lowest value at 0.19. Regarding maximum absolute loading, Swelling ratio (%) and Isocyanate (mmol NCO) demonstrated the highest values at 0.77 and 0.73 respectively. Co-polyol type (PTHF) showed moderate importance in this metric at 0.55, while Tin(II) octoate and Ratio had lower maximum loadings at 0.52 and 0.47 respectively.

Given the variability between these two loading metrics, we opted to complement our analysis with a wrapper method for feature selection. We retained *Lignin (wt%)*, *Co-polyol type (PTHF)*and *Ratio* as a base set ("*mandatory*" features), then systematically added and alternated the remaining features ("*optional*" features).

We identified the top five performing models and observed several aspects in the features selection (see Table 4). For example, two exceptions (ranks 4 and 5) included *Isocyanate type* in place of *Isocyanate (wt%)*, while one exception (rank 2) excluded *Co-polyol (wt%)* entirely.

Co-polyol (wt%) was present in four of the five top-performing models; despite its relatively low importance in the PCA analysis, its inclusion consistently improved model performance, suggesting it captures variance not fully reflected by the principal components. The model that did not contain this feature (rank 2) returned a slightly higher MAE for the validation dataset.

Additionally, the *Isocyanate (wt%)*, which is related to the Ratio feature, is also part of the datasets returning the lowest MAE metrics.

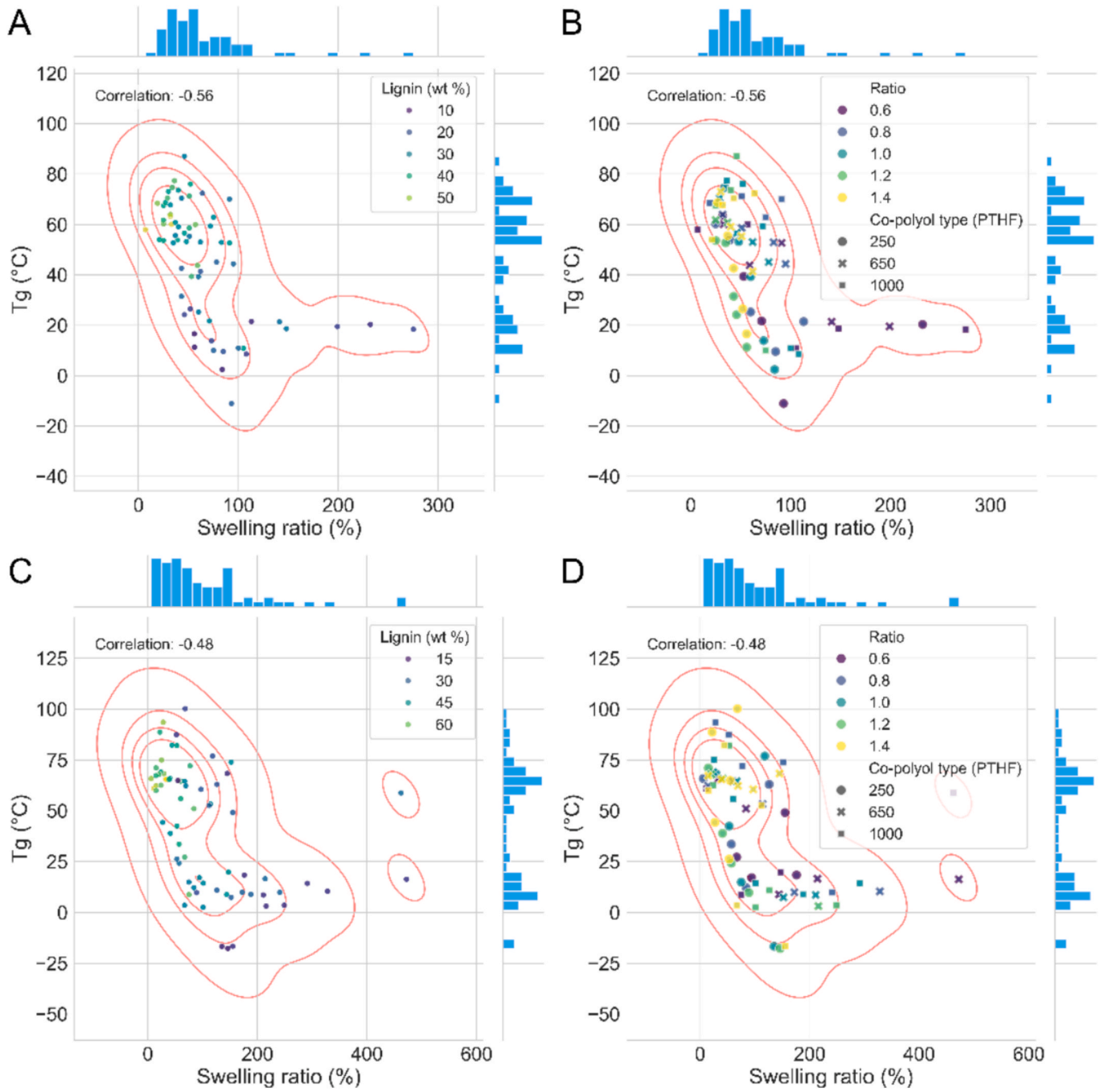

Fig. 4. Partial dependence plots of three representative features: Lignin (wt%), Co-polyol Type (PTHF) and the Ratio. The dataset was divided into two groups with (A) and (B) plots showing 90 samples, where HDIt is present. (C) and (D) is the dataset composed of 90 samples with HDI in the composition.

The *Isocyanate (wt%)* and *Ratio* have shown a moderate positive correlation (see Fig. 3), and in addition to the proportional factor that Ratio provides, the *Isocyanate (wt%)* contains exact weight percentage values, which positively influences the model's understanding of the dataset. Notably, the models that utilized *Isocyanate type* instead of *Isocyanate (wt %)* (Ranks 4 and 5) returned higher MAE values on the validation dataset. Although *Isocyanate (wt%)* and *Isocyanate type* are strongly correlated, the use of continuous numerical values over categorical representations contributes to better model accuracy. The *Co-polyol type (PTHF)* is also consistently part of all five top-performing datasets, and as evidenced by the PCA's Maximum Absolute Loading Value, it does have an important role on model performance. Based on this metric, it was the third most important feature.

After analyzing the dataset through the three methods: correlation, PCA, and the wrapper method applied within the ensemble model, it is evident that the features emphasized in the stacking ensemble are highly valuable. These features provide compelling insights that can be leveraged to develop a robust final model for predicting the T_g of lignin PUs.

3.3. Model performance as a function of base estimators

In Table 4, the performances of the ensembles showed to fluctuate based on the number of base estimators present in the tree-based algorithms. In this section, we investigated this performance from 1 to 1000 base estimators within intervals described earlier in section 2.2.4. We used the dataset comprising the features we determined above.

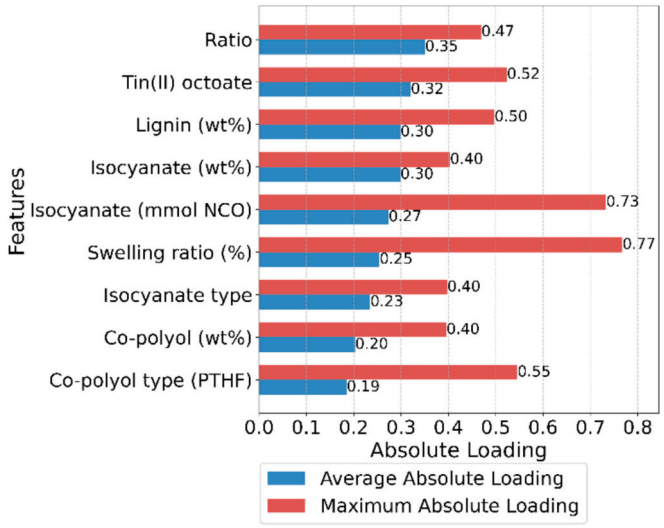

Fig. 5. Features ranked by average and maximum absolute loading based on 4 PCs explaining 97% of data. T_g was excluded from the dataset when applying PCA.

Fig. 6 shows the metrics across the base estimators of the ensemble. The train R^2 values steadily increase from 0.942 to 0.997 as the number of base estimators grows, and the test R^2 similarly improves from 0.457 to 0.765 (Fig. 6A). However, the validation R2 slightly decreases from 0.494 to 0.428, indicating that while the model's ability to explain variance improves on training and test data, this improvement does not fully generalize to the validation set. The widening gap between the train and validation curves indicates overfitting, especially at higher numbers of base estimators, where the stacked model performs substantially better on the training data than on unseen data. The train MSE decreases markedly from 41.936 to 2.479 with increasing base estimators (Fig. 6B). The test MSE also generally decreases; however, the validation MSE, after an initial decrease to approximately 338.368, increases to 443.225 at higher estimator counts. This divergence between train and validation MSE further confirms the overfitting trend. Similarly, the train MAE decreases from 3.918 to 0.982 and the test MAE from 15.17 to 10.814 (Fig. 6C). In contrast, the validation MAE increases from 13.41 to 15.168, reinforcing the observation that additional estimators primarily benefit training and test performance at the expense of validation generalization.

This trend is consistent with prior polyurethane ML studies on small datasets, where stronger apparent fit on the training set did not necessarily translate into improved generalization. Menon *et al.* [36] specifically highlighted that small and chemically diverse PU datasets are difficult to model using tree-based approaches alone, and Ding *et al.* [37] showed that predictive performance improved substantially only after reducing intrinsic dataset diversity through benchmark curation.

Although there is evidence of overfitting in the stacked models, they still display better performances than the individual algorithmic models across validation and test metrics. As shown in Table 5, the stacking ensemble outperforms all individual base models on the validation dataset, achieving the lowest validation MAE of 13.41 compared to the best individual model (GBR) at 15.99, and the highest validation R2 of 0.49 compared to GBR's 0.45. On training data, the improvement is even more pronounced, with the stacking ensemble achieving an MAE of 3.92 and R^2 of 0.94, substantially surpassing all base models. This suggests that the meta-features generated by the base models, together with the application of Ridge regression as a meta-model, have successfully captured generalizable patterns in the data while maintaining good

Table 4
Results of the validation dataset during feature selection using the wrapper method.

Rank	R^2		MSE		MAE		# base estimators	Features
	Test	Validation	Test	Validation	Test	Validation		
1	0.55	0.49	338.37	391.95	15.17	13.41	10	Lignin (wt%), Co-polyol type (PTHF), Ratio, **Co-polyol (wt%),** Isocyanate (wt%)
2	0.20	0.46	595.12	417.46	17.77	13.54	10	Lignin (wt%), Co-polyol type (PTHF), Ratio, Isocyanate (wt%)
3	0.61	0.49	292.55	392.19	14.26	13.55	100	Lignin (wt%), Co-polyol type (PTHF), Ratio, **Co-polyol (wt%),** Isocyanate (wt%)
4	0.63	0.43	273.85	439.07	10.99	14.06	100	Lignin (wt%), Co-polyol type (PTHF), Ratio, **Co-polyol (wt%)**, Isocyanate type
5	0.67	0.43	245.82	439.48	10.21	14.08	500	Lignin (wt%), Co-polyol type (PTHF), Ratio, **Co-polyol (wt%)**, Isocyanate type

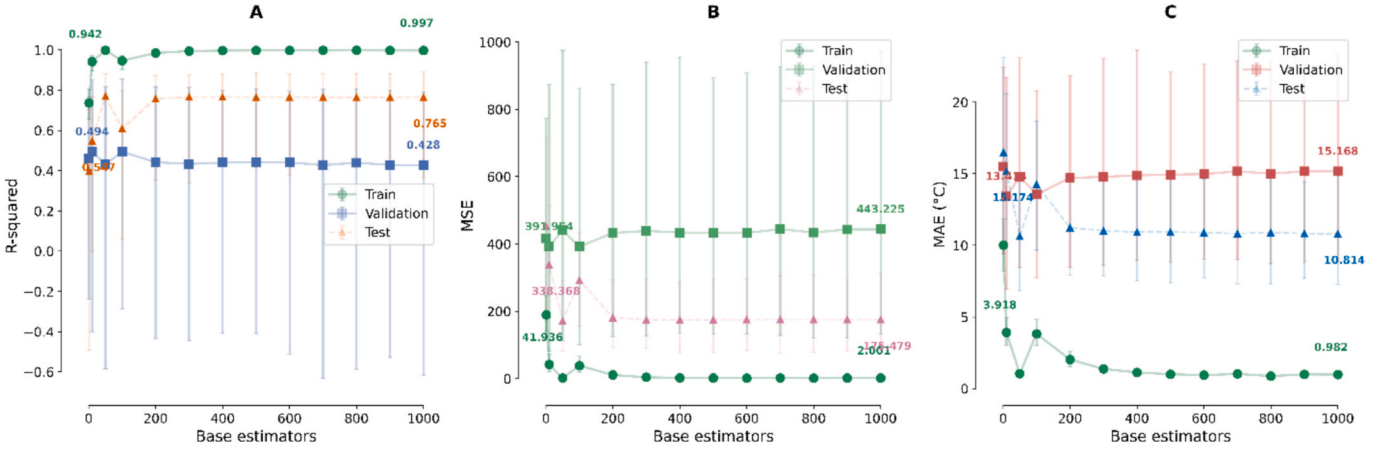

Fig. 6. Performance metrics of the Stacked Ensemble with stratified sampling at various base estimators from 1 to 10 to 50, 100 and up to 1000 of the tree-base algorithms.

Table 5
Performance metrics of the individual models and the stacked ensemble of the selected model. Generalizability represents the difference between the Validation MAE and the Training MAE, where a negative value means it underfits and a positive value it overfits.

MODEL	R^2			*MSE*			*MAE*		
	Train	**Test**	**Val**	**Train**	**Test**	**Val**	**Train**	**Test**	**Val**
GBR	0.81	0.53	0.45	134.18	348.84	429.07	10.00	15.26	15.99
RF	0.74	0.51	0.29	189.89	364.11	550.17	9.78	15.74	17.58
SVR	0.40	0.50	0.27	432.04	370.88	562.71	15.22	15.45	17.94
LASSO	0.31	0.32	0.24	501.53	507.15	589.55	18.44	19.09	19.59
ELASTICNET	0.39	0.47	0.30	437.72	399.07	543.65	16.04	16.58	18.51
STACKING ENSEMBLE	0.94	0.55	0.49	41.94	338.37	391.95	3.92	15.17	13.41

performance across both training and validation sets. The stacking approach has effectively leveraged the strengths of each individual model, combining their predictions in a way that mitigates their individual weaknesses. There is also the question of the optimal number of base estimators and the trade-off between performance gains and computational cost. Hence, it can be deduced that the stacked ensemble method regularly yielded the best predictions, as evaluated by MAE on the validation dataset.

When compared with the broader literature, the present validation performance should be interpreted as moderate rather than state-of-the-art. Polyurethane-specific and polymer-informatics studies have reported higher R^2 values, often in the range of approximately 0.71–0.91 or above, but those studies generally relied on either more curated benchmark datasets, alternative target properties, or substantially larger recipe libraries. Accordingly, the present performance is better viewed as realistic for a heterogeneous lignin-based polyurethane dataset rather than directly comparable to best-case results from larger or more homogenized studies [37–41].

Worth mentioning is that we also considered a different architecture in the building of the ensemble, namely an out-of-fold (OOF) cross-validation technique, in which base models generate predictions on held-out folds through nested cross-validation, ensuring the *meta*-learner is trained exclusively on OOF predictions to prevent data leakage. Under this scheme, the best-performing model configuration, using the same five-feature input, achieved an MAE of 15.71 °C and an R2 of 0.34, compared to the stratified fixed-split approach which yielded a validation MAE of 13.41 °C and R2 of 0.49 for the same feature set (see Section 3.2 in the Supporting Information for more results). The OOF framework produced notably weaker predictive performance, likely attributable to the limited dataset size, where repeated fold-based splitting reduces the effective training set per iteration and inflates variance in the *meta*-learner's training signal.

In the case of reference [18], 1000 base estimators were used with the aim of getting sufficiently smooth statistics. Based on the current results, where the validation MAE is lowest at lower estimator counts (~10) and begins to increase beyond that point, a smaller number of base estimators may have sufficed, though it should also be noted that the dataset used in the earlier study was significantly smaller, with fewer than 50 data points. The trend observed in Fig. 6 (C), suggests an asymptotic effect, with the MAE approaching zero. However, the validation MAE trend indicates diminishing returns and potential overfitting beyond a certain number of estimators. Additionally, adding more base estimators also increases training time and demands greater computational resources.

This interpretation is also in line with the general trade-off reported in ensemble learning for polymer-property prediction: beyond a certain model complexity, gains in apparent fit may mostly reflect variance reduction on the seen data rather than a true increase in transferable chemical insight. In small-data polyurethane systems, simpler ensemble settings may therefore be preferable when validation performance rather than training smoothness is the selection criterion [36,40].

3.4. Evaluation of model fit and residual distribution

Based on the model performance study in section 3.3, we selected the best performing stacked model that included 10 base estimators for the tree-based algorithms, and evaluated the model fit and residual distribution. Fig. 7 (A) shows a moderate linear relationship between the predicted and actual values for the T_g, with a Pearson correlation coefficient of 0.73. While the general trend follows the ideal fit line, there is considerable scatter, with several points deviating notably from the diagonal, notably in the mid-range values. From a model perspective, the overall alignment of points along the ideal line suggests the model has captured the general underlying pattern in the data, though not without notable prediction errors for certain datapoints. There is no obvious systematic bias visible, as points are distributed above and below the line across the range. The model appears to perform across a range of values, from approximately −10 °C to 75 °C. Fig. 7 (B) provides additional insights into the model's performance. The residuals appear roughly distributed around the zero line, which indicates the model is not strongly systematically over- or under- predicting. However, some larger positive residuals are observed at lower predicted values, suggesting the model tends to underpredict in certain cases within the range. Additionally, the spread of residuals appears somewhat larger at lower predicted values compared to higher ones, suggesting mild heteroscedasticity that warrants consideration.

3.5. Mapping data

Bivariate kernel density estimation (KDE) plots were used to analyze the mapping data (Fig. 8). In these plots, high-density regions indicate confident, well-supported predictions, whereas low-density regions suggest sparse training coverage or model uncertainty. The resulting density patterns are characteristic of ensemble methods trained on imbalanced datasets, where the majority T_g group dominates the estimation [42].

Fig. 8 (A) shows the relationship between *Lignin (wt%)* and T_g. High-density predictions shift progressively upward with increasing lignin content: 15–20 °C at 0–10 wt% lignin, 40–50 °C at 15–30 wt%, and 50–60 °C at 30–50 wt%, the latter representing the most prominent cluster in the entire mapping. Above 50 wt%, the density persists around 55–65 °C with reduced intensity. Prediction confidence is highest in the 30–50 wt% range, where the variance narrows to approximately 20 °C (50–70 °C), compared to ~ 40 °C at lower lignin contents. From a structure–property perspective, the progressive T_g increase is attributed to higher crosslink density and π-π stacking interactions introduced by the rigid aromatic lignin backbone that reduces chain mobility. The optimal processing window for maximum T_g enhancement lies between 30–50 wt% lignin, whole 50–70 wt% range provides more consistent thermal properties albeit with declining prediction confidence. Sparse

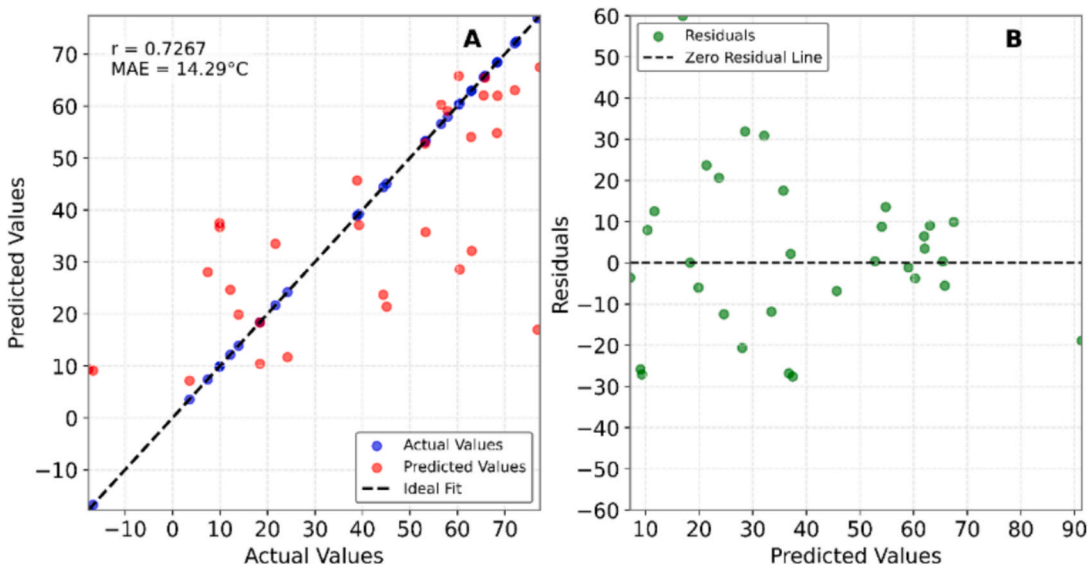

Fig. 7. Predicted versus actual T_g °C. The ensemble combines five base models (GBR, Random Forest, SVR, Lasso, ElasticNet) with 10 estimators each (for tree-based models), trained on the optimal feature combination [Lignin (wt%), Co-polyol type (PTHF), Ratio, Co-polyol (wt%), Isocyanate (wt%)]. (A) correlation between predicted and actual T_g values for validation and test datasets (n = 32), with a Pearson correlation coefficient of r = 0.73 and mean absolute error of 14.29 °C. (B) residual analysis showing prediction errors randomly distributed around the zero line, indicating no systematic bias in model predictions.

regions at low lignin contents and the discretization boundary around 10–15 wt% reflect sampling limitations in the training dataset.

In Fig. 8 (B) the *Co-polyol type (PTHF)* analysis reveals distinct prediction density clusters that shift non-monotonically across molecular weights: the highest density is centered at approximately 43–45 °C for PTHF 250, at 50–55 °C and 60–62 °C (bimodal) for PTHF 650, and at 53–55 °C for PTHF 1000 with a secondary region around 20 °C. Longer PTHF chains act as flexible soft segments that increase free volume and chain mobility, resulting in lower T_g. The broad vertical spread within each cluster (spanning ~ 10–70 °C) confirms that T_g is co-determined by other formulation parameters, particularly lignin content and crosslink density. The non-monotonic peak shift, from 45 °C (PTHF 250) to 60 °C (PTHF 650) and back to 55 °C (PTHF 1000), and the bimodal distributions for PTHF 650 and 1000 indicate multiple thermally distinct prediction domains, suggesting the model has captured fundamental relationships between co-polyol chain length and phase behavior.

In Fig. 8 (C), the *Ratio* exhibits two distinct horizontal prediction bands: a primary band at 50–55 °C and a secondary band around 20 °C, revealing a bimodal T_g distribution across the stoichiometric range. The highest prediction density is concentrated at Ratio = 0.55–0.85, with localized peaks approaching 60–63 °C near Ratio ~ 0.75–0.85. Above Ratio = 1.0, prediction confidence decreases progressively, indicating greater model uncertainty in the excess-isocyanate regime. The bimodal structure—with the upper band reflecting efficient crosslinking and the lower band (~20 °C) representing under- or over-crosslinked networks, indicates that the [NCO]/[OH] ratio acts as a threshold-sensitive parameter rather than a continuously linear predictor. Notably, the sub-stoichiometric range consistently produces more confident and higher T_g predictions, suggesting that excess hydroxyl groups contribute more favorably to network formation in these lignin-based PU systems.

Fig. 8 (D) shows continuous density distributions predominantly in the 45–60 °C range, distributed broadly across the entire 0–20 wt% isocyanate range. Unlike the other features, the absence of a clear concentration-dependent shift in peak T_g position indicates that Isocyanate (wt%) primarily modulates the breadth of achievable T_g values rather than their central tendency, with the thermal behavior being strongly co-determined by other formulation parameters.

In Fig. 8 (E), the co-polyol content exhibits a pronounced U-shaped dependence: a high-density region at 0 wt% centered around 60 °C transitions sharply to a minimum at 25–35 wt% (prediction density concentrated around 15–20 °C), before recovering above 50 wt% with a second high-density region at 60–70 wt% around 55–65 °C. This behavior reflects competing effects: at low co-polyol content, the rigid lignin-rich network dominates, yielding higher T_g; at intermediate concentrations, the flexible soft-segment phase dilutes the hard-segment network, depressing T_g to its minimum; while at high co-polyol content, increased chain entanglement density restores higher thermal properties.

In addition, we can also see the tendency of the model predictions in the overall target frequency distribution (Fig. S6). This data reveals a multimodal distribution with the strongest peaks concentrated in the mid-to-high temperature range, where frequencies reach approximately 10.928 predictions at the highest point. Several prominent peaks are observed in close proximity, with frequencies of 10.539, 10.381 and 10.331, forming a dominant cluster, followed by a secondary peak of 9.883. The distribution is notably right skewed, with predictions gradually building from very low frequencies, below −100 °C, passing through moderate-frequency region around −50 °C to 0 °C (where frequencies reach approximately 7.700), and culminating in the highest frequency region. Predictions at the lower and upper extremes of the temperature range remain relatively sparse. The same histogram demonstrates that the model makes the most predictions in the mid-range of predicted T_g values, with a gradual decline in frequency towards both lower and higher temperatures. The frequency drops significantly for predictions in the deeply negative range and shows lower confidence at extreme temperatures on both ends of the spectrum. To address this issue, future studies could focus on correcting the imbalance or expanding the dataset through synthetic augmentation.

Taken together, the mapping trends align with established polymer physics principles and translate into actionable formulation design guidelines. Lignin content is the primary T_g lever: formulations targeting high T_g (>50 °C) should employ 30–50 wt% lignin, where aromatic crosslinking and π–π stacking interactions are maximized. Co-polyol molecular weight provides secondary thermal tuning, with PTHF 650 offering the broadest T_g design space due to its bimodal distribution. The [NCO]/[OH] ratio should be maintained in the sub-stoichiometric (0.55–0.85) to stoichiometric range for maximum T_g and prediction confidence, as excess isocyanate reduces both. Co-polyol content should either be minimized (<25 wt%, favoring rigid lignin-rich networks) or increased above 50 wt% (where entanglement effects restore thermal

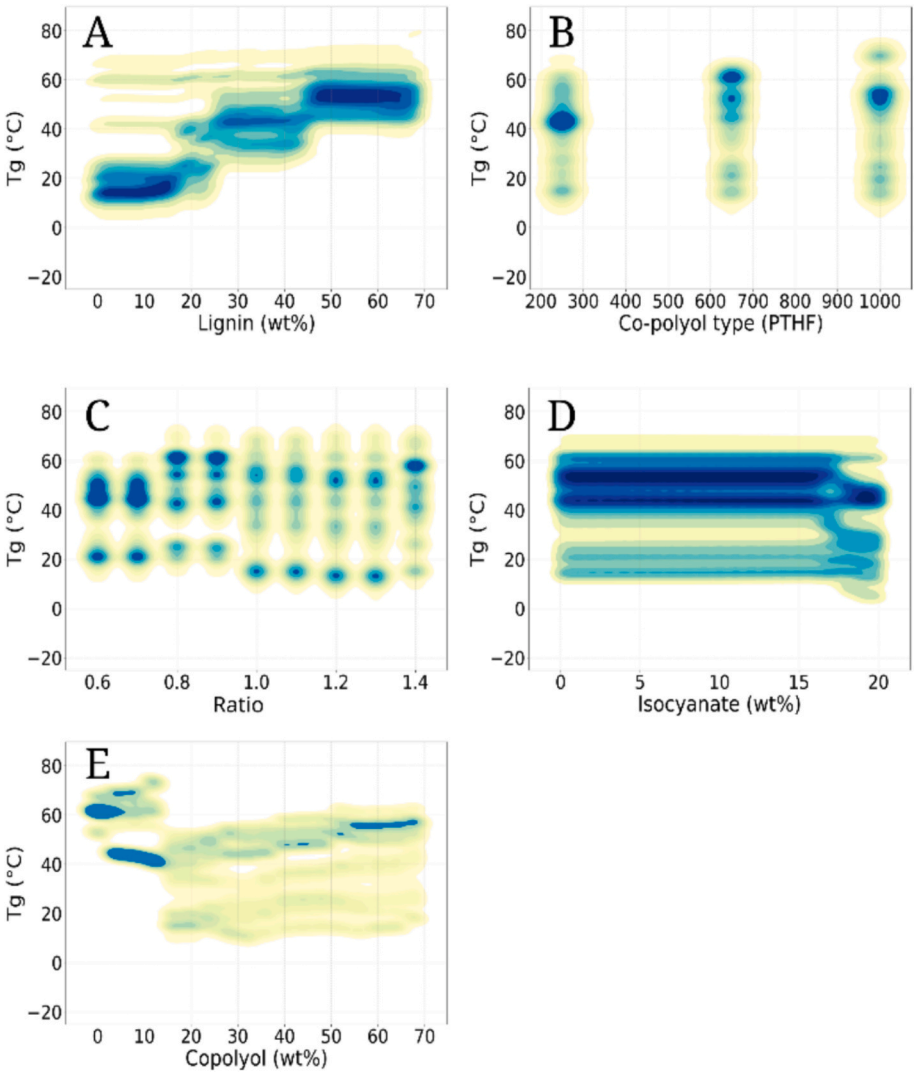

Fig. 8. Kernel density estimation (KDE) with a Gaussian kernel and 10 contour levels was used to approximate the probability density function of the mapping data as a function of (A) lignin wt%, (B) co-polyol type, (C) ratio, (D) isocyanate wt%, (E) copolyol wt%.

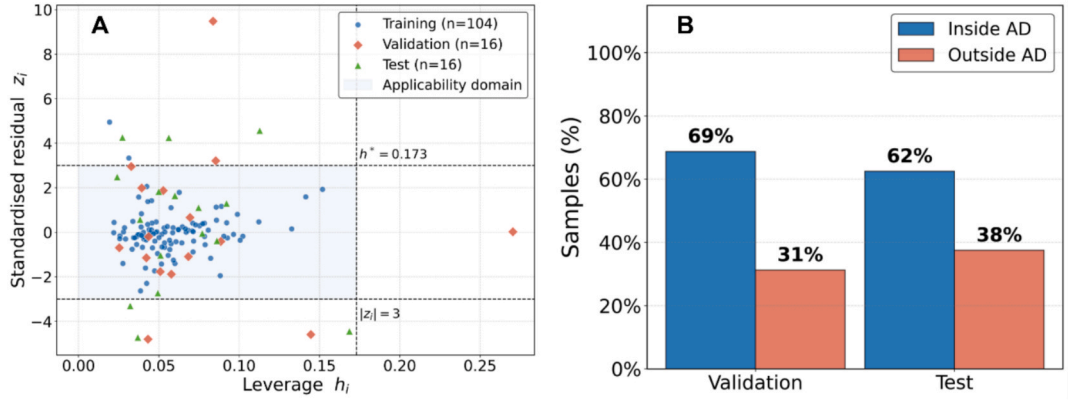

Fig. 9. Applicability domain (AD) analysis of the stacking ensemble model. (A) Williams plot showing the leverage (h_i) versus standardized residual (z_i) for training (filled circles, n = 104), validation (filled diamonds, n = 16), and test (filled triangles, n = 16) samples. The vertical dashed line denotes the leverage warning threshold $h^* = 0.173$ [$h^* = 3(k + 1)/n$; k = 5, n = 104] and horizontal dashed lines indicate the ± 3 reliability criterion for standardized residuals. The shaded region marks the applicability domain. (B) Percentage of validation and test samples classified as inside (blue) or outside (red) the AD. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

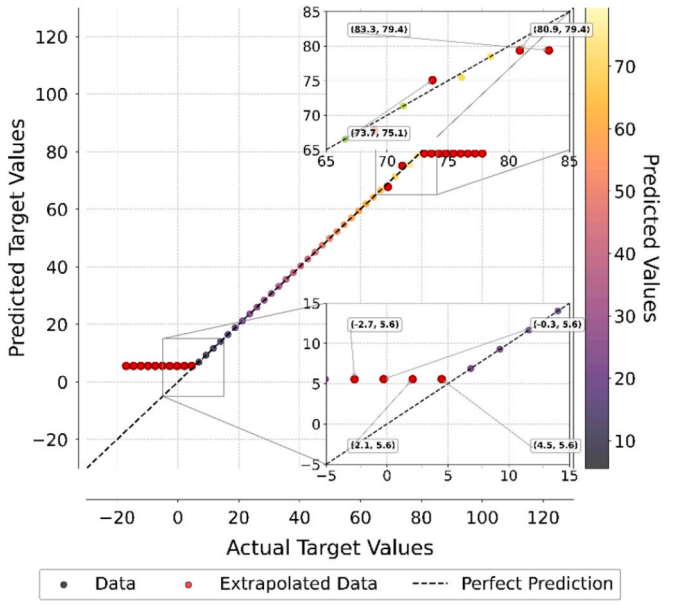

Fig. 10. Extrapolation of T_g between −17 °C and 100 °C. The bottom range predicts target until the 5 °C data point. In the upper range the model extends accurate predictions up to 99 °C. The original dataset had four negative T_g values, (−11 °C, −16 °C x2, −17 °C), and the highest positive T_g values being 88 °C, 93 °C and 100 °C.

performance), while the intermediate range (25–35 wt%) should be avoided when high T_g is desired. These guidelines, derived from the systematic exploration of over 4 million formulation combinations, provide a rational framework for the design and optimization of lignin-based polyurethane architectures with targeted thermal properties.

3.6. Applicability domain analysis

To delineate the chemical space within which the stacking ensemble produces reliable predictions, a leverage-based applicability domain (AD) analysis was performed. Leverage values (h_i) were computed from the HAT matrix of the scaled training set, and standardised residuals ($z_i = e_i/\sigma$) were derived using the training residual standard deviation. The warning threshold was set at $h^* = 3(k + 1)/n = 0.173$ ($k = 5$ features, $n = 104$ training samples), and a sample was considered reliable if it simultaneously satisfied $h_i \leq h^*$ and $|z_i| \leq 3$. The outcome is visualised in the Williams plot (Fig. 9 (A)). Of the training samples, 98.1% (102 of 104) fall within the AD, confirming that the model is internally well-calibrated. For the validation and test sets, 68.8% (11 of 16) and 62.5% (10 of 16) of samples lie within the AD, respectively (Fig. 9 (B)). Samples outside the AD are characterised by either atypically high leverage, reflecting structural dissimilarity from the training centroid, or by large, standardised residuals, which is consistent with the scatter visible in Fig. 7 (A). This finding contextualises the prediction errors discussed in Section 3.5: the model performs reliably within its defined chemical space, while formulations at the periphery of the training distribution carry increased prediction uncertainty.

To confirm that the ensemble has captured genuine structure–property relationships rather than spurious correlations inherent to a small dataset, a permutation test was conducted. The T_g labels of the training set were randomly shuffled (1000 permutations), and the validation MAE was re-evaluated for each permutation while preserving the fitted base models. As shown in Fig. S7 (Supporting Information), the true validation MAE of 13.41 °C lies substantially below the fifth percentile of the permuted distribution (22.01 °C), corresponding to $p < 0.001$. This result provides statistical confirmation that the predictive performance of the stacking ensemble is not attributable to chance, offering additional evidence for its validity within the defined applicability domain.

3.7. Extrapolation

Beyond analyzing our model's performance within the known temperature range, we tested its ability to predict T_g in the boundary regions of our dataset (Fig. 10). Using an adaptive grid search method that refined broad intervals into precise results, we explored predictions from −17 °C to 100 °C. As shown in the regression scatter plot (Fig. 7), the model performed well within the majority of the trained range. However, extrapolation led to a noticeable decline in performance beyond certain thresholds at both ends of the range. At the lower boundary, the model's predictions plateaued at approximately 5.6 °C for actual values ranging from −2.7 °C to 4.5 °C and below, as highlighted in the inset of Fig. 10. Similarly, at the upper boundary, predictions converged toward approximately 79.4 °C for actual values beyond roughly 80 °C, with transitioning beginning around 73.7 °C where the predicted value (75.1 °C) already began deviating from the ideal fit. The model retained accuracy for points closest to the training range boundaries; but as predictions extended further beyond these boundaries, its output became increasingly constant rather than tracking the actual values, reflecting characteristic flattening behavior of tree-based ensemble models during extrapolation. These limitations are well-documented in the literature, even with other models and architectures such as neural network and linear regression models [43].

3.8. User interface

The development of a model capable of accurately predicting the T_g represents a significant advancement in lignin-based PU formulations. However, a notable challenge persists: the accessibility and usability of such models for practicing chemists. To address this gap, we propose implementing a parallel coordinates plot as an intuitive and effective user interface. This visualization technique enables chemists to interact with the ensemble model and interpret data generated through the adaptive grid search algorithm.

The parallel coordinates visualization technique presents several key advantages for analyzing T_g formulation spaces. This methodology enables the simultaneous representation of multiple dimensional parameters, providing researchers with a comprehensive visualization of the complex variable interactions that influence T_g behavior. Through dynamic filtering capabilities, researchers can systematically explore specific parameter ranges, thereby facilitating the rational identification of optimal formulation conditions. Furthermore, the parallel coordinates representation elucidates both direct and inverse correlations between formulation parameters, offering valuable insights into the underlying physicochemical relationships that govern the glass transition phenomenon. This visualization approach proves valuable for understanding the multifaceted interplay between composition, processing conditions, and resulting thermal properties within the formulation space.

In Fig. 11, we implemented data points with T_g values between 0 °C and up to 80 °C. This decision was based on observed reduced accuracies in extrapolated predictions beyond these boundaries. By constraining the model to this range, we ensured more reliable predictions within the most practically relevant temperature span.

Having presented all the results and returning to the central question of this study, we can confidently conclude that ML can effectively predict the synthesis formulation of lignin PUs. We have demonstrated this capability specifically for the T_g.

Moreover, this method can be applied to a wide array of materials science and chemistry, such as predicting the mechanical strength, thermal conductivity or electrical properties of novel materials [44,45]. This is a particularly useful alternative to deep-learning techniques,

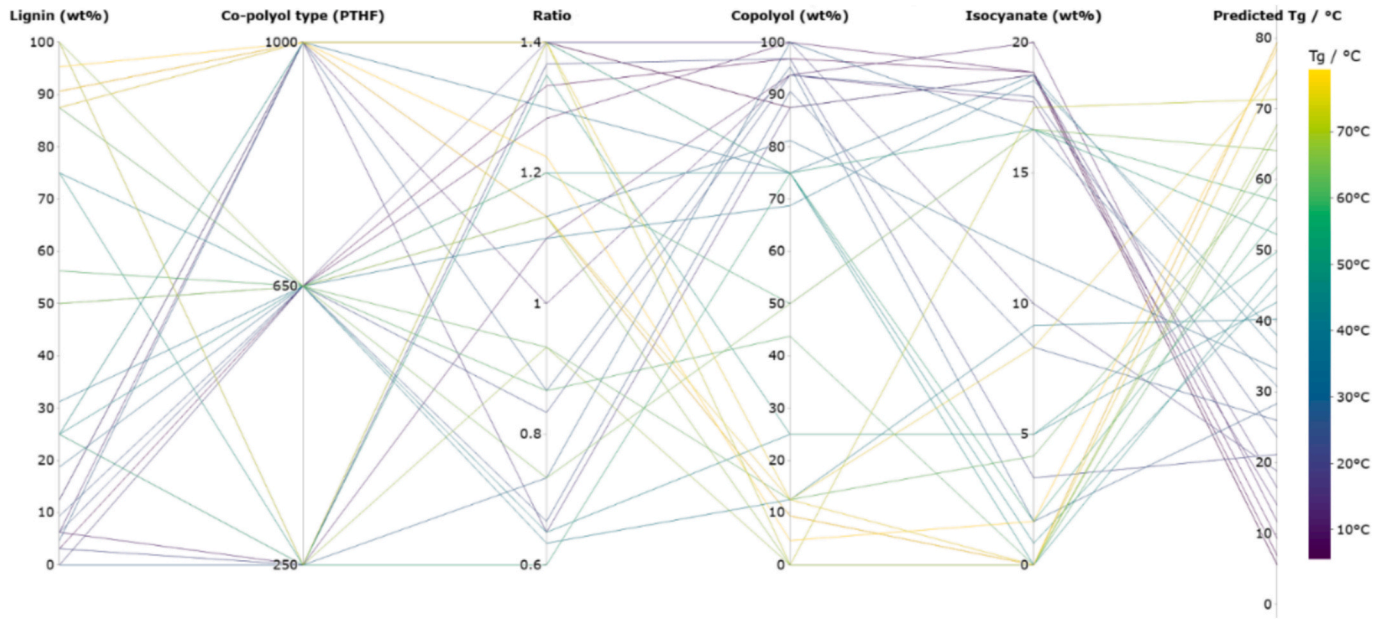

Fig. 11. Parallel coordinates plot using the predictions generated by the ensemble ML model from the adaptive grid search study predicting T_g of 100 wt% Co-polyol.

which are commonly used in this domain, but fall short on low-data regimes [46].

4. Conclusion

Our analysis revealed that lignin is a major contributor to enhancing the T_g of the lignin PUs. Interestingly, co-polyol content exhibited a complex relationship with T_g, showing a negative correlation overall but demonstrating a non-monotonic, U-shaped dependence in the mapping analysis, where T_g was depressed at intermediate co-polyol concentrations (~25–35 wt%) yet recovered at higher contents (~60–70 wt%) due to competing effects between soft-segment dilution and chain entanglement density. The swelling ratio emerged as a critical characterization parameter, exhibiting the highest contribution in terms of maximum absolute loading in PCA. Nevertheless, as a post-synthesis characterization measurement rather than a controllable formulation variable, swelling ratio was ultimately not included in the final predictive model.

The best-performing stacking ensemble was built using five input features. Ensemble learning techniques demonstrated varying performance based on the number of estimators, with lower estimator counts (~10) yielding the best validation performance, while higher counts primarily improve training and test metrics at the expense of validation generalization. The best-performing stacking ensemble model achieved a validation MAE of 13.41 °C and a test MAE of 15.17 °C with a combined validation and test MAE of 14.29 °C (Pearson r = 0.73). The generalizability gap (validation MAE minus training MAE) was 9.49 °C, indicating a moderate overfitting of the data. Nevertheless, the stacking ensemble consistently outperformed all individual base models, with the best individual model (GBR) achieving a validation MAE of only 15.99 °C. The selected model demonstrated restricted extrapolation capabilities beyond its training domain. With predictions plateauing at approximately 5.6 °C at the lower boundary and converging toward 79.4 °C at the upper boundary. This limitation emphasizes the importance of careful consideration when applying these models to underrepresented conditions or scenarios outside their training range. For visualization, parallel coordinates proved to be a meaningful technique for this dataset, given the relatively small number of features, and the user interface was constrained to the reliable prediction of 0 °C to 80 °C.

Returning to the central question of this study, we demonstrated that the ML model can predict and optimize the formulations of lignin PUs using limited experimental datasets with reasonable accuracy. Despite the promising results, several limitations of this study must be acknowledged. The small experimental dataset (136 samples after preprocessing from an initial 180) may have constrained the models' ability to capture the full complexity of lignin PU synthesis. The limited extrapolation range restricts the models' applicability to a narrow temperature window outside the training data. Additionally, the observed overfitting – evidenced by the widening gap between training and validation metrics with increasing model complexity – suggests that further optimization of ML techniques for small datasets is necessary. Future research should focus on expanding the experimental dataset, incorporating a wider range of lignin sources, synthesis conditions, and characterization parameters to improve model robustness and generalization. Advanced feature engineering techniques, including non-linear transformations and interaction terms, should be investigated to capture complex relationships within the data. The integration of physics-based models with ML techniques could leverage domain knowledge and potentially improve extrapolation capabilities. From a practical perspective this model is suitable for initial screening designs, due to its MAE values, and further optimizations could be proceeding from there through techniques such as Bayesian optimization to identify the T_g's global optimum. Incorporating additional algorithms such as Gaussian Process Regressors into the ensemble could be an area worth trying for expanding the extrapolation abilities of the ensemble model.

CRediT authorship contribution statement

Silviu Florin Acaru: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Marc Comí:** Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Panagiotis Falireas:** Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Danny E.P. Vanpoucke:** Writing – review & editing, Resources, Project administration, Funding acquisition. **Richard Vendamme:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization. **Katrien V. Bernaerts:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Research Foundation – Flanders (FWO), Belgium under grant G0E0223N digiLignin, funded by the European Union – NextGenerationEU.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.matdes.2026.116265.

Data availability

Data will be made available on request.

References

[1] A. Delavarde, G. Savin, P. Derkenne, M. Boursier, R. Morales-Cerrada, B. Nottelet, J. Pinaud, S. Caillol, Sustainable polyurethanes: toward new cutting-edge opportunities, Prog. Polym. Sci. 151 (101805) (2024).
[2] J. Brzoska, J. Smorawska, E. Głowińska, J. Datta, A green route for high-performance bio-based polyurethanes synthesized from modified bio-based isocyanates, Ind. Crop. Prod. 222 (2024) 119542.
[3] M. Comí, M. Fernández, A. Santamaría, G. Lligadas, J.C. Ronda, M. Galià, V. Cádiz, Carboxylic acid ionic modification of castor-oil-based polyurethanes bearing amine groups: chemically tunable physical properties and recyclability, Macromol. Chem. Phys. 218 (2017) 1700379.
[4] M. Comí, G. Lligadas, J.C. Ronda, M. Galià, V. Cádiz, Synthesis of castor-oil based polyurethanes bearing alkene/alkyne groups and subsequent thiol-ene/yne post-modification, Polymer 103 (2016) 163–170.
[5] M. Comí, G. Lligadas, J.C. Ronda, M. Galià, V. Cádiz, Adaptive bio-based polyurethane elastomers engineered by ionic hydrogen bonding interactions, Eur. Polym. J. 91 (2017) 408–419.
[6] E. Pichon, D. De Smet, K. Freulings, A. Pich, K.V. Bernaerts, Bio-based non-isocyanate polyurethane(urea) waterborne dispersions for water resistant textile coatings, Mater. Today Chem. 34 (2023) 101822.
[7] E. Pichon, J. Verstappen, S. Stepanova, A. Pich, K.V. Bernaerts, Bio-based non-isocyanate poly(urea)-PEG hybrids for VOC-free waterborne dispersions and fast UV-curable film application, Eur. Polym. J. 217 (2024) 113309.
[8] E. Pichon, A. Fordham, A. Pich, K.V. Bernaerts, Green synthesis of biobased NIPUrea-acrylate hybrids for versatile fast-curing hot-melt coating development, ACS Appl. Mater. Interfaces 17 (51) (2025) 69960–69969.
[9] B.G. Laycock, C.M. Chan, P.J. Halley, A review of computational approaches used in the modelling, design, and manufacturing of biodegradable and biobased polymers, Prog. Polym. Sci. 157 (2024) 101874.
[10] C. Kim, R. Batra, L. Chen, H. Tran, R. Ramprasad, Polymer design using genetic algorithm and machine learning, Comput. Mater. Sci 186 (2021) 110067.
[11] I.P. Malashin, V.S. Tynchenko, V.A. Nelyub, A.S. Borodulin, A.P. Gantimurov, Estimation and prediction of the polymers' physical characteristics using the machine learning models, Polymers 16 (2024) 115.
[12] A.U. Hassan, S.S.A. Shah, H.M. Abo-Dief, S. Naeem, N. Khan, E. Alzahrani, Z.M. El-Bahy, Polymer design using machine learning: a quest for high glass transition temperature, Synth. Met. 307 (2024) 117659.
[13] A. Alesadi, Z. Cao, Z. Li, S. Zhang, H. Zhao, X. Gu, W. Xia, Machine learning prediction of glass transition temperature of conjugated polymers from chemical structure, Cell Rep. Phys. Sci. 3 (2022) 100911.
[14] J. Leem, Y. Jiang, A. Robinson, Y. Xia, X. Zheng, Data-driven approach to tailoring mechanical properties of a soft material, Adv. Funct. Mater. 33 (38) (2023) 2304451.
[15] B. Dou, Z. Zhu, E. Merkurjev, L. Ke, L. Chen, J. Jiang, Y. Zhu, J. Liu, B. Zhang, G. W. Wei, Machine learning methods for small data challenges in molecular science, Chem. Rev. 123 (13) (2023) 8736–8780.
[16] J.M. Weber, Z. Guo, C. Zhang, A.M. Schweidtmann, A.A. Lapkin, Chemical data intelligence for sustainable chemistry, Chem. Soc. Rev. 50 (2021) 12013–12036.
[17] L. Tao, V. Varshney, Y. Li, Benchmarking machine learning models for polymer informatics: an example of glass transition temperature, J. Chem. Inf. Model. 61 (11) (2021) 5395–5413.
[18] D.E.P. Vanpoucke, O.S.J. van Knippenberg, K. Hermans, K.V. Bernaerts, S. Mehrkanoon, Small data materials design with machine learning: when the average model knows best, J. Appl. Phys. 128 (5) (2020) 054901.
[19] D.E.P. Vanpoucke, M.A.F. Delgove, J. Stouten, J. Noordijk, N. De Vos, K. Matthysen, G.G.P. Deroover, S. Mehrkanoon, K.V. Bernaerts, A machine learning approach for the design of hyperbranched polymeric dispersing agents based on aliphatic polyesters for radiation-curable inks, Polym. Int. 71 (8) (2022) 966–975.
[20] R.Q. Albuquerque, F. Rothenhäusler, H. Ruckdäschel, Designing formulations of bio-based, multicomponent epoxy resin systems via machine learning, MRS Bull. 49 (2024) 59–70.
[21] H. Esmaeili, R. Rizvi, An accelerated strategy to characterize mechanical properties of polymer composites using the ensemble learning approach, Comput. Mater. Sci 229 (2023) 112432.
[22] A. Milad, S.H. Hussein, A.R. Khekan, M. Rashid, H. Al-Msari, T.H. Tran, Development of ensemble machine learning approaches for designing fiber-reinforced polymer composite strain prediction model, Eng. Comput. 38 (2022) 3625–3637.
[23] Y. Zhao, R.J. Mulder, S. Houshyar, T.C. Le, A review on the application of molecular descriptors and machine learning in polymer design, Polym. Chem. 14 (2023) 3325–3346.
[24] S. Zhan, W. Huang, C. Dong, Q. Chen, H. Zhao, P. Duan, A. Hu, Q. Li, Y. Li, J. Liu, L. Zhang, Predicting glass transition temperature of polymers by combining molecular dynamics simulations and machine learning techniques, Mater. Today Commun. 40 (2024) 110181.
[25] S. Shirazian, T. Huynh, S.M. Sarkar, M. Habibi Zare, Development and optimization of machine learning models for estimation of mechanical properties of linear low-density polyethylene, Polym. Test. 137 (2024) 108525.
[26] H. Abdi, L.J. Williams, Principal component analysis, WIREs Comput. Stat. 2 (4) (2010) 433–459.
[27] S.F. Acaru, R. Abdullah, D.T.C. Lai, R.C. Lim, Hydrothermal biomass processing for green energy transition: insights derived from principal component analysis of international patents, Heliyon 8 (9) (2022) e10738.
[28] T.G. Dietterich, Ensemble Methods in Machine Learning, Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2000, pp. 1-15.
[29] I.D. Mienye, Y. Sun, A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects, IEEEAccess 10 (2022) 99129-99149.
[30] P. Ortiz-Serna, M. Carsí, M. Culebras, M.N. Collins, M.J. Sanchis, Exploring the role of lignin structure in molecular dynamics of lignin/bio-derived thermoplastic elastomer polyurethane blends, Int. J. Biol. Macromol. 158 (2020) 1369–1379.
[31] P. Jutrzenka Trzebiatowska, A. Santamaria Echart, T. Calvo Correas, A. Eceiza, J. Datta, The changes of crosslink density of polyurethanes synthesised with using recycled component. Chemical structure and mechanical properties investigations, Prog. Org. Coat. 115 (2018) 41–48.
[32] J.R. Gouveia, R.R. de Sousa Júnior, A.O. Ribeiro, S.A. Saraiva, D.J. dos Santos, Effect of soft segment molecular weight and NCO:OH ratio on thermomechanical properties of lignin-based thermoplastic polyurethane adhesive, Eur. Polym. J. 131 (2020) 109690.
[33] M. Wadekar, W. Eevers, R. Vendamme, Influencing the properties of LigninPU films by changing copolyol chain length, lignin content and NCO/OH mol ratio, Ind. Crop. Prod. 141 (2019) 111655.
[34] J. Shi, T. Zheng, Y. Zhang, B. Guo, J. Xu, Cross-linked polyurethane with dynamic phenol-carbamate bonds: Properties affected by the chemical structure of isocyanate, Polym. Chem. 12 (2021) 2421–2432.
[35] F.E. Levine, John, J. La Scala, Effect of isocyanate to hydroxyl index on the properties of clear polyurethane films, Progress in Organic Coatings 74(3) (2012) 572-581.
[36] A. Menon, J.A. Thompson-Colón, N.R. Washburn, Hierarchical machine learning model for mechanical property predictions of polyurethane elastomers from small datasets, Front. Mater. 6 (2019).
[37] F. Ding, L.-Y. Liu, T.-L. Liu, Y.-Q. Li, J.-P. Li, Z.-Y. Sun, Predicting the mechanical properties of polyurethane elastomers using machine learning, Chin. J. Polym. Sci. 41 (3) (2022) 422–431.
[38] J.A. Pugar, C. Gang, I. Millan, K. Haider, N.R. Washburn, Machine learning of polyurethane prepolymer viscosity: a comparison of chemical and physicochemical approaches, Digital Discovery 4 (12) (2025) 3652–3661.
[39] C. Joo, H. Park, H. Kwon, J. Lim, E. Shin, H. Cho, J. Kim, Machine learning approach to predict physical properties of polypropylene composites: application of MLR, DNN, and random forest to industrial data, Polymers (Basel) 14 (17) (2022).
[40] I.P. Malashin, V.S. Tynchenko, V.A. Nelyub, A.S. Borodulin, A.P. Gantimurov, Estimation and prediction of the polymers' physical characteristics using the machine learning models, Polymers (Basel) 16 (1) (2023).
[41] E. Kazemi-Khasragh, J.P. Fernández Blázquez, D. Garoz Gómez, C. González, M. Haranczyk, Facilitating polymer property prediction with machine learning and group interaction modelling methods, Int. J. Solids Struct. (2024) 286–287.
[42] S. Das, S.S. Mullick, I. Zelinka, On supervised class-imbalanced learning: an updated perspective and some key challenges, IEEE Trans. Artif. Intell. 3 (2022) 973–993.
[43] A. Babbar, S. Ragunathan, D. Mitra, A. Dutta, T.K. Patra, Explainability and extrapolation of machine learning models for predicting the glass transition temperature of polymers, J. Polym. Sci. 62 (6) (2024) 1175–1186.
[44] X. Hong, Q. Yang, K. Liao, J. Pei, M. Chen, F. Mo, H. Lu, W.B. Zhang, H. Zhou, J. Chen, L. Su, S.Q. Zhang, S. Liu, X. Huang, Y.Z. Sun, Y. Wang, Z. Zhang, Z. Yu, S. Luo, X.F. Fu, S.L. You, AI for organic and polymer synthesis, Sci. China Chem. 67 (8) (2023) 2461–2496.
[45] H. Jang, D. Ryu, W. Lee, G. Park, J. Kim, Machine learning-based epoxy resin property prediction, Mol. Syst. Des. Eng. 9 (2024) 959–968.
[46] D. van Tilborg, H. Brinkmann, E. Criscuolo, L. Rossen, R. Özçelik, F. Grisoni, Deep learning for low-data drug discovery: hurdles and opportunities, Curr. Opin. Struct. Biol. 86 (2024) 102818.