Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2025, **4**, 3652

Machine learning of polyurethane prepolymer viscosity: a comparison of chemical and physicochemical approaches

Joseph A. Pugar, †[a] Calvin Gang,[b] Isabelle Millan,[a] Karl Haider[c] and Newell R. Washburn *[ab]

Polyurethane prepolymers are essential intermediates in the production of polyurethane foams, films, and elastomers, with viscosity playing a critical role in formulation, processing, and manufacturing. Despite its importance, there are no models that quantitatively predict the viscosity of a given polymer as a function of monomer chemistry. Traditional empirical models can effectively capture viscosity trends but often require extensive experimental datasets and provide limited interpretability, particularly when applied to novel formulations. Here, we explored regression options for representing polymer chemistry and for modeling the form of the temperature dependence. Monomers can be represented as a formulation in which they are labeled by monomer name or in a physicochemical framework where they are labeled by molecular characteristics derived from experimental measurements or computational methods. The overall form of the temperature-dependent viscosity can be modeled through a generic regressor function or by assuming the empirical form of the Andrade equation. A 39-sample training library was used to evaluate both approaches, with machine learning models achieving a coefficient of determination (r^2) of 0.71 for the chemical model testing data and 0.86 in predicting the Andrade equation parameter, which provides interpretable access to the continuous viscosity-temperature curve, for previously untested compositions. While chemically defined models offer a direct path to high-accuracy predictions within known compositional spaces, physicochemical informed models provide deeper insight into structure–property relationships, facilitating extrapolation to novel materials. This work underscores the tradeoffs between empirical and physics-informed modeling strategies and offers a structured approach to integrating domain knowledge into predictive frameworks for complex material systems.

Received 28th June 2025
Accepted 7th September 2025

DOI: 10.1039/d5dd00287g

rsc.li/digitaldiscovery

1 Introduction

Polyurethane prepolymers are essential intermediates in the synthesis of polyurethane (PUR) elastomers, serving as the foundational components in the two-step polyaddition process. In this process, an oligomeric polyol reacts with a diisocyanate to form the prepolymer, which then undergoes chain extension to produce the final elastomer product. The prepolymer route offers significant advantages over direct polymerization approaches, such as the targeted formation of soft and hard segments in separate steps, allowing for fine-tuning of mechanical properties and processability.[1,2] Among these characteristics, viscosity plays a critical role in dictating the practical handling and processing of the prepolymer melt, as well as the quality of the final elastomer. The viscosity of polyurethane prepolymers is highly dependent on both temperature and shear rate, creating challenges in predicting processing behavior across a wide range of industrial conditions. Industrially, the viscosity of the prepolymer must remain within a controlled range during blending and subsequent chain extension, as excessive viscosity may impede mixing and reduce product homogeneity.[3–6] However, understanding the underlying physicochemical principles that govern the temperature-dependent viscosity behavior remains a significant challenge. Traditional empirical models, while capable of capturing trends, often demand extensive experimental data and struggle to generalize to novel chemical formulations.[7,8] On the other hand, theoretical approaches or fully atomistic coarse-grained models could be used to predict polymer chain mobility and segmental relaxation across temperature ranges, yielding highly detailed viscosity predictions.[9–11] However, such simulations are expensive in both time and computational resources, often requiring days or weeks of high-performance computing for

[a]*Department of Materials Science and Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA. E-mail: washburn@andrew.cmu.edu*

[b]*Department of Chemistry, Carnegie Mellon University, Pittsburgh, PA 15213, USA*

[c]*Covestro AG, Pittsburgh, PA 15205, USA*

† Present address: Department of Surgery, University of Chicago, Chicago, IL 60615.

a single formulation. In contrast, existing data libraries capturing rheological and compositional information can be leveraged to build a variety of chemically or physically informed machine learning models.[12–15] These models require little to no additional resources to train and validate, yet can be rapidly scaled to support formulation optimization and accelerate materials discovery when constructed with appropriate rigor.[16]

The need to accurately predict prepolymer viscosity has motivated the development of both composition (chemical) and physics-informed (physicochemical) modeling approaches. Recent advances in machine learning–assisted modeling have demonstrated the potential to predict viscosity evolution in polyurethane systems across a range of chemistries and processing routes, from composite molding to modified asphalt binders.[17–20] These studies illustrate how data-driven models, often coupled with rheokinetic analysis, can accelerate formulation optimization and process control. Composition models, driven primarily by formulation-specific data, offer practical and accurate predictions within known compositional spaces. Furthermore, design-of-experiments (DOE) approaches—such as factorial designs and response-surface methodologies—provide a systematic framework for exploring key formulation factors and their interactions with minimal experimental runs.[21,22] However, empirical models lack interpretability and fail to generalize beyond the data used in training, limiting their utility in deeply studying the material library or for generalizing the findings to discover novel materials. In contrast, physics-informed models leverage molecular-scale features (some of which can be directly borrowed from existing empirical or physical models), such as hydrogen bonding interactions and molecular weight distributions, to establish a more mechanistic understanding of structure–property relationships.[23,24] By integrating physicochemical insights into DOE, one can both guide the selection of experimental factors and interpret complex interaction effects, thereby reducing the experimental burden while enhancing the mechanistic validity of the resulting models. Building on this foundation, the present work extends these concepts to prepolymer viscosity prediction by comparing both composition-driven and physicochemical modeling frameworks.

These models hold potential for extrapolating viscosity predictions to novel formulations, as illustrated in the schematic in Fig. 1, albeit at the cost of greater complexity in data acquisition and feature engineering. This duality presents a fundamental trade-off in the predictive modeling of complex systems, demonstrated in the following analysis with PUR prepolymer viscosity. Balancing these contrasting objectives requires strategic consideration of the intended application—whether the goal is formulation optimization within a known space or the discovery of new material combinations with improved performance characteristics. In this study, we aim to address this balance by presenting a dual modeling framework that leverages both chemical and physicochemical perspectives, offering insights into how these approaches can be strategically integrated to optimize both prediction accuracy and generalizability.

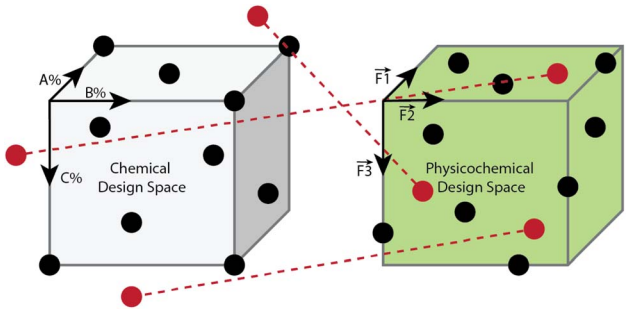

Fig. 1 Schematic representation of the limitations of chemical (A–C%) design spaces *versus* transformed physical parameter (Force 1–3) spaces for models to learn from. By defining samples with continuous physical parameters, what was an extrapolation task in the chemical space (new chemical reagents not defined by the existing axes/model features) can be re-framed as an interpolation task in the physical space.

2 Methodology

2.1 Empirical and physical modeling of PURs

A significant factor influencing prepolymer viscosity is the ratio of reactive groups NCO/OH during synthesis. Prepolymers typically have an NCO/OH ratio of 1.5–3, forming intermediates with narrow molecular weight distributions and well-defined end-functionalities.[25] In some cases, it is advantageous to intentionally increase this ratio beyond 3, producing semi-prepolymers characterized by a higher weight percent of free NCO groups (Fig. 2). These semi-prepolymers are advantageous due to their lower viscosity, which results from a combination of unreacted diisocyanate monomer acting as a solvent, end-capping of chains with isocyanate functionalities, and minimized chain extension during prepolymer formation. The

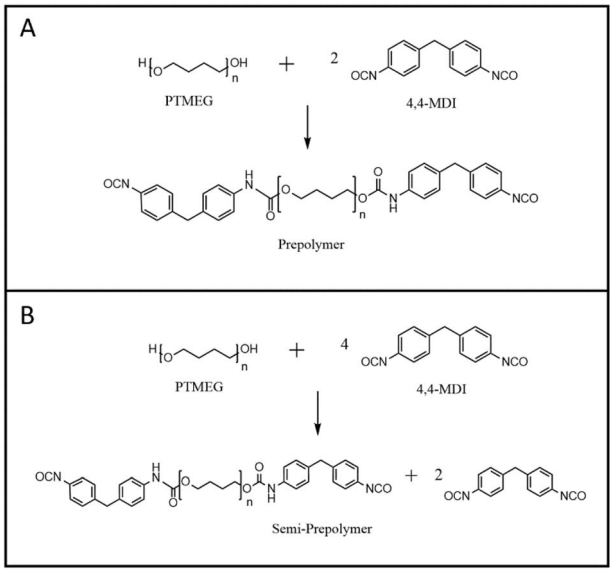

Fig. 2 Urethane reaction for varying diisocyanate amounts. (A) Prepolymer reaction with 2 : 1 NCO : OH stoichiometry. (B) Semi-prepolymer reaction with 4 : 1 NCO : OH stoichiometry.

stoichiometric imbalance of the functional groups (practically the percent by weight of NCO, also referred to as prepolymer index) along with the chemical and physical properties of the polyol and diisocyanate determine the behavior of the viscous prepolymer liquid. The viscosity is of critical importance because often the PU elastomer preparation involves mixing the NCO-terminated prepolymer with a variety of other reagents (*e.g.*, chain extenders, crosslinkers, catalysts) and subsequently transferring the reacting mixture for final processing. In practice, the prepolymer viscosity over a range of typical processing temperatures (40–80 °C) and shear rates (1–50 s^{-1}) is a critical processing variable.[26,27]

The chemical diversity of polyurethane prepolymers further complicates viscosity prediction. Common polyol types (*e.g.*, polyether, polyester, and polycarbonate) differ substantially in their hydrogen bonding potential, crystallinity, and hydrolysis resistance, impacting the mechanical and thermal performance of the final elastomer.[1,24,28,29] Polyesters are most often used in elastomer formulations where the mechanical toughness that results from their ability to form strong inter-chain hydrogen bonds and their crystallinity is desired. Polyethers are amorphous and do not participate as significantly in hydrogen bonding, which results in a more compliant material, but they have the advantage of being more resistant to hydrolysis than polyester based PU elastomers. Polycarbonate polyols offer the combination of mechanical performance and hydrolysis resistance and therefore are used as building blocks for high-toughness PUR elastomers. Other processing considerations stem from polyesters and polycarbonates having high glass transition temperatures (T_g) due to backbone ordering and high polarity while polyethers have much lower T_g. During polymerization, the diisocyanates connect the polyols *via* urethane linkages to extend the chains to higher molecular weights. While doing so, thermodynamic and electronic dissimilarities between the alternating segments can alter chain dynamics with effects from intermolecular length scales to bulk network mobility.[2,30–32] These same factors that govern the physical properties of the solid materials are also expected to influence melt properties, such as thermodynamics, dipole forces, and hydrogen bonding.

The temperature-dependent viscosities of polymer melts have been studied and modeled in a variety of different approaches. For moderate to high molecular weight polymers, complex viscoelastic responses dependent on shear history and shear rate complicate the creation of models that can capture behavior over a broad range of processing temperatures.[33,34] For low molecular weight (less than a few thousand g mol^{-1}) and under low-shear conditions, a first approximation can be made with an Arrhenius law or the related Andrade equation describing the viscosity (η) of liquids as a function of temperature (eqn (1)).

$$\eta = A \exp\left(\frac{B}{T}\right) \tag{1}$$

In the equation, A is a material constant and B is the activation energy normalized by a gas or Boltzmann constant. Subsequently, the Andrade equation was expanded to account for the effects of free volume and configuration in higher molecular weight systems. The third parameter C in eqn (2) introduces a temperature reference to the threshold of configurational entropy required for viscous flow behavior (*i.e.*, the glass transition temperature T_g) in polymers.

$$\eta = A \exp\left(\frac{B}{T-C}\right) = A \exp\left(\frac{-B}{T_g - T}\right) \tag{2}$$

Similarly, the incorporation of the T_g into the temperature-dependent viscosity is also present in the WLF equation, which relates viscosity at the glass transition temperature (η_{T_g}) to any temperature (T) with a shift factor (a_T). The constants (C_1, C_2) are specific material constants usually found experimentally. The relationship is outlined in eqn (3).

$$\log a_T = \log \eta(T) - \log \eta_{T_g}, \log a_T = \frac{-C_1(T - T_g)}{C_2 + (T - T_g)} \tag{3}$$

A further extension of both the Andrade and WLF equations for polymer solution viscosity is the Mark–Houwink (MH) equation (eqn (4)), which uses the molecular weight of the polymer (M) and fitted parameters (K, α) defined by the chemical environment. When tabulated, the two parameters are often accompanied by the temperature at which the experiment was conducted and thus provide an approximate temperature range for the accuracy of using the values to predict viscosity.

$$\eta = KM^{\alpha} \tag{4}$$

The empirical and analytical models used to describe PUR viscosity commonly rely on descriptors such as free volume, molecular configuration, and molecular size. However, the vast chemical diversity of PURs, encompassing numerous polyol backbones (*e.g.*, polyethers, polyesters, and polycarbonates) and a wide range of diisocyanate structures, introduces substantial complexity into viscosity modeling. Even among industrially standard formulations, the number of variables influencing viscosity is considerable, making empirical curve-fitting for each new formulation inefficient.

To address this challenge, we explored two distinct but complementary modeling strategies. The first approach leverages composition vectors to represent each unique sample: a chemical or "black-box" approach more traditional to formulation optimization. These models require minimal domain knowledge and provide high accuracy within the compositional space on which they are trained. However, they lack physical interpretability and cannot extrapolate to different monomers. In contrast, the second approach aims to capture the underlying physicochemical drivers of viscosity by explicitly modeling the temperature-dependent response (*e.g.*, *via* Andrade equations) and relating the fitted parameters to molecular-scale features. Here, variables derived from thermodynamics, electronic structure, and molecular geometry—such as dipole moment, polarizability, and topological surface area—are used to construct interpretable, mechanistically grounded models. This approach is more time-intensive and requires

significant feature engineering, but it offers a clearer link between chemical structure and macroscopic behavior, enabling greater generalizability to untested formulations. Together, these approaches illustrate a strategic tradeoff between predictive convenience and mechanistic insight. The chemical model is well-suited for rapid screening and optimization within known design spaces, while the physicochemical model provides a pathway for materials discovery and formulation guidance when experimental data are sparse or unavailable. Bridging these perspectives offers a robust framework for both empirical deployment and scientific understanding.

The goal of the training library design was to maximize the range of polarity of the polyol backbone, the variety of chemical structures of the diisocyanate, and the range of stoichiometries used to prepare the prepolymer leading to a distribution of different prepolymer chain lengths, all while staying within the compositional bounds of industrial applications. One of each common polyol chemistries (ether, ester, carbonate) was incorporated into the library design. Secondly, the diisocyanate chemical structure was varied to incorporate differing thermodynamic incompatibilities throughout the dataset as well as the reactivity of functional groups. Lastly, for a fixed chemical structure combination, a series of different %NCO by weight formulations were made to create a series of viscosities within a suitable range. The final library had a total of 39 unique prepolymers. Table 1 contains a list of each prepolymer formulated.

Samples were made using the first of the 2-step synthesis procedure for PUR polymerization. The macro polyols were heated in an 110 °C oil bath for 10 minutes before the stoichiometrically predetermined amount of diisocyanate was introduced to the mixture *via* syringe. The prepolymer reaction continued for 2 hours under the same temperature conditions and constant mechanical stirring at 200 RPM. After the allotted reaction time, each prepolymer liquid was poured into a polypropylene centrifuge tube, capped, and placed on the bench top.

After one week of ambient storage, each sample's temperature-dependent shear viscosity was measured with a DHR Rheometer. Each sample was heated to 40 °C and poured onto a Peltier plate which was subsequently heated to 80 °C with a (2 °C) per min temperature ramp. While the sample was heated, a 40 mm diameter 1° cone geometry sheared the sample at a constant 1 s^{-1} shear rate with viscosity and temperature data points acquired every 10 seconds. This low shear rate was deliberately chosen to isolate effects driven by chemical structure and thermophysical interactions, minimizing the confounding effects of shear-thinning behavior or entanglement typical of high molecular weight polymers. Each unique formulation was synthesized and tested once to generate a single viscosity–temperature curve, reflecting the common constraint of limited material availability in industrial screening and formulation environments.

Table 1 Sample library organized by polyol composition, diisocyanate composition, and percent excess NCO. The asterisk (*) denotes samples withheld for physicochemical model validation

Sample IDs	Polyol	Diisocyanate	%NCO(s)
P_44M_4 (6, 8, 10)	PTMEG	4,4-MDI	4, 6, 8, 10
P_MLQ_4 (6, 8, 10)	PTMEG	Mondur MLQ	4, 6, 8, 10
D_44M_4 (6, 8, 10)	Desmophen 2000	4,4-MDI	4, 6, 8, 10
D_MLQ_4 (6, 8, 10)	Desmophen 2000	Mondur MLQ	4, 6, 8, 10
C_44M_6 (8, 10)	Desmophen-C 2202	4,4-MDI	6, 8, 10
C_MLQ_4 (6, 8, 10)	Desmophen-C 2202	Mondur MLQ	4, 6, 8, 10
P_TD80_4 (7, 10)	PTMEG	TD80	4, 7, 10
D_TD80_7 (10)	Desmophen 2000	TD80	7, 10
P_TDS_4 (7, 10)	PTMEG	TDS	4, 7, 10
P_I_5 (9)	PTMEG	Desmodur I	5*, 9*
P_W_5 (9)	PTMEG	Desmodur W	5*, 9*
D_I_5 (9)	Desmophen 2000	Desmodur I	5*, 9*
D_W_5 (9)	Desmophen 2000	Desmodur W	5*, 9*

2.2 Modeling approaches, feature engineering, and selection

Machine learning (ML) has now become a ubiquitous tool for modeling response surfaces when a host of candidate variables are hypothesized to relate to the response. ML models can be adapted with feature-selection algorithms, or more generally regularization, that help identify a sparse set of independent variables to use in the parametric or nonparametric fitting of the response. For the chemical model, the dataset was expanded from 39 to 117 data points by extracting each formulation's viscosity value in Pa s from 40–80 °C in intervals of 10 °C and modeled with Random Forest Regression (RF) and Gaussian Process Regression (GPR). The model was trained to predict the viscosity of a formulation provided its temperature, %NCO, and a composition vector containing the respective amounts of polyol and diisocyanate reagents used during synthesis. RF models were fit using a grid search which spanned the number of estimators, the max depth of the decision tree(s), and the max features used in ensembles. The GPR covariance matrix was fit with a radial basis function (RBF) kernel to embed similarity information into the prior distribution of the dataset. RF and GPR were chosen as the modeling approaches because they represent a practical middle ground in model complexity; more expressive than linear regression but more interpretable and easier to tune than deep learning approaches. GPR's probabilistic nature offers meaningful uncertainty estimates, and its mathematical transparency makes it particularly suitable for materials scientists with limited machine learning experience. All models presented were fit validated using K-fold ($K = 5$) cross-validation and performance was evaluated using either a 20% testing set (chemical model) or a validation set (physicochemical model). All model performance metrics, namely the parity coefficient of determination r^2 and the root mean squared error (RMSE), reported and represented within figures are evaluated using the best estimator from the cross-validation procedure. Throughout the analysis, the GPR models consistently yielded better testing/validation set metrics and less evidence of overfitting, justifying its use as the primary model in this study. Full tables of modeling results are available in the SI material to this text.

Subsequently, each formulation's raw viscosity-temperature data was fit using the Andrade expression (eqn (2)), in which

A represents the viscosity of the material as $T \rightarrow \infty$, B represents the activation threshold for the material's viscous behavior, and C represents the reference temperature for viscous behavior or significant free volume change in the polymer material (T_g). The B-parameter for each unique sample was found by minimizing the chi-squared statistic between the raw data collected from the rheometer and an Andrade expression when C was set equal to each sample's base polyol T_g *i.e.*, polytetrahydrofuran (PTMEG) = −80 °C, poly(ethylene adipate) diol (PEAD, commercial name: Desmophen 2000) = −73 °C, poly(-hexamethylene carbonate) diol (PHMCD, commercial name: Desmophen-C 2202) = −64 °C. The A parameter was found to be randomly distributed values about a 0.001 Pa s mean and was not further considered during the modeling. Each error minimization resulted in an $r^2 \approx 0.99$ fit or better for the Andrade expression fittings. The collection of B-parameter values was therefore used to build GPR models trained to predict this viscosity activation energy parameter provided a physicochemical feature space modeled from the known chemical structure and stoichiometry of the prepolymer sample. To model the physicochemical feature space, stoichiometry and relative reactivities of the two isocyanate functional groups on the diisocyanate (provided by Covestro LLC) were used to convert percent weight NCO into the average degree of polymerization and average prepolymer molecular weight features. Density Functional Theory (DFT) calculations and cheminformatics computations were used to generate electronic, size, and shape descriptors of the prepolymers.

As discussed, molecular weight is the hallmark descriptor for viscosity in macromolecular fluids and with any condensation polymerization, the degree of polymerization given the stoichiometry of the motifs can be approximated with eqn (5) using equivalent weight ratios, assuming complete conversion. The average degree of polymerization of the prepolymers can also be used to approximate the molecular weight of the average prepolymer (PPMW) chain with eqn (6). Either of these transformations could be suitable proxies for chain length.

$$\mathrm{DP} = \frac{1 + \mathrm{eq_{OH}}/\mathrm{eq_{NCO}}}{1 - \mathrm{eq_{OH}}/\mathrm{eq_{NCO}}} \quad (5)$$

$$\mathrm{PPMW} = \mathrm{DP}(M_{\mathrm{polyol}} + M_{\mathrm{diisocyanate}}) \quad (6)$$

Quantum chemical calculations of ground state electronic parameters of the repeat units of synthesized polymers have been a popular methodology for describing the electronic properties of chain segments.[35] Motivated by the substantial literature on the roles of phase segregation and hydrogen bonding given the polarity of polyol backbone and hard segment structure, each motif's dipole moment (μ), polarizability (α), hyperpolarizability (β), and electronic energy (EE) were targets of DFT calculations. The repeat unit structures were drawn directly into the AMPAC 10.1 software, and the energy calculations were run with the Gaussian 16W engine. The B3LYP/6–311++G(2d,p) basis set was used for all calculations, performed in the gas phase with no implicit solvent model, using default tight SCF convergence criteria. No empirical dispersion correction was applied. All DFT calculations were executed using Gaussian 16 (Rev. C.01), with typical CPU time per optimization under 10 minutes on a 6-core processor.[36] Electronic, shape and size cheminformatics descriptors were also calculated with the RDKit package.[37] Both DFT and cheminformatics-based calculations were performed on either the polyol repeat units or the urethane-terminated prepolymer structures (Fig. 2). The calculated features were extracted from their respective output files and tabulated alongside the polyol T_g, DP, and PPMW in the candidate feature space (Table 2). Lastly, the reactivities of the first and second isocyanate groups were provided by collaborators at Covestro LLC.

Lastly, to reduce the dimensionality of the models built from the physicochemical feature space (Table 2) and improve model interpretability, we removed co-linear features and only kept the top ranking of the remaining after a SHapley Additive Explanations (SHAP) analysis. The co-linear removal consisted of taking every pair of features with > 0.8 Pearson scores and discarding the one with the lower co-linearity to the response variable. The SHAP analysis was subsequently performed on the remaining dataset to extract feature importance within a fit model. SHAP values provide a model-agnostic method for quantifying the marginal contribution of each feature to the prediction of the output parameter by computing the average effect of including each feature across all permutations of the feature set.[38] This analysis allowed us to rank features by their global importance and identify the subset most strongly influencing model predictions. Using this approach, we iteratively down-selected the original high-dimensional feature library down to just the top three dominant physicochemical descriptors. These selected features captured the essential structure–property relationships governing viscous flow activation in the prepolymers and were sufficient to retain high model performance while enabling interpretation grounded in molecular behavior. For completeness, SHAP analysis was also performed on the chemical model to evaluate the contribution of temperature and formulation-specific components (*e.g.*, specific polyol identity) to the prediction task. While these results are inherently less interpretable due to categorical input encoding, they are provided in the SI.

3 Results

The chemical model was trained using RF and GPR on an expanded dataset of 117 data points, each representing the viscosity of a formulation at one of five temperatures ranging from 40 °C to 80 °C. Each input vector encoded the formulation's %NCO value, temperature, and a chemically defined composition vector corresponding to the quantities of polyol and diisocyanate reagents. The best GPR estimator achieved an r^2 test value of 0.71 and a RMSE of 17.27 Pa s, Fig. 3 (RF r^2 = 0.57, RMSE = 20.82 Pa s. The model consistently tracked the temperature-dependent viscosity across the composition space, with poor model performances observed in isolated high/low viscosity regimes ($r^2 = -0.61$ η > 25 Pa s, $r^2 = 0.68$ η < 25 Pa s) and isolated high/low temperature regimes ($r^2 = 0.31$ T > 60 °

Table 2 Candidate feature space utilized during GPR modeling. All DFT and cheminformatics (ChemI) features were calculated for both polyol and diisocyanate repeat units

Feature	Description	Source
T_g	Glass transition temperature of the base polyol	Data sheet
DP	Degree of polymerization of the prepolymer	Stoichiometry
PPMW	Molecular weight of the prepolymer	Stoichiometry
μ	Dipole moment	DFT
α	Polarizability	DFT
β	Hyperpolarizability	DFT
EE	Electronic energy	DFT
TPSA	Topological polar surface area	ChemI
ISF	Inertial shape factor	ChemI
R_G	Radius of gyration	ChemI
M_V	Molar volume	ChemI
Log P	Partition coefficient	ChemI
k_1	Reactivity of the first isocyanate group	Covestro LLC
k_2	Reactivity of the second isocyanate group	Covestro LLC

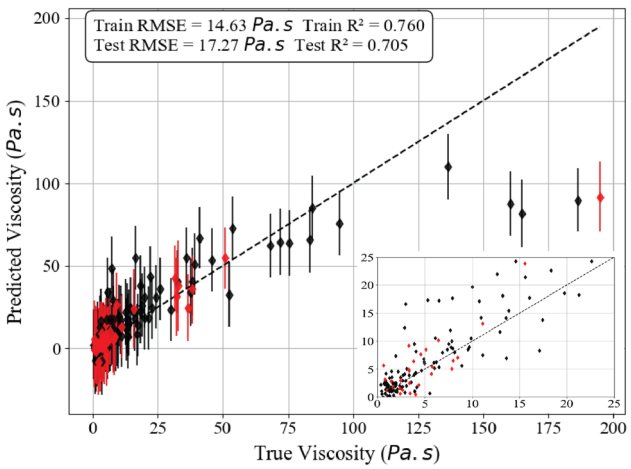

Fig. 3 Predicted *versus* measured viscosity values for the GPR chemical model using 5-fold cross-validation. Each point represents a unique formulation–temperature combination across a 39-sample library expanded to 117 data points. The input features to the model include %NCO, temperature, and a composition vector encoding the identity and composition of polyol and diisocyanate reagents. Black and red (training and testing data, respectively) vertical bars indicate one standard deviation of the predicted distribution for each point, reflecting the Bayesian uncertainty provided by the GPR model. The inset highlights model resolution and uncertainty behavior in the low-viscosity regime (≤25 Pa s), which is particularly relevant to processing conditions.

C, $r^2 = 0.29T < 60$ °C). The figure inset highlights the resolution of the model at lower viscosities when training data is sampled from all available data, which are most relevant to processing conditions and where the majority of the data resides. Dimensionality reduction was also used to obtain results for polyol-only and diisocyanate-only chemical spaces, yielding poor performances. Full results of the sparse chemical models are in the SI materials in Table A1.

In a complementary approach, we introduced the physicochemical feature space into the model framework to attempt to capture mechanistic structure–property relationships with the $\eta(T)$ response. Following feature selection (Fig. A1) and 5-fold cross-validation, this model achieved a moderate testing performance $r_{test}^2 = 0.73$, RMSE = 16.58 Pa s (RF $r_{test}^2 = 0.60$, RMSE = 20.12 Pa s) utilizing temperature, PPI, PolyLogP, k_2, and IsoDM suggesting viscosity behavior can be captured with the right combination of temperature, size, polyol and isocyanate physicochemical features. However, it generalized poorly to withheld validation sets (H12MDI and IPDI), yielding an r_{val}^2 of −0.60, RMSE = 9.04 Pa s (RF r_{val}^2 of −0.33, RMSE = 8.22 Pa s). This discrepancy reflects a key limitation of models trained directly on discrete viscosity data: they risk overfitting and fail to capture the latent physical and chemical drivers of viscous flow. Motivated by this, we next turn to modeling a material's activation energy which provides a more generalizable and interpretable framework for predicting viscosity across diverse chemistries.

To develop a more generalizable model, the raw viscosity–temperature profiles of each of the 39 unique prepolymers were fit to a general exponential express of Andrade form. The B parameter, from the least squares fit of raw viscosity-temperature data, corresponding to the activation threshold for viscous flow, was used as the regression target. The models were trained to predict this B value using a sparse set of three physicochemical features identified *via* SHAP analysis (Fig. 4). The beeswarm–style plot indicates the importance of the average prepolymer molecular weight PPMW, the topological polar surface area of the polyol repeat unit Polyol TPSA, and the inertial shape factor of the isocyanate motif Iso ISF. At this time during modeling all other features were disregarded due to sufficient correlation with other features or because their presence did not increase model performance.

Performance of the physicochemical model is summarized in Fig. 5. On the training set, composed of prepolymers excluding those based on IPDI and H12MDI, the model achieved an $r_{train}^2 = 0.96$ RMSE = 24.76 °C (RF $r_{train}^2 = 0.98$ RMSE = 15.71 °C). On the validation set, which included previously unobserved IPDI- and H12MDI-based formulations, the model retained strong performance $r_{val}^2 = 0.86$ RMSE = 65.89 °C (RF $r_{val}^2 = 0.78$ RMSE = 83.53 °C), indicating successful

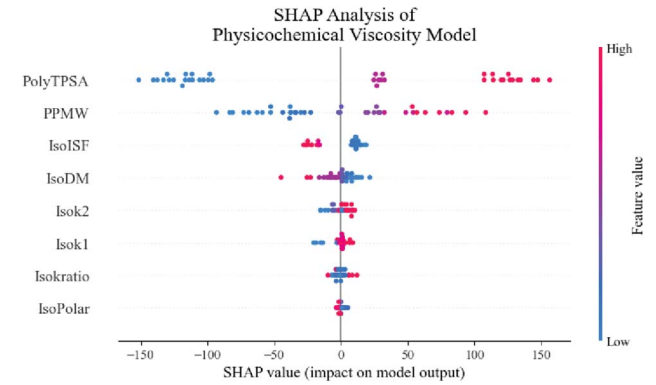

Fig. 4 SHAP beeswarm plot showing the impact of physicochemical features on the viscosity model output. Each point represents a unique data point and prepolymer formulation; color indicates the feature value from low (blue) to high (red). Features are sorted by overall importance (top to bottom). Positive SHAP values indicate a feature pushes predictions higher, while negative values indicate a downward effect.

extrapolation beyond the training distribution. The model's sparse physical feature set provides interpretable insight into the molecular mechanisms underlying viscosity activation. The dominant role of PPMW is consistent with empirical relationships such as the Mark-Houwink-Sakurada equation, which links molecular weight to viscosity through entropic constraints on chain mobility. Polyol TPSA captures the contribution of polar, hydrogen-bonding surface area to interchain interactions, effectively differentiating polyether, polyester, and polycarbonate polyols (further demonstrated by the clustering in this feature in Fig. 4). Iso ISF reflects differences in diisocyanate geometry (between aromatic and aliphatic diisocyanate structures) and polarity that may influence segmental rigidity and inter-chain orientation.

4 Discussion

This study demonstrates a dual modeling framework for predicting the temperature-dependent viscosity of polyurethane (PUR) prepolymers, balancing empirical accuracy with physicochemical interpretability. Chemistry-based models trained directly on compositional vectors and temperature offer high predictive power within known design spaces. These models are ideal for formulation optimization, where the goal is to interpolate within a well-characterized dataset demonstrated by high training and testing scores in the chemical model. However, physicochemical models abstract away from categorical identifiers and instead learn from molecular-scale descriptors, enabling predictions grounded in chemical theory and mechanistic understanding. This duality reflects a broader tension in materials informatics: should models prioritize practical accuracy or generalizability?

The chemical model, trained using GPR and RF on an expanded dataset of 117 data points, achieved a high degree of accuracy with no feature engineering and minimal preprocessing of the dataset. Its success highlights the value of structured composition vectors and serves as a benchmark for rapid screening applications. However, such models are fundamentally limited in their ability to extrapolate to new chemistries or uncover mechanistic trends (Fig. 1). To address these limitations, we transformed the modeling task using the Andrade equation, reducing each prepolymer's viscosity–temperature profile to a single activation parameter B, which encodes the energy barrier for flow. This scalar response variable, grounded in physical chemistry, was then modeled using a sparse set of three physicochemical descriptors selected using SHAP-based feature importance ranking from an initial high-dimensional library that included DFT- and cheminformatics-derived descriptors.

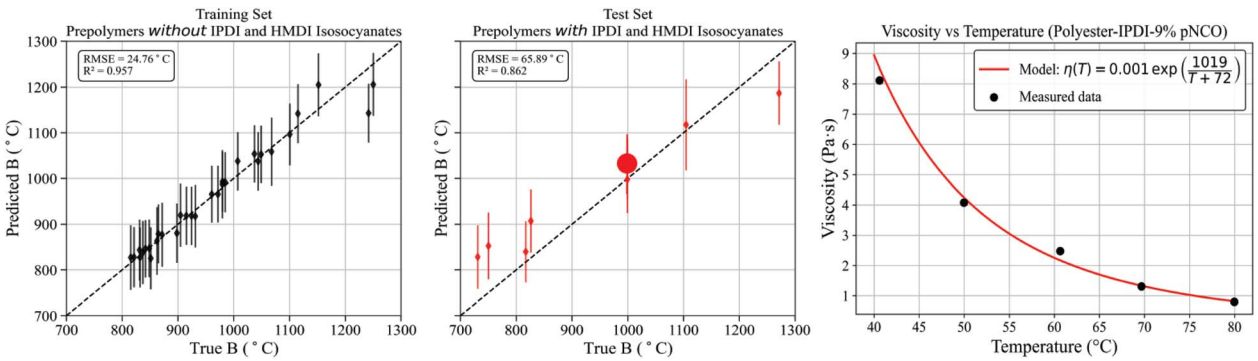

Fig. 5 Parity plots for GPR-predicted B values from the Andrade model using the sparse physicochemical feature set. (Left) Training set performance on samples excluding IPDI- and H12MDI-based diisocyanates. (Middle) validation set predictions for previously unseen IPDI and H12MDI prepolymers. Each B value corresponds to the activation energy extracted from fitting the viscosity–temperature profile of a unique prepolymer using the Andrade equation. The GPR model was trained on a three-dimensional feature space: the prepolymer molecular weight PPMW, the topological polar surface area of the polyol Polyol TPSA, and the inertial shape factor of the isocyanate Iso ISF. Error bars indicate the predictive uncertainty estimated by the GPR posterior. The high parity across both sets highlights the model's ability to interpolate and extrapolate from physicochemical structure to macroscopic viscous activation behavior. The final panel (right) shows the model's continuous prediction output for $\eta(T) = f(B)$ and the measured viscosity data for an example formulation from the validation set; Polyester Polyol-IPDI-pNCO 9%. The same formulation can be seen as a red scatter point in the validation parity plot in the middle panel.

Each final model feature carries physical significance. PPMW captures chain size and entropic resistance to flow and aligns with classical scaling laws such as the Mark–Houwink–Sakurada equation, where polymer viscosity scales sublinearly with molecular weight ($0.5 \leq \alpha \leq 0.8$). Polyol TPSA, a geometric proxy for polar surface area, captures hydrogen bonding potential and inter-chain interactions, distinguishing polyether, polyester, and polycarbonate architectures. Iso ISF, a shape-based descriptor, quantifies the anisotropy of the diisocyanate motif and reflects how steric factors and segmental symmetry influence packing, rigidity, and ultimately viscous resistance. Collectively, these three features span chain length, intermolecular cohesion, and molecular geometry—dimensions that directly influence viscous activation.

The resulting GPR model trained on this physicochemical feature space demonstrated strong generalizability, achieving $r^2 = 0.96$ on the training set and $r^2 = 0.86$ on a validation set comprising previously untested diisocyanate chemistries specifically H12MDI and IPDI. Its ability to interpolate across untrained regions of this defined chemical space suggests that the model captures the latent structure–property relationships governing viscous flow in aliphatic and cycloaliphatic diisocyanates. However, other chemical diversity and physicochemical gaps may still exist in the model *e.g.*, capturing the nature of highly polar or highly flexible polyol backbones in PDMS-based or natural oil-based polyols are likely to result in predictive failure. The SHAP-based feature selection further emphasizes this boundary: polarizability, hydrogen bonding capacity, and backbone rigidity emerged as dominant factors influencing viscosity, highlighting the need to sample more chemically diverse formulations. This limitation represents a clear opportunity for future dataset expansion aimed at improving model robustness across broader formulation spaces.

By comparing these two approaches—the empirical chemical model and the physicochemical model—we highlight the trade-offs surrounding generalizability. The chemical model is quick to deploy, requires no feature engineering, and performs well within the bounds of known compositions. In contrast, the physicochemical model requires domain knowledge, descriptor curation, and greater upfront effort, but yields models that not only generalize beyond the initial library but also offer mechanistic insight into how structure governs flow, all the while providing the investigator with a fitted parametric equation for the prepolymer viscosity–temperature dependence $\eta(T) = f(B)$.

5 Conclusions

This work illustrates how integrating empirical machine learning with physically informed modeling can offer a more holistic framework for materials prediction. Rather than positioning empirical and physics-based strategies as mutually exclusive, we demonstrate how they can compliment one another, serving distinct but complementary roles in the development and new formulation screen in complex datasets. By constraining statistical learning to physical regimes, and by re-framing noisy experimental data through canonical thermodynamic models, we show that even small, sparse datasets can be leveraged to uncover meaningful structure–property relationships that extend beyond the training set. This dual approach offers a scalable, interpretable pathway forward for predictive modeling in complex chemical systems. In the context of these broader developments, our results contribute a complementary perspective by explicitly uniting empirical and physics-informed strategies for prepolymer viscosity prediction. While prior work has successfully applied machine learning to related polyurethane systems, our dual-framework approach provides a pathway to generalize across diverse formulations while retaining interpretability grounded in rheological principles.

While promising, this study also carries important limitations. Each formulation was synthesized and tested once, resulting in a small sample size with pseudo-replication introduced by temperature increments. The viscosity measurements were performed at a constant low shear rate, restricting applicability to Newtonian flow regimes. Additionally, although our library included chemically diverse diisocyanates it did not include more complex polyols, chain extenders, or additives/fillers that are often encountered in industrial formulations. These constraints reflect the realities of early-stage formulation and screening, where material throughput is limited, but they also define clear directions for expanding and validating this dual-modeling framework in future work.

Author contributions

Conceptualization: J. P., K. H., N. W. Data curation: J. P., C. G., I. M. Formal analysis: J. P. Funding acquisition: N. W. Investigation: J. P., C. G. methodology: J. P., N. W. Project administration: N. W. Resources: N. W. supervision: N. W. Validation: J. P. Writing – original draft: J. P. Writing – review & editing: J. P., N. W.

Conflicts of interest

There are no conflicts to declare.

Data availability

Data for this article, including modeling code are available on GitHub at **https://github.com/joepugar/viscosity-modeling**.

Supplementary information: additional tables and figures. See DOI: **https://doi.org/10.1039/d5dd00287g**.

Acknowledgements

The authors thank Prof. Lynn Walker and Junghyun Ahn for access to and help with experimental equipment. J. A. P. and N. R. W. gratefully acknowledge support from a Covestro Science Award.

Notes and references

1 G. Oertel, Polyurethane Handbook: Chemistry, Raw Materials, Processing, Application, *Properties*, 2nd edn, Hanser, 1994.
2 M. F. Sonnenschein, Polyurethanes: Science, Technology, Markets, and Trends, 2nd edn, *Wiley Series on Polymer Engineering and Technology*, John Wiley & Sons, 2020.
3 M. E. Hyun and S. C. Kim, A Study on the Reactive Extrusion Process of Polyurethane, *Polym. Eng. Sci.*, 1988, **28**(11), 743–757.
4 S. Coudray, J. P. Pascault and M. Taha, Acrylated Polyurethanes by Reactive Extrusion, *Polym. Bull.*, 1994, **32**(5), 605–610.
5 Y. Liu, Z. Meng, Y. Wang, P. Li and Y. Sun, Analysis and Modeling of Viscosity for Aqueous Polyurethane Dispersion as a Function of Shear Rate, Temperature, and Solid Content, *ACS Omega*, 2020, **5**(40), 26237–26244.
6 N. D. Polychronopoulos and J. Vlachopoulos, Polymer Processing and Rheology, ed. Mondal M. I. H., in *Cellulose-Based Superabsorbent Hydrogels*, Springer International Publishing, 2018, pp. 1–47.
7 C. K. Chen, I. P. Liang, H. T. Chang, W. Y. Chen, I. M. Chen, M. H. Wu, *et al.*, Impact on Outcomes by Measuring Tortuosity with Reporting Standards for Thoracic Endovascular Aortic Repair, *J. Vasc. Surg.*, 2014, **60**(4), 937–944.
8 M. A. Bohn, The Connection Between the Parameters of WLF Equation and of Arrhenius Equation, *Propellants, Explos., Pyrotech.*, 2019, **44**(6), 696–705.
9 Y. Camberlin and J. P. Pascault, Phase Segregation Kinetics in Segmented Linear Polyurethanes: Relations Between Equilibrium Time and Chain Mobility and Between Equilibrium Degree of Segregation and Interaction Parameter, *J. Polym. Sci., Polym. Phys. Ed.*, 1984, **22**(10), 1835–1844.
10 J. Frenkel, The Viscosity of Liquids, *Nature*, 1930, **125**, 581–582.
11 M. L. Williams, R. F. Landel and J. D. Ferry, The Temperature Dependence of Relaxation Mechanisms in Amorphous Polymers and Other Glass-forming Liquids, *J. Am. Chem. Soc.*, 1955, **77**(14), 3701–3707.
12 E. Yildirim and M. Yurtsever, The Role of Diisocyanate and Soft Segment on the Intersegmental Interactions in Urethane and Urea Based Segmented Copolymers: A DFT Study, *J. Mol. Struct.*, 2013, **1035**, 28–38.
13 E. Yildirim, M. Yurtsever, G. L. Wilkes and I. Yilgör, Effect of Intersegmental Interactions on the Morphology of Segmented Polyurethanes with Mixed Soft Segments: A Coarse Grained Simulation Study, *J. Mol. Model.*, 2023, **29**(12), 204.
14 Y. He, D. Qiu and Z. Yu, Multiscale Investigation on Molecular Structure and Mechanical Properties of Thermal-treated Rigid Polyurethane Foam under High Temperature, *J. Appl. Polym. Sci.*, 2021, **138**(44), e51302.
15 Y. Li, Z. Ren, M. Zhao, H. Yang and B. Chu, Multiphase Structure of Segmented Polyurethanes: Effects of Hard-Segment Flexibility, *Macromolecules*, 1993, **26**(4), 612–622.
16 C. M. Childs and N. R. Washburn, Embedding Domain Knowledge for Machine Learning of Complex Material Systems, *Matter*, 2023, **9**(3), 806–820.
17 Y. Dong, Y. Li, F. He, J. Deng and Y. Zhang, Applications of Machine Learning in Polymeric Materials: A Review, *Polymers*, 2021, **13**(11), 1768.
18 P. Lu, S. Huang, C. Zhou, Z. Xu and Y. Wu, Preparation Process and Performance of Polyurethane Modified Bitumen Investigated Using Machine Learning Algorithm, *Constr. Build. Mater.*, 2022, **356**, 129231.
19 J. A. Cruz, S. C. Amico and O. Bianchi, Effect of the Aramid Pulp on the Physicochemical, Viscoelastic Properties and Rheokinetics of Polyurethanes, *J. Polym. Res.*, 2023, **30**(12), 410.
20 J. Ad Cruz, Jr H. L. Ornaghi, S. C. Amico and O. Bianchi, Predicting Viscosity in Polyurethane Polymerization for Liquid Composite Molding Using Neural Networks and Surface Methodology, *Polymers*, 2024, **16**(15), 2283.
21 G. Pilania, C. Wang, X. Jiang, S. Rajasekaran and R. Ramprasad, Accelerating Materials Property Predictions Using Machine Learning, *Sci. Rep.*, 2013, **3**(1), 2810.
22 A. Jha, A. Chandrasekaran, C. Kim and R. Ramprasad, Impact of Dataset Uncertainties on Machine Learning Model Predictions: The Example of Polymer Glass Transition Temperatures, *Model. Simul. Mater. Sci. Eng.*, 2019, **27**(2), 024002.
23 Y. Berezkin and M. Urick, Modern Polyurethanes: Overview of Structure–Property Relationship, in *Polymers for Personal Care and Cosmetics*, ed. Schlossman M., vol. 1148, ACS Symposium Series. American Chemical Society, 2013. pp. 65–81.
24 *Szycher's Handbook of Polyurethanes*, ed. Szycher M., 2nd edn, Boca Raton, CRC Press, 2012.
25 T. J. Dearlove and G. A. Campbell, Synthesis and Characterization of Isocyanate-Terminated Polyurethane Prepolymers, *J. Appl. Polym. Sci.*, 1977, **21**(6), 1499–1509.
26 A. M. Heintz, D. J. Duffy, S. L. Hsu, W. Suen, W. Chu and C. W. Paul, Effects of Reaction Temperature on the Formation of Polyurethane Prepolymer Structures, *J. Appl. Polym. Sci.*, 1990, **36**(8), 2695–2704.
27 S. Clauß, D. J. Dijkstra, J. Gabriel, O. Kläusler, M. Matner, W. Meckel, *et al.*, Influence of the Chemical Structure of PUR Prepolymers on Thermal Stability, *Int. J. Adhes. Adhes.*, 2011, **31**(6), 513–523.
28 J. O. Akindoyo, M. D. H. Beg, S. Ghazali, M. R. Islam, N. Jeyaratnam and A. R. Yuvaraj, Polyurethane Types, Synthesis and Applications – A Review, *RSC Adv.*, 2016, **6**(115), 114453–114482.
29 P. Król, Synthesis Methods, Chemical Structures and Phase Structures of Linear Polyurethanes. Properties and Applications of Linear Polyurethanes in Polyurethane Elastomers, Copolymers and Ionomers, *Prog. Mater. Sci.*, 2007, **52**(6), 915–1015.

30 T. K. Kwei, Phase Separation in Segmented Polyurethanes, *J. Appl. Polym. Sci.*, 1982, **27**(8), 2891–2899.
31 Z. S. Petrovíc and I. Javni, The Effect of Soft-Segment Length and Concentration on Phase Separation in Segmented Polyurethanes, *J. Polym. Sci., Part B: Polym. Phys.*, 1989, **27**(3), 545–560.
32 Y. Li, W. Kang, J. O. Stoffer and B. Chu, Effect of Hard-Segment Flexibility on Phase Separation of Segmented Polyurethanes, *Macromolecules*, 1994, **27**(2), 612–614.
33 U. Šebenik and M. Krajnc, Influence of the Soft Segment Length and Content on the Synthesis and Properties of Isocyanate Terminated Urethane Prepolymers, *Int. J. Adhes. Adhes.*, 2007, **27**(7), 527–535.
34 E. Głowińska and J. Datta, A Mathematical Model of Rheological Behavior of Novel Bio-based Isocyanate-Terminated Polyurethane Prepolymers, *Int. J. Adhes. Adhes.*, 2015, **60**, 123–129.
35 Y. Zhao, R. J. Mulder, S. Houshyar and T. C. Le, A Review on the Application of Molecular Descriptors and Machine Learning in Polymer Design, *Polym. Chem.*, 2023, **14**(29), 3325–3346.
36 P. Xu, T. Lu, L. Ju, L. Tian, M. Li and W. Lu, Machine Learning Aided Design of Polymer with Targeted Band Gap Based on DFT Computation, *J. Phys. Chem. B*, 2021, **125**(2), 601–611.
37 *RDKit: Open-source Cheminformatics*, **https://www.rdkit.org**.
38 S. M. Lundberg and S. I. Lee, A Unified Approach to Interpreting Model Predictions, in *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*. Curran Associates, Inc., 2017, pp. 4765–4774.