Chinese Journal of
POLYMER SCIENCE

RESEARCH ARTICLE

https://doi.org/10.1007/s10118-022-2838-6
Chinese J. Polym. Sci. **2023**, *41*, 422–431

Predicting the Mechanical Properties of Polyurethane Elastomers Using Machine Learning

Fang Ding[a,b], Lun-Yang Liu[a], Ting-Li Liu[a,b], Yun-Qi Li[a,b,c*], Jun-Peng Li[d*], and Zhao-Yan Sun[a,b*]

[a] *State Key Laboratory of Polymer Physics and Chemistry, Changchun Institute of Applied Chemistry, Chinese Academy of Sciences, Changchun 130022, China*
[b] *School of Applied Chemistry and Engineering, University of Science and Technology of China, Hefei 230026, China*
[c] *Department of Polymer Materials and Engineering, College of Materials and Metallurgy, Guizhou University, Guiyang 550025, China*
[d] *State Key Laboratory of Advanced Technologies for Comprehensive Utilization of Platinum Metals, Sino-Platinum Metals Co. Ltd., Kunming 650106, China*

Electronic Supplementary Information

Abstract Bridging the gap between the computation of mechanical properties and the chemical structure of elastomers is a long-standing challenge. To fill the gap, we create a raw dataset and build predictive models for Young's modulus, tensile strength, and elongation at break of polyurethane elastomers (PUEs). We then construct a benchmark dataset with 50.4% samples remained from the raw dataset which suffers from the intrinsic diversity problem, through a newly proposed recursive data elimination protocol. The coefficients of determination (R^2s) from predictions are improved from 0.73–0.78 to 0.85–0.91 based on the raw and the benchmark datasets. The fitting of stress-strain curves using the machine learning model shows a slightly better performance than that for one of the well-performed constitutive models (*e.g.*, the Khiêm-Itskov model). It confirmed that the black-box machine learning models are feasible to bridge the gap between the mechanical properties of PUEs and multiple factors for their chemical structures, composition, processing, and measurement settings. While accurate prediction for these curves is still a challenge. We release the raw dataset and the most representative benchmark dataset so far to call for more attention to tackle the long-standing gap problem.

Keywords Mechanical properties; Stress-strain curves; Polyurethane elastomers; Machine learning; Benchmark dataset

Citation: Ding, F.; Liu, L. Y.; Liu, T. L.; Li, Y. Q.; Li, J. P.; Sun, Z. Y. Predicting the mechanical properties of polyurethane elastomers using machine learning. *Chinese J. Polym. Sci.* **2023**, *41*, 422–431.

INTRODUCTION

Density, Poisson ratio, Young's modulus, constitutive models, *etc.* are essential inputs for finite element analysis and computer-aided design & engineering, such as commercial software Abaqus, Comsol Multiphysics, Ansys Fluent, *etc.*, in the computation of mechanical properties for polymer materials. It is a great and long-standing challenge to bridge the gap between these inputs with molecular descriptors, especially for polymeric elastomers. Intensive efforts to bridge the gap are impeded by many factors, including the unclear understanding of hyper-elastic properties at the molecular level, lacking of parameters to quantify the deformation and energy dissipation at multi-scales, the absence of a benchmark dataset that can support the development of new methods, and the inherent limitations of the human brain to find patterns from multiple non-linear, non-monotonous and non-orthogonal correlations. The emergence of the data-driven method, *i.e.*, machine learning (ML), provides a new perspective to facilitate the calculation and prediction of the mechanical properties of polymeric elastomer materials. ML has shown significant advantages in the quantification of separation and mechanical properties for variant polymeric membranes,[1–3] the optimization of compositions for epoxy resins,[4] inverse molecular design for given properties including proton conductivity, methanol permeability, tensile modulus, *etc.*, properties.[5] Polyurethane elastomers (PUEs) can be a good model system that has widely distributed and steadily tunable mechanical properties and is believed to have strong correlations between structure and mechanical properties. It is interesting to know whether applying ML study is a feasible method to bridge the gap.

PUEs are a class of elastomers that have broad applications, massively used in structural and infrastructural engineering,[6] electromechanical actuators,[7] biomedical packages and devices,[8,9] electronics and sensors,[10,11] and so forth. They are normally linear multi-block copolymers, composed of hard segments made up of diisocyanate (DI) with optional chain extender (CE, *e.g.*, diol, diamine, and thiol), and soft segments with polyol (PO). Hard segments (HS) normally have ur-

* Corresponding authors, E-mail: yunqi@ciac.ac.cn (Y.Q.L.)
E-mail: lijunpeng@ipm.com.cn (J.P.L)
E-mail: zysun@ciac.ac.cn (Z.Y.S.)

Received June 13, 2022; Accepted July 6, 2022; Published online October 8, 2022

ethane (—NHCOO—), urea (—NHCONH—) or thiourethane (—NHCOS—) groups and soft segments (SS) contain carbonate (—OCOO—), ester (—COO—) or ether (—O—) connections. The regulation of the content of hard segments (CHS) and the fuzzy soft-hard interface is crucial to making PUEs have plastic and elastic mechanical properties.[12] Important properties including Young's modulus (YM, MPa), tensile strength (TS, MPa), and elongation at break (EB, %), can be steadily extracted from uniaxial tensile stress-strain curves. Theoretically, there are dozens of constitutive models have been proposed to quantify the stress-strain curves.[13,14] In our previous work,[13] we found three models that can be used to quantitatively fit the stress-strain curves of most PUEs. The distributions of model parameters through well-fitting are non-Gaussian, which reflect structural dispersity. In addition, the distributions of mechanical properties are broad and non-Gaussian, indicating the dispersive structure-properties correlations of PUEs. Establishing datasets with conserved structure-property correlations, *e.g.* benchmark dataset, can aid in the development of high-performance materials. Benchmark datasets can be established by either filling or removing sparse spaces in the dataset. By using the filling method, Ma and Luo *et al.*[15] built a polymer benchmark dataset based on the PolyInfo database.[16] The newly generated dataset significantly populates regions where PolyInfo data are sparse, and intrinsic properties of polymers including density, glass transition temperature, melting temperature, and dielectric constant can be accurately predicted. In our work, it is expected to construct a benchmark dataset for PUEs which can provide conserved structure-properties correlations.

The core to bridge the gap is the quantitative composition-processing-structure-mechanical properties relationship, where the structure contains the chemical structure of monomers and the aggregated structure. It can be determined from a set of experimental techniques, typically mass spectrum, nuclear magnetic resonance, infrared spectrum, *etc.*, to characterize chemical structures, and atomic force microscopy, small- and wide-angle scattering, electron microscopy, *etc.*, to analyze the aggregation structure. For PUEs, their inherent structural characteristics include the degree of hydrogen bonding attributed to the existence of —NHCO— groups, the microphase separation between hard and soft segments, and the distribution of crystallite. The degree of hydrogen bonding is usually qualitatively measured by the Fourier transform infrared spectroscopy.[17,18] Similarly, the quantitative parameters in the description of microphase-separated morphologies or crystallite, such as the average size and fraction of domains, the phase interfaces, *etc.*, are isolated from physical studies and material reports. The missing of clear structure information is the main challenge in the construction of quantitative composition-processing-structure-mechanical properties relationship for PUEs and is ubiquitous for other polymer materials. While owing to strong correlations between chemical structure and mechanical properties for PUEs, the construction of a quantitative chemical structure-composition-processing-mechanical properties relationship (CCPMr) using ML is practicable.

Here, the predictive targets are the three mechanical properties: YM, TS, EB, as well as the original stress-strain curves from uniaxial tensile tests for PUEs. The contents are ordered in (1) perform exhaustive data mining from accessible sources focusing on the mechanical properties for PUEs; (2) digitalize the chemical structure, processing, and measurement settings through feature engineering; (3) build predictive models for the mechanical properties distributed in the raw dataset; (4) construct a benchmark dataset for PUEs that have a conserved CCPMr; (5) predict and fit the stress-strain curves.

DATA AND METHODS

Workflow

A brief introduction to the workflow is shown in Fig. 1. Section 1 is to create an exhaustive raw dataset through data mining from diverse sources, mainly from academic literature about the synthesis, characterization, and mechanical measurements for PUEs. Sections 2 and 3 are aimed to build a set of features that bear the most informed correlations with mechanical properties through feature engineering, then interactively and iteratively construct and validate predictive models by using the extreme gradient boosting tree (XGB) algorithm.[19] Schematic diagrams of the predictions for three mechanical properties by using the XGB algorithm in section 6 are shown in section 7. Based on the set of most informed features, the target of sections 4 and 5 is to construct a benchmark dataset that has a consistent and conserved CCPMr, though a newly proposed recursive data elimination (RDE) protocol integrated with multiple ML algorithms as shown in section 6. The predictions of three mechanical properties in the benchmark dataset are also presented in section 7. Further, the stress-strain curves in both raw and benchmark datasets are predicted and fitted by using XGB, as shown in sections 6 and 7.

Feature Engineering

Features to describe the chemical structure, interaction between hard and soft segments, composition, processing, and measurements are presented here. Chemical structures are encoded using the Simplified Molecular-Input Line-Entry System (SMILES), and a set of features are calculated using RDKit 2021.[20] Features include constitutional (count of atoms, groups, and bonds), connective (Chi indices), topological (BalabanJ), MOE-type (such as EState_VSA series), and molecular properties descriptors (TPSA). They constitute 200 features for each monomer, hard or soft segment. Then the corresponding features for a PUE are calculated through the molar weighted average according to the group additive principle.[21]

The interaction between hard (HS) and soft segments (SS) is expressed by the Flory-Huggins interaction χ, computed through

$$\chi = \frac{(\delta_{HS} - \delta_{SS})^2 V_m}{RT} \quad (1)$$

where δ_{HS} and δ_{SS} are the Hildebrand solubility parameters for HS and SS, V_m is the molar volume of the equivalent monomer for PUE, R is the gas constant, and T is the measured temperature. The Hildebrand solubility is the square root of the cohesive energy density (CED), and the CED is calculated following the Fedors method[22] based on the group additive principle.[23] The composition of PUE consists of molecular weight (MW) of PO, mass fraction of HS (CHS), the molar ratio of CE (n_{CE}), and the isocyanate index (IsoIndex), which is the molar

https://doi.org/10.1007/s10118-022-2838-6

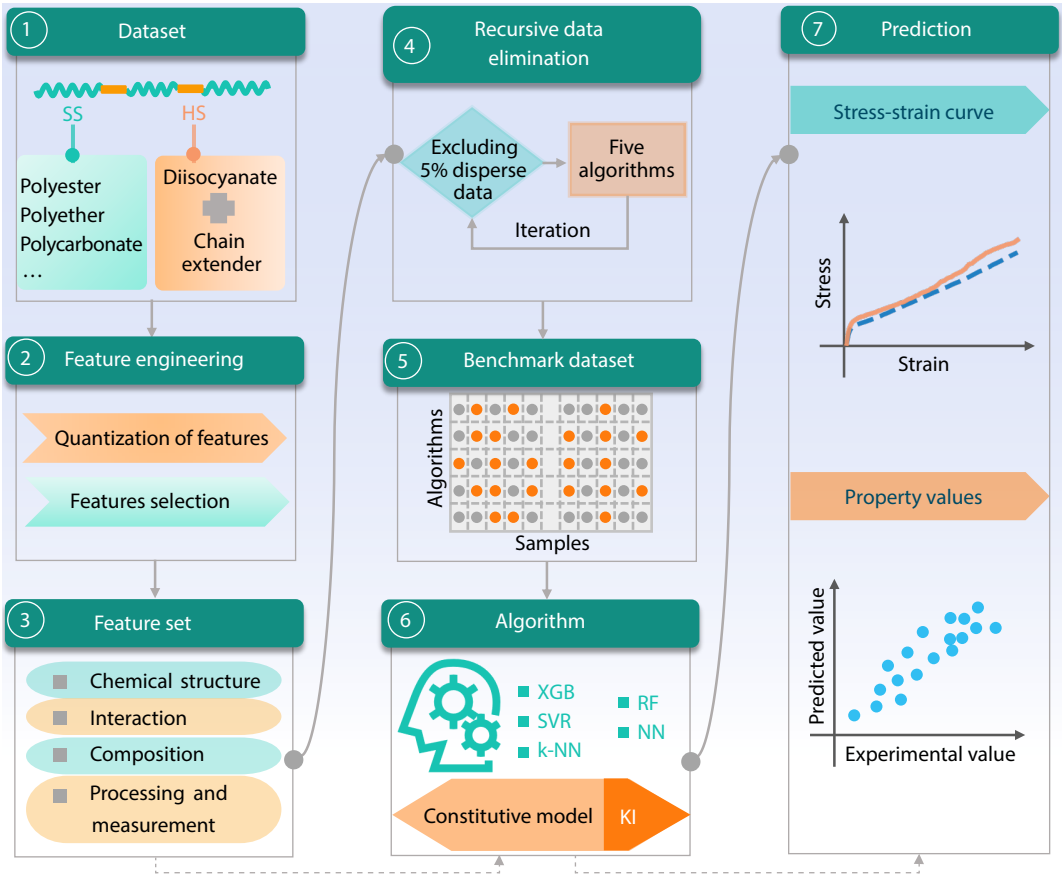

Fig. 1 Workflow to predict the mechanical properties of PUEs. Sections from 1 to 7 cover data mining, feature and data engineering, tuning algorithms, and construction of predictive models.

ratio of the —NCO and the —OH groups.

For processing settings, PUEs are polymerized through either one-step or two-step strategies at up to two reaction temperatures (Tr1, Tr2), and tuned by the time for solvent addition (including ts0, ts1 and ts2, which represents that no solvent is added, the solvent is added at the 1st step, and the solvent is added at the 2nd step, respectively) and the time for catalyst addition (including tc0, tc1 and tc2, which represents that no catalyst is added, the catalyst is added at the 1st step, and the catalyst is added at the 2nd step, respectively). By default, good solvents are used in the polymerization process such as dimethylformamide, tetrahydrofuran, dimethylacetamide, and catalysts generally are tin compounds. Hence, the chemical structures of solvent and catalyst are not considered in this work. The forming methods (FM) of the spline can be grouped into solution casting, melt casting, hot pressing, spin-casting, and microinjection. The measurements record the shape (dumbbell or rectangle-shaped), the gauge length (mm), and the cross-sectional area (CSArea, mm^2) of splines, associated with the elongation rate (mm/min) and the strain rate (the elongation rate normalized by the gauge length, min^{-1}) in the tensile test. Totally 629 features are generated to record chemical structure, interaction, composition, processing, and measurements.

Based on the raw dataset, the collection of features is filtered using the L_{sig} criteria[1,5] at 0.90 confidence level to remove insufficient or redundant features. The remaining 204 features are further experienced a non-linear and non-monotonous recursive feature elimination[24] using the XGB algorithm through 5-fold cross-validation. In the cross-validation, all data are randomly split into 5 sets, 4 sets are assembled to be the train set (80%) and the remaining one is regarded as the test set (20%). It repeats 5 times till each sample is tested once. The predicted values are from the test set which is unseen in the construction of the predictive model. The optimal combination to predict mechanical properties (YM, TS, and EB) has 20 features, and the detail is listed in Table S1 (in the electronic supplementary informmation, ESI).

Construction of Predictive Models

Predictive models of three mechanical properties and stress-strain curves are constructed by using the XGB algorithm (details can be seen in ESI). The hyper-parameters for these models are optimized under Bayesian inference and the final model is to achieve minimized mean squared error in 5-fold cross-validation. The models are constructed following previous strategies for the predictions of discrete points of mechanical properties, and continuous stress-strain curves.[2,3] To evaluate the performance of the predictive models, two metrics including the coefficient of determination (R^2) and the root mean squared error (RMSE) is calculated. They are defined as:

https://doi.org/10.1007/s10118-022-2838-6

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

$$\text{RMSE} = \left[\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\right]^{1/2} \quad (3)$$

where y_i and $\hat{y}_i$ are the experimental value and the predicted value associated with the i^{th} sample, $\bar{y}_i$ is the average value of all samples, and N is the number of all samples.

Recursive Data Elimination

Since PUEs are developed for diverse applications in the form of plastic and elastic products, the chemical structures, composition, processing methods, and mechanical properties are broadly distributed. Here we propose a ranking-based recursive data elimination (RDE) protocol to construct a benchmark dataset by using five machine learning algorithms, *i.e.*, XGB, random forest (RF),[25] support vector regression (SVR),[26] neural network (NN),[27] *k*-nearest neighbors regression (k-NN).[28,29] The resulted benchmark dataset may share a consistent and conserved chemical structure-composition-processing-mechanical properties relationship (CCPMr) with the enclosure of as many samples as possible. The RDE procedure utilizes the fixed combination of features and hyper-parameters (Table S2 in ESI) trained from the raw dataset. Then the samples with large weighted predictive error scores (WPES) are recursively dropped, and the drop-off ratio is fixed at 5% that allowing 20 iterations to obtain a coherent CCPMr for the remained samples. Here, the WPES is defined as:

$$\text{WPES} = \sum_{i}^{3} N_{\text{index},i} \times \frac{\text{RE}_i - \overline{\text{RE}_i}}{\overline{\text{RE}_i}} \quad (4)$$

where $N_{\text{index},i}$, RE_i, and $\overline{\text{RE}_i}$ are the index number of the sorted relative predictive error in ascending order, the relative error of the prediction, and the mean of relative error of the remained samples, respectively. The subscript i represents the i^{th} mechanical properties, which include YM, TS, and EB. To eliminate the impact of the combination of samples, 100 replicates with different random seeds for data split in the RDE procedure are computed. Then the RDE score for each sample can be labeled. It has a max of 20 which means the sample has the most conserved CCPMr and is kept at the last iteration of drop-off, and the min of 1 means the sample has the most distinctive CCPMr and is dropped at the first iteration. From the convergence of 5 machine learning algorithms, the candidate benchmark dataset is determined based on the truncation of a given RDE score.

RESULTS AND DISCUSSION

Predictive Models for Mechanical Properties

The 20 optimal features listed in Table S1 (in ESI) are obtained through feature engineering. They cover information on the chemical structure, interaction between soft and hard segments, composition, and experimental settings, which are the optimal combination of all features. The predicted mechanical properties in the test set from the 5-fold cross-validation using these 20 selected features against the corresponding experimental values are shown in Fig. 2. It can be seen that the predicted values are evenly distributed on both sides of the experimental values. R^2 is 0.73, 0.78, and 0.76 in the prediction of YM, TS, and EB, and their corresponding RMSE is

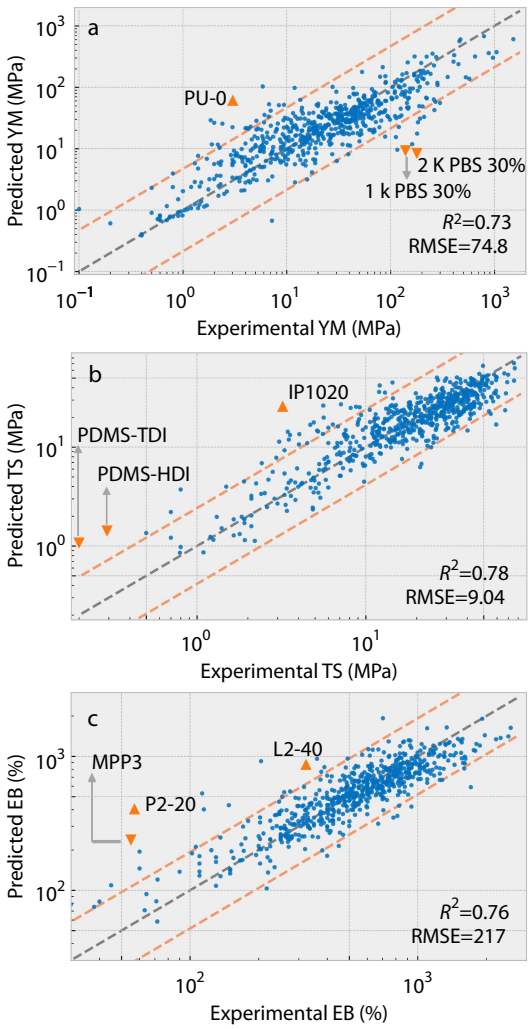

Fig. 2 Performance of predictive models for YM (a), TS (b), and EB (c) based on the raw dataset. The light orange dash lines in the plot are the upper and the lower bounds at 95% confidence level. The triangles label typical samples with large predictive errors, and their names are from original reports.

74.8 MPa, 9.04 MPa, and 217%, respectively. It is noted that the RMSE for the three properties is slightly large, but these values are within an acceptable range compared with the mean and standard deviation (Std) of the experimental values of properties (Table 1). Furthermore, R^2 in the test set is higher than 0.70 which can be considered to obtain meaningful predictions from the machine learning model based on a statistical benchmark.[30] Hence, it can be concluded that all the models are accurate to deliver robust predictions for the three mechanical properties of PUEs in the raw dataset, suggesting that constructed models can help researchers to design new PUE with high performance through virtual experiments.

We then investigate the feature importance to explore information for 20 selected features. The rankings of feature importance based on the raw dataset are shown in Figs. S1(a)–S1(c) (in ESI), presented by the gain value calculated using the XGB algorithm, the Pearson (R_P) and the Spearman (R_{Sp}) correlation coefficients between features and mechanical properties. The logarithm of molecular weight for the polyol (log_PO_MW) is negatively correlated with YM because the

Table 1 Summary of predictive models for three mechanical properties of samples in the raw and benchmark datasets.

Property	Raw dataset (*N*=643)				Benchmark dataset (*N*=326)			
	Mean	Std	RMSE	R^2	Mean	Std	RMSE	R^2
YM (MPa)	57.2	118	74.8	0.73	67.3	101	60.5	0.89
TS (MPa)	23.5	15.4	9.04	0.78	26.4	14.8	6.80	0.91
EB (%)	629	373	217	0.76	624	306	131	0.85

increase in the molecular weight of polyol normally causes the decrease of CHS,[31] and the latter is dominant for YM. The chemical structure and the corresponding electron distribution over van der Waals surface[32] (SS_VSA_Estate) in the soft segment is negatively and strongly correlated with YM, TS, and EB, which indicates the enclosure of such as aromatic ring, conjugated groups, *etc.*, in the soft segment are deleterious for all three mechanical properties of PUEs.[33] The SS_TPSA_N has a strong positive correlation with YM and TS, suggesting that the enrichment of polar groups in the soft segment can strengthen the hard-soft segmental interface[34] and hereby enhance the mechanical properties of PUEs. Similarly, improving SS_PEOE_VSA,[35] *i.e.*, the long-range electrostatic attraction in the soft segment, allows the tolerance of large deformation, which is the most important feature that positively correlates with EB. The molar volume (log_Vm) and the cohesive energy density (CED) are obtained based on the group additive principle, which plays an important role in predicting YM and TS. Especially, the negative correlation of log_Vm versus YM and TS is owing to the packing of larger monomers normally leading to a higher fraction of free volume and loose aggregated structures. However, on the contrary, CED tends to maintain the integrity of domains. It positively correlates with YM and TS, satisfying its role in the correlation with the glass transition temperature for linear PUEs[36] and the mechanical properties in a small PUE dataset.[37] Experimental settings-based features are crucial but difficult to set up exact correlations over reports from different groups, making the ranking in feature importance off the top location. Overall, for features listed in Table S1 (in ESI), those chemical structure-, interaction-, composition- and experimental settings-based features are not highly correlated with the three properties but are vital for the prediction of the mechanical properties.[38]

We further analyze some individual cases from these "poor" predictions (with errors outside 95% confidence level) as marked in Fig. 2. The samples with large predictive errors, such as "PU-0",[18] "2K PBS 30%"[39] and "1K PBS 30%"[39] for YM, "IP1020",[40] "PDMS-TDI"[41] and "PDMS-HDI"[41] for TS, "P2-20",[42] "L2-40"[42] and "MPP3"[43] for EB, are labeled. To make clear the change of a single feature derived from the original records in the academic literature, other samples in the raw dataset with controlled variables (*i.e.*, different in only one feature) in chemical structure, composition, processing,

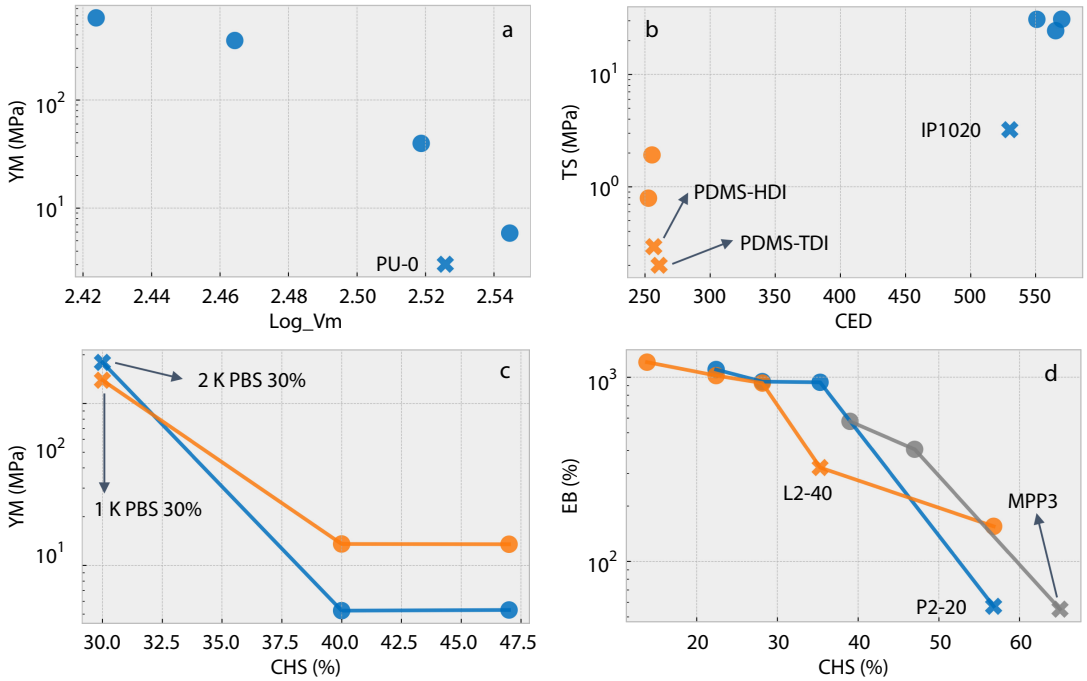

Fig. 3 Individual cases in the construction of the CCPMr for PUEs: samples with different chain extenders (CEs) (a), diisocyanates (DIs) (b), and content of hard segment (CHS) (c, d). Differences in precursors in panels (a) and (b) are described by using calculated features log_Vm and CED, which are the most important features for the predictions of YM and TS. Samples with the same color (controlled samples) can be compared with each other in each panel, and the mark "O" and "X" represents well- and poorly-predicted samples, respectively.

https://doi.org/10.1007/s10118-022-2838-6

or measurement are also enclosed. Figs. 3(a) and 3(b) show the poorly predicted samples are out of the linearity or monotonicity in the set of controlled samples. It may be due to poorly predicted samples having strong phase mixing in morphology, which is different from other controlled samples.[40,41] Fig. 3(c) shows the break of common sense for PUEs, where a normally higher CHS leads to stronger YM and TS. This is probably caused by the crystallization of soft segment with regular chemical structure at low hard segment content.[39] While the EB versus CHS for "L2-40", "P2-20" and "MPP3" seem reasonable in the controlled samples (Fig. 3d), other unrecorded aspects such as the defects in testing samples may overwhelm such consistency.[44] Based on the above discussion, it can be concluded that the morphology and mechanical properties of a small number of samples in the raw dataset are deviated from those of other highly similar samples. It is also confirmed that CCPMr of samples in the raw dataset is highly dispersed.

On the other hand, the three mechanical properties of PUEs in the raw dataset are broadly distributed up to four magnitudes, which is slightly broader than those from our recent work (Fig. S2 in ESI).[13] The distributions of mechanical properties deviate from the Gaussian distribution, reflecting the diversity in the structure-property correlations.[44,45] Generally, most PUE materials share an intrinsic conserved CCPMr. However, since the diversity of the raw dataset, a set of samples with intrinsic conserved CCPMr are difficult to obtain by studying relations between a single feature and properties. Alternatively, we can build a benchmark dataset with intrinsic conserved CCPMr driven by many coupled features using implicit machine learning.

Construction of Benchmark Dataset

The results from the recursive data elimination (RDE) procedure are shown in Fig. 4 and Fig. S3 (in ESI). In principle, the average R^2 increases with the reduction of samples, where samples with more conserved CCPMr remain. Most of these profiles in the prediction of YM, TS or EB, using one of the XGB, RF, SVR, NN or k-NN algorithms follow this expectation. While the standard deviations along the recursive drop-off may fluctuate, suggesting that the remained samples still hold some degree of diversity. The evaluation of the RDE score against the sorted samples by the five ML algorithms shows an almost overlapped decay curve (Fig. 4d). These curves have a coincident cross at ~300 samples with an RDE score of 15, we then pick this RDE score to select candidates for the benchmark dataset. As shown in Fig. 4(e), we label each sample that is retained by each of the 5 machine learning algorithms, then count the number of well-

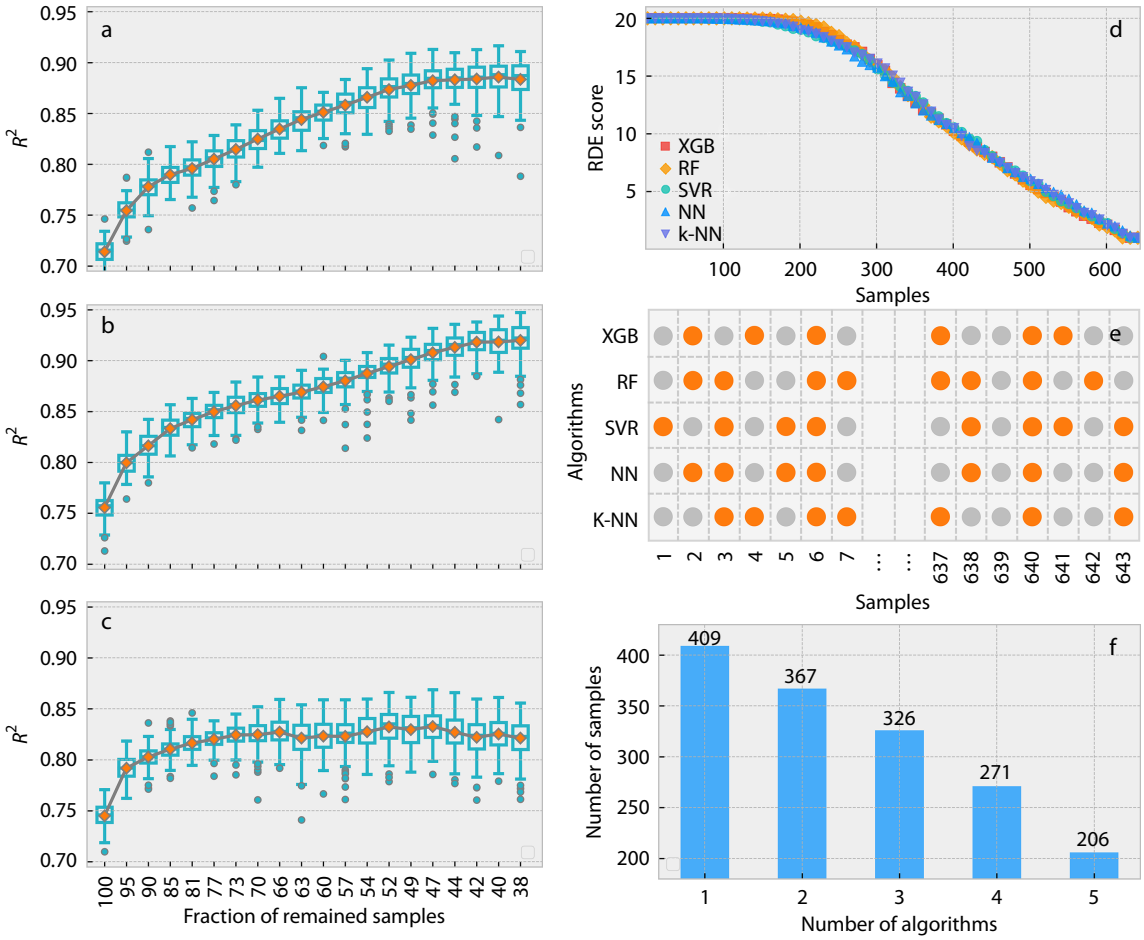

Fig. 4 Distribution of R^2 in the RDE protocol using the XGB algorithm for YM (a), TS (b), and EB (c); RDE score for sorted samples using five machine learning algorithms (d); Scheme for the remained samples selected by each of the five ML algorithms (e); Number for samples with conserved CCPMr based on the number of algorithms (f).

https://doi.org/10.1007/s10118-022-2838-6

predicted algorithms. From the intersection of these labels, the number of samples as a function of the number of well-predicted algorithms is shown in Fig. 4(f). With the number of algorithms increasing from 1 to 5, the number of satisfied samples decreases from 409 to 206. We then select the conservancy based on 3 algorithms, which are close to most of the level-off points of R^2 in RDE, and achieve a balance between the coverage and the convergency, as the final criteria for the benchmark dataset. It gets 326 out of the 643 samples in the raw dataset (remained 50.4%).

In the benchmark dataset, the chemical structures of 326 samples cover 20 different polyols and 66 hard segments composed of 13 diisocyanates and 39 chain extenders. To explore the differences in chemical structures between the benchmark and raw datasets, we calculate the similarity between paired chemical structures of PUEs using Euclidean distance based on the MACCS fingerprint (Fig. S4 in ESI).[46,47] The distributions are very similar, and only tiny differences can be seen. This indicates that the benchmark dataset may share a consistent and conserved CCPMr with the raw dataset. Further, we compare the distribution of three mechanical properties, as well as the Pearson correlations (R_P) between features and properties, between the benchmark dataset and the raw dataset. Compared with broad non-Gaussian distributions of YM, TS, and EB in the raw dataset, these properties in the benchmark dataset have slightly narrow and more Gaussian-like distributions, as shown in Fig. S2 (in ESI). A comparison of R_P between features and the mechanical properties (as shown in Fig. 5) for the raw and the benchmark datasets should provide more tuition. It can be found that almost all features get stronger correlations for YM and TS in the benchmark dataset. These results support that the benchmark dataset obtained based on the RDE protocol has more conserved CCPMr than the raw dataset. In addition, it is also noteworthy that the R_P between features and EB for the benchmark dataset did not increase or even decrease, compared with the R_P for the raw dataset. This may be due to the inherent non-linear, non-monotonous, and non-orthogonal correlations between these features and EB. A similar correlation has been reported for Acrylonitrile-Butadiene-Styrene (ABS) resins, where the group additive principle[21] is followed by YM and TS but is deviated by EB.[48]

To further investigate whether the benchmark dataset has an intrinsic conserved CCPMr, here we construct predictive models for YM, TS, and EB based on the benchmark dataset. The predictions for these three properties are shown in Fig. 6, and the comparison with those for the raw dataset is summarized in Table 1. It is found that the predicted values are more closely distributed on both sides of the experimental values. The predicted R^2 is significantly improved while the RMSE is declined through the RDE protocol. In 500 repeats with different random seeds in the splitting of data for the 5-fold cross-validation, the R^2s from the train and test sets are also distributed in conserved regions with reasonable in-between differences. It verifies the initial prospection of the RDE to screen out a benchmark dataset that has an intrinsic conserved CCPMr.

Predicting Stress-Strain Curves

The predictions for YM, TS, and EB show sufficient accuracy and robustness, we then challenge the prediction of the stress-strain curves in the benchmark dataset and raw dataset. The 156 curves in the benchmark dataset (BC) and 386 curves in the raw dataset (including curves in the "BC" and "Not BC") are shown in Fig. 7(a). The curves in the benchmark and the raw datasets share similar profiles and distribution regions, with clear signals for yield, necking, strain hardening, *etc*. The prediction of these curves using the 21 features (20 features used above and the strain) and the distribution of R^2 is shown in Fig. 7(b). It can be seen that the fraction of samples with high R^2 is significantly improved from the raw to the benchmark dataset. For example, the fraction of samples with R^2 above 0.90 increases from 14.8% to 30.1%. It also confirms that the CCPMr in the benchmark dataset is more conserved than that in the raw dataset. In our previous work, we found that the Khiêm-Itskov (KI)[52] constitutive model can well fit the stress-strain curves for PUEs.[13] From the train set of the curves, the prediction becomes the fitting problem. The fitting using the XGB algorithm gets an average R^2 of 0.98, which is slightly higher than the fitting using the KI model with an average R^2 of 0.96. It strongly suggests that applying an ML study is feasible to bridge the gap between the mechanical properties of PUEs and their chemical structures, composition, processing, and

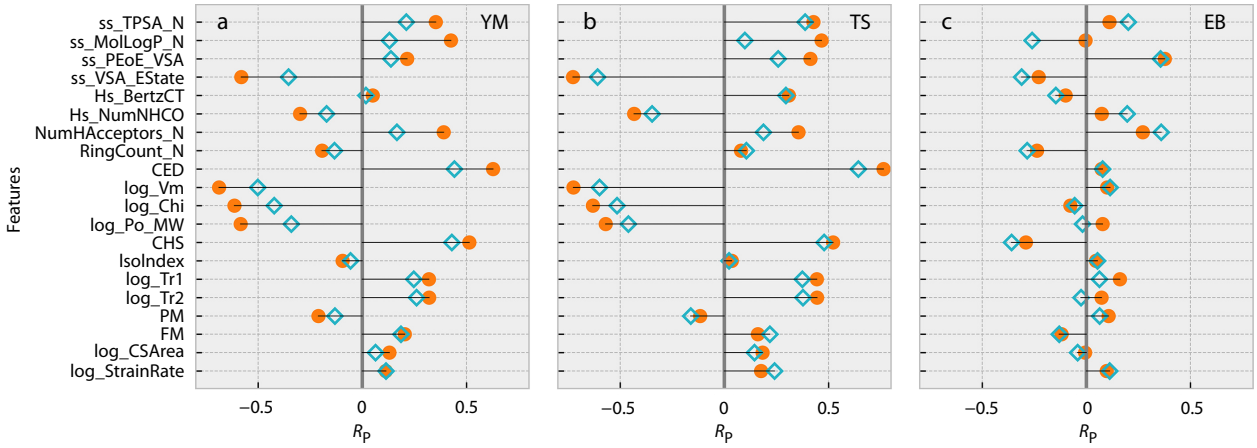

Fig. 5 Comparison of R_P between features and mechanical properties for samples in the benchmark dataset (orange circle) and the raw dataset (cyan diamond).

https://doi.org/10.1007/s10118-022-2838-6

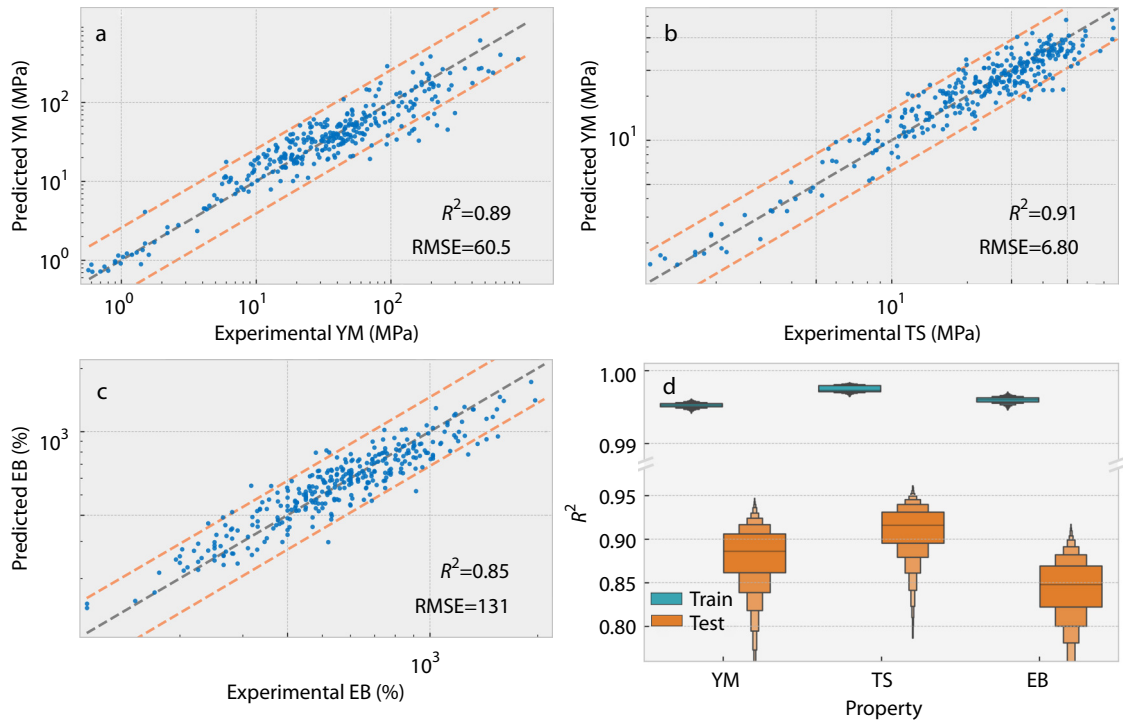

Fig. 6 Predictive models for YM (a), TS (b), and EB (c) for the samples in the benchmark dataset, and the distributions of R^2 from the train and the test sets based on 500 repeats of 5-fold cross-validation (d).

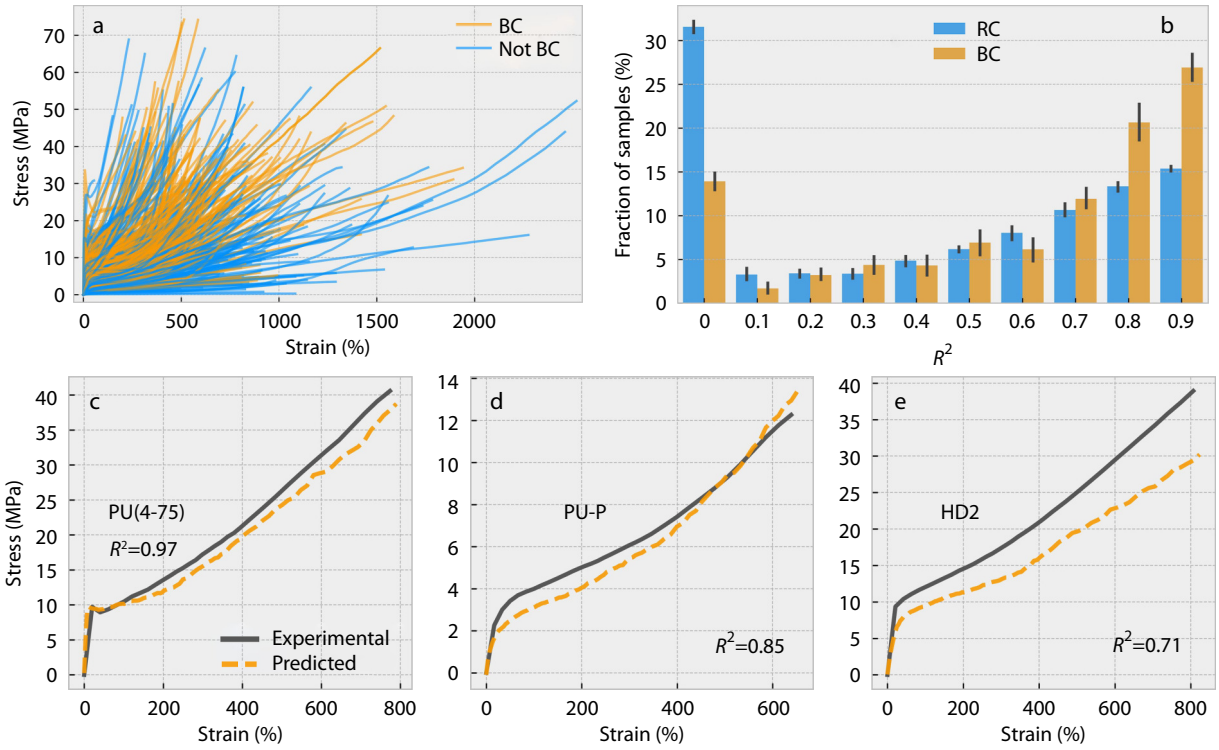

Fig. 7 The distribution of stress-strain curves: "BC" and "Not BC" represent those curves in the benchmark dataset and not in the benchmark dataset, respectively (a). The distribution of R^2: "RC" represents those curves in the raw dataset (b). Typical examples for the prediction of stress-strain curves (c–e). Samples names in the original report are labeled in (c),[49] (d)[50] and (e).[51]

measurements. However, accurate prediction for these stress-strain curves is still a challenge, even for those PUE samples with high predicted R^2 (as shown in Figs. 7c–7e). Hence, we release the raw dataset and the most representative benchmark dataset

https://doi.org/10.1007/s10118-022-2838-6

covering the detailed experimental and calculated information and call for more interdisciplinary efforts to tackle the long-standing gap problem. In addition, these datasets can serve as the public dataset to facilitate the development of polymer informatics or finite element analysis.

CONCLUSIONS

In this work, the broadly distributed mechanical properties of polyurethane elastomers (PUEs) are investigated. It is a good model system to tackle the long-standing challenge to bridge the gap between the mechanical properties and molecular descriptors. Suffering from the diverse CCPMr and the non-Gaussian distributions for the mechanical properties, we propose a recursive data elimination protocol construct a benchmark dataset from a raw dataset with 50.4% samples remained. The R^2s for the predictions of YM, TS, and EB are improved from 0.73–0.78 for the raw dataset to 0.85–0.91 for the benchmark dataset. The result shows that samples in the benchmark dataset have an intrinsic conserved CCPMr, and their distributions of mechanical properties and profiles of stress-strain curves are similar to those in the raw dataset. Furthermore, the fitting of stress-strain curves utilizing machine learning is more accurate and more robust than that using a well-performed constitutive model, *i.e.*, Khiêm-Itskov model. The challenge to bridge chemical structures, interaction, composition, processing, and measurement with mechanical curves at the macroscale needs further effort, accompanying this work we release the raw dataset and the most representative benchmark dataset to data with detailed information in this work to call for more attention to tackle this long-standing gap problem. It is worth noting that the sequence of monomers, the aggregated structures, and even the after-treatment process may also influence the mechanical properties of PUEs, which were not considered in the present study. Therefore, a more accurate model considering the molecular detail, phase morphology and processing method will be the focus of future studies for polymer systems.

NOTES

The authors declare no competing financial interest.

Electronic Supplementary Information

Electronic supplementary information (ESI) is available free of charge in the online version of this article at http://doi.org/10.1007/s10118-022-2838-6.

ACKNOWLEDGMENTS

The work was financially supported by the National Natural Science Foundation of China (Nos. 51988102 and 22173094), CAS Key Research Program of Frontier Sciences (No. QYZDY-SSW-SLH027). We are grateful to Network and Computing Center, Changchun Institute of Applied Chemistry for essential support. The authors also appreciate the financial support of Major Science and Technology Project in Yunnan Province (No. 202002AB080001-1).

REFERENCES

1 Liu, T.; Liu, L.; Cui, F.; Ding, F.; Zhang, Q.; Li, Y. Predicting the performance of polyvinylidene fluoride, polyethersulfone and polysulfone filtration membranes using machine learning. *J. Mater. Chem. A* **2020**, *8*, 21862–21871.
2 Liu, L.; Chen, W.; Li, Y. A statistical study of proton conduction in nafion®-based composite membranes: prediction, filler selection and fabrication methods. *J. Membr. Sci.* **2018**, *549*, 393–402.
3 Liu, L.; Liu, T.; Ding, F.; Zhang, H.; Zheng, J.; Li, Y. Exploration of the polarization curve for proton-exchange membrane fuel cells. *ACS Appl. Mater. Interfaces* **2021**, *13*, 58838–58847.
4 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. Des.* **2020**, *194*, 108932.
5 Liu, L. Y.; Chen, W. D.; Liu, T. L.; Kong, X. X.; Zheng, J. F.; Li, Y. Q. Rational design of hydrocarbon-based sulfonated copolymers for proton exchange membranes. *J. Mater. Chem. A* **2019**, *7*, 11847–11857.
6 Somarathna, H. M. C. C.; Raman, S. N.; Mohotti, D.; Mutalib, A. A.; Badri, K. H. The use of polyurethane for structural and infrastructural engineering applications: a state-of-the-art review. *Constr. Build. Mater.* **2018**, *190*, 995–1014.
7 Opris, D. M. Polar elastomers as novel materials for electromechanical actuator applications. *Adv. Mater.* **2018**, *30*, 1703678.
8 Xiao, R.; Huang, W. M. Heating/solvent responsive shape-memory polymers for implant biomedical devices in minimally invasive surgery: current status and challenge. *Macromol. Biosci.* **2020**, *20*, e2000108.
9 Shi, R.; Chen, D.; Liu, Q.; Wu, Y.; Xu, X.; Zhang, L.; Tian, W. Recent advances in synthetic bioelastomers. *Int. J. Mol. Sci.* **2009**, *10*, 4223–4256.
10 Utrera-Barrios, S.; Verdejo, R.; Lopez-Manchado, M. A.; Santana, M. H. Evolution of self-healing elastomers, from extrinsic to combined intrinsic mechanisms: a review. *Mater. Horizons* **2020**, *7*, 2882–2902.
11 Ma, Z. P.; Li, H.; Jing, X.; Liu, Y. J.; Mi, H. Y. Recent advancements in self-healing composite elastomers for flexible strain sensors: materials, healing systems, and features. *Sensors Actuators A-Phys.* **2021**, *329*, 112800.
12 Sui, T.; Baimpas, N.; Dolbnya, I. P.; Prisacariu, C.; Korsunsky, A. M. Multiple-length-scale deformation analysis in a thermoplastic polyurethane. *Nat. Commun.* **2015**, *6*, 6583.
13 Ding, F.; Liu, T.; Zhang, H.; Liu, L.; Li, Y. Stress-strain curves for polyurethane elastomers: a statistical assessment of constitutive models. *J. Appl. Polym. Sci.* **2021**, *138*, e51269.
14 He, H.; Zhang, Q.; Zhang, Y.; Chen, J.; Zhang, L.; Li, F. A comparative study of 85 hyperelastic constitutive models for both unfilled rubber and highly filled rubber nanocomposite material. *Nano Mater. Sci.* **2021**, *4*, 64–82.
15 Ma, R.; Luo, T. Pi1m: A benchmark database for polymer informatics. *J. Chem. Inf. Model.* **2020**, *60*, 4684–4690.
16 Otsuka, S.; Kuwajima, I.; Hosoya, J.; Xu, Y.; Yamazaki, M. Polyinfo: Polymer database for polymeric materials design. *Proceedings of the 2011 International Conference on Emerging Intelligent Data and Web Technologies* **2011**, 22–29.
17 Nozaki, S.; Masuda, S.; Kamitani, K.; Kojio, K.; Takahara, A.; Kuwarnura, G.; Hasegawa, D.; Moorthi, K.; Mita, K.; Yamasaki, S. Superior properties of polyurethane elastomers synthesized with aliphatic diisocyanate bearing a symmetric structure. *Macromolecules* **2017**, *50*, 1008–1015.

18 Hu, J.; Mo, R.; Sheng, X.; Zhang, X. A self-healing polyurethane elastomer with excellent mechanical properties based on phase-locked dynamic imine bonds. *Polym. Chem.* **2020**, *11*, 2585–2594.
19 Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **2016**, 785–794.
20 Landrum, G. Rdkit: Open-source cheminformatics software. **2021**. https://rdkit.org/
21 Van Krevelen, D. W.; Te Nijenhuis, K., in *Properties of polymers (fourth edition)*, Elsevier. Amsterdam, **2009**, 189–227.
22 Fedors, R. F. A method for estimating both the solubility parameters and molar volumes of liquids. Supplement. *Polym. Eng. Sci.* **1974**, *14*, 472–472.
23 Zhang, H.; Ding, F.; Liu, T. L.; Liu, L. Y.; Li, Y. Q. Additivity of the mechanical properties for acrylonitrile-butadiene-styrene resins. *J. Appl. Polym. Sci.* **2022**, *139*, e51923.
24 Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **2002**, *46*, 389–422.
25 Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
26 Smola, A. J.; Scholkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222.
27 Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536.
28 Fix, E.; Hodges, J. L. Discriminatory analysis-nonparametric discrimination - consistency properties. *Int. Stat. Rev.* **1989**, *57*, 238–247.
29 Altman, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.* **1992**, *46*, 175–185.
30 Pugar, J. A.; Gang, C.; Huang, C.; Haider, K. W.; Washburn, N. R. Predicting young's modulus of linear polyurethane and polyurethane–polyurea elastomers: bridging length scales with physicochemical modeling and machine learning. *ACS Appl. Mater. Interfaces* **2022**, *14*, 16568–16581.
31 Ertem, S. P.; Yilgor, E.; Kosak, C.; Wilkes, G. L.; Zhang, M. Q.; Yilgor, I. Effect of soft segment molecular weight on tensile properties of poly(propylene oxide) based polyurethaneureas. *Polymer* **2012**, *53*, 4614–4622.
32 Cordero, J. A.; He K.; Janya K.; Echigo S.; Itoh S. Predicting formation of haloacetic acids by chlorination of organic compounds using machine-learning-assisted quantitative structure-activity relationships. *J. Hazard. Mater.* **2021**, *408*, 124466.
33 Yang, S.; Wang, S.; Du, X.; Du, Z.; Cheng, X.; Wang, H. Mechanically robust self-healing and recyclable flame-retarded polyurethane elastomer based on thermoreversible crosslinking network and multiple hydrogen bonds. *Chem. Eng. J.* **2020**, *391*, 123544.
34 Prasanna, S.; Doerksen, R. J. Topological polar surface area: A useful descriptor in 2d-qsar. *Curr. Med. Chem.* **2009**, *16*, 21–41.
35 Labute, P. A widely applicable set of descriptors. *J. Mol. Graph. Model.* **2000**, *18*, 464–477.
36 Pugar, J. A.; Childs, C. M.; Huang, C.; Haider, K. W.; Washburn, N. R. Elucidating the physicochemical basis of the glass transition temperature in linear polyurethane elastomers with machine learning. *J. Phys. Chem. B* **2020**, *124*, 9722–9733.
37 Menon, A.; Thompson-Colon, J. A.; Washburn, N. R. Hierarchical machine learning model for mechanical property predictions of polyurethane elastomers from small datasets. *Front. Mater.* **2019**, *6*, 87.
38 He, Y.; Xie, D.; Zhang, X. The structure, microphase-separated morphology, and property of polyurethanes and polyureas. *J. Mater. Sci.* **2014**, *49*, 7339–7352.
39 Sonnenschein, M. F.; Guillaudeu, S. J.; Landes, B. G.; Wendt, B. L. Comparison of adipate and succinate polyesters in thermoplastic polyurethanes. *Polymer* **2010**, *51*, 3685–3692.
40 Shin, J.; Matsushima, H.; Chan, J. W.; Hoyle, C. E. Segmented polythiourethane elastomers through sequential thiol-ene and thiol-isocyanate reactions. *Macromolecules* **2009**, *42*, 3294–3301.
41 Falco, G.; Simonin, L.; Pensec, S.; Dalmas, F.; Chenal, J. M.; Bouteiller, L.; Chazeau, L. Linear and nonlinear viscoelastic properties of segmented silicone-urea copolymers: Influence of the hard segment structure. *Polymer* **2020**, *186*, 122041.
42 Rogulska, M.; Kultys, A.; Pikus, S. Studies on thermoplastic polyurethanes based on new diphenylethane-derivative diols. Iii. The effect of molecular weight and structure of soft segment on some properties of segmented polyurethanes. *J. Appl. Polym. Sci.* **2008**, *110*, 1677–1689.
43 Kim, H. D.; Lee, T. J.; Huh, J. H.; Lee, D. J. Preparation and properties of segmented thermoplastic polyurethane elastomers with two different soft segments. *J. Appl. Polym. Sci.* **1999**, *73*, 345–352.
44 Liao, T.; Yang, X.; Zhao, X. T.; Tang, Y. J.; Jiang, Z. Y.; Men, Y. F. Gaussian and non-gaussian distributions of fracture properties in tensile stretching of high-density polyethylene. *Macromolecules* **2021**, *54*, 8860–8874.
45 Tang, H.; Cui, F.; Liu, L.; Li, Y. Predictive models for tyrosinase inhibitors: challenges from heterogeneous activity data determined by different experimental protocols. *Comput. Biol. Chem.* **2018**, *73*, 79–84.
46 Cereto-Massague, A.; Ojeda, M. J.; Valls C.; Mulero M.; Garcia-Vallve S.; Pujadas G. Molecular fingerprint similarity search in virtual screening. *Methods* **2015**, *71*, 58–63.
47 Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of mdl keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.
48 Zhang, H.; Ding, F.; Liu, T.; Liu, L.; Li, Y. Additivity of the mechanical properties for acrylonitrile-butadiene-styrene resins. *J. Appl. Polym. Sci.* **2021**, *139*, e51923.
49 Kim, B. K.; Lee, S. Y.; Xu, M. Polyurethanes having shape memory effects. *Polymer* **1996**, *37*, 5781–5793.
50 Rahmawati, R.; Nozaki, S.; Kojio, K.; Takahara, A.; Shinohara, N.; Yamasaki, S. Microphase-separated structure and mechanical properties of cycloaliphatic diisocyanate-based thiourethane elastomers. *Polym. J.* **2018**, *51*, 265–273.
51 Oprea, S.; Timpu, D.; Oprea, V. Design-properties relationships of polyurethanes elastomers depending on different chain extenders structures. *J. Polym. Res.* **2019**, *26*, 117.
52 Khiem, V. N.; Itskov, M. Analytical network-averaging of the tube model: rubber elasticity. *J. Mech. Phys. Solids* **2016**, *95*, 254–269.

https://doi.org/10.1007/s10118-022-2838-6