THE JOURNAL OF PHYSICAL CHEMISTRY B

pubs.acs.org/JPCB Article

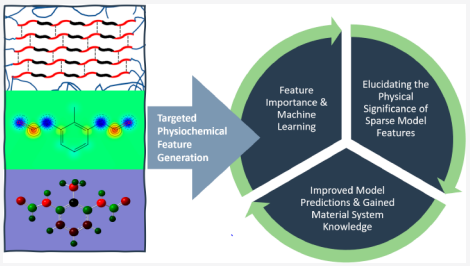
Elucidating the Physicochemical Basis of the Glass Transition Temperature in Linear Polyurethane Elastomers with Machine Learning

Published as part of The Journal of Physical Chemistry virtual special issue "Machine Learning in Physical Chemistry".

Joseph A. Pugar, Christopher M. Childs, Christine Huang, Karl W. Haider, and Newell R. Washburn*

Cite This: https://dx.doi.org/10.1021/acs.jpcb.0c06439 Read Online

ABSTRACT: The glass transition temperature (T_g) is a fundamental property of polymers that strongly influences both mechanical and flow characteristics of the material. In many important polymers, configurational entropy of side chains is a dominant factor determining it. In contrast, the thermal transition in polyurethanes is thought to be determined by a combination of steric and electronic factors from the dispersed hard segments within the soft segment medium. Here, we present a machine learning model for the T_g in linear polyurethanes and aim to uncover the underlying physicochemical parameters that determine this. The model was trained on literature data from 43 industrially relevant combinations of polyols and isocyanates using descriptors derived from quantum chemistry, cheminformatics, and solution thermodynamics forming the feature space. Random forest and regularized regression were then compared to build a sparse linear model from six descriptors. Consistent with empirical understanding of polyurethane chemistry, this study indicates the characteristics of isocyanate monomers strongly determine the increase in T_g. Accurate predictions of T_g from the model are demonstrated, and the significance of the features is discussed. The results suggest that the tools of machine learning can provide both physical insights as well as accurate predictions of complex material properties.



1. INTRODUCTION

By the 1950s, the wide ranging properties of thermoplastic polyurethanes (TPUs) were being explored on the basis of their considerable chemical diversity.[1,2] Such diversity led to their uses in applications such as adhesives, elastomers, films, and foams, all based on the same fundamental urethane linkage. As the understanding and use of TPUs broadened, tailoring individual material properties to serve specific applications became increasingly important.

TPUs are typically prepared from three components: a macrodiol (polyol), a diisocyanate, and a low molecular weight diol (chain extender). The relative amount and nature of these three major components dramatically influence the thermal and mechanical properties of the resulting polymers. TPUs are known to segregate into soft and hard segments, where the macro-diols (soft segments) and the reaction product of the chain extender and diisocyanate (hard segments) are in separate domains. The chemical interactions and morphology of these domains lead to different thermal and mechanical properties in the TPU. The differences give rise to the tunability of the chemical, mechanical, thermal, and morphological properties and, therefore, the bulk polyurethane material properties. The hard segments have high degrees of hydrogen bonding and polarity and form ordered structures within the soft segments, which have low polarity and contribute elastically to the bulk material. Material design is then approached by controlling chemical structure and molecular weight affecting chain−chain interactions, the degree of polymerization, and hard segment content, primarily dictated by the functional NCO/OH index. However, due to the diverse chemical structures of the constituents accessible to polyurethanes, the challenges to accurately predict and therefore tune the material property outcomes are complex.

One of the most fundamental of these properties is the glass transition temperature (T_g). The importance of this property is due to the dramatic change in almost all other properties as the thermal threshold is reached. Its control is necessary for processing the material in both the solid and liquid phases, and it is a critical factor in determining the mechanical properties, making it important to understand what drives it and build

Received: July 14, 2020
Revised: August 26, 2020
Published: September 8, 2020

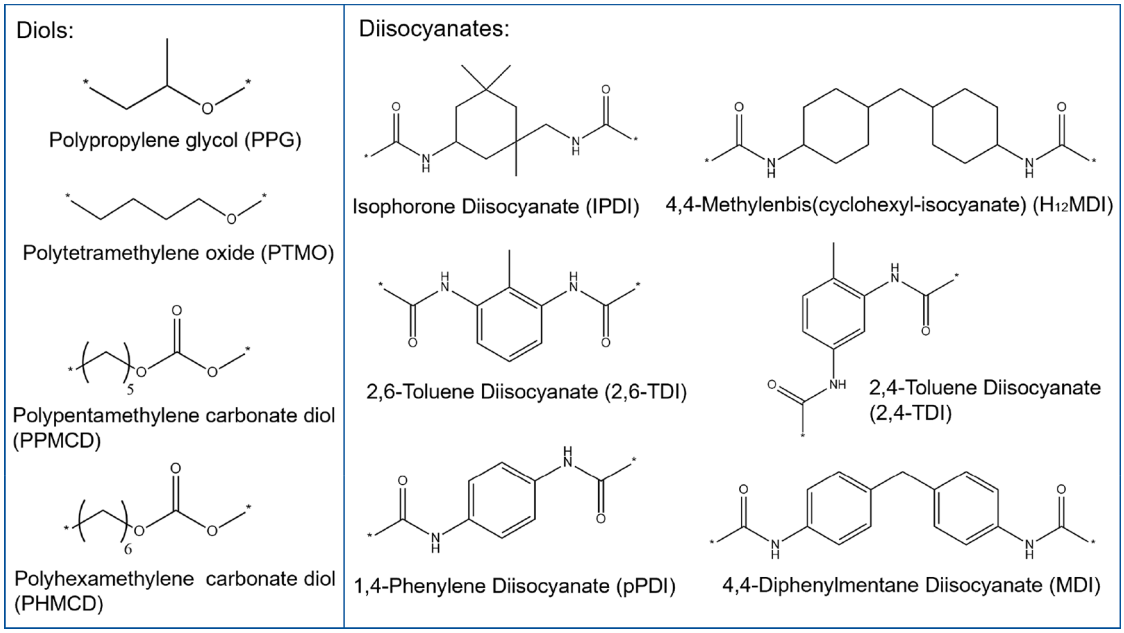


Figure 1. Two-dimensional chemical structures of 4 unique soft segments and 6 unique diisocyanates. Each "*" termination identifies the end of the block or repeat unit and is modeled assuming a hydrogen atom.

predictive models based on this. While T_g for polymeric materials is broadly understood to be a function of molecular weight and steric and electronic factors[3−6] for TPUs, these models have limitations. Many of these models have been built and tested on polyolefin, polyacrylate, and polyester-based systems using free volume parameters but have different fundamental chain network architectures than TPUs.[7−9] The pendant functional groups that tune the free volume and configurational entropy, which increase the T_g with increasing size, are absent in the most common TPU materials. Instead, the network structure based on hard segment chemical structure and weight fraction strongly influence T_g of the soft segment in phase segregated TPUs.[10] The elucidation of these competing forces remains a significant challenge.

Early computational work developing structure−property relations was based on group contribution methods (GCM).[3,11] When one sums individual terms representing the physico-chemical contributions of chemical components to the material property, an estimation of the property for the bulk material can be made. However, major limitations exist due to the neglect of intermolecular interactions, which may have a significant impact on T_g. Intermolecular interactions are explicitly included in molecular dynamics simulations, but these are computationally expensive and offer less insight into the effects of the electronic structure of the monomers.[12] A more recent approach leverages computational and informatics techniques employing data-driven models.[13−15] For example, Pilania et al.[16] have combined cheminformatics data and feature engineering techniques to build machine learning algorithms for polyhydroxyalkanoate-based polymers. The approach resulted in the predicted root mean squared error (RMSE) of less than 5 K in some cases, and their models learned physical trends for the effects of side-chain length and aromaticity on T_g. Computational and molecular modeling of the polyurethane system as a micromechanical homogenization of the two varying segments has also been explored.[17,18]

In similar predictive modeling studies, polymers have been described using physically relevant features and quantitative structure−property relationship (QSPR) variables to capture the underlying forces in the material system.[3] The works of computational modeling and prediction of polymer properties have influenced the use of domain knowledge as modeling data.[19,20] However, ultimately restricting the extrapolation of these models across different chemical spaces is the model dependence on the input data set. In a specific analysis of T_g in these model types, Jha et al.[21] demonstrate how slight variations in input temperatures, for the historically error-associated reported value, can result in major changes in the model. Lastly, some hierarchical machine learning approaches have been developed to accommodate systems with small data sets.[22] These models utilize an array of middle layer variables based on material system domain knowledge and underlying forces fixated between the model input and response variables. The model quantifies and minimizes the variance when parametrized by latent variables to predict the output trends through statistical learning. Data-driven approaches such as these provide insight into structure and patterns in material data. These structures and patterns describe physical relationships within the material system, which allows the modeling tool to capture physically realizable trends while interpolating across the chemical space.

With the efforts of prior work as a basis, here, we further expand the use of machine learning methodologies as a valuable tool for developing a fundamental understanding of complex material systems. TPU sample data were taken from the literature[23−27] to build a 43-sample data set having a diversity of technologically relevant monomers and identical preparation and testing conditions with reported T_g values ranging from 193 to 375 K (Supporting Information). In addition to atomic group counts from repeat unit chemical structure, physically meaningful features were selected to capture the significant contributing factors to T_g in TPUs. These were calculated from the chemical structures of the repeat units using quantum chemistry, cheminformatics, and traditional computation of solubility factors. Cross-validation and collinearity analysis were performed to determine the features that captured variance in measured properties. The remaining features were subjected to

https://dx.doi.org/10.1021/acs.jpcb.0c06439
J. Phys. Chem. B XXXX, XXX, XXX−XXX

Table 1. List and Descriptions of the Calculated Features Used in Model Creation

name	acronym/ symbol	description	calc. method
C groups	C	the number of quaternary carbon structural units in the repeat unit	counts
CH groups	CH	the number of tertiary carbon structural units in the repeat unit	counts
CH_2 groups	CH2	the number of secondary carbon structural units in the repeat unit	counts
CH_3 groups	CH3	the number of primary carbon structural units in the repeat unit	counts
C_6H_6 groups	C6H6	the number of aromatic ring structural units in the repeat unit	counts
O groups	O	the number of oxygens in the repeat unit	counts
maximum absolute atomic charge	MPC	maximum absolute partial Gasteiger atomic charge	RDKit
topological polar surface area	TPSA	contributions from N and O atoms to the polar surface area of the molecule	RDKit
normalized number of hydrogen acceptors	HBA/MolWt	the number of hydrogen-bond acceptors for a molecule divided by the molecular weight of the molecule	RDKit
normalized number of hydrogen donors	HBD/MolWt	the number of hydrogen-bond donors for a molecule divided by the molecular weight of the molecule	RDKit
partition coefficient	LogP	the measure of relative hydrophobicity or lipophilicity of the molecule	RDKit
molar volume	MV	the molecular volume of the molecule	RDKit
radius of gyration	R_G	Arteca's radius of gyration of the molecule	RDKit
normalized number of rotatable bonds	NRB/MolWt	the number of rotatable bonds in the molecule divided by the molecular weight of the molecule	RDKit
dipole moment	DM	the electric dipole moment to capture a molecule's polarity	DFT
HOMO/LUMO energy gap	HL gap	the energy gap between the highest occupied molecular orbital and the lowest unoccupied molecular orbital	DFT
solubility parameter	Sol	Hildebrand solubility parameter calculated by group contribution values provided by Fedors	GCM
molecular weight	HS/SS MolWt	the number-average molecular weight of the corresponding component	
density	HS/SS density	the intrinsic molecular density of the corresponding component	
soft segment T_g	SS T_g	the glass transition temperature of the macrodiol component	
hard segment weight percentage	HS wt %	the percentage by weight of diisocyanate and chain extender in the final polyurethane sample	

importance studies throughout modeling, and specific latent variables capturing the underlying physical trends are discussed. Accurate predictions of the change in glass transition temperature from the pure soft segment to that of the TPUs were made after feature selection by regularized linear regression, and random forest modeling was performed. Validation was also performed on withheld samples, showing similar accuracy and predicted error to the test sets. The modeling strategies are shown to be effective tools in capturing the change from that of the soft segment T_g to the elevated value in the TPU system using small data sets.

2. METHODS

2.1. Data Collection. The glass transition data compiled for this study were gathered from experimental values reported in the literature. In total, 43 unique TPU compositions were collected for analysis.[23−27] The chemical diversity of the sample space spans 6 different diisocyanates and 4 macrodiols, with all samples incorporating 1,4-butanediol as a chain extender (Figure 1).

The functional index, which represents the molar ratio of NCO to OH groups, of all the samples in the data set was NCO/OH ≈ 1. Generally, if only a single index was reported for a given sample, the OH contributions to the ratio come from both the macrodiol and chain-extender hydroxyl functional groups. A similar functional group index for the prepolymer controls the hard segment weight percentage of the system, which was systematically increased across the sample set. As the prepolymer index increases, the hard segment weight percentage increases due to the higher fraction of diisocyanate and chain extender used in the synthesis. The prepolymer index was not used in the initial training set because of its significant correlation to the hard segment weight fraction. The reported number-average molecular weight of the samples in the data set were all on the order of 10 000−70 000 g/mol. Some sources did not report the final molecular weight but used similar monomer molecular weight reactants, functional indices, and reaction conditions. In addition to the chemical structure of the building blocks, initially organized as functional group counts, the densities, macrodiol molecular weights, hard segment weight fraction, and soft segment T_g for each sample were tabulated alongside the reported TPU T_g. The transition temperatures were experimentally measured from either differential scanning calorimetry (DSC) or dynamic mechanical analysis (DMA), where an example of a DMA sweep is shown in Figure S1. All collected data utilized heating rates of 2−20 °C/min and strain frequencies of 1 Hz. These measurement criteria were used to ensure an accurate comparison between all data sources as polymer thermal transitions are a strong function of scan rate.[28] To further ensure consistency among all data sources, all polymerizations were carried out using the two-step prepolymer method with similar reaction conditions to ensure there were no major deviations in the step-growth polymerizations. The reaction schemes contained <0.1% by weight of the catalyst dibutyltin dilaurate or were synthesized in catalyst-free reactions. The reported T_g values range from 193 to 375 K throughout the collected data set and include systematic chemical and compositional changes.

2.2. Computational Methods. *2.2.1. Descriptor Generation.* The glass transition is a kinetic phenomenon in which the loss of free volume as a function of temperature results in a divergence of relaxation times observed near T_g, leading to vitrification of amorphous phases and characterized by a significant increase in the storage modulus and reduction of viscous flow.[29] At a microscopic level, polymer T_g is associated with loss of configurational entropy as caging at short length scales, such as side chains and other substituents (β relaxation), becomes the dominant mode of stress−relaxation as chain-scale

https://dx.doi.org/10.1021/acs.jpcb.0c06439
J. Phys. Chem. B XXXX, XXX, XXX−XXX

reorganization (α relaxation) is suppressed, leading to a strong correlation between free volume and T_g.[30] Thus, in broadly used materials, such as polyolefins, polyacrylates, and polyhydroxyalkanoates, the incorporation of alkyl side chains is a straightforward strategy to reduce T_g, such as in poly(*n*-butyl acrylate) (219 K) vs poly(methyl acrylate) (283 K).

Even in homopolymers, this dynamic transition is heterogeneous, and glassy domains on the order of the Kuhn volume form.[31,32] These domains are generated through local energetic barriers that prevent reorganization,[33] and their incipient formation leads to a rubbery plateau that extends ca. 50 K above the measured T_g. In polymer−polymer and polymer−solvent blends, the microstructure becomes more complex, and the breadth of the glass transition depends strongly on the thermodynamic compatibility of the components. For material systems with low polarity, the barriers that inhibit thermally activated reorganization are due primarily to excluded volume interactions, making the molecular shape of central importance. However, it is expected that strongly interacting multiphase systems, such as TPUs, will have a diversity of steric and electronic factors that govern the glass transition. In the following paragraphs, the descriptor generation methodologies are discussed. A full list of the 40 descriptors considered in this study is tabulated in Table 1. All descriptors besides soft segment T_g and hard segment weight fraction were calculated for each soft segment and diisocyanate.

2.2.1.1. Functional Group Counts. Each of the 10 unique monomer structures was represented by a vector of integers describing the number of functional units that varied between them. The functional units specified were the number of primary, secondary, ternary, and quaternary carbons, along with aromatic rings and oxygens. For the polyhexamethylene-pentamethylene carbonate diol (PHMPMCD), a weighted average of the length of the hydrocarbon backbone in the repeat unit was taken.

2.2.1.2. Quantum Mechanical. Density-functional theory (DFT) calculations were carried out for each of the 10 possible composition units. Two features describing electrostatic interaction potential were extracted from the calculation output file: electric dipole moment and HOMO/LUMO energy gap. All molecule structures were drawn in Ampac 10.1 software and saved as Gaussian input files. DFT calculations were run in the Gaussian 16W program package at 298.15 K.[34] The B3LYP/6-31G* basis set was chosen because of its use previously in quantum chemical calculations of polyurethanes.[35,36] The calculations were run with the "polar" keyword input advised in the Gaussian guidelines.[37] The Gaussian output file was then accessed through Ampac, and the two variables were extracted for each unique chemical structure calculation.

2.2.1.3. Solubility. A Hildebrand solubility parameter for each unique repeat unit was added to the data set to capture the incompatibility between the soft and hard segments. To calculate the Hildebrand parameter, a group additivity approach pioneered by Van Krevelen was used for the calculation with the following equation.[11] In eq 1, δ is the solubility parameter, e_{coh} is the cohesive energy density, and $E_{coh,i}$ and $V_{m,i}$ are cohesive energy and molar volume contributions for the i^{th} structural unit. The group contribution values used for $E_{coh,i}$ and $V_{m,i}$ were taken from published works from Fedors,[38] which cover an extensive list of structural groups.

$$\delta = \sqrt{e_{coh}} = \sum E_{coh,i} / \sum V_{m,i} \quad (1)$$

2.2.1.4. Cheminformatics. The completion of the data set is compiled from physical chemistry-based features computed in RDKit.[39] The open-source software can compute hundreds of molecular descriptors from molecules represented by the simplified molecular-input line-entry system (SMILES). SMILES strings were generated for each of the monomers represented and are shown in the Supporting Information data set.

2.2.2. Machine Learning. In total, 43 candidate polyurethanes were tabulated in a spreadsheet (Supporting Information) for machine learning (ML) analysis. For each of the 43 unique samples, the reported T_g was accompanied by a 40-component feature vector. The elements of the feature vectors within the data set had their column orders preserved and were centered at zero and standardized for all analyses. A set of six data points composed of the soft segment monomer polypropylene glycol (PPG) was withheld as a validation set. As PPG had no remaining candidate molecules for training, this allowed for the analysis of the generalization capability of the ML model. The remaining 37 polyurethanes were randomly split into training and test sets at an 80% to 20% ratio, respectively. All data analysis was performed in Python with all ML algorithms utilized found in the Scikit-Learn package.[40]

With a training set of 29 polyurethanes and 40 features, the number of features needed to be reduced to less than the number of training points to avoid overfitting. Two feature selection methodologies were performed, analyzed, and compared to determine the best feature space in the prediction of T_g. The first was a sparse linear model predicted through the utilization of the least absolute shrinkage and selection operator (Lasso).[41] Lasso utilizes the least-squares regression cost function with an added penalty term in the form of the L_1-Norm as shown in eq 2:

$$\min_\beta(\|y - \beta x\|_2^2 + \alpha \|\beta\|_1) \quad (2)$$

In this equation, y represents a vector of the response variable, T_g. The feature space of chemical descriptors was encoded in x, while the β parameters are the model coefficients corresponding to the best model. The hyperparameter α was selected through cross-validation on the training set. For this analysis, leave-one-out cross-validation (LOOCV) was performed. A mean squared error (MSE) for each value of α for all folds was found. The selection of the value of α corresponding to the minimized MSE drives nonimportant β parameters and the corresponding features to zero. When one discards features that correlate with variations in T_g, the dimensionality of the system can be reduced.

In the random forest (RF), to optimize model complexity, various hyperparameters can be tuned through cross-validation. A random search grid was utilized with a 5-fold-cross-validation. The hyperparameters corresponding to the minimized MSE were selected for model analysis. These hyperparameters that can be tuned include:

1. Number of estimators: This describes the number of individual decision trees composing each forest. Values from 200 to 2000 were searched.
2. Each tree can be optimized by either utilizing or not bootstrapping. Bootstrapping is a technique of resampling subsets of the entire data set many times (equal to the number of estimators).[42] If bootstrapping is not utilized, the entire data set is utilized to train each individual tree. A

https://dx.doi.org/10.1021/acs.jpcb.0c06439
J. Phys. Chem. B XXXX, XXX, XXX−XXX

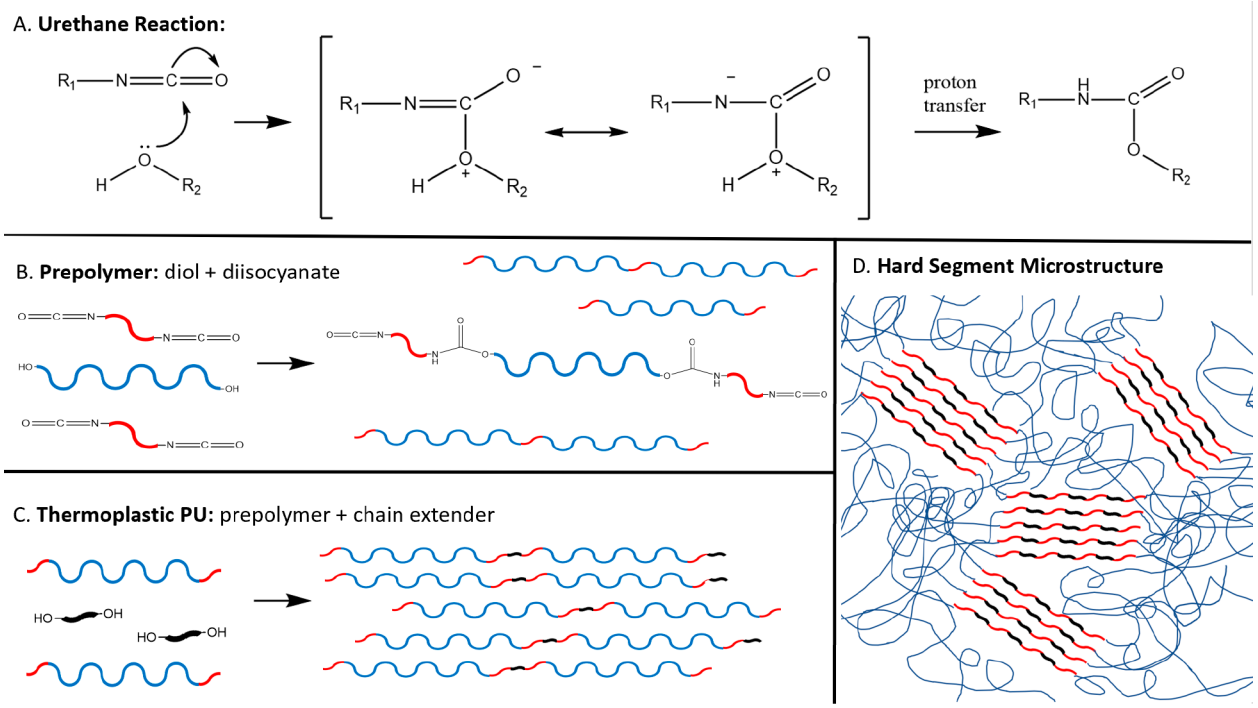


Figure 2. (A) General reaction mechanism pathway for a diol and diisocyanate creating a urethane bond without a catalyst. (B) The first of the two-step polymerization method reacting the diol and diisocyanates through urethane chemistry. The diisocyanate was in stoichiometric excess, i.e., NCO/OH > 1 for each sample to terminate each growing chain with an NCO functional group. (C) The second step of the two-step polymerization method in which short-chain diols extended the chains into linear networks. (D) Schematic representation of the microstructure depicting the ordered hard segment domains within the soft segment medium.

binary true/false operation was utilized in the hyperparameter search grid.

3. The number of features to split on is another parameter that can be optimized through cross-validation techniques. These were set to be bagging models, where the number of features is equal to the random subset of features, or the square root of the number of features was set as the random subset of features.
4. A collection of parameters to control the maximum depth of the tree is also optimized through cross-validation. These include controlling the maximum depth of the tree where values between 10 and fully grown trees were searched, and the number of samples required to be classified as a leaf in the decision tree where one, two, and four were screened.

To determine dominant features, permutation importance as implemented in Sci-Kit Learn based on Breiman[43] was performed. Permutation importance works through randomly shuffling each feature's value k times, in this case, 20. The average change in model performance at each shuffle, as determined through the r^2 value, is subtracted from the r^2 of the base model trained on all features. A plot of the average change in the r^2 score for each feature is created, where larger values are indicative of that feature having a higher importance toward the model. Permutation importance can be calculated on the training set itself, along with the test set to differentiate between factors contributing to overfitting and those capable of generalization.

3. RESULTS AND DISCUSSION

3.1. Training Set. In the literature data selected for this analysis, the TPU samples were all prepared using a two-step polymerization method comparable to previously developed synthesis routes.[44,45] In both reaction steps, urethane linkages were created between NCO and OH functional groups. The chemical mechanism of the catalyst-free reaction can be seen in Figure 2A. In all instances, the prepolymer was made by mixing liquid macrodiol and liquid diisocyanate in stoichiometric excess (Figure 2B). Next 1,4-butanediol was added to the heated reaction vessel as chain extender, which through analogous urethane chemistry, creates a linear network between the previously NCO-terminated prepolymer chains (Figure 2C).

The predictive models for T_g in polyolefins, polyacrylates, and polyhydroxyalkanoates depend strongly on monomer shape as a feature. However, most of the industrially relevant polyols used in TPU production lack side chains and therefore were not included in the training set. Descriptors were selected for the data set to describe electrostatic, hydrogen-bonding induced, and interchain interactions in the two-phase system that forms the superstructure in Figure 2D. All of the following descriptors were calculated using the 2D chemical structures shown in Figure 1.

Unit counts were performed to quantify the presence of various functional groups in each structure. Specifically, the relative amount of aromaticity and the free volume associated with different functional groups have been shown to capture morphological trends in polymers.[46−48] The single energy value dipole moment was chosen to describe the dielectric polarizability, which has been shown to relate to the two-phase segregation in polyurethanes.[49] The HOMO/LUMO energy gap was chosen because of the potential to describe the relative strengths of the interactions between the units.[50] Hildebrand solubility parameters have been shown to accurately predict the morphology of TPUs.[18,19] The larger the difference between the

https://dx.doi.org/10.1021/acs.jpcb.0c06439
J. Phys. Chem. B XXXX, XXX, XXX−XXX

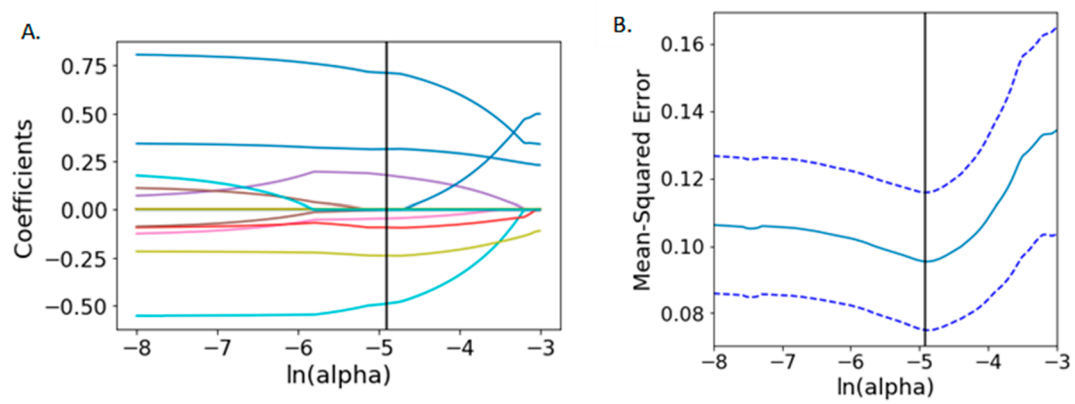


Figure 3. (A) The coefficient plot shows the variation of the model β parameters over the possible selections of α. (B) The selected value of α (0.0074) is indicated by the black vertical line. This corresponds to the minimized MSE determined through cross-validation. It should be noted that the Lasso coefficient path plot contains multiple instances of entry/exit/re-entry for certain coefficients. This can be attributed to multicollinearity in the feature set, which can contribute toward uncertainty in the proper feature selection. The resulting Lasso, however, selected a sparse set of seven features as shown in eq 3.

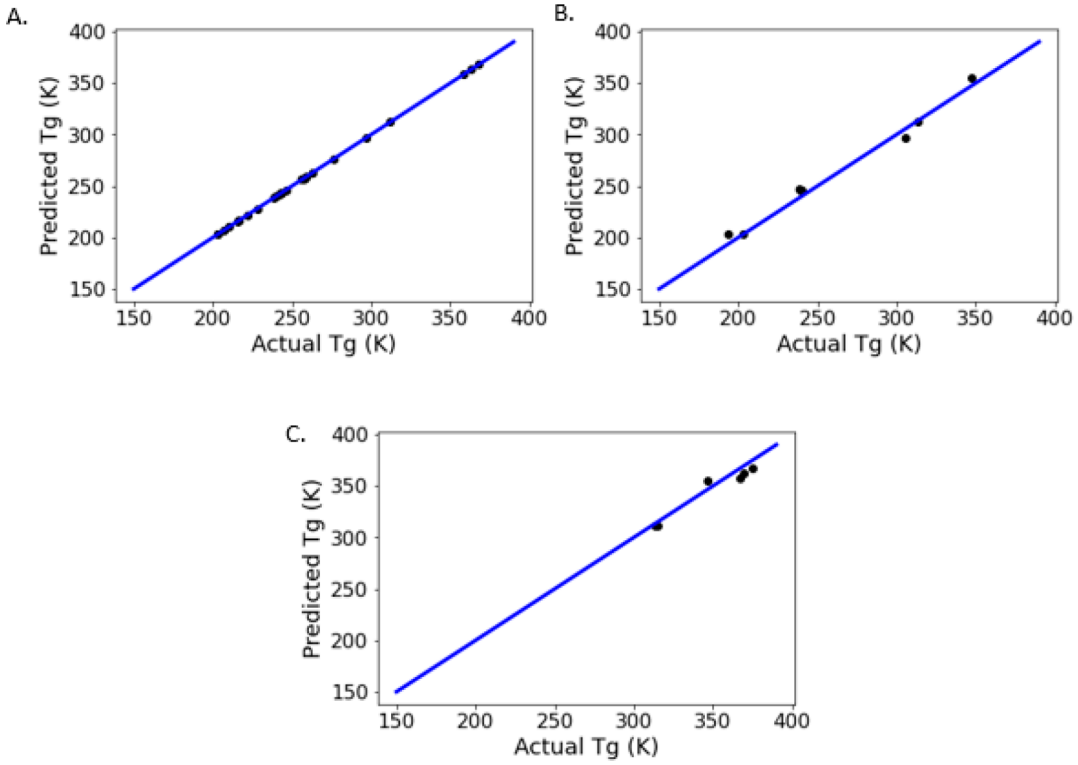


Figure 4. Plots showing the actual T_g compared to the T_g predicted through RF regression on the full data set for the (A) training set, (B) test set, and (C) validation set. A perfect prediction would show all points along the blue line.

parameters for soft and hard segments, the more likely the two phases were to be immiscible. The RDKit descriptors chosen were influenced by the same supporting domain knowledge to describe electrostatic interactions: maximum partial atomic charge (MPC) and topological polar surface area (TPSA), hydrogen bonding, the number of hydrogen-bond acceptors (HBA/MolWt) and donors (HBD/MolWt) and partition coefficient (LogP), and shape as represented by molecular volume (MV), the radius of gyration (R_G), and the number of rotatable bonds (NRB/MolWt) for both the hard and soft segment monomers. The number count descriptors were normalized by repeat unit molecular weight.

The resulting data set, after the addition of the domain-specific descriptors, contained 40 features. For each unique sample, the calculated features for the soft segment and reacted diisocyanate repeat unit involved were made individually available; i.e., no unique combinations of features were calculated before the machine learning methodology began. We note that all calculated chemical structure-based values in our data set were performed on postpolymerization (i.e., urethane chemistry complete) structures and each of these structures was passivated with hydrogens. Structure unit count, DFT, and RDKit descriptors were combined with soft segment molecular weight, soft segment glass transition temperature,

https://dx.doi.org/10.1021/acs.jpcb.0c06439
J. Phys. Chem. B XXXX, XXX, XXX−XXX

hard segment weight percentage, and component densities to create the full feature space for ML analysis.

3.2. Lasso Feature Selection Analysis. The full feature set was utilized to train and discover important descriptors through Lasso. The identification of forces that minimize the error in determining the glass transition temperature is done through regularization. Through this process, feature coefficients that do not best represent the true properties of the prediction goal are reduced to have little or no impact on the model. Figure 3 shows the coefficient plots and corresponding MSE determined through cross-validation across the range of the α hyperparameter.

$$T_g = 0.314 \times \mathrm{HSwt} + 0.178 \times \mathrm{SS}\frac{\mathrm{NRB}}{\mathrm{MolWt}} - 0.046 \times \mathrm{SS}\frac{\mathrm{HBA}}{\mathrm{MolWt}} - 0.491 \times \mathrm{HSlogP} + 0.711 \times \mathrm{HSDM} - 0.093 \times \mathrm{isodensity} - 0.240 \times \mathrm{chHS} \quad (3)$$

Figure S2 shows the resulting plots, and Table S1 shows the statistical analysis of the regression, which exhibit poor performance on the validation set, having an RMSE of 27.62 K. The entry/exit trade-off could cause the model to not appropriately include important features and is a reason more robust techniques, such as RF feature selection, will be explored further. To further evaluate the Lasso selected features, an RF model (Figure S3 and Table S2) was also applied to these seven features showing an RMSE of 14.94 K on the validation set, around one-half of that determined through Lasso.

3.3. Random Forest Feature Selection Analysis. The same procedure for feature selection was repeated, starting with the entire feature set, but this time performing feature selection through permutation importance rankings. A nonbootstrapped, unbagged ensemble consisting of 400 trees was found to provide the RF model corresponding to the lowest MSE. The train, test, and validation set performance are shown in Figure 4 and Table 2.

Table 2. r^2 and Root-Mean-Squared Error (RMSE) Scores for the RF Regression on the Complete Data Set

data set	r^2	RMSE (K)
train	1.00	0.00
test	0.98	7.03
validation	0.93	6.85

Regression performed on the complete data set had a 5 K lower RMSE on the test set and 8 K lower RMSE on the validation set than the RF performed Lasso selected features; however, it should be noted that the number of features was still greater than the number of data points.[51] Feature selection utilizing permutation importance was then compared, and underlying factors affecting T_g were learned. Figure 5 shows the permutation importance of the training set and test sets for the most important features. A complete figure is shown in Figure S4.

3.4. Optimal Features and Feature Contributions. After the feature dimensionality reductions were performed, a final collinearity analysis was used to remove redundant descriptors. A Pearson correlation matrix was calculated on the reduced feature space. When similar collinearity criteria to other feature selection methodologies were used,[52,53] two features were considered redundant if a greater than ±0.8 correlation was shared between them. Of the two correlated features, the one with a higher correlation to the T_g was retained in the model, while the other was discarded from the feature space. Although the Pearson correlation criteria only describe linear redundancies in the feature space, careful attention was paid to the prediction metrics as each feature was removed to ensure minimal information was being removed from the prediction space. The resulting heat plot shown in Figure 6 identifies three instances in which the nine highest importance features are strongly correlated.

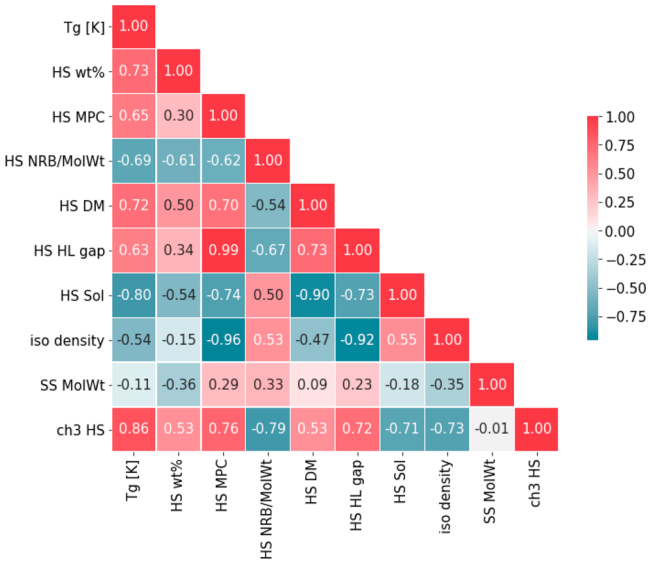


Figure 6. Pearson correlation coefficient heat map. A complete list of values can be found in Table S4.

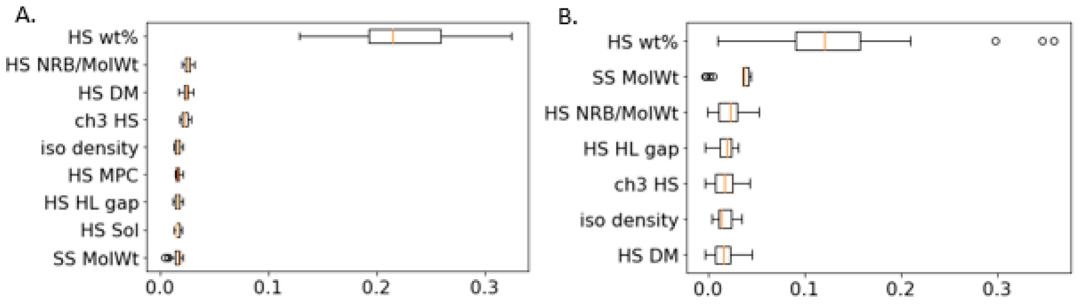


Figure 5. Permutation feature importance as a box and whisker plot, sorted by the mean value of importance, for (A) the training set and (B) the test set. The top features between both the training and test sets, which had a mean permutation importance score of greater than 0.015, are shown. These nine features were chosen as the reduced feature space for predicting T_g.

https://dx.doi.org/10.1021/acs.jpcb.0c06439
J. Phys. Chem. B XXXX, XXX, XXX−XXX

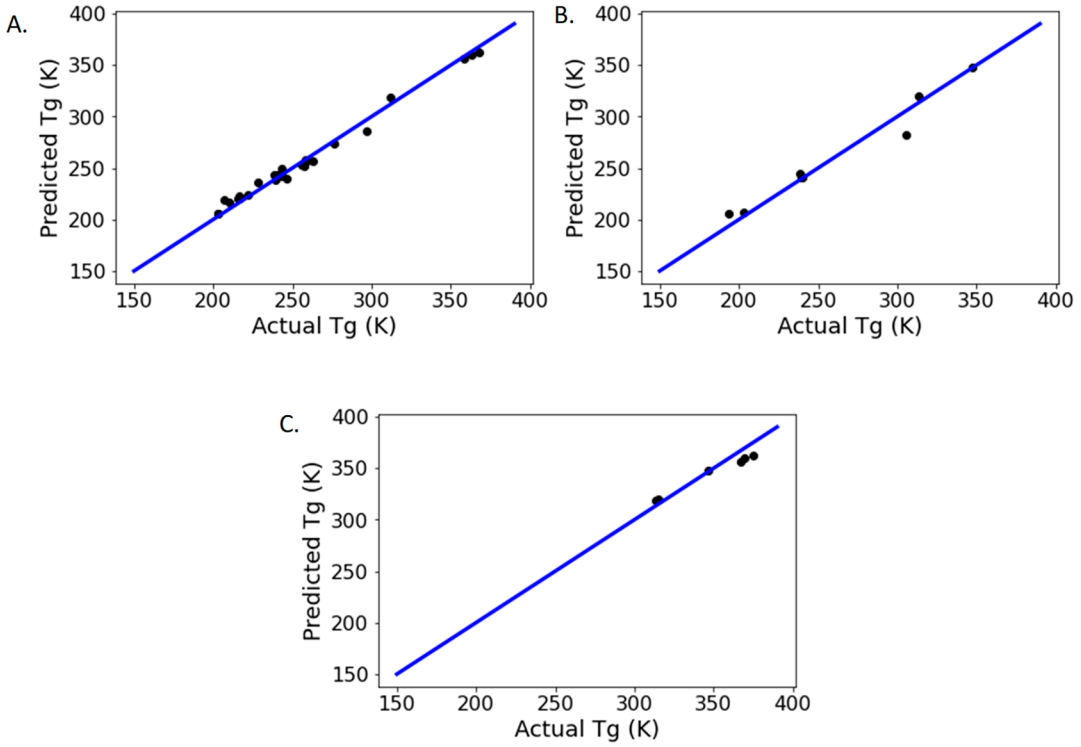


Figure 7. Plots showing the actual T_g compared to the T_g predicted through RF regression on the feature selected data set determined by RF and correlation analysis for the (A) training set, (B) test set, and (C) validation set. A perfect prediction would show all points along the blue line.

First, the density of the diisocyanate (g/cm^3) was shown to have a high negative correlation with the HS HOMO/LUMO gap and MPC. The increased structure density and, therefore, increased electron density of the repeat unit are expected to directly impact the corresponding molecular orbitals. By increasing election density, the energy gap between the highest occupied and lowest unoccupied molecular orbitals decreases, resulting in the negative correlation between the two properties. As such, the intrinsic density property was removed from the model for further analysis. Second and unsurprisingly, the HS HOMO/LUMO gap is shown to have a high positive correlation with the HS MPC. The MPC directly reflects the corresponding molecular orbits in the repeat unit and describes the electrostatic potential of the hard segment. The HS MPC had a marginally higher correlation with T_g, so the HS HOMO/LUMO gap was removed from the feature space. Finally, HS solubility and HS DM are found to have a high negative correlation. Both features describe the insoluble and polar nature of the hard segment within the soft segment medium in the form of phase segregation. Their negative relationship is also unsurprising: As the solubility parameter of the diisocyanate increases, it becomes less compatible with the soft segment, and phase segregation increases. As the dipole moment increases, a greater polar attraction is present between the two phases, and phase segregation decreases. The HS DM was removed from the data set due to its lower correlation with T_g, and it was found that this resulted in minimal reduction in r^2 for the overall model.

The six remaining features were retrained in an RF algorithm, and a bootstrapped, unbagged ensemble consisting of 1000 trees provided the RF model corresponding to the lowest MSE. The results are presented in Figure 7 and Table 3.

Despite eliminating 34 features, this RF model performed similar to a RF trained on the entire feature set, indicating that these six features are those that had the highest information content for predicting T_g in a sparse linear model. Provided in eq 4, the standardized coefficients for an ordinary linear regression model prediction on the six selected features from RF are shown to examine the effect each feature has on the T_g, and plots and statistical analysis are shown in Figure S5 and Table S3. In comparison, the linear regression had a test r^2 = 0.85, RMSE = 19.88 K and validation r^2 = 0.51, RMSE = 17.74 K. In the further discussion of the methodologies and conclusions, this model's features and characteristics will be discussed specifically.

Table 3. r^2 and Root-Mean-Squared Error (RMSE) Scores for the RF Regression on the Feature Selected Data Set Determined by RF

data set	r^2	RMSE (K)
train	0.99	5.17
test	0.96	10.04
validation	0.89	8.37

$$T_g = 0.265 \times \text{HS wt\%} + 0.05 \times \text{HS MPC} + 0.177\,\text{HS NRB/MolWt} - 0.394 \times \text{HS Sol} - 0.142 \times \text{SS MolWt} + 0.566 \times \text{ch3 HS} \quad (4)$$

The validation set accuracy and prediction error were similar to their respective test sets using the RF models. The withheld validation set of 6 unique PPG-soft segment TPUs of unseen chemical structure had an r^2 = 0.93, RMSE = 6.85 K in the full feature space RF model and a comparable r^2 = 0.89, RMSE = 8.37 K in the sparse RF model. The similarity of these scores to the test scores indicates the generalization of the model outside of the explicitly trained chemical space. The sparse RF model achieved good performance, which further supports the model's utilization of hard segment features to drive the increase in T_g from the pure soft segment value. Those features and their

https://dx.doi.org/10.1021/acs.jpcb.0c06439
J. Phys. Chem. B XXXX, XXX, XXX−XXX

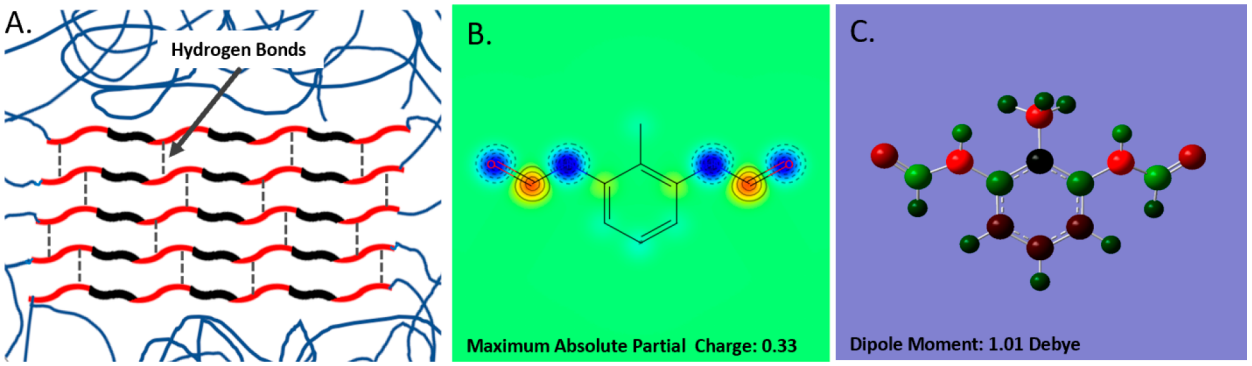


Figure 8. (A) Schematic of hard segment microstructure with dashed gray lines designating hydrogen bond sites between the N−H group of one repeat unit and the carbonyl group of another. (B) RDKit rendering of the maximum partial charge (MPC) distribution of the hard segment. The highest absolute charge value (red-highlighted carbons) are assigned to MPC. (C) Ampac rendering of the hydrogen-terminated hard segment. Quantum chemical calculations of the structures resulted in molecular charge distributions and quantum chemical descriptors.

physical significance to the material system are discussed in the following section.

3.5. Physical Insight into Selected Features. As discussed in Section 3.2, LASSO regularization and random forest permutation importance feature selection analysis were performed on the initial feature space composed of 40 descriptors. The significantly better performing RF model returned 5 features related to hard segment descriptors: weight percentage, maximum partial atomic charge (MPC), solubility parameter (Sol), number of CH_3 structural units, and normalized number of rotatable bonds (NRB/MolWt) in addition to the soft segment molecular weight. The most accurate sparse model developed was the RF trained on permutation importance-derived features; yielding a test of $r^2 = 0.96$, MSE = 10.04 K and validation of $r^2 = 0.89$, MSE = 8.37 K.

Of the original data provided by the literature sources, two features remained in the model: the soft segment molecular weight and hard segment weight fraction. These two features had high relative importance in T_g predictions, which is consistent with prior research.[54−56] It has been well understood that typical phenomenological models relating glass transition to molecular weight, such as the Flory-Fox equation,[57] do not capture the two-phase nature of polyurethanes. Rather, as the molecular weight of the soft segment increases, so does phase separation and therefore a reduction in the hard/soft segment interaction. This decrease in interaction leads to more viscous behavior and decreased T_g. Throughout the data set, the hard segment content increases for groups of samples with fixed chemical structure. This results in a varying hard segment dispersed in the soft segment medium, thus increasing the T_g differently. These two features describing material morphology support the classic representation of the hard segment dispersed in a soft segment medium.[58] Additionally, the lack of the soft segment T_g in the final feature space is initially surprising. Indeed, the model aimed to capture the increased shift in the thermal transition from this original value of the pure soft segment to the T_g for the two-phase material. This lack of representation is explained by the homogeneity of the soft segment T_g in the sample space and the resulting low Pearson's correlation coefficient to the TPU T_g ($r^2 = -0.17$). The segment T_g values only ranged from ∼190 to 210 K, leading to the decreased importance of this minimal value for the output prediction. The final model establishes a minimum of the temperature prediction space without it explicitly defined as a feature and predicts the TPU Tg from the variations in the 6 remaining features on a learned temperature scale.

The dynamic and thermally activated mechanical properties of TPUs have been shown to follow nonlinear behaviors influenced by hard segment content, chemical structure, and multiple temperature transitions contributed separately to the bulk by the individual phases.[17] Other RF features of importance further indicate the significant role of two-phase segregation and hard/soft segment interaction as the key forces being balanced. The lack of all other soft segment features, in general, suggests that, among standard and nondiverse choices for the soft segment in industrially relevant TPUs, the hard segment microstructures drive the accurate prediction of T_g.

The hard segment features used to describe the free-volume, electrostatic, and hydrogen-bonding interactions make up the remaining factors in the sparse model. In the RF feature set, the number of CH_3 structural methyl functional groups and the number of rotatable bonds normalized by the molecular weight of the hard segment (HS NRB/MolWt) are present. Of the 6 diisocyanate chemical structures used in this study, 3 of them had CH_3 structural groups (2,4-TDI, 2,6-TDI, IPDI) and 3 did not (MDI, pPDI, H_{12}MDI). The latter 3 structures also have higher degrees of symmetry, which indicates the formation of higher-ordered hard segment microstructures (Figure 8A), whereas the presence of CH_3 groups and less symmetry in the former 3 structures indicate greater amounts of free volume and configurational entropy in the microstructure. Less ordered structures lead to less phase segregation and an increase in T_g. Additionally, the HS NRB/MolWt remained as a feature and suggests that the conformational degrees of freedom to form the microstructure domains is important. We hypothesize that these features are capturing hard segment microstructure formation trends.

Similarly, the CH_3 structural unit has a role in hard/soft segment insolubility as well. In the solubility calculation through GCM performed in this study, the CH_3 group has the largest discrepancy between the cohesive energy contributed and the molar volume occupied, i.e., decreased cohesive energy density, which leads to a lower hard segment solubility, more compatibility and interaction with the soft segment, and a higher T_g. The last feature associated with physical forces within the molecular units' electrostatic force fields is HS MPC (Figure 8B). The MPC and electric dipole moment values from quantum chemical calculations (HS DM, Figure 8C) alike support the importance of capturing the polar nature of the two

https://dx.doi.org/10.1021/acs.jpcb.0c06439
J. Phys. Chem. B XXXX, XXX, XXX−XXX

segments (HS DM was removed from the feature space due to high collinearity with HS MPC.) The interfacial polarizability between the hard and soft segments in the polyurethane morphology is a key factor in capturing the electrostatic interactions within the complex material system.[49,59] Evident from the remaining features in the sparse model is the importance of hard segment microstructure and compatibility within the soft segment medium.

3.6. Future Directions. Future studies on the material properties of TPUs will expand the diversity of the training set. The sample set included many prototypical monomers used in the study of thermoplastic polyurethanes as well as having 1,4-butanediol as a universally conventional chain extender. Given a small sample set of 43 unique compositions, the modeling approach was to supplement chemical structure data with meaningful latent variables. The utilization of these domain-influenced descriptors targets underlying forces in the material space to result in physically meaningful predictions.[60] The methodology additionally presents a data-driven and reinforced understanding of the driving forces in the complex material system, but it can be susceptible to bias from this selection process. Similar to prior research, the emphasis is put on deriving generalizable models to capture trends from phenomenological data using a hierarchy of descriptors from different sources of domain knowledge.[61−63]

To further expand the capabilities of predictive models, the combination of hierarchical training data and machine learning models should be employed. However, the suitable approaches offer a variety of different methods that change the research tool by compromising certain characteristics of the model for the enhancement of others. This challenge is known as the accuracy versus generalizability trade-off and is exceptionally existent in the field of new materials discovery. In the recent work of Kopal et al.,[64,65] the dynamic thermomechanical responses in TPUs have been modeled using artificial neural networks (ANNs). Specifically, both the linear and nonlinear behavior of the dynamic and thermal transition-dependent loss and storage modulus (E'' and E', respectively) and the damping factor (tan δ) were accurately predicted from the results of the DMA experiments. The input data for the models is the thermal history of the material $T(t)$, and the predicted output is the corresponding values of E'', E', and tan δ. The predictive capabilities of the model built are critical to tailoring short-term viscoelastic responses in the system. However, the model was built upon a single unique material data set, and the insights from the ANN can be difficult to extract and even further difficult to interpret, ultimately not making the approach ideal for extrapolating knowledge and trends learned from a more general model to a new TPU composition. Models built around a narrow chemical feature space are often uninterpretable and lack generalizability.

Second and specific to complex material properties such as T_g, the robustness of the models will be critical in materials design. Jha et al.[21] demonstrate how sensitive models are to experimental measurement noise in the prediction of thermal transitions in polymers. They modeled for glass transition temperature prediction using a previously published fingerprinting hierarchy of data.[61] However, training on the different median, mean, minimum, and maximum T_g values resulted in four significantly different models. Values of $1 - r^2 = 13-21\%$ across the four models demonstrated the significance of variations in training data. It is important to note that their data set used significantly more diverse sample structures and syntheses, whereas this analysis had a more controlled and reproducible data set. Experimental error should remain an important consideration of model development, especially when extrapolating capabilities for the discovery of unobserved trends and new materials. New materials discovery will require models that are robust and can extrapolate outside the limits of the training space to include variations in experimental techniques. The inherent scarcity of data available in niche material classes inspires the combination of data from multiple sources to train models. Transfer learning, in which information from one data set supplements another for improved training, is a methodology increasingly being researched as a means to improve modeling of sparse data sets.[66] The results of this study suggest the importance of cheminformatic and quantum chemical data in the prediction of phase interaction in TPUs. A pretrained model built from large data sets of cheminformatics and quantum chemical variables could support transfer learning. These models describing the small molecule motifs used in polyurethane systems could further identify nontrivial relations to their corresponding polymerized macromolecules.[67] Furthermore, large feature engineered data sets of relevant descriptor combinations using mathematical operation expansion algorithms have been shown to be effective and could be utilized in transfer learning.[53] Coupling the computational power of black-box models with domain-derived and interpretable features in this manner motivated the work presented here.

In future work, we intend to generalize the model further to capture responses to a larger diversity of monomers. To construct tools for new materials discovery and improvement in polyurethanes, a holistic discovery of new high-performance or plant-derived polyols will become a research focus. We aim to expand the current model, which accurately predicts the change in T_g of a TPU from a biased soft segment starting point to a model that captures the physiochemical system as a whole.

4. CONCLUSION

This work demonstrated machine learning as a valuable and versatile tool for modeling the complex behavior of polyurethanes. When phenomenological but theory-based features are leveraged, the models have prediction capabilities within the chemical space of TPUs with varying chemical structures, chain architectures, soft/hard segment weight fractions, and a 1,4-butanediol chain extender. Predictions made were both accurate (RMSE on the order of 10 K) and generalizable to a validation set of chemical structures not present in the training set. The methodologies used to develop these models couple the separate, but synergistic, efficacies of using domain knowledge and statistical learning to describe a material system. A comprehensive data set creation approach was used to gather physically meaningful features spanning the monomer's chemical structure, DFT quantum chemical calculations, solubility parameters, and cheminformatics variables. After feature and collinearity reductions were performed, the remaining 6 features formed the low-dimensional feature space. Discussion of the relative importance of the remaining features suggests approaches to further refinement of the model as well as future directions for building more extensive models. Of the 6 remaining features, 5 of them described the hard segment: weight percentage, maximum partial atomic charge, number of rotatable bonds, solubility parameter, and number of methyl groups. For the narrow training set used here, the only feature corresponding to the soft segment was the molecular weight. This is consistent with how TPUs are tuned in practice

https://dx.doi.org/10.1021/acs.jpcb.0c06439
J. Phys. Chem. B XXXX, XXX, XXX−XXX

using the narrow range of polyols commonly found in industrially relevant materials, suggesting the importance of electrostatic, hydrogen bonding, and microstructural impacts of the hard segment in the modeling of these systems. This sparse model captured physically meaningful parameters from small data sets and provided an accurate prediction of a complex material behavior.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcb.0c06439.

Example DMA curve, complete ML r^2 and MSE values for all regressions performed, complete permutation importance plots, and complete correlation table (PDF)

Complete data set citing sources and complete listing of SMILES strings for all modeled monomers (XLSX)

Data set utilized for analysis (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Newell R. Washburn − *Department of Materials Science and Engineering, Department of Chemistry, and Department of Biomedical Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, United States;* orcid.org/0000-0001-7843-8860; Email: washburn@andrew.cmu.edu

Authors

Joseph A. Pugar − *Department of Materials Science and Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, United States*

Christopher M. Childs − *Department of Chemistry, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, United States;* orcid.org/0000-0001-8739-5997

Christine Huang − *Department of Chemistry, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, United States*

Karl W. Haider − *Covestro LLC, Pittsburgh, Pennsylvania 15205, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.0c06439

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Professor David Yaron (CMU) for guidance on quantum chemical calculations and thank Dr. Karen Stuart (Covestro, LLC) for input on the manuscript and for providing representative dynamic mechanical analysis data. C.M.C. was supported by ARPA-E (DE-AR0001138). J.A.P. and N.R.W. gratefully acknowledge support from the Covestro Science Award.

■ REFERENCES

(1) Oertel, G. *Polyurethane Handbook*, 2nd ed.; Oertel, G., Ed.; 1994.

(2) Petrović, Z. S. Polyurethanes. In *Handbook of Polymer Synthesis*, 2nd ed.; 2004.

(3) Weyland, H. G.; Hoftyzer, P. J.; Van Krevelen, D. W. Prediction of the Glass Transition Temperature of Polymers. *Polymer* **1970**, *11* (2), 79−87.

(4) DiBenedetto, A. T. Prediction of the Glass Transition Temperature of Polymers: A Model Based on the Principle of Corresponding States. *J. Polym. Sci., Part B: Polym. Phys.* **1987**, *25* (9), 1949−1969.

(5) Murugan, R.; Grendze, M. P.; Toomey, J. E.; Katritzky, A. R.; Karelson, M.; Lobanov, V.; Rachwal, P. Predicting Physical Properties from Molecular Structure. *Chemtech* **1994**, *24*, 17−23.

(6) Katritzky, A. R.; Rachwal, P.; Law, K. W.; Karelson, M.; Lobanov, V. S. Prediction of Polymer Glass Transition Temperatures Using a General Quantitative Structure-Property Relationship Treatment. *J. Chem. Inf. Comput. Sci.* **1996**, *36* (4), 879−884.

(7) Matsuoka, S. Relationship between Structure and Mechanical Properties of Polyolefins. *Polym. Eng. Sci.* **1965**, *5*, 142.

(8) Krause, S.; Gormley, J. J.; Roman, N.; Shetter, J. A.; Watanabe, W. H. Glass Temperatures of Some Acrylic Polymers. *J. Polym. Sci., Part A: Gen. Pap.* **1965**, *3*, 3573.

(9) Grieveson, B. M. The Glass Transition Temperature in Homologous Series of Linear Polymers. *Polymer* **1960**, *1*, 499.

(10) Wu, L.; Van Gemert, J.; Camargo, R. E. *Rheology Study in Polyurethane Rigid Foams*; 2008.

(11) Van Krevelen, D. W. *Cohesive Properties and Solubility*; 1997.

(12) Oguz, O.; Koutsoumpis, S. A.; Simsek, E.; Yilgor, E.; Yilgor, I.; Pissis, P.; Menceloglu, Y. Z. Effect of Soft Segment Molecular Weight on the Glass Transition, Crystallinity, Molecular Mobility and Segmental Dynamics of Poly(Ethylene Oxide) Based Poly(Urethane-Urea) Copolymers. *RSC Adv.* **2017**, *7*, 40745.

(13) Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine Learning in Materials Informatics: Recent Applications and Prospects. *npj Comput. Mater.* **2017**, *3*, 54.

(14) Mueller, T.; Kusne, A. G.; Ramprasad, R. Machine Learning in Materials Science: Recent Progress and Emerging Applications. In *Reviews in Computational Chemistry*; 2016.

(15) Pilania, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R. Accelerating Materials Property Predictions Using Machine Learning. *Sci. Rep.* **2013**, *3*, 1−6.

(16) Pilania, G.; Iverson, C. N.; Lookman, T.; Marrone, B. L. Machine-Learning-Based Predictive Modeling of Glass Transition Temperatures: A Case of Polyhydroxyalkanoate Homopolymers and Copolymers. *J. Chem. Inf. Model.* **2019**, *59* (12), 5013−5025.

(17) Huh, D. S.; Cooper, S. L. Dynamic Mechanical Properties of Polyurethane Block Polymers. *Polym. Eng. Sci.* **1971**, *11*, 369.

(18) Ginzburg, V. V.; Bicerano, J.; Christenson, C. P.; Schrock, A. K.; Patashinski, A. Z. Theoretical Modeling of the Relationship between Young's Modulus and Formulation Variables for Segmented Polyurethanes. *J. Polym. Sci., Part B: Polym. Phys.* **2007**, *45*, 2123.

(19) Bicerano, J. *Prediction of Polymer Properties*; CRC Press, 2002.

(20) Bicerano, J. *Computational Modeling of Polymers*; CRC Press, 1992.

(21) Jha, A.; Chandrasekaran, A.; Kim, C.; Ramprasad, R. Impact of Dataset Uncertainties on Machine Learning Model Predictions: The Example of Polymer Glass Transition Temperatures. *Modell. Simul. Mater. Sci. Eng.* **2019**, *27* (2), 024002.

(22) Menon, A.; Thompson-Colón, J. A.; Washburn, N. R. Hierarchical Machine Learning Model for Mechanical Property Predictions of Polyurethane Elastomers From Small Datasets. *Front. Mater.* **2019**, *6*, 87.

(23) Klinedinst, D. B.; Yilgör, I.; Yilgör, E.; Zhang, M.; Wilkes, G. L. The Effect of Varying Soft and Hard Segment Length on the Structure-Property Relationships of Segmented Polyurethanes Based on a Linear Symmetric Diisocyanate, 1,4-Butanediol and PTMO Soft Segments. *Polymer* **2012**, *53* (23), 5358−5366.

(24) Eceiza, A.; et al. Thermoplastic Polyurethane Elastomers Based on Polycarbonate Diols With Different Soft Segment Molecular Weight and Chemical Structure: Mechanical and Thermal Properties. *Polym. Eng. Sci.* **2008**, *48*, 297−306.

(25) Wang, C.-S.; Kenney, D. J. Effect of Hard Segments on Morphology and Properties of Thermoplastic Polyurethanes. *J. Elastomers Plast.* **1995**, *27*, 182−199.

(26) Kim, H. Do; Huh, J. H.; Kim, E. Y.; Park, C. C. Comparison of Properties of Thermoplastic Polyurethane Elastomers with Two Different Soft Segments. *J. Appl. Polym. Sci.* **1998**, *69* (7), 1349−1355.

(27) Schneider, N. S.; Sung, C. S. P.; Matton, R. W.; Illinger, J. L. Thermal Transition Behavior of Polyurethanes Based on Toluene Diisocyanate. *Macromolecules* **1975**, *8* (1), 62−67.
(28) Schawe, J. E. K. Analysis of Non-Isothermal Crystallization during Cooling and Reorganization during Heating of Isotactic Polypropylene by Fast Scanning DSC. *Thermochim. Acta* **2015**, *603*, 85.
(29) Napolitano, S.; Glynos, E.; Tito, N. B. Glass Transition of Polymers in Bulk, Confined Geometries, and near Interfaces. *Rep. Prog. Phys.* **2017**, *80*, 036602.
(30) White, R. P.; Lipson, J. E. G. Free Volume in the Melt and How It Correlates with Experimental Glass Transition Temperatures: Results for a Large Set of Polymers. *ACS Macro Lett.* **2015**, *4*, 588.
(31) Lodge, T. P.; McLeish, T. C. B. Self-Concentrations and Effective Glass Transition Temperatures in Polymer Blends. *Macromolecules* **2000**, *33*, 5278.
(32) Lipson, J. E. G.; Milner, S. T. Multiple Glass Transitions and Local Composition Effects on Polymer Solvent Mixtures. *J. Polym. Sci., Part B: Polym. Phys.* **2006**, *44*, 3528.
(33) Hodge, I. M. Adam-Gibbs Formulation of Enthalpy Relaxation near the Glass Transition. *J. Res. Natl. Inst. Stand. Technol.* **1997**, *102*, 195.
(34) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; et al. *Gaussian 16*, Revision C.01; 2016.
(35) Lempesis, N.; In'T Veld, P. J.; Rutledge, G. C. Atomistic Simulation of a Thermoplastic Polyurethane and Micromechanical Modeling. *Macromolecules* **2017**, *50*, 7399.
(36) Demir, P.; Akman, F. Molecular Structure, Spectroscopic Characterization, HOMO and LUMO Analysis of PU and PCL Grafted onto PEMA-Co-PHEMA with DFT Quantum Chemical Calculations. *J. Mol. Struct.* **2017**, *1134*, 404.
(37) *Polar*; https://gaussian.com/polar/.
(38) Fedors, R. F. A Method for Estimating Both the Solubility Parameters and Molar Volumes of Liquids. *Polym. Eng. Sci.* **1974**, *14*, 472.
(39) *RDKit: Open-source cheminformatics*; http://www.rdkit.org.
(40) Pedregosa, F.; Michel, V.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Vanderplas, J.; Cournapeau, D.; Pedregosa, F.; Varoquaux, G.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(41) Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267.
(42) AA; Mooney, C. Z.; Duval, R. D. Bootstrapping: A Non-parametric Approach to Statistical Inference. *J. Am. Stat. Assoc.* **1994**, *89*, 1150.
(43) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5−32.
(44) Dearlove, T. J.; Campbell, G. A. Synthesis and Characterization of Isocyanate-terminated Polyurethane Prepolymers. *J. Appl. Polym. Sci.* **1977**, *21*, 1499.
(45) Chang, E. Y. C.; Kaizerman, S. Thermoplastic Polyurethane Elastomers. US3503927A, 1968.
(46) Schollenberger, C. S. Thermoplastic Polyurethane Elastomer Structure-Thermal Response Relations. In *Advances in Chemistry*; ACS: Washington, DC, 1979.
(47) Clough, S. B.; Schneider, N. S. Structural Studies on Urethane Elastomers. *J. Macromol. Sci., Part B: Phys.* **1968**, *2*, 553.
(48) Prisacariu, C.; Prisacariu, C. Structural Studies on Polyurethane Elastomers. In *Polyurethane Elastomers*; 2011.
(49) North, A. M.; Reid, J. C.; Shortall, J. B. Some Physical Properties Associated with the Urethane Group—II. *Eur. Polym. J.* **1969**, *5*, 565.
(50) Winiwarter, S.; Ax, F.; Lennernäs, H.; Hallberg, A.; Pettersson, C.; Karlén, A. Hydrogen Bonding Descriptors in the Prediction of Human in Vivo Intestinal Permeability. *J. Mol. Graphics Modell.* **2003**, *21* (4), 273−287.
(51) Rogers, J.; Gunn, S. Identifying Feature Relevance Using a Random Forest. *Lecture Notes in Computer Science* **2006**, *3940*, 173.
(52) Dormann, C. F.; Elith, J.; Bacher, S.; Buchmann, C.; Carl, G.; Carré, G.; Marquéz, J. R. G.; Gruber, B.; Lafourcade, B.; Leitão, P. J. Collinearity: A Review of Methods to Deal with It and a Simulation Study Evaluating Their Performance. *Ecography* **2013**, *36*, 27.
(53) Ouyang, R.; Curtarolo, S.; Ahmetcik, E.; Scheffler, M.; Ghiringhelli, L. M. SISSO: A Compressed-Sensing Method for Identifying the Best Low-Dimensional Descriptor in an Immensity of Offered Candidates. *Phys. Rev. Mater.* **2018**, *2* (8), 083802.
(54) Seefried, C. G.; Koleske, J. V.; Critchfield, F. E. Thermoplastic Urethane Elastomers. I. Effects of Soft-segment Variations. *J. Appl. Polym. Sci.* **1975**, *19*, 2493.
(55) Seefried, C. G.; Koleske, J. V.; Critchfield, F. E. Thermoplastic Urethane Elastomers. II. Effects of Variations in Hard-segment Concentration. *J. Appl. Polym. Sci.* **1975**, *19*, 2503.
(56) Šebenik, U.; Krajnc, M. Influence of the Soft Segment Length and Content on the Synthesis and Properties of Isocyanate-Terminated Urethane Prepolymers. *Int. J. Adhes. Adhes.* **2007**, *27*, 527.
(57) Flory, P. J. *Principles of Polymer Chemistry*; Cornell University Press, 1953.
(58) Foreman, J. P.; Porter, D.; Pope, D.; Jones, F. R. Predicting the Material Properties of a Polyurethane Matrix (a Composite within a Composite). In *ECCM15 - 15th European Conference on Composite Materials*, Venice, Italy, June 24−28, 2012.
(59) North, A.M.; Reid, J.C.; Shortall, J.B. Some physical properties associated with the urethane group-II: Dielectric relaxation in thermoplastic polyurethane elastomers. *Eur. Polym. J.* **1969**, *5*, 565−573.
(60) Childs, C. M.; Washburn, N. R. Embedding Domain Knowledge for Machine Learning of Complex Material Systems. *MRS Commun.* **2019**, *9* (2), 806−820.
(61) Kim, C.; Chandrasekaran, A.; Huan, T. D.; Das, D.; Ramprasad, R. Polymer Genome: A Data-Powered Polymer Informatics Platform for Property Predictions. *J. Phys. Chem. C* **2018**, *122* (31), 17575−17585.
(62) Kim, C.; Pilania, G.; Ramprasad, R. From Organized High-Throughput Data to Phenomenological Theory Using Machine Learning: The Example of Dielectric Breakdown. *Chem. Mater.* **2016**, *28*, 1304−1311.
(63) Yang, Z.; Al-Bahrani, R.; Reid, A. C. E.; Papanikolaou, S.; Kalidindi, S. R.; Liao, W. K.; Choudhary, A.; Agrawal, A. Deep Learning Based Domain Knowledge Integration for Small Datasets: Illustrative Applications in Materials Informatics. In *Proceedings of the International Joint Conference on Neural Networks*; 2019.
(64) Kopal, I.; Harničárová, M.; Valíček, J.; Kušnerová, M. Modeling the Temperature Dependence of Dynamic Mechanical Properties and Visco-Elastic Behavior of Thermoplastic Polyurethane Using Artificial Neural Network. *Polymers (Basel, Switz.)* **2017**, *9*, 519.
(65) Kopal, I.; Harničárová, M.; Valíček, J.; Krmela, J.; Lukáč, O. Radial Basis Function Neural Network-Based Modeling of the Dynamic Thermo-Mechanical Response and Damping Behavior of Thermoplastic Elastomer Systems. *Polymers (Basel, Switz.)* **2019**, *11*, 1074.
(66) Hutchinson, M. L.; Antono, E.; Gibbons, B. M.; Paradiso, S.; Ling, J.; Meredig, B. Overcoming Data Scarcity with Transfer Learning. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, 2017; pp 1−10.
(67) Yamada, H.; Liu, C.; Wu, S.; Koyama, Y.; Ju, S.; Shiomi, J.; Morikawa, J.; Yoshida, R. Predicting Materials Properties with Little Data Using Shotgun Transfer Learning. *ACS Cent. Sci.* **2019**, *5*, 1717−1730.

https://dx.doi.org/10.1021/acs.jpcb.0c06439
J. Phys. Chem. B XXXX, XXX, XXX−XXX