Quantitative structure-property relations for polyester materials via statistical learning

Stephen McCoy[1], Damilola Ojedeji[2], Brendan Abolins[3], Cameron Brown[3], Manolis Doxastakis[2], and Ioannis Sgouralis[1]

[1]Department of Mathematics, University of Tennessee Knoxville, Knoxville, TN
[2]Department of Chemical and Biomolecular Engineering, University of Tennessee Knoxville, Knoxville, TN
[3]Eastman Chemical Company, Kingsport, TN

Abstract

We employ statistical learning to present a principled framework for the establishment of quantitative structure-property relationships (QSPR). We focus on property predictions of industrial polymers formed by multiple reagents and at varying molecular weights. We develop a theoretical description of QSPR as well as a rigorous mathematical method for the assimilation of experimental data. Results show that our methods can perform exceptionally well at establishing QSPR for glass transition temperature and intrinsic viscosity of polyesters.

Keywords: QSPR, polymer, polyester, glass transition temperature, intrinsic viscosity, statistical learning

1 Introduction

Material design is an important aspect of modern Chemistry and Chemical Engineering, especially for industrial applications [1]. Material design is particularly challenging when it comes to the design of polymers [42, 2]. Due to the complexity of modern polymer materials, the variety of available resources, and a high sensitivity on the raw materials, the design pipeline of polymers often starts with computational studies that can narrow down a small set of candidate compositions which, after appropriate processing, may give rise to materials of the desired properties [29]. Due to time and cost constraints, such studies most often rely on quantitative structure-property relations (QSPR) that are either well-established or tentative [43].

By QSPR we mean a mathematical relationship between material properties of interest and a set of structural descriptors of the molecules of the target materials desired to have such properties [32, 13]. QSPR methods are flexible as to what descriptors and mathematical models are used, but in general, QSPR maintain the formulation of their relationships as a function that maps a set of molecular descriptors to one or more numerical properties of interest [10, 31].

To date, the process of establishing a QSPR is largely achieved empirically [38]. In particular, to establish a QSPR, first a suitable set of data relevant to the targeted properties and materials needs to be identified based on expert opinion and manual curation of historical measurements. Then, an appropriate set of mathematical relations that summarizes the data needs to be identified. Subsequently, values of various unknown parameters in these relations are adjusted during parameter fitting. Finally, the pipeline is terminated with a validation stage in which the developed QSPR is tested or cross-validated against real-life data outside these used in its development. In practice, several mathematical methods are employed in each of the stages and these involve a combination of heuristics derived from experience [14] ad hoc algorithms [37] or, more recently, Artificial Intelligence implemented via Machine Learning [16].

Like most data-driven methods [18], the performance and practicality of the resulting QSPR depends critically not only on the quality of the supplied data, but also on the mathematical methods employed in its establishment [26, 39]. For this reason, it is important that all stages of QSPR development are carried out within the same framework and under a single computational process. For this, a unifying framework of data representation, structure-property relation, and parameter estimation via assimilation of experimental measurements is required. Such a framework needs (i) to be flexible enough to support arbitrarily complex compositions (for instance materials created by more than a few reagents) and (ii) to allow for parameter estimation that is computationally efficient and robust even in cases of an excessively high number of involved parameters.

Here, we present such a unifying framework for QSPR. Our framework builds on chemical intuition to motivate mathematical forms for the QSPR functions and applies the principles of Statistical Learning [44, 30] to adjust the values of the unknown parameters in them. As a result, we present a comprehensive mathematical framework that can speed up QSPR development and increase its robustness. Built on statistical principles, our framework does not only provide property predictions but also estimates the uncertainty associated with them, making it particularly useful in cases of limited, corrupted, or highly complex data.

The remaining of this study is structured as following. In Methods we first set basic terminology, describe our data representation scheme, QSPR forms, and parameter estimation methods. In Results we demonstrate the use of our framework by key applications in polyester materials. Finally, in Discussion we present an overview of the advantages and disadvantages of our QSPR framework.

Acids	Reactants	Abbreviations	Types	Functionalities	Functional groups
	Cap acid	-	V	1	f_1
	Terephthalic acid	TPA	A	2	f_2, f_3
	Isophthalic acid	IPA	A	2	f_4, f_5
	1,4-Cyclohexanedicarboxylic acid	1-4-CHDA	A	2	f_6, f_7
	Adipic acid	ADP	A	2	f_8, f_9
	Sebacic acid	SA	A	2	f_{10}, f_{11}
	Hexahydrophthallic anhydride	HHPA	A	2	f_{12}, f_{13}
	Trimellitic anhydride	TMA	A	3	f_{14}, f_{15}, f_{16}
	1,12-Dodecanedioic acid	DDDA	A	2	f_{17}, f_{18}
	Azelaic acid	AZL	A	2	f_{19}, f_{20}
Alcohols	Reactants	Abbreviations	Types	Functionalities	Functional groups
	Cap alcohol	-	V	1	f^1
	1,6-Hexane diol	HDO	A	2	f^2, f^3
	Neopentyl glycol	NPG	A	2	f^4, f^5
	Trimethylolpropane	TMP	A	3	f^6, f^7, f^8
	1,4-Cyclohexanedimethanol	1,4-CHDM	A	2	f^9, f^{10}
	1,3-Cyclohexanedimethanol	1,3-CHDM	A	2	f^{11}, f^{12}
	2,2,4,4-tetramethyl-1,3-cyclobutanediol	TMCD	A	2	f^{13}, f^{14}
	2-Methyl-2,4 Pentanediol	MPD	A	2	f^{15}, f^{16}
	4,8-Bis(hydroxymethyl)tricyclo[5.2.1.02,6]decane	TCDDM	A	2	f^{17}, f^{18}

Table 1: List of the acid and alcohol reactants we consider in our polyesters. Types A and V distinguish our reactants into actual and virtual ones. The former are those used in the preparation of a polyester; while the latter are those introduced for computational purposes only.

2 Methods

2.1 QSPR schemes

Each polyester resin created has its own numerical properties, such as glass transition temperature or intrinsic viscosity, and its own composition of raw ingredients such as moles of reactants. Here, our goal is to develop a QSPR scheme that uses polyester composition to predict polyester resin properties.

2.1.1 Polyester ingredients

In this study, we consider QSPR developed from polyester data that consist of compositions and experimentally measured properties. For concreteness, we consider only polyesters formed by the acid and alcohol reactants shown in table 1. These are polyfunctional monomers that react to form acid-alcohol dimers which, in turn, are our polyester's repeat units. We classify our reactants as actual and virtual. The former are those used in a polyester's preparation; while, the latter are those needed in our framework only for computational reasons. In particular, we include acid and alcohol caps which are monofunctional virtual monomers that can simulate the formulation of finite molecules and so to allow for the study of polyesters at varying molecular weight including those at the low end [14].

To represent each functional group of our reactants, we use the scheme shown in table 1. This scheme uses f_i with *subscripts* $i = 1, \ldots, I$ to represents acid groups and f^j with *superscripts* $j = 1, \ldots, J$ to represent alcohol

groups. In this particular study $I = 20$ and $J = 18$. Similarly, we use $h^{1:I}$ and $h_{1:J}$ to denote the moles of each group present in a polyester. For the actual monomers, this is obtained by the moles of the corresponding reactant in a polyester *multiplied* by its functionality. To obtain the moles of each cap reactant, we assume that the moles of every acid and every alcohol group in a resin are equal, so a polyester's formation terminates when both its acid and alcohol free groups are exhausted [14]. Our assumption results in the formulas

$$h_1 = \max\left(0, \sum_{j\neq 1} h^j - \sum_{i\neq 1} h_i\right), \qquad h^1 = \max\left(0, \sum_{i\neq 1} h_i - \sum_{j\neq 1} h^j\right).$$

Further, we use f_i^j to denote a bond formed in a polyester by the reaction of groups f_i and f^j. Similarly, we use h_i^j to denote the bond's moles within a polyester. Assuming no direct bond formation between cap monomers, we readily obtain $h_1^1 = 0$. Further, we assume well-mixed conditions and so no preferential reactions between the groups forming each actual reactant. This results in moles of bonds obtained by

$$h_i^j = \frac{h_i h^j}{\sum_i h_i} = \frac{h_i h^j}{\sum_j h_j}.$$

Here, the denominators are equal to each other due to the inclusion of caps as described above.

Because our polyesters are formed by polycondensation reactions, we use the formula

$$m_i^j = m_i + m^j - m_{\text{condensate}}$$

to obtain a weight m_i^j characteristic of the bond f_i^j. For groups of actual reactants, m_i and m^j are the molecular weights of the monomers carrying the groups f_i and f^j *divided* by their functionalities. For groups of virtual reactants, these are by definition $m_1 = m^1 = m_{\text{condensate}}$ where $m_{\text{condensate}}$ is the weight of the molecule lost from the bond reaction. For our polyesters this is a single water molecule.

Next, using the bond weights m_i^j and a polyester's mole composition h_i^j, we calculate the weight fraction ω_i^j of our bond ingredients by

$$\omega_i^j = \frac{m_i^j h_i^j}{\sum_i \sum_j m_i^j h_i^j}.$$

Our QSPR scheme, detailed below, models the properties of a polyester material as a function of characteristic contributions by its ingredients. To represent contributions, we denote with p_i^j the contribution to a polyester material's property caused by the bond f_i^j. We build the contributions of each ingredient p_i^j using characteristic primitives q_i^j as follows

$$p_i^j = p_{\text{ref}} \cdot \exp\left(q_i^j\right). \tag{1}$$

Here, p_{ref} is a reference value that allows for matching of the units and the scaling factor $\exp\left(q_i^j\right)$ to adjust for ingredient-specific characteristics.

To encode structural descriptors in the selection of $p_{1:I}^{1:J}$ or their primitives $q_{1:I}^{1:J}$, we employ the simplified molecular-input line-entry system (SMILES) which is a standard way to turn a molecular representation into processable strings for chemoinformatics analysis [40]. Specifically, we denote with σ_i^j the SMILES string that is characteristic of the bond f_i^j. We obtain our SMILES σ_i^j using the string corresponding to the resulting dimer

created by the specified bond-forming reaction. We note that for an actual reactant monomer combined with a virtual monomer, such for in f_1^j or f_i^1, we ignore the virtual monomer and use simply the SMILES string corresponding to the actual monomer alone. Also we leave undefined any SMILES capturing virtual-to-virtual reactions, such as f_1^1, since, as we mention earlier, we exclude such ingredients from our compositions. To generate our strings, we use the cheminformatic methods in the software package RDKit [23].

2.1.2 Polyester properties

Mathematically, our QSPR is a function that maps ingredient contributions, encoded via $p_{1:I}^{1:J}$, and ingredient compositions, encoded via $\omega_{1:I}^{1:J}$, to resin properties, denoted with $P^{\text{predicted}}$. We represent this relationship with a function $P^{\text{predicted}} = G(\omega_{1:I}^{1:J}, p_{1:I}^{1:J})$. Such function corresponds to a mixing rule [39] which we model by a non-linear average

$$G\left(\omega_{1:I}^{1:J}, p_{1:I}^{1:J}\right) = g^{-1}\left(\sum_i \sum_j \omega_i^j g(p_i^j)\right) \tag{2}$$

induced by an invertible function $g(p)$. Here, $g^{-1}(p)$ denotes the inverse of $g(p)$ which satisfies $g^{-1}(g(p)) = p$. According to our formulation, the contribution p_i^j is the property of a polyester material exclusively consisting of the ingredient f_i^j.

2.2 Statistical learning

Given a mixing rule, induced by an appropriate function $g(p)$, our QSPR contains uncharacterized ingredient contributions $p_{1:I}^{1:J}$. Since, unlike $\omega_{1:I}^{1:J}$, these are not polyester specific, their values can be obtained via formal parameter estimation. From now on, we employ Statistical Learning [18, 30, 20] to describe a mathematical method that uses experimentally obtained measurements to derive values for $p_{1:I}^{1:J}$ or, equivalently, values for their primitives $q_{1:I}^{1:J}$.

For clarity, we consider the assimilation of experimentally acquired data and use subscripts $r = 1, \ldots, R$ to label the individual polyester materials in our dataset. Our dataset contains R polyester assessments and consists of property measurements, denoted with P_r^{measured}, and reactant moles, denoted with $(h_{1:I})_r$ and $(h^{1:J})_r$. To denote the *entire* collection of measured properties we use $P_{1:R}^{\text{measured}}$.

To process our data, we follow the Bayesian methodology [30, 36, 24]. This requires a likelihood that assumes values for the unknowns $q_{1:I}^{1:J}$ and links predicted with measured properties. In addition, it requires a prior on the unknowns $q_{1:I}^{1:J}$ that penalizes nonphysical values. Below, we detail our choices.

2.2.1 Likelihood

Our likelihood is a probability distribution that penalizes deviations of property predictions $P_r^{\text{predicted}}$ from property measurements P_r^{measured} based on assumed values of the primitives $q_{1:I}^{1:J}$. Because, in this study, we consider only properties such as glass transition temperature or intrinsic viscosity, which are naturally positive quantities, we employ a likelihood in the form

$$P_r^{\text{measured}} | \tau, q_{1:I}^{1:J} \sim \text{LogNormal}\left(P_r^{\text{predicted}}, \tau\right). \tag{3}$$

Here, $P_r^{\text{predicted}}$ is the property prediction resulting from the primitives $q_{1:I}^{1:J}$ via eq. (1) and the mixing rule in eq. (2). Our probability distributions, such as LogNormal, are parameterized as detailed in appendix B.

2.2.2 Priors

Our priors are probability distributions that penalize nonphysical values for the unknown parameters in the likelihood. For this purpose, on the likelihood's precision τ we use

$$\tau \sim \text{Gamma}\left(\phi, \tau_{\text{ref}}/\phi\right),$$

which excludes negative values. Our choice is conditionally conjugate to eq. (3), which allows for efficient computational implementation of our method as described below.

Finally, on the primitives $q_{1:I}^{1:J}$ we apply a joint multivariate normal prior [8]

$$q_{1:I}^{1:J} \sim \text{Normal}_I^J\left(0, S\right)$$

and use the covariance matrix S to incorporate structural information represented by the SMILES $\sigma_{1:I}^{1:J}$. Specifically, we use chemical similarity [4], encoded via the ingredients' SMILES, to setup a matrix S such that it balances an ingredient's contribution relative to another's. Our covariance matrix S is a square positive definite matrix with entries $S_{i,i'}^{j,j'}$ that are defined via the sum of a white noise and a square exponential kernel [6, 18, 30, 41]. This results to

$$S_{i,i'}^{j,j'} = \lambda^2 \left(a\, \delta_{i,i'}^{j,j'} + (1-a) \exp\left[-\frac{1}{2} \left(\frac{1 - \mathcal{T}\left(\mathcal{M}\left(\sigma_i^j\right), \mathcal{M}\left(\sigma_{i'}^{j'}\right)\right)}{\ell} \right)^2 \right] \right) \tag{4}$$

where a, ℓ, and λ are tunable parameters and $\delta_{i,i'}^{j,j'}$ is the Kronecker delta function. In turn, our covariance kernel is activated by the Tanimoto coefficient [3] and the Morgan fingerprint [7] denoted with $\mathcal{T}(\mu, \mu')$ and $\mathcal{M}(\sigma)$, respectively. We compute these via RDKit [23].

Our choice of the covariance in eq. (4) preserves the comparison coefficient of similarity or difference but can scale the relationship to reflect how the similarity in structures relates to the similarity in measurements [41, 11]. In essence, as we explain below, with our combination of Morgan fingerprint and Tanimoto coefficient, we have a comparison metric that quantifies the intensity of similarity or dissimilarity between two SMILES strings [3]. We use this in eq. (4) as a basis to make inferences on the correlation of the contributions associated with our QSPR ingredients. Put simply, our covariance S fits similar contributions p_i^j to ingredients f_i^j with similar SMILES σ_i^j and vice versa dissimilar contributions to ingredients with dissimilar SMILES.

Chemical information on the Morgan fingerprint $\mu = \mathcal{M}(\sigma)$ comes in the form of a bit vector $\mu = (\mu_1, \mu_2, \dots)$ with each bit μ_k indicating whether an attribute is present or not in that spot. To calculate the Morgan fingerprint we also need to specify the *radius* (number of atoms used to create a mapping) and the *nBits* (the number of unique combination pathways to represent features). Once a radius and a bit size are chosen, the Morgan fingerprint is created by (i) identifying each atom in the assembly, (ii) updating each atom's identifiers based on its neighbors (to the size of the radius), (iii) removing duplicates, and (iv) collapsing down into a single-bit vector.

With the Morgan fingerprint representation $\mathcal{M}(\sigma)$ of a SMILES σ in hand, we then define a metric for comparison of our SMILES. We opt for the conventional choice of the Tanimoto similarity coefficient [34]. This is defined by

$$\mathcal{T}(\mu, \mu') = \frac{\mu \cdot \mu'}{||\mu||^2 + ||\mu'||^2 - \mu \cdot \mu'}$$

for $\mu = \mathcal{M}(\sigma)$ and $\mu' = \mathcal{M}(\sigma')$, note that for self-comparison we get $\mathcal{T}(\mu, \mu) = 1$.

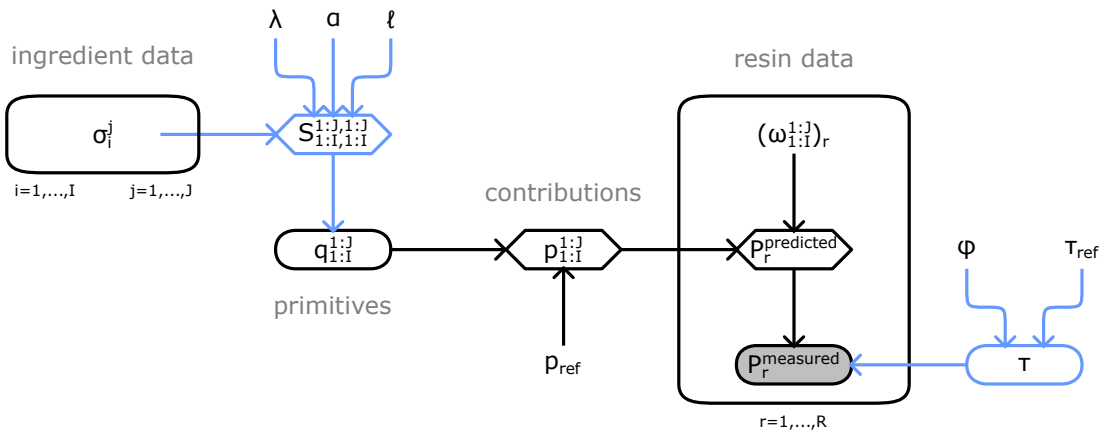


Figure 1: Schematic representation of our learning framework: Our QSPR modeling scheme includes ingredient primitives $q_{1:I}^{1:J}$, ingredient contributions $p_{1:I}^{1:J}$, weight fractions $(\omega_{1:I}^{1:J})_{1:R}$, and property measurements $P_{1:R}^{\text{measured}}$. The latter two are polyester specific; while, the former two are not. The SMILES of our ingredients $\sigma_{1:I}^{1:J}$ are used to generate a covariance matrix $S_{1:I,1:I}^{1:J,1:J}$, with hyperparameters λ, α, ℓ that controls the similarity between the fitted primitives $q_{1:I}^{1:J}$. In turn, fitted primitives $q_{1:I}^{1:J}$ determine the contributions $p_{1:I}^{1:J}$ and, via the mixing rule, eventually the predicted polyester properties $P_{1:R}^{\text{predicted}}$. Following the common convention: we denote random variables with circles, deterministic variables with diamonds, and we leave open hyperparameters with set values. Measurements are shaded and arrows indicate the dependencies between our variables. Finally, we highlight in color the dependencies mediating our priors.

2.3 Computational inference

Our statistical learning framework is summarized in fig. 1 and it is fully detailed in appendices A.1 and A.2. We use our scheme to conclude on unknown parameter values via its marginal posterior probability distribution $\mathcal{P}\left(q_{1:I}^{1:J}|P_{1:R}^{\text{measured}}\right)$. Using Bayes' theorem [36], the posterior is formally derived by

$$\mathcal{P}\left(q_{1:I}^{1:J}|P_{1:R}^{\text{measured}}\right) \propto \mathcal{P}\left(q_{1:I}^{1:J}\right) \int_{\tau} d\tau\, \mathcal{P}(\tau) \prod_{r} \mathcal{P}\left(P_{r}^{\text{measured}}|\tau, q_{1:I}^{1:J}\right), \tag{5}$$

where $\mathcal{P}(q_{1:I}^{1:J})$ and $\mathcal{P}(\tau)$ are our prior distributions and $\mathcal{P}\left(P_{r}^{\text{measured}}|\tau, q_{1:I}^{1:J}\right)$ our likelihood. Because of the complexity of eq. (5), we cannot obtain a closed form. Instead, we characterize our posterior using Markov chain Monte Carlo (MCMC) sampling by drawing samples of $q_{1:I}^{1:J}|P_{1:R}^{\text{measured}}$ [30, 35, 25]. Because our model attains the form of a latent Gaussian model, that is a multivariate Gaussian prior and a complicated likelihood function that ties the unknown variables to the observed data, we develop our MCMC scheme by means of an elliptic slice sampler [27].

2.4 Data acquisition

To generate the results shown later in this study, we relied on polyester data. Our data are acquired by polycondensation and characterization procedures as described next.

2.4.1 Reagents

Isophthalic acid (IPA), terephthalic acid (TPA), 1,4-cyclohexanedicarboxlyic acid (1,4-CHDA), 1,4- cyclohexane dimethanol (1,4-CHDM), 1,3-cyclohexane dimethanol (1,3-CHDM), 2,2,4,4- tetramethylcyclobutanediol (TMCD), hexahydrophthallic anhydride (HHPA), and neopentyl glycol (NPG) were provided by Eastman Chemical Company.

Adipic acid (AD), sebacic acid (SA), trimellitic anhydride (TMA), 1,12-dodecanedioic acid (DDDA), azelaic acid (AZL), 2-methyl-1,3-propanediol (MP Diol), trimethylolpropane (TMP), 1,6-hexane diol (HDO), and 4,8-bis(hydroxymethyl)tricyclo[5.2.1.0 ^{2,6}]decane (TCDDM) were purchased from Sigma Aldrich. Aromatic 150 was purchased from ExxonMobile and Fascat 4100 was purchased from PMC Organometallix Inc.

2.4.2 Polyester resin synthesis

The polyols were produced using either a resin kettle reactor setup via solvent-assisted polycondensation or a resin rig reactor setup via melt polycondensation, both of which were controlled with automated control software. The solvent-assisted resins were produced on a 3.5 mole scale using a 2 L kettle with overhead stirring and a partial condenser topped with total condenser and Dean Stark trap. Approximately 10 wt % (based on reaction yield) azeotroping solvent of high boiling point (A150 and A150ND) was used to both encourage egress of the water condensate out of the reaction mixture and keep the reaction mixture viscosity at a reasonable level using the standard paddle stirrer. Chemical reagents were added to the kettle, which was then completely assembled. The Fascat 4100 (monobutyltin oxide) catalyst was added via the sampling port after the reactor had been assembled and blanketed with nitrogen for the reaction. Additional A150/A150ND solvent was added to the Dean Stark trap to maintain the $\sim$10 wt % solvent level in the reaction kettle. The reaction mixture was heated without stirring from room temperature to 150°C using a set output controlled through the automation system. Once the reaction mixture was fluid enough, the stirring was started to encourage even heating of the mixture. At 150°C, the control of heating was switched to automated control and the temperature was ramped to 230°C over the course of 4 h. The reaction was held at 230°C for 1 h and then heated to 240°C over the course of 1 h. The reaction was then held at 240°C and sampled every 1-2 h upon clearing until the desired acid value was reached. The $\sim$90%-solids resins were ground into 6 mm pellets and thoroughly dried in a vacuum oven at 150°C for 24 h prior to characterization.

The melt polycondensation resins were produced on a 0.5 mole scale. A 500 mL, one-neck, round- bottom flask was carefully charged with all chemical reagents and Fascat 4100 (monobutyltin oxide) catalyst. The flask was equipped with a polymer head adapter with stainless steel mechanical stirrer and securely clamped to the polymerization rig. To the polymer head, a distillation side arm and Erlenmeyer flask were attached. The automated-controlled vacuum system was attached to the flask side arm to allow for a reduction in pressure of the reaction vessel. The Belmont metal bath was preheated to 20°C above the recipe starting temperature (180°C). The apparatus was subjected to two iterations of a nitrogen (N_2) purge to remove oxygen and then dunked into the metal bath to begin. The flask was held at 180°C for 10 minutes to melt the starting materials and then stirring was started to encourage even heating on the mixture. The flask was heated to 240°C over 4 hours and then held there for an additional hour. Pressure in the reaction flask was reduced to 1.5 torr over 45 minutes and then the reaction was held at 1.5 torr until the final acid value was reached. Dry ice was used to ensure that the solvent traps were sufficiently cold to prevent any solvent/organic matter from going to the vacuum pump. After completion, the flask was slowly brought back to atmospheric pressure and removed from the hot metal bath. Upon solidification, the polymer was pulled from the round bottom flask by partially melting the edges, then the glass flask was broken with a hammer to give the solid polymer "lollipop" on the stir rod. The polymer was cooled in dry-ice, removed from the stir-rod, and ground into 6 mm pellets prior to characterization.

2.4.3 Polyester resin characterization

The Inherent Viscosity ($[\eta]$) of all polymers was determined in 0.5 wt% PM 95 (60/40 phenol/1,1,2,2- tetrachloroethane) solution at 25°C. Molecular weights were determined by gel permeation chromatography (GPC) with 95/5 methylene chloride/HFIP mobile phase and calibration curves for polystyrene standards. Monomer composition was determined via gas chromatography (GC) hydrolysis. The glass transition temperature (T_g) was determined using differential scanning calorimeter (DSC) at 20°C / min. ramp rate with a N_2 sweep. The T_g was based on second heat thermograms.

3 Results

In this section, we demonstrate how our statistical learning framework applies in the establishment of QSPR schemes for polyester resins. Our raw data assess polyesters created by 9 acid and 8 alcohol reactants with different functionalities resulting in $I = 20$ and $J = 18$ different functional groups, forming 360 different bonds, as shown in table 1. The calculated Tanimoto coefficients and associated prior covariance matrix S is detailed in appendix D and the sensitivity of our method to the hyperparameters is detailed in appendix E.

For each polyester resin, our data contain reactant mole amounts $h_{1:I}, h^{1:J}$ and measurements of glass transition temperature T_g, intrinsic viscosity $[\eta]$, and number average molecular weight M_n. From these, T_g^{measured} and $[\eta]^{\text{measured}}$ are our targeted properties. Overall, we use $R = 137$ resin data points with T_g^{measured} measurements, $R = 79$ resin data points with $[\eta]^{\text{measured}}$ measurements, and $R = 79$ resin data points with both T_g^{measured} and $[\eta]^{\text{measured}}$ measurements.

For all cases shown below, we obtain the maximum a posteriori estimator $\hat{q}_{1:I}^{1:J}$ of our primitives by selecting the MCMC sample with the highest $\mathcal{P}(q_{1:I}^{1:J}|P_{1:R}^{\text{measured}})$ computed via eq. (5) [30, 6, 36, 18]. This estimator represents the single best choice of parameter values generated by our learning framework. Using this estimator, we obtain our contributions $\hat{p}_{1:I}^{1:J}$ via eq. (1) and our QSPR predictions $\hat{P}_r^{\text{predicted}}$ for the r^{th} resin using our mixing rule

$$\hat{P}_r^{\text{predicted}} = G\left((\omega_{1:I}^{1:J})_r, \hat{p}_{1:I}^{1:J}\right).$$

3.1 Glass transition temperature

First, we demonstrate how our method applies in the case of establishing a QSPR for T_g. For our QSPR development, we use $p_{\text{ref}} = 65.30$°C and $g(p) = 1/p$. Specifically, fig. 2 shows maximum a posteriori predictions $\hat{T}_g^{\text{predicted}}$ and compares them with T_g^{measured}. The credible intervals result from taking the whole set of sampler predictions calculated under our $q_{1:I}^{I:J}$ values and summarizing them by their 5$^{\text{th}}$ and 95$^{\text{th}}$ percentiles.

For comparing our $(\hat{T}_g^{\text{predicted}})_{1:R}$ with $(T_g^{\text{measured}})_{1:R}$ we use the coefficient of determination [17]. This is denoted with $\hat{R}^2$ and detailed in appendix C.1. The resulting values of $\hat{R}^2$ quantify how well properties from the data are reproduced by our QSPR. In particular, $\hat{R}^2 = 0$ indicates no ability to make a prediction and $\hat{R}^2 = 1$ indicates a perfect replication of the data [9]. We emphasize that our learning scheme does not rely on $\hat{R}^2$, and therefore $\hat{R}^2$ is not used in our learning process. We employ it here only to facilitate a presentation of how well our model has been able to match the data it was supplied with.

To further assess the performance of our framework, we explore leave-one-out cross-validation of our model by holding one of the resin data points for our T_g data out of our learning process and then using the contribution predictions of our model to generate predictions for the T_g^{measured} of the withheld resin. Here we report the experimental data values as well as the 5$^{\text{th}}$, 50$^{\text{th}}$, and 95$^{\text{th}}$ percentiles denoted $((T_g^{\text{predicted}})_{5\%}, (T_g^{\text{predicted}})_{50\%}, (T_g^{\text{predicted}})_{95\%})$

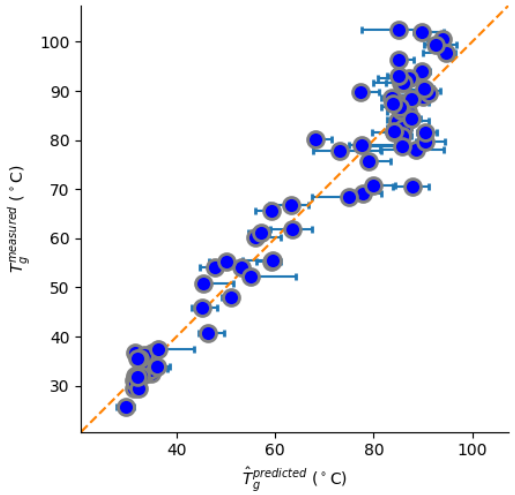


Figure 2: Quantitative comparison of our QSPR method using the maximum a posteriori estimator $\hat{T}_g^{\text{predicted}}$, plotted against the experimentally determined T_g^{measured} from the resin data. In this figure, points represent predicted and measured values. Points along the dashed line represent the perfect agreement between our prediction and measurements data and any deviation from this line is factored in the $\hat{R}^2$ value. For this figure, we get the value $\hat{R}^2 = .96$.

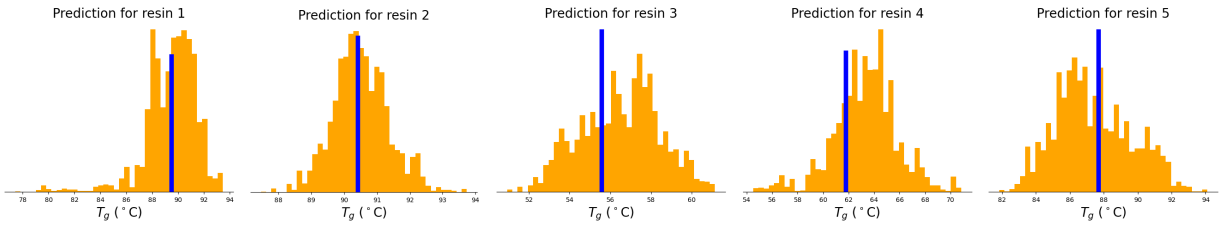


Figure 3: Leave-one-out cross-validation comparison of $T_g^{\text{predicted}}$ with T_g^{measured}: Histograms summarize suggested property values and vertical lines compare with their measured ones. For a numerical comparison see table 2 below.

as well as the error of our percentiles compared to the measured value $\left(T_g^{\text{err}_{5\%}}, T_g^{\text{err}_{50\%}}, T_g^{\text{err}_{95\%}}\right)$ see appendix C.1 for the formulas used to calculate the error. As the selected results in fig. 3 indicate, the experimental values typically fall between the $((T_g^{\text{predicted}})_{5\%}, (T_g^{\text{predicted}})_{95\%})$ interval and often get close to our median $(T_g^{\text{predicted}})_{50\%}$ prediction.

To demonstrate how effective our approach is, we test it against a well-validated method on the same data for T_g^{measured}. Specifically, we compare our methods against the formulation of Bicerano [5] as shown in fig. 4. We also note that with Bicerano's prediction method, we found it necessary to correct the infinite molecular weight predictions with additional molecular weight data, namely the number average molecular weight M_n. Our method for $T_g^{\text{predicted}}$ does not require this information, but is limited to predictions for resins in the same molecular weight neighborhood. For further explanation of the methods used to generate the results, we compared our methods against see appendix F.

3.2 Intrinsic viscosity

Because previous studies [33, 12] suggest a strong dependence of $[\eta]$ on molecular weight that is captured by the empirical Mark–Houwink equation [39], we use our QSPR scheme to model improvements to the Mark–Houwink

	T_g^{measured}	$T_g^{\text{predicted}}$			Comparison		
Resin		$(T_g^{\text{predicted}})_{5\%}$	$(T_g^{\text{predicted}})_{50\%}$	$(T_g^{\text{predicted}})_{95\%}$	$T_g^{\text{err}_{5\%}}$	$T_g^{\text{err}_{50\%}}$	$T_g^{\text{err}_{95\%}}$
1	89.52	86.11	89.74	92.13	-3.80	0.25	2.92
2	90.42	89.16	90.47	92.10	-1.39	0.06	1.85
3	55.58	53.24	56.60	59.72	-4.20	1.84	7.45
4	61.80	59.11	63.46	67.47	-4.35	2.68	9.17
5	87.70	84.02	87.02	91.30	-4.19	-0.77	4.11
	°C	°C	°C	°C	%	%	%

Table 2: Leave-one-out cross-validation comparison of $T_g^{\text{predicted}}$ with T_g^{measured} results: A comparison of our predictions with experimental measurements reveals good agreement. This table shows experimental values against characteristic quantiles values of our predictive probability distribution. Characteristically relative error sticks below 10 percent.

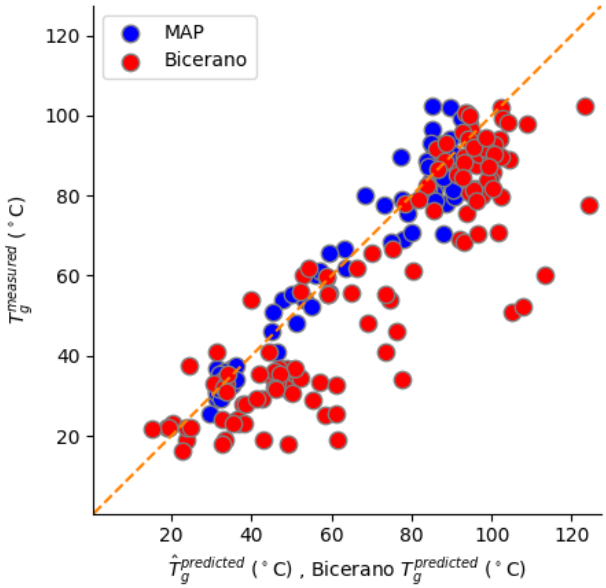


Figure 4: Quantitative comparison of QSPR developed using our framework via the maximum a posteriori estimator $\hat{T}_g^{\text{predicted}}$ and the $T_g^{\text{predicted}}$ produced by Bicerano's methods against the experimentally determined T_g^{measured}. In this figure, points represent predicted and measured values. Points along the dashed line represent the perfect agreement between our prediction and measurements and any deviation from this line is factored in the $\hat{R}^2, R^2$ values. For this figure, we get the value $\hat{R}^2 = .96$ and for Bicerano's method $R^2 = .83$.

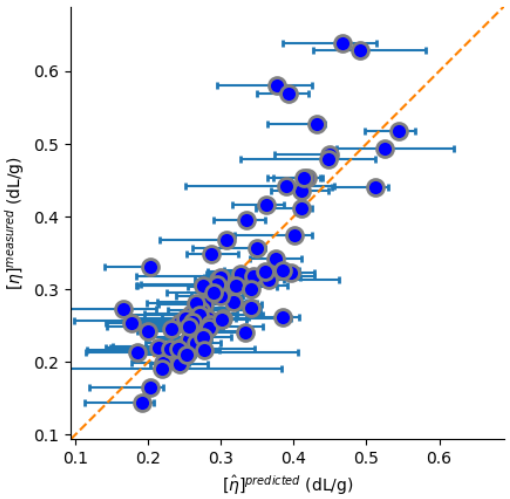


Figure 5: Quantitative comparison of our QSPR method using the maximum a posteriori estimator $[\hat{\eta}]^{\text{predicted}}$, plotted against the experimentally determined $[\eta]^{\text{measured}}$ from the resin data. In this figure, we see blue points that represent the value for each resin data point $r = 1, \ldots, R$ from both the data and our prediction. Points along the orange dashed line represent the perfect agreement between our prediction and the experimental data and any deviation from this line is factored in the $\hat{R}^2$ value. For this figure, we get the value $\hat{R}^2 = .71$.

equation instead of modeling $[\eta]$ directly. Specifically, to develop a QSPR for $[\eta]$ we target an artificial numerical property given by the ratio

$$P = \frac{[\eta]}{\kappa M_n^\alpha}$$

that is insensitive to molecular weight. Subsequently, we convert our predictions $P^{\text{predicted}}$ for P to predictions $[\eta]^{\text{predicted}}$ for $[\eta]$ according to $[\eta]^{\text{predicted}} = \kappa M_n^\alpha P^{\text{predicted}}$. Here, κM_n^α is the Mark–Houwink prediction [39] with the parameter values κ, α adjusted to best match our data set. For our QSPR development, we use $p_{\text{ref}} = .0197$ (dL/g) and obtain the mixing rule also by $g(p) = 1/p$.

As earlier, for comparing our $[\hat{\eta}]^{\text{predicted}}_{1:R}$ with $[\eta]^{\text{measured}}_{1:R}$ we use the coefficient of determination [17], denoted with $\hat{R}^2$ and detailed in appendix C.2. We compare the prediction of our QSPR method for $[\eta]^{\text{predicted}}$ against the measurements. As earlier, we show the maximum a posteriori predictions $[\hat{\eta}]^{\text{predicted}}$ and compare them with $[\eta]^{\text{measured}}$ in fig. 5.

We explore leave-one-out cross-validation of our model by holding one of the resin data points for our $[\eta]$ data out of our sampling and then using the repeat unit property predictions of our model to generate a collection of $[\eta]^{\text{predicted}}$ values of the withheld resin. As earlier, here we report the 5$^{\text{th}}$, 50$^{\text{th}}$, and 95$^{\text{th}}$ percentiles denoted $([\eta]^{\text{predicted}}_{5\%}, [\eta]^{\text{predicted}}_{50\%}, [\eta]^{\text{predicted}}_{95\%})$ as well as the error of our percentiles compared to the measured value $([\eta]^{\text{err}}_{5\%}, [\eta]^{\text{err}}_{50\%} [\eta]^{\text{err}}_{95\%})$ see appendix C.2 for the formulas used to calculate the error.

3.3 Multiproperty QSPR

Having demonstrated how our method performs in the simple context of a single property, either T_g or $[\eta]$, we now move to the more demanding setting, where with our QSPR we target both properties simultaneously. Our

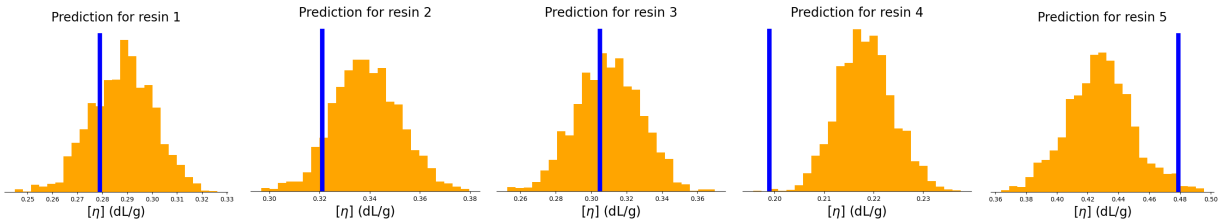


Figure 6: Leave-one-out cross-validation comparison of $[\eta]^{\text{predicted}}$ with $[\eta]^{\text{measured}}$: Histograms summarize suggested property values and vertical lines compare with their measured ones. For a numerical comparison see table 3 below.

	$[\eta]^{\text{measured}}$	$[\eta]^{\text{predicted}}$			Comparison		
Resin		$[\eta]^{\text{predicted}}_{5\%}$	$[\eta]^{\text{predicted}}_{50\%}$	$[\eta]^{\text{predicted}}_{95\%}$	$[\eta]^{\text{err}}_{5\%}$	$[\eta]^{\text{err}}_{50\%}$	$[\eta]^{\text{err}}_{95\%}$
1	0.279	0.267	0.289	0.309	-4.16	3.51	10.66
2	0.321	0.318	0.338	0.361	-0.82	5.37	12.48
3	0.305	0.280	0.310	0.340	-8.21	1.54	11.28
4	0.199	0.209	0.218	0.228	5.13	9.65	14.48
5	0.479	0.391	0.428	0.466	-18.31	-10.66	-2.73
	dL/g	dL/g	dL/g	dL/g	%	%	%

Table 3: Leave-one-out cross-validation comparison of $[\eta]^{\text{predicted}}$ with $[\eta]^{\text{measured}}$ results: This table shows experimental values against characteristic percentile values of our predictive probability distribution. We have a higher relative error compared with our T_g fits which is to be expected. As the selected values suggest we tend to see a good agreement of our median prediction when the value is around the average of our data but our prediction struggles as we move higher or lower.

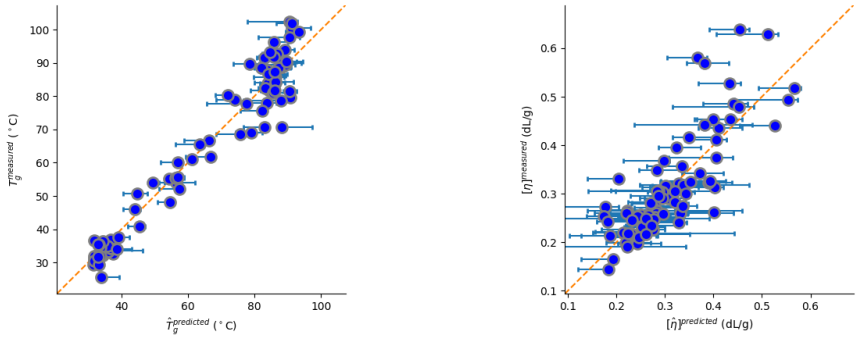
learning scheme extends naturally to this setting as detailed in appendix A.2. Namely, we now consider a bivariate property with $P_r = (T_g, [\eta])_r$. Now, in fig. 7, we demonstrate how the fits work across T_g and $[\eta]$ simultaneously.

4 Discussion

In this study, we presented a statistical learning approach to the development of QSPR for polyester properties and, to demonstrate characteristic predictions, we assimilated data from experimentally characterized poly-condensation reactions. To model how the structural contributions of the repeat unit ingredients combined to give rise to the recorded properties we developed a novel group contribution model in the style of Van Krevelen's foundational work [39] as well as taking a statistical learning approach to adjust the arising parameters with unknown values [30, 18]. In this study we focused on polyester resins, however our methodology is trans-formative and can be extended to other types of polyester products and even other chemical materials.

Our key contribution is a method that is grounded in the statistical learning methodology as well as the recent progress of empirical methods in the field of polymers. In particular, we introduce a probabilistic formulation for the known and unknown variables involved in the development of a QSPR. This allows us to take advantage of Bayesian concepts [36, 30] and encode our entire scheme in a single posterior $\mathcal{P}(\text{QSPR}|\text{data})$.

We also took the new approach of encoding the structural difference based on the molecular footprints into the prior's co-variances of our repeat unit property predictions in the statistical learning scheme we applied to the QSPR model. From these contributions, we were able to assembly a modeling procedure that generates estimates of contributions from our polymer ingredients to give rise to accurate property prediction. We used two key properties from the experimental data and showed good agreement on both properties when fitted both apart and



(a) T_g $\hat{R}^2 = .95$ *(b)* $[\eta]$ $\hat{R}^2 = .67$

Figure 7: Quantitative comparison of our both of our QSPR methods coupled together using the maximum a posteriori estimator $\hat{T}_g^{\text{predicted}}$ *and* $[\hat{\eta}]^{\text{predicted}}$*, plotted against the experimentally determined* T_g^{measured} *and* $[\eta]^{\text{measured}}$ *from the resin data. In both figures, we see blue points that represent the value for each resin data point* $r = 1, \ldots, R$ *from both the data and our prediction. Points along the orange dashed line represent the perfect agreement between our prediction and the experimental data and any deviation from this line is factored in the* $\hat{R}^2$ *value. For the figure, we get the values* $\hat{R}^2 = .95$ *for* T_g *(left) and* $\hat{R}^2 = .67$ *for* $[\eta]$ *(right).*

together.

Our methods are not without their limitations, we are required to integrate new structural information into a statistical formulation. Also approach is more designer and requires more mathematical ability. To keep our method simple we ignore the cap molecules in our Tanimoto. Our method is computationally expensive compared to naive methods like a pre-trained Bicerano model, taking a couple hours on a standard research desktop with running on a single core to yield good results. And up to ten to twelve hours on the same equipment to give excellent results. However, in the context of some molecular dynamics simulations this time frame is quick. With this in mind, our main computational bottleneck is the Metropolis-Hastings scheme we use to sample our target distribution. Methods to parallelize this procedure or simplify the proposal setup could give significant speed-ups and will be explored in addition to our other extensions.

In addition to computational or algorithmic improvement. Future research can try to investigate how statistical learning can be applied to non-polyester polymers. It also remains to show if statistical learning can be applied to properties beyond T_g or $[\eta]$. These may include properties like the melting point (T_{m}) or equilibrium melting point (T_{m}°) [39, 19].

Acknowledgements

This work is based upon raw data provided by Eastman Chemical Company, Kingsport, TN, USA. The authors thank Erik Burke for help in synthesizing a majority of the polyester resins used and acknowledge support from Eastman via grant No R011052266.

Conflict of interest

SM, DO, MD, and IS declare no competing interest. BA and CB declare no competing non-financial interest but competing financial interest as employees at Eastman Chemical Company.

Author contributions

All authors conceived research and wrote the manuscript. CB contributed experimental data. DO contributed the software implementation of Bicerano's methods. SM contributed software implementation of the developed algorithms for statistical learning and produced synthetic data. SM, DO, MD, IS developed theoretical methods. IS and MD oversaw all aspects of this work.

References

[1] Michael F Ashby and David CEBON. Materials selection in mechanical design. *Le Journal de Physique IV*, 3(C7):C7–1, 1993.

[2] Debra J Audus and Juan J de Pablo. Polymer informatics: Opportunities and challenges. *ACS macro letters*, 6(10):1078–1082, 2017.

[3] Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7(1):1–13, 2015.

[4] Andreas Bender and Robert C Glen. Molecular similarity: a key technique in molecular informatics. *Organic & biomolecular chemistry*, 2(22):3204–3218, 2004.

[5] Jozef Bicerano. *Prediction of polymer properties*. cRc Press, 2002.

[6] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[7] Adrià Cereto-Massagué, María José Ojeda, Cristina Valls, Miquel Mulero, Santiago Garcia-Vallvé, and Gerard Pujadas. Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58–63, 2015.

[8] Zexun Chen, Bo Wang, and Alexander N Gorban. Multivariate gaussian and student-t process regression for multi-output prediction. *Neural Computing and Applications*, 32:3005–3028, 2020.

[9] Davide Chicco, Matthijs J Warrens, and Giuseppe Jurman. The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *PeerJ Computer Science*, 7:e623, 2021.

[10] Viviana Consonni and Roberto Todeschini. Molecular descriptors. *Recent advances in QSAR studies: methods and applications*, pages 29–102, 2010.

[11] Andreas C Damianou, Michalis K Titsias, and Neil D Lawrence. Variational inference for latent variables and uncertain inputs in gaussian processes. *Journal of Machine Learning Research*, 17:1–62, 2016.

[12] Andrea Dodero, Silvia Vicini, Marina Alloisio, and Maila Castellano. Rheological properties of sodium alginate solutions in the presence of added salt: An application of kulicke equation. *Rheologica Acta*, 59:365–374, 2020.

[13] Guillaume Fayet and Patricia Rotureau. How to use qspr-type approaches to predict properties in the context of green chemistry. *Biofuels, Bioproducts and Biorefining*, 10(6):738–752, 2016.

[14] Paul J Flory. *Principles of polymer chemistry*. Cornell university press, 1953.

[15] Thomas G Fox. Influence of diluent and of copolymer composition on the glass temperature of a polymer system. *Bull. Am. Phs. Soc.*, 1:123, 1952.

[16] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.

[17] Stanton A Glantz, Bryan K Slinker, and Torsten B Neilands. *Primer of applied regression & analysis of variance*, volume 654. McGraw-Hill, Inc., New York, 2001.

[18] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[19] Kazuhiko Ishikiriyama. Polymer informatics based on the quantitative structure-property relationship using a machine-learning framework for the physical properties of polymers in the athas data bank. *Thermochimica Acta*, 708:179135, 2022.

[20] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. *An introduction to statistical learning*, volume 112. Springer, 2013.

[21] LB Kier and LH Hall. Molecular connectivity in chemistry and drug research, 1976.

[22] Lemont Burwell Kier and Lowell H Hall. Molecular connectivity in structure-activity analysis. *Journal of Pharmaceutical Sciences*, 1986.

[23] Greg Landrum et al. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 2013.

[24] Peter M Lee. *Bayesian statistics*. Oxford University Press London, 1989.

[25] Jun S Liu and Jun S Liu. *Monte Carlo strategies in scientific computing*, volume 75. Springer, 2001.

[26] Erlita Mastan, Xiaohui Li, and Shiping Zhu. Modeling and theoretical development in controlled radical polymerization. *Progress in Polymer Science*, 45:71–101, 2015.

[27] Iain Murray, Ryan Adams, and David MacKay. Elliptical slice sampling. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 541–548. JMLR Workshop and Conference Proceedings, 2010.

[28] James S Peerless, Nina JB Milliken, Thomas J Oweida, Matthew D Manning, and Yaroslava G Yingling. Soft matter informatics: current progress and challenges. *Advanced Theory and Simulations*, 2(1):1800129, 2019.

[29] Robert Pollice, Gabriel dos Passos Gomes, Matteo Aldeghi, Riley J Hickman, Mario Krenn, Cyrille Lavigne, Michael Lindner-D'Addario, AkshatKumar Nigam, Cher Tian Ser, Zhenpeng Yao, et al. Data-driven strategies for accelerated materials design. *Accounts of Chemical Research*, 54(4):849–860, 2021.

[30] Steve Pressé and Ioannis Sgouralis. *Data Modeling for the Sciences: Applications, Basics, Computations*. Cambridge University Press, 2023.

[31] Owen Queen, Gavin A McCarver, Saitheeraj Thatigotla, Brendan P Abolins, Cameron L Brown, Vasileios Maroulas, and Konstantinos D Vogiatzis. Polymer graph neural networks for multitask property learning. *npj Computational Materials*, 9(1):90, 2023.

[32] Bakhtiyor Rasulev and Gerardo Casanola-Martin. Qsar/qspr in polymers: Recent developments in property modeling. *International Journal of Quantitative Structure-Property Relationships (IJQSPR)*, 5(1):80–88, 2020.

[33] Jean Carlo Rauschkolb, Bruna Caroline Ribeiro, Thais Feiden, Bruno Fischer, Thiago André Weschenfelder, Rogério Luis Cansian, and Alexander Junges. Parameter estimation of mark-houwink equation of polyethylene glycol (peg) using molecular mass and intrinsic viscosity in water. *Biointerface Res. Appl. Chem*, 12:1778–1790, 2021.

[34] John W Raymond and Peter Willett. Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2d chemical structure databases. *Journal of computer-aided molecular design*, 16:59–71, 2002.

[35] Christian P Robert, George Casella, and George Casella. *Monte Carlo statistical methods*, volume 2. Springer, 1999.

[36] Devinderjit Sivia and John Skilling. *Data analysis: a Bayesian tutorial*. OUP Oxford, 2006.

[37] EA Smolenskii, GV Vlasova, D Yu Platunov, and AN Ryzhov. Ad hoc optimal topological indices for qspr. *Russian chemical bulletin*, 55:1508–1515, 2006.

[38] John G Topliss and Robert J Costello. Chance correlations in structure-activity studies using multiple regression analysis. *Journal of Medicinal Chemistry*, 15(10):1066–1068, 1972.

[39] Dirk Willem Van Krevelen and Klaas Te Nijenhuis. *Properties of polymers: their correlation with chemical structure; their numerical estimation and prediction from additive group contributions*. Elsevier, 2009.

[40] David Weininger. Smiles, a chemical language and information system. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

[41] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[42] Pengcheng Xu, Huimin Chen, Minjie Li, and Wencong Lu. New opportunity: machine learning for polymer materials design and discovery. *Advanced Theory and Simulations*, 5(5):2100565, 2022.

[43] Mengxian Yu, Yajuan Shi, Xiao Liu, Qingzhu Jia, Qiang Wang, Zheng-Hong Luo, Fangyou Yan, and Yin-Ning Zhou. Quantitative structure-property relationship (qspr) framework assists in rapid mining of highly thermostable polyimides. *Chemical Engineering Journal*, 465:142768, 2023.

[44] Yun Zhang and Xiaojie Xu. Machine learning glass transition temperature of polymers. *Heliyon*, 6(10):e05055, 2020.

A Summary of statistical learning

Here, we summarize our statistical learning schemes. For a detailed presentation, see main text of this study.

A.1 Uniproperty statistical learning framework

For QSPR with a single scalar property P, our learning framework consists of

$$\begin{aligned} q_{1:I}^{1:J} &\sim \mathsf{Normal}_I^J\left(0, S\right) \\ \tau &\sim \mathrm{Gamma}\left(\phi, \tau_{\mathrm{ref}}/\phi\right) \\ P_r^{\mathrm{measured}}|\tau, q_{1:I}^{1:J} &\sim \mathrm{LogNormal}\left(P_r^{\mathrm{predicted}}, \tau\right) \qquad r = 1\ldots, R \end{aligned}$$

Contributions $p_{1:I}^{1:J}$ are obtained from primitives $q_{1:I}^{1:J}$ using

$$p_i^j = p_{\mathrm{ref}} \cdot \exp(q_i^j) \qquad i = 1, \ldots, I \qquad j = 1, \ldots, J$$

and the mixing rule is given by

$$G\left(\omega_{1:I}^{1:J}, p_{1:I}^{1:J}\right) = g^{-1}\left(\sum_i \sum_j \omega_i^j g(p_i^j)\right).$$

A.2 Multiproperty statistical learning framework

For QSPR with two scalar properties $\dot{P}$ and $\ddot{P}$, our learning framework consists of

$$\begin{aligned} q_{1:I}^{1:J} &\sim \mathsf{Normal}_I^J\left(0, S\right) \\ \dot{\tau} &\sim \mathrm{Gamma}\left(\phi, \dot{\tau}_{\mathrm{ref}}/\phi\right), \\ \ddot{\tau} &\sim \mathrm{Gamma}\left(\phi, \ddot{\tau}_{\mathrm{ref}}/\phi\right), \\ \dot{P}_r^{\mathrm{measured}}|\dot{\tau}, q_{1:I}^{1:J} &\sim \mathrm{LogNormal}\left(\dot{P}_r^{\mathrm{predicted}}, \dot{\tau}\right) \qquad r = 1\ldots, R \\ \ddot{P}_r^{\mathrm{measured}}|\ddot{\tau}, q_{1:I}^{1:J} &\sim \mathrm{LogNormal}\left(\ddot{P}_r^{\mathrm{predicted}}, \ddot{\tau}\right) \qquad r = 1\ldots, R \end{aligned}$$

Contributions $\dot{p}_{1:I}^{1:J}$ and $\ddot{p}_{1:I}^{1:J}$ are obtained from primitives $q_{1:I}^{1:J}$ using

$$\begin{aligned} \dot{p}_i^j &= \dot{p}_{\mathrm{ref}} \cdot \exp(q_i^j) \qquad i = 1, \ldots, I \qquad j = 1, \ldots, J \\ \ddot{p}_i^j &= \ddot{p}_{\mathrm{ref}} \cdot \exp(q_i^j) \qquad i = 1, \ldots, I \qquad j = 1, \ldots, J \end{aligned}$$

and the mixing rules are given by

$$\begin{aligned} \dot{G}\left(\omega_{1:I}^{1:J}, \dot{p}_{1:I}^{1:J}\right) &= \dot{g}^{-1}\left(\sum_i \sum_j \omega_i^j \dot{g}(\dot{p}_i^j)\right), \\ \ddot{G}\left(\omega_{1:I}^{1:J}, \ddot{p}_{1:I}^{1:J}\right) &= \ddot{g}^{-1}\left(\sum_i \sum_j \omega_i^j \ddot{g}(\ddot{p}_i^j)\right). \end{aligned}$$

B Definitions of probability distributions

Here we present the parameterization of the probability distributions we use throughout this study.

The *Log Normal* has probability density given by

$$\text{LogNormal}\left(x; y, z\right) = \frac{1}{x}\sqrt{\frac{z}{2\pi}}\exp\left(-\frac{z}{2}\left(\log\frac{x}{y}\right)^2\right)$$

with y being the median and z being the precision.

The *Gamma* has probability density given by

$$\text{Gamma}\left(\tau; \phi, \psi\right) = \frac{1}{\Gamma\left(\phi\right)\psi^{\phi}}\tau^{\phi-1}\exp\left(-\frac{\tau}{\psi}\right)$$

with ϕ being the shape and $\phi\psi$ the mean.

The *Multivariate Normal* has probability density given by

$$\text{Normal}_I^J\left(x; \mu, S\right) = \frac{1}{\sqrt{(2\pi)^{IJ}|S|}}\exp\left(-\frac{1}{2}(x-\mu)^T S^{-1}(x-\mu)\right)$$

with μ being the mean and S the covariance.

C Performance metrics

C.1 Glass transition temperature

For calculating $\hat{R}^2$ for our T_g predictions we have

$$\hat{R}^2 = 1 - \frac{\sum_r \left((T_g^{\text{measured}})_r - (\hat{T}_g^{\text{predicted}})_r\right)^2}{\sum_r \left((T_g^{\text{measured}})_r - \bar{T}_g^{\text{measured}}\right)^2}$$

where $\bar{T}_g^{\text{measured}}$ is the average of $(T_g^{\text{measured}})_{1:R}$. For the error calculations shown in the last section of table 2 we use the following formulation

$$T_g^{\text{err}_{5\%}} = \left(\frac{(T_g^{\text{predicted}})_{5\%}}{T_g^{\text{measured}}} - 1\right)100\%,$$
$$T_g^{\text{err}_{50\%}} = \left(\frac{(T_g^{\text{predicted}})_{50\%}}{T_g^{\text{measured}}} - 1\right)100\%,$$
$$T_g^{\text{err}_{95\%}} = \left(\frac{(T_g^{\text{predicted}})_{95\%}}{T_g^{\text{measured}}} - 1\right)100\%,$$

these formulas give us the percentage of error in each chosen percentile compared to the measured value for T_g.

C.2 Intrinsic viscosity

We calculate $\hat{R}^2$ for our $[\eta]$ predictions in a similar way

$$\hat{R}^2 = 1 - \frac{\sum_r \left([\eta]_r^{\text{measured}} - [\hat{\eta}]_r^{\text{predicted}}\right)^2}{\sum_r \left([\eta]_r^{\text{measured}} - [\bar{\eta}]_r^{\text{measured}}\right)^2}$$

where $[\bar{\eta}]_r^{\text{measured}}$ is the average of $[\eta]_{1:R}^{\text{measured}}$. For the error calculations shown in the last section of table 3 we use the following formulation

$$[\eta]_{5\%}^{\text{err}} = \left(\frac{[\eta]_{5\%}^{\text{predicted}}}{[\eta]^{\text{measured}}} - 1\right) 100\%,$$
$$[\eta]_{50\%}^{\text{err}} = \left(\frac{[\eta]_{50\%}^{\text{predicted}}}{[\eta]^{\text{measured}}} - 1\right) 100\%,$$
$$[\eta]_{95\%}^{\text{err}} = \left(\frac{[\eta]_{95\%}^{\text{predicted}}}{[\eta]^{\text{measured}}} - 1\right) 100\%,$$

these formulas give us the percentage of error in each chosen percentile compared to the measured value for $[\eta]$.

D Prior's covariance matrix and ingredient indexing

The calculated Tanimoto coefficients, computed under a radius 3, that quantify the chemical similarity between our ingredients are shown in fig. S.2 and we also show how the covariance matrix S responds to changes in our hyperparameters fig. S.3.

To convert from tensor indexing of the ingredients f_i^j to vector indexing, i.e. from a double index scheme to a single index scheme, as required in fig. S.2 and fig. S.3, we use the convention below. Our convention starts by forming each pair of f_i, f^j, resulting to every f_i^j, and subsequently enumerating linearly according to the ordering shown in table 1.

E Sensitivity analysis

Here we only consider T_g for the analysis of our parameters, we use the sensitivity of $\hat{R}^2$ to our parameter changes as it is a value that corresponds to the performance of our methods on the actual data. The hyperparameters a, λ, ℓ influence the training and prediction ability of our QSPR scheme. We would like to explore how responsive our model is to each of these parameters so we do sensitivity analysis. Here we test the sensitivity of our model by fixing two of the three parameters and varying the third by a different percentage increase or decrease and examine the resulting $\hat{R}^2$ from the changes, see fig. S.1 in appendix E for the results. We then conclude that our methods are more sensitive to a but that the model response is not overly sensitive to any of the parameters. Overall our methods are robust to changes in our parameters as the $\hat{R}^2$ value does not respond that intensely to changes in the parameters.

F Bicerano's method

Bicerano's correlations for property predictions were primarily developed for homopolymers (single repeat unit) of infinite molecular weight [5]. The main structural and topological descriptors employed are connectivity indices as inspired by graph theory and further described in pioneering studies by Kier and Hall [21, 22]. These connectivity

indices 1-95	
1	f_2^2
⋮	⋮
19	f_{20}^2
20	f_2^3
⋮	⋮
38	f_{20}^3
39	f_2^4
⋮	⋮
57	f_{20}^4
58	f_2^5
⋮	⋮
76	f_{20}^5
77	f_2^6
⋮	⋮
95	f_{20}^6

indices 96-190	
96	f_2^7
⋮	⋮
114	f_{20}^7
115	f_2^8
⋮	⋮
133	f_{20}^8
134	f_2^9
⋮	⋮
152	f_{20}^9
153	f_2^{10}
⋮	⋮
171	f_{20}^{10}
172	f_2^{11}
⋮	⋮
190	f_{20}^{11}

indices 191-285	
191	f_2^{12}
⋮	⋮
200	f_{20}^{12}
210	f_2^{13}
⋮	⋮
228	f_{20}^{13}
229	f_2^{14}
⋮	⋮
247	f_{20}^{14}
248	f_2^{15}
⋮	⋮
266	f_{20}^{15}
267	f_2^{16}
⋮	⋮
285	f_{20}^{16}

indices 286-360	
286	f_2^{17}
⋮	⋮
304	f_{20}^{17}
305	f_2^{18}
⋮	⋮
323	f_{20}^{18}
324	f_2^1
⋮	⋮
342	f_{20}^1
343	f_1^2
⋮	⋮
359	f_1^{18}
360	f_1^1

Table S.1: Indexing of ingredients is explained in further detail in this table. Here for example the index 1 refers to the dimer made by the reaction of the functional group f^2 of the alcohol 1,6-Hexane diol (HDO) and the functional group f_2 of Terephthalic acid (TPA) which we call ingredient f_2^2.

indices include the simple connectivity index, δ and valence connectivity index δ^v, which fully describe the electronic environment and the bonding configuration of atom in a hydrogen-suppressed graph representing the repeat unit. Bond indices, β and β^v, are obtained from the product of simple connectivity indices of directly connected edges. Zeroth-order (atomic) ($^0\chi$ and $^0\chi^v$), and first order ($^1\chi$ and $^1\chi^v$) connectivity for the entire repeat unit are calculated from simple connectivity indices and bond indices. For polymers, the procedure utilizes solely the repeat unit but includes one of the two bonds that links the unit to the main repeating sequence [5]. Furthermore, in order to treat locally anisotropic properties, it is necessary to first identify the chain backbone and distinguish between the contributions made to the connectivity indices by groups along the chain from those atoms that belong to side groups.

In our study, we first carried out *in silico* condensation reactions as shown in fig. S.4a that take a list of alcohols and acids to produce samples of each resin in the 137 polyester dataset using a Python open source cheminformatics library, RDKit. A numerical approach was designed, where the acid and alcohol pairs were selected with replacement based on composition. We repeated the reaction multiple times to create a statistical sample of 500 molecules for each resin. Subsequent reactions for each sample created a trimer as presented in fig. S.4b. This was necessary to avoid any finite size effect when calculating connectivity indices for the central repeat unit, since a single bond linking the unit with the preceding one needs to be accounted for [5]. Groups participating in the condensation scheme were selected randomly using SMARTS pattern matching, while leaving multifunctional units partially reacted as a first approximation in our property prediction process, as a result, branching was not considered. If necessary, the effect of branching (just as the impact of finite molecular weight in the modeled resins was described later in the text) could be considered *a posteriori* by additional correlations.

The next step was then to select the backbone atoms as shaded in grey in fig. S.4b. The backbone atoms are

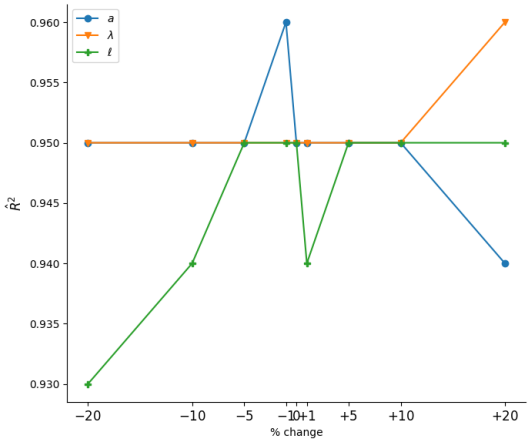


Figure S.1: Sensitivity analysis of the parameters that control our model. The y-axis is the $\hat{R}^2$ for the statistical learning model evaluated at the parameter values. The x-axis is the percentage increased or decreased from the baseline (0% change). The blue dots represent the results from changing a by the indicated percentage, the orange triangles represent the results from changing λ by the indicated percentage, and the green plus symbols represent the results from changing ℓ by the indicated percentage. Overall we see that our methods are robust and do not respond drastically to parameter changes.

defined as any atom in all possible paths (where no bond in the path is retraced) between the OH and COOH functional terminal groups. Atoms outside of the backbone were defined to be the side groups. The central atoms that make up a repeat unit are selected within a trimer as shown by the square brackets in fig. S.4b.

All indices, δ, δ^v, β, β^v, ${}^0\chi$, ${}^0\chi^v$, were calculated for each sample. The molar volume at room temperature was then computed from these indices. Two molar volume correlations were presented by Bicerano [5], with both resulting in predictions that are very similar for our data set. The solubility parameters, which quantifies the strength of the physical links between the components of a material, were computed from the cohesive energy and molar volume according to $\delta_s = \sqrt{E_{coh}/V}$. Solubility parameter estimations, while laborious and often difficult to validate, have been shown to systematically improve glass transition temperature (T_g) predictions over correlations that rely solely on chain stiffness [5].

The parameter N_{Tg} was calculated using eq. (6) from structural parameters x_1 to x_{12} [5] that refer to chemical architecture of the repeat unit and incorporate characteristics such as flexible, semi-flexible and rigid rings:

$$N_{Tg} = 15x_1 - 4x_2 + 23x_3 + 12x_4 - 8x_5 - 4x_6 - 8x_7 + 5x_8 + 11x_9 + 8x_{10} - 11x_{11} - 4x_{12} \tag{6}$$

Bicerano presented a general expression as obtained using linear regression on data from 320 polymers of varying structural features eq. (7) [5]. This expression accounts for the structure of a polymer, as reflected in chain stiffness and the cohesive forces between different chains.

$$T_{g,\infty} \equiv 351.00 + 5.63\delta_s + 31.68(N_{Tg}/N) - 23.95x_{13} \tag{7}$$

This expression was employed as described to predict $T_{g,\infty}$, which represented the limiting value of the T_g in the

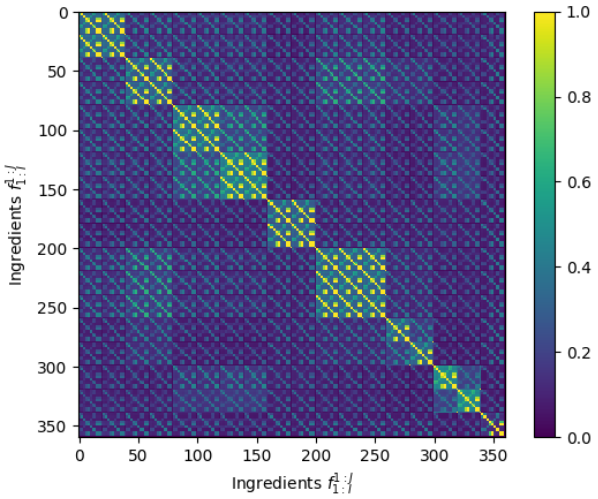


Figure S.2: Tanimoto matrix: This heat map shows the values of the Tanimoto coefficients for each pairwise comparison of ingredients f_i^j formed by the reactants f_i and f^j of table 1.

absence of chain-end effects (infinite molecular weight). To account for compositional variations, a value for each of the 500 samples for each of the 137 resins was estimated. A numerical scheme inspired by Fox's equation [15], as shown in eq. (8) was used to average the T_g for each resin across the 500 samples

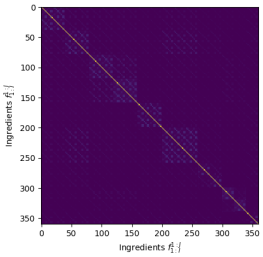
$$\frac{1}{T_{g,\infty}} = \frac{\sum_{m=1}^{500} \frac{M_m}{T_{gm,\infty}}}{\sum_{m=1}^{500} M_m}. \tag{8}$$

where M is the molecular weight of each sample's repeat unit and m is the number of samples.

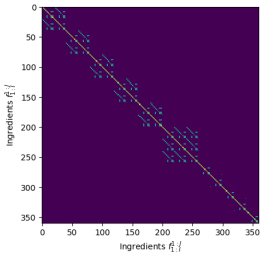
We then corrected for molecular weight (M_n) using the Flory-Fox equation eq. (9)

$$T_g^{\text{predicted}} = T_{g,\infty} - \frac{K_g}{M_n} \tag{9}$$

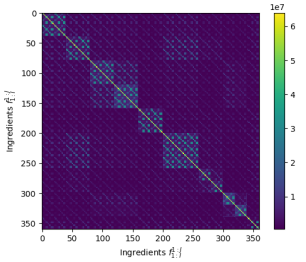
where $K_g \cong 0.002715 T_{g,\infty}^3$ following the recommendations by Bicerano [5]. It is important to emphasize that a large source of error can be identified in estimates of K_g [5]. Recent studies based on extensive data from large polymer databases such as PoLyInfo have underlined challenges present (such as different methods of molecular weight determination, crystallinity etc.) in structure-property relationships targeting K_g [28].



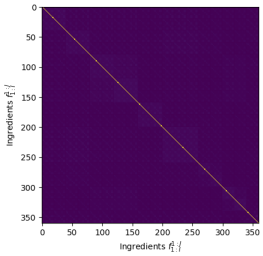
(a) a *increased* 80%. $\ell = \mathit{std}(\mathcal{T})$



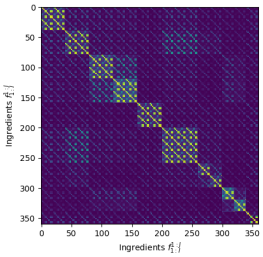
(b) $a = .5$. ℓ *decreased* 99.9%



(c) $a = .5 . \ell = \mathit{std}(\mathcal{T})$

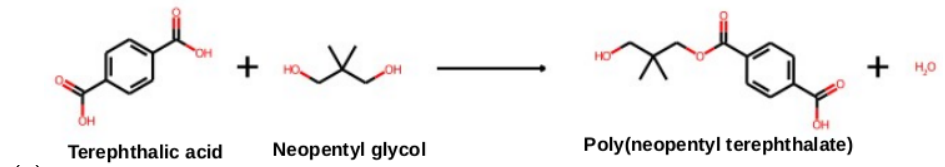


(d) $a = .5$ ℓ *increased* 1000%

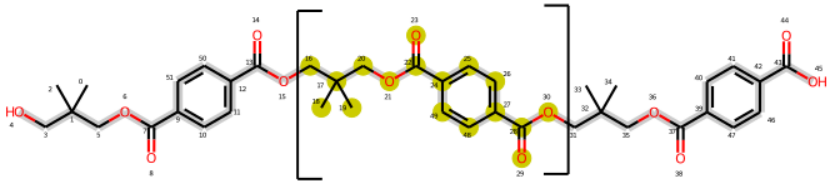


(e) a *decreased* 80%. $\ell = \mathit{std}(\mathcal{T})$

Figure S.3: Covariance matrix: Here we compare heat maps of the covariance matrix as defined by eq. (4) *and we show characteristic low and high values of the parameters* a, ℓ *to demonstrate how our hyperparameter values affect our statistical scheme.*



(a)



(b)

Figure S.4: (a) Condensation reaction between terephthalic acid and neopentyl glycol to produce poly(neopentyl terephthalate) (b) A trimer of poly(neopentyl terephthalate). The yellow shaded region enclosed in square brackets shows the central atoms