Macromolecules

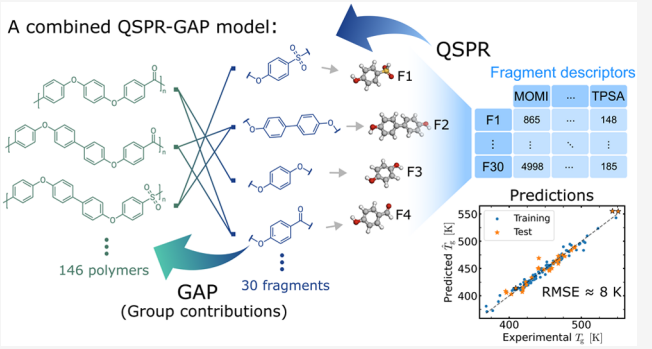
pubs.acs.org/Macromolecules
Article

Polymer Informatics Method for Fast and Accurate Prediction of the Glass Transition Temperature from Chemical Structure

Sebastian Brierley-Croft, Peter D. Olmsted, Peter J. Hine, Richard J. Mandle, Adam Chaplin, John Grasmeder, and Johan Mattsson*

Cite This: https://doi.org/10.1021/acs.macromol.5c00178
Read Online

ABSTRACT: We present a new polymer informatics framework that successfully predicts the glass transition temperature T_g of polymers based on their chemical structure. The framework combines ideas from group additive properties (GAP) and quantitative structure−property relationship (QSPR) methods, where GAP (or group contributions) assumes that submonomer motifs contribute additively to T_g, and QSPR links T_g to the physicochemical properties of the structure through a set of molecular descriptors. By integrating these methodologies, our combined QSPR−GAP framework overcomes limitations inherent in using either method independently. We demonstrate its application on a data set of 146 linear homo- and copolymers of the poly(aryl ether ketone) (PAEK) family, achieving a median root mean square error of 8 K for T_g, representing a significant improvement over standalone QSPR or GAP models. Moreover, using a genetic algorithm, we identify two molecular descriptors that predominantly drive T_g predictions. The QSPR−GAP framework can be readily adapted to forecast other physical properties and activity (QSAR) or transferred to other polymer families, including conjugated and biopolymers.



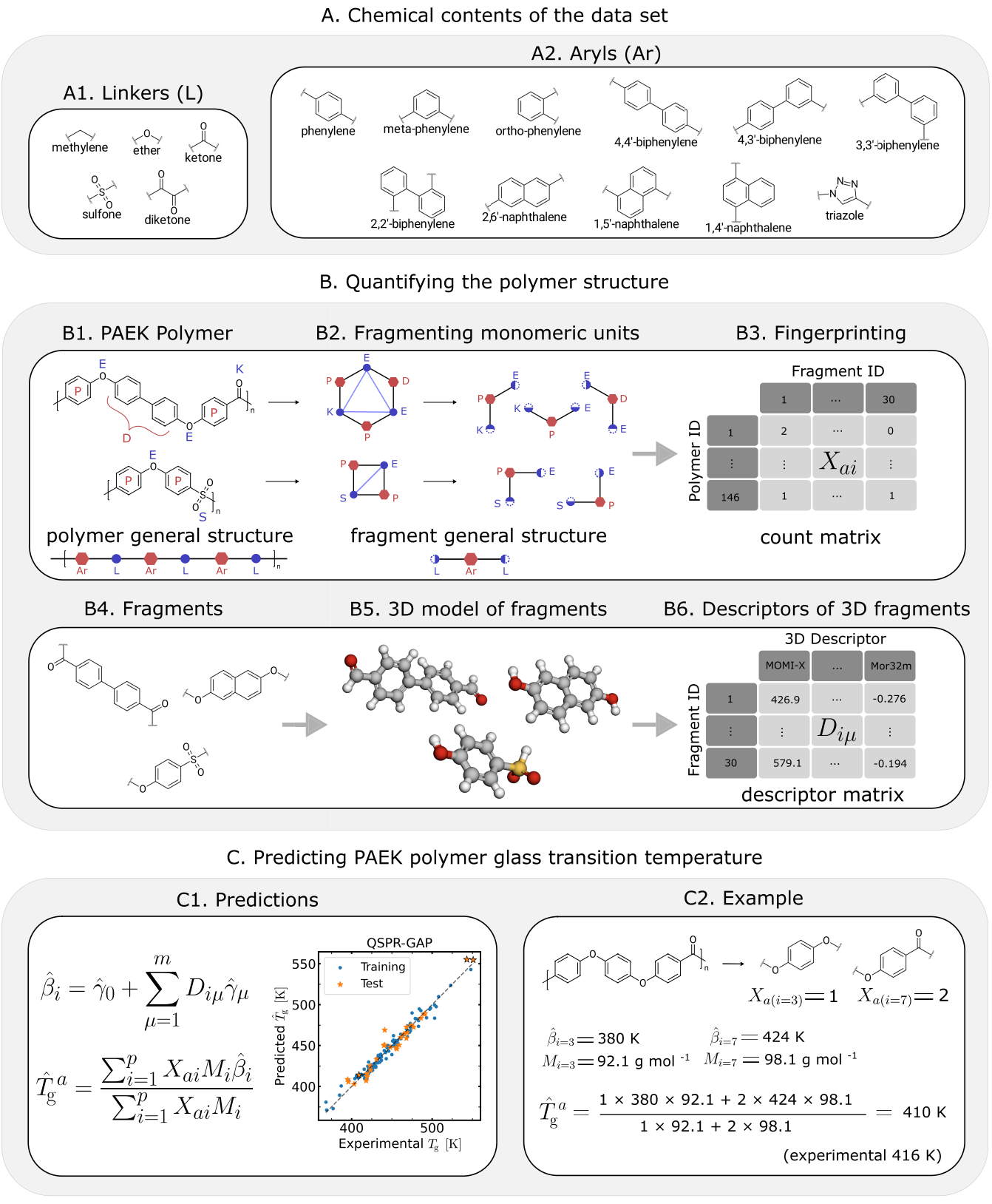
INTRODUCTION

Polymers are remarkably versatile materials, and the combined control of monomer chemistry and chain length allows for superior tunability of physical properties. As a polymer melt is cooled, the timescale τ_α characterizing its structural (α) relaxation increases dramatically, and in the absence of crystallization, the structure freezes into an amorphous solid, a glass, at the glass transition temperature T_g.[1] Since molecular motions are controlled by T_g, this is a key parameter for understanding and predicting material behavior, and it is thus essential to develop methods for accurately predicting T_g directly from the chemical structure.

For long-chain polymers, T_g is molecular weight (M)-independent[2−5] but strongly affected both by intramolecular dihedral barriers[6,7] (chain flexibility) and intermolecular packing effects, both of which are chemistry-specific.[5] Importantly, it has been shown that the α relaxation, which defines T_g, is linked to relaxations on a relatively 'local' submonomer length scale,[8−14] which in turn suggests that models that predict T_g from monomer structure should be achievable. In this paper, we present such a model and apply it to the poly(aryl ether ketone) (PAEK) family of polymers.

Predictive models that relate structure-based properties and T_g and are suitable for small data sets with low chemical variability have been proposed for polymers.[5,15−21] For instance, an approximate correlation has been found between T_g and monomer-scale properties such as the molecular weight per conformational (or flexible) degree of freedom of the chain (M_ϕ),[5,15−18] where M_ϕ captures both chain flexibility and chain bulkiness (reflecting molecular packing). As one example, Schut et al.[18] correlated T_g with the mass per flexible bond for a data set divided into three polymer classes by introducing flexible groups into both the main chain and the side chains; an out-of-sample mean absolute error (MAE) for the T_g of $\lesssim 6$ K (per polymer class) was obtained. In another example, Xie et al.[19] assigned an ad-hoc mobility factor to each atom based on the chemical group it belongs to (e.g., alkyl, phenyl, or thiophene). The monomer's mobility was then averaged over the atomic contributions, followed by a regression of T_g on the monomer mobility. For a family of 32 conjugated polymers, an RMSE $\simeq$ 13 K was attained for in-sample T_g predictions. These methods are easily applicable and intuitive, e.g., by linking a relevant physical property, such as molecular weight or volume,

Received: January 20, 2025
Revised: May 2, 2025
Accepted: June 4, 2025



$$\hat{\beta}_i = \hat{\gamma}_0 + \sum_{\mu=1}^{m} D_{i\mu}\hat{\gamma}_\mu$$

$$\hat{T}_g^a = \frac{\sum_{i=1}^{p} X_{ai} M_i \hat{\beta}_i}{\sum_{i=1}^{p} X_{ai} M_i}$$

Figure 1. QSPR−GAP analysis: predicting T_g from the polymer structure. (A) Chemical building blocks (flexible linker and aryl moieties) of the PAEK polymer data set. (B) Quantifying the monomeric structure: a step-by-step illustration: Fragments are extracted from the repeating units while recording their occurrences in the monomer using a count-based fingerprinting scheme. 3D molecular models are generated from the fragments, and descriptors are computed from the 3D fragments. (C) Calculations required to obtain a predicted T_g from the descriptor and count matrix, where an example illustrates the latter calculation on poly(ether ether ketone) (PEEK). Values of $\hat{\beta}_i$ and M_i are listed in Table S5, SI.

to each 'flexible bond', where ad-hoc rules are often introduced to quantify the influence of different bonds. However, the approaches are typically tailored to specific data sets and are not generalizable to a wider set of polymer structures.[20]

https://doi.org/10.1021/acs.macromol.5c00178
Macromolecules XXXX, XXX, XXX−XXX

Conversely, a more generalizable approach is the so-called group contribution or the group additive properties (GAP) method.[22−24] It assumes that a polymer property can be expressed by a composition-weighted average over contributions from submonomer motifs (fragments). The fragment contributions can be determined directly from the data by a linear regression. van Krevelen[24] applied GAP to predict various polymer properties, such as transition temperatures; solubility; and mechanical, optical, and electrical properties, while Weyland et al.[23] quoted in-sample MAE $\simeq$ 10 K for predictions of T_g. Despite their broad applicability, a fundamental flaw of GAP models is that they cannot be used to make predictions for polymers containing fragments outside of the data sample.[21,25,26]

A method that addresses some shortcomings of GAP models is the so-called quantitative structure−property relationship (QSPR) approach. QSPR-based methods use molecular descriptors,[27,28] which quantify electronic, topological, or geometric properties that are calculated from atomistic representations of molecules. For polymers, QSPR methods are normally applied either to the monomer[21,25,29,30] or to oligomers consisting of a few monomers,[31−33] and statistical or machine learning (ML) techniques are used to determine the relationship between the descriptors and the investigated property (such as T_g).[30,34,35] For QSPR methods applied to T_g predictions, RMSEs typically vary from $\simeq$4 to 35 K,[21,25,36−38] depending on the chemical variation within the data set. Models on larger data sets,[39,40] with higher chemical variation, typically yield prediction errors exceeding 25 K.[36,41] A significant drawback of QSPR models is that accurate descriptor calculations can be computationally costly, especially for large monomers or oligomers.

GAP and QSPR methods have usually been applied separately.[19−21,29] However, Hopfinger et al.[26] proposed a linear regression-based model for predicting T_g based on a GAP-like averaging scheme, combined with associating physical properties (conformational entropy and mass) with individual bonds. Inspired by this approach, we suggest extending QSPR methods to a smaller structural scale than the monomer unit, assuming interactions between these submonomer motifs negligibly contribute to the property of interest.

Here, we resolve the shortcomings of both the GAP and standard QSPR models by developing a hybrid QSPR−GAP method: a molecule is divided into submonomer fragments for which molecular descriptors are calculated, and various linear regression methods are used to link T_g to the fragment structure. The QSPR−GAP method provides more accurate predictions than either of the standalone methods, significantly faster descriptor calculations compared with QSPR, and accurate predictions of polymers containing fragments *outside* of the training set (where GAP fails).

We apply our new QSPR−GAP method to a data set of 146 linear homo- and copolymers of poly(aryl ether ketone) (PAEK)—an important class of linear polymers characterized by alternating stiff (aryls such as phenyls or biphenyls) and flexible linker (such as ethers or ketones) moieties, as shown in Figure 1A. The properties of PAEK polymers are highly tunable by varying these moieties, making them suitable for a wide range of applications including smartphone speakers, electrical insulation, automotive gears, medical implants, and aircraft components.[42] To design PAEK polymers with optimized properties for specific applications, reliable structure−property relationships are essential. Recent work[43] investigated a similar class of polymers (poly(aryl ethers)), predicting T_g using a purely QSPR-based approach where descriptors are calculated on the monomer units (referred to as "repeat units" by the authors), achieving an RMSE of $\simeq$17−19 K.

Our alternative QSPR−GAP method predicts T_g from the chemical structure with an RMSE of $\simeq$5−12 K (out-of-sample). Moreover, by identifying the molecular descriptors most important for predicting T_g, we reach new insights into how the local molecular structure relates to the glass transition temperature in polymers. Our findings offer a pathway to predict the properties of highly complex polymer structures using small data sets, thus circumventing the need for more elaborate ML methods that typically require larger data sets. Our method is readily generalizable to both a wider range of polymer properties (such as mechanical, optical, or electrical properties) and different classes of polymers.

RESULTS AND DISCUSSION

Characterization of PAEK Polymers. Our QSPR−GAP model is applied to a data set of 77 PAEK homopolymers and 69 copolymers, sourced from both the literature and experimental measurements conducted by Victrex R&D. We ignore any minor effects of chain length on T_g[5] and assume that all measured $T_g \equiv T_g^\infty$ (the long-chain limit); this limit is reached for PAEKs with $M_w \gtrsim 25$ kg/mol.[44] We note that many of the PAEKs investigated are commercial-grade, and although we do not have supporting molecular weight data, the manufacturing process generally does not allow access to molecular weights lower than this limit.

The monomer of a PAEK polymer (see examples in Figure 1,B1) is a sequence of alternating rigid aryl *Ar* (Figure 1,A2) and flexible linker *L* (Figure 1,A1) moieties, where the alternating arrangement

$$\ldots L_1 - Ar_1 - L_2 - Ar_2 \ldots L_N - Ar_N \ldots \quad (1)$$

is simple, yet different choices of *Ar* and *L* moieties lead to diverse material behavior, as illustrated by the T_g range of 375−550 K for the present polymer data set (Figure S1, SI).

We divide the monomer structure into unique submonomer "fragments" that constitute all PAEK monomers in the data set. Many fragment choices are possible, including *L*−*Ar*, *L*−*Ar*−*L*, *Ar*−*L*−*Ar*, or even longer sections. However, we mainly focus on *L*−*Ar*−*L* since the calculation of descriptors (see details below) requires the addition of hydrogens to the two ends of the fragments, and *L*−*Ar*−*L* is the only candidate that retains the uniqueness of the fragments once end-capped with hydrogens (Figure S20, SI). The data set of 146 polymers comprises 30 unique *L*−*Ar*−*L* fragments (Figure S3, SI), and as two examples, Figure 1,B1 illustrates how the monomers of poly(ether sulfone) (PES) and poly(ether biphenyl ether ketone) (PEDEK) are divided into *L*−*Ar*−*L* fragments.

We also benchmark QSPR−GAP against both the pure GAP and pure QSPR frameworks. For the GAP and QSPR−GAP methods, each homopolymer is parametrized by its *count matrix* **X**, where X_{ai} is the (integer) number of occurrences of fragment *i* in homopolymer *a*'s monomer (see the illustration in Figure 1,B3). Correspondingly, for copolymer *a*, we define $X_{ai} = \sum_{\xi=1}^{l} w_\xi X_{\xi ai}$, where w_ξ is the molecular weight fraction of comonomer ξ and $X_{\xi ai}$ is the count of fragment *i* in copolymer *a*'s comonomer ξ.

https://doi.org/10.1021/acs.macromol.5c00178
Macromolecules XXXX, XXX, XXX−XXX

For the QSPR−GAP and QSPR models, the molecular descriptors are calculated on the submonomer fragments and the monomer repeating units, respectively. Apart from the starting motifs (*i.e.*, the fragments vs the monomer repeat units), the procedure for calculating the descriptors is identical in both QSPR−GAP and QSPR frameworks: (i) add hydrogen atoms to the ends of each motif, (ii) generate an energy-minimized 3D representation of each motif (Figure 1,B5) using the Merck Molecular Force Field (MMFF)[45] via `RDKit`,[46] and (iii) calculate the values of molecular descriptors using `Mordred`.[47] Using this procedure, we calculate $m = 213$ descriptors for every unique motif.

The descriptors consist of six types: 41 charged partial surface area (CPSA) descriptors, 4 geometrical indices, 4 gravitational indices,[27,28] 160 3D-MoRSE descriptors,[48−50] 3 moment of inertia descriptors, and 1 plane of best fit (PBF) descriptor.[51] For the QSPR−GAP method, these $\mu = 1, 2, ..., m$ descriptors encode the $i = 1, ..., p$ fragments constituting the *descriptor matrix* $\mathbf{D}$, where $D_{i\mu}$ provides the value of descriptor μ for fragment i (see Figure 1,B6). The same $\mu = 1, 2, ..., m$ descriptors apply to the pure QSPR method, but now they encode the $a = 1, ..., A$ polymers, thus constituting a polymer-based descriptor matrix, where $D_{a\mu}$ gives the value of descriptor μ for polymer a. In this case, we note that copolymer descriptors were obtained by $D_{a\mu} = \sum_{\xi=1}^{l} w_\xi D_{\xi a\mu}$, where $D_{\xi a\mu}$ is the descriptor value μ in copolymer a's comonomer ξ. Finally, to maintain independence with the size of the repeating unit, all descriptors were normalized by the mass of the repeating unit (averaged for copolymers), resulting in the full set of inputs for the QSPR regression models (noting that this step only applies to the pure QSPR method).

The generation of 3D molecules and subsequent descriptor calculations are significantly faster for the QSPR−GAP method compared to the pure QSPR approach, with CPU times of a few seconds compared to 140 min. The speedup is primarily driven by two factors. First, the MMFF energy optimization requires fewer conformational degrees of freedom for the smaller fragment molecules compared to the entire (flexible) repeating monomer unit. Second, the QSPR−GAP method requires the energy optimization of 30 fragment molecules, whereas the QSPR method requires the energy optimization of 83 unique monomeric repeating units. For the full details on the 3D molecular optimization process; see SI Sec. S-IB.

GAP and QSPR−GAP Approaches. The predicted glass transition temperature for the ath polymer $\hat{T}_g^a$ is represented as a molar mass-weighted average of the estimated T_g-contribution $\hat{\beta}_i$ from each ith fragment

$$\hat{T}_g^a = \frac{\sum_{i=1}^{p} X_{ai} M_i \hat{\beta}_i}{\sum_{i=1}^{p} X_{ai} M_i} \equiv \sum_{i=1}^{p} \bar{X}_{ai} \hat{\beta}_i \tag{2}$$

where i indexes the fragments (as labeled in Figure S3, SI), M_i is the molar mass of the ith fragment, and thus $\bar{X}_{ai}$ is the mass-weighted composition of fragment i in polymer a. The polymer T_g is thus modeled by its composition-weighted constituent fragment contributions, β_i, where β_i corresponds to T_g of a long-chain homopolymer composed entirely of the ith fragment. Since β_i is unknown, it is estimated; note that we denote an estimated (or predicted) value by a hat $\hat{}$.

We estimate β_i in two different ways: (i) as a benchmark, we use a GAP approach based simply on the identity of the fragment or (ii) a novel combined QSPR−GAP approach based on the molecular features of each fragment encoded in the descriptors. In the GAP approach, the count matrix $\mathbf{X}$ is molar mass-normalized (eq 9), giving the composition matrix $\bar{\mathbf{X}}$ with elements $\bar{X}_{ai}$, and β_i is estimated from the experimentally available T_g values by ordinary least-squares (OLS) regression against $\bar{\mathbf{X}}$ (eq 11).

In the QSPR−GAP approach, the key distinction from the GAP method lies in the parametrization of β_i (and consequently, T_g) by a set of molecular descriptors that encode the structure of each fragment (see the Methods section for a detailed description of how β_i is estimated). The T_g contribution of fragment i is expressed in terms of the values of the molecular descriptors $D_{i\mu}$, according to

$$\beta_i = \gamma_0 + \sum_{\mu=1}^{m} D_{i\mu}\gamma_\mu \tag{3}$$

Here, the regression coefficient γ_μ parametrizes the influence of molecular descriptor μ on T_g, and γ_0 is a constant, both of which are estimated by the regression methods explained below. Since the inputs of eq 3 are physical molecular descriptors rather than occurrences of a given fragment, the QSPR−GAP model can also be used to predict the T_g value of polymers that contain a jth fragment that does not exist within the data sample.

Regression Methods. Our data set of $n = 146$ polymers with corresponding T_g values was divided into $p = 30$ unique $L-Ar-L$ fragments. The benchmark GAP analysis was performed using OLS to estimate the T_g contributions, β_i, of fragments $i = 1, ..., p$. For the QSPR−GAP analysis, in turn, the information about the $p = 30$ fragments was encoded into $m = 213$ molecular descriptors, and four linear regression methods were used to determine $\hat{\gamma}_0$ and $\hat{\gamma}_\mu$ (for each descriptor, $\mu = 1, ..., m$): principal component regression (PCR), ridge regression, lasso regression,[52] and partial least-squares (PLS) regression[53] (see SI Sec. S−II for a brief discussion of each). The benchmark QSPR models were based on descriptors derived directly from the $n = 146$ polymers instead of from the $p = 30$ fragments. For this series of models, we applied the linear regression methods already mentioned and an additional nonlinear model: kernel ridge regression (KRR) with a radial basis function (RBF) kernel (SI Sec. S−II).

These regression methods were chosen due to their robustness against overfitting, which would otherwise occur since the number of fit parameters in eq 3 ($m + 1$) exceeds the number of observed data points (i.e., the n polymers). The regression methods also account for the multicolinearity among the molecular descriptors (see Figures S5 and S6, SI) by penalizing the size of the estimated coefficients $\hat{\gamma}_\mu$, resulting in many fewer 'effective' regression coefficients.

As an alternative implementation of QSPR−GAP, a genetic algorithm (GA) was applied to select the subset of m_{GA} descriptors (out of all $m = 213$ descriptors) that best predict T_g by linear regression (see SI Sec. S−II for more details). Ten GA models were investigated, here termed "QSPR−GAP GAm_{GA}" ($m_{GA} = 1, ..., 10$), each resulting in different estimates for coefficients γ_0 and $\gamma_\mu = 1, ..., m_{GA}$, for the m_{GA} descriptors chosen.

Performance of the QSPR−GAP Model. To assess how well a model generalizes to new (or unseen) data, it is essential to perform an external validation. Often, external validation is performed on a reserved test set used only for this purpose, while model selection and/or tuning is performed on the training set (the remaining part of the data set) during an

https://doi.org/10.1021/acs.macromol.5c00178
Macromolecules XXXX, XXX, XXX−XXX

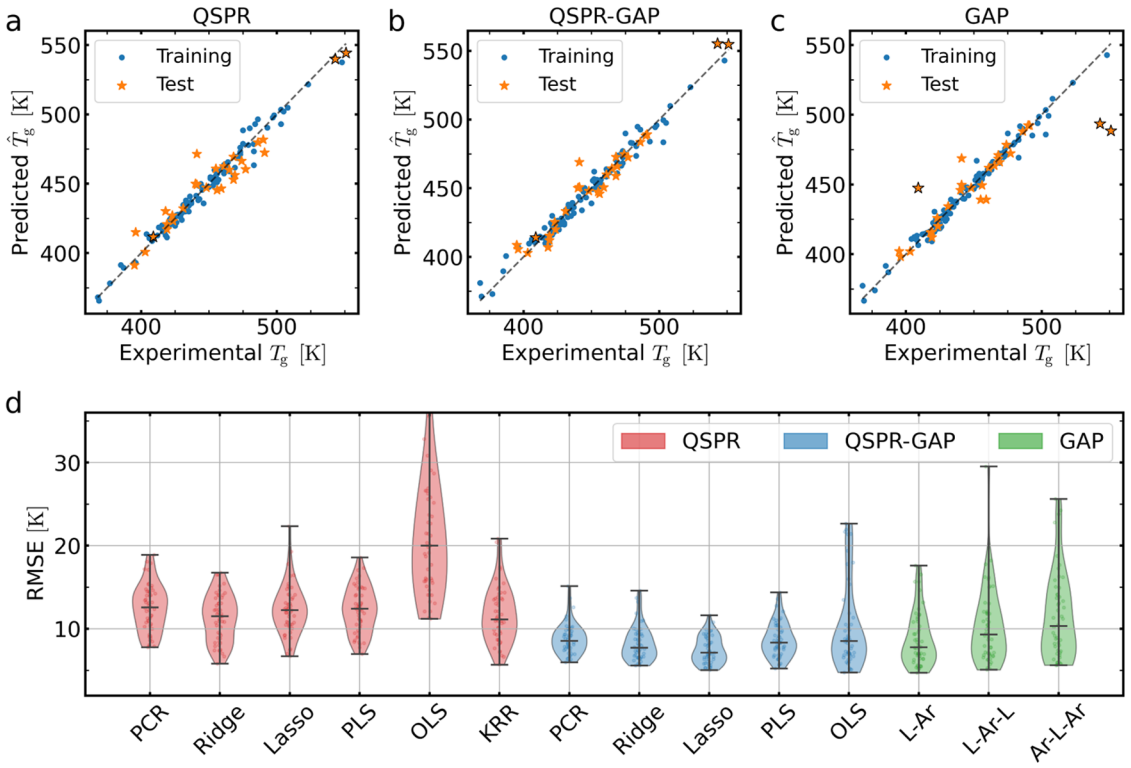


Figure 2. Benchmarking the QSPR−GAP model. (a−c) Comparing QSPR, QSPR−GAP, and GAP models from a single training−test split, where four polymers (indicated by the black marker edge color) in the test set contain at least one out-of-sample (L−Ar−L) fragment: (a) QSPR lasso model, (b) QSPR−GAP lasso model, and (c) GAP L−Ar−L model. (d) All models are compared by the distributions of their root mean square error (RMSE) during external validation, which involves predictions of the test set for 50 different training−test splits. RMSE distributions are represented using violin plots, where the outer envelope represents a kernel density estimation of the data (which is faintly displayed as points within the envelope). The black centerline represents the median, while the ends mark the extrema of the data. We note that the full distribution for the QSPR OLS goes beyond the scale and is truncated for a clearer comparison between the other models—its upper extrema extends as high as 59 K.

internal validation.[20,21,32,54] A drawback in selecting a dedicated test set is possible selection bias, i.e., bias due to random fluctuations in smaller data sets. To avoid this, we iteratively select different test sets such that all of the data points are eventually used in a test set. Full details of the external and internal validations are outlined in SI Sec. S−III.

Briefly, the external validation was performed using a repeated five-fold cross validation (five-fold CV), where the full data set was shuffled randomly and subsequently partitioned into five distinct subsets. A test set was iteratively selected from the five subsets, and in each iteration, the remaining four subsets were combined into a single training set. Internal validation was performed on the training set at each iteration to tune the model "hyperparameters" (e.g., the number of principal components in PCR and PLS or the degree of shrinkage in ridge and lasso; see SI Sec. S−II). This procedure was repeated 10 times, leading to 50 different training−test splits, each with a unique combination of polymers. One important aim of this procedure is to ensure that many test sets contain polymers with fragment IDs absent from the training set, which enables efficient probing of the robustness of our proposed QSPR−GAP approach. Such out-of-training set fragment occurrences, in the following referred to as "out-of-sample fragments", were identified 34 times for the 50 different training−test splits (Figure S14, SI).

Our proposed QSPR−GAP method is benchmarked against the pure GAP and QSPR methods. As mentioned above, the GAP models were fit using the OLS; however, the coefficients corresponding to the out-of-sample fragments were modified *a posteriori* to represent the mean of the coefficients associated with the in-sample fragments. This ad-hoc modification was performed to improve the robustness of the GAP model's out-of-sample fragment coefficient estimates, which would otherwise be zero based on the least-squares fit. Again, the need for this ad-hoc approach highlights the fundamental problem with the GAP method.

Figure 2a−c displays the predicted $\hat{T}_g$ versus experimental T_g for a representative training−test split across the three different methods as representative examples. In these figures, blue circles represent the training set data, while orange stars indicate the test set data. We include results for the QSPR lasso model in panel (a), the QSPR−GAP lasso model in panel (b), and the GAP L−Ar−L model in panel (c). For this particular partition (training−test split) of the data, fragments with i = 8 and 30 do not exist in the training set. For the GAP model, these two out-of-sample fragments manifest as three clear outlier polymers, which the model can not handle (outlined orange stars in Figure 2a−c). Since these fragments are absent in the training set, the corresponding $\hat{\beta}_i$ values are not known, and we thus set $\hat{\beta}_8$ and $\hat{\beta}_{30}$ to the means of the in-sample fragment contributions, leading to the outlier polymers.

The advantages of the QSPR−GAP and QSPR approaches become obvious when results from the same data partitions involve the same out-of-sample fragments. The QSPR (lasso) model in Figure 2a and QSPR−GAP (lasso) model in Figure 2b demonstrate a significantly more robust prediction than the

https://doi.org/10.1021/acs.macromol.5c00178
Macromolecules XXXX, XXX, XXX−XXX

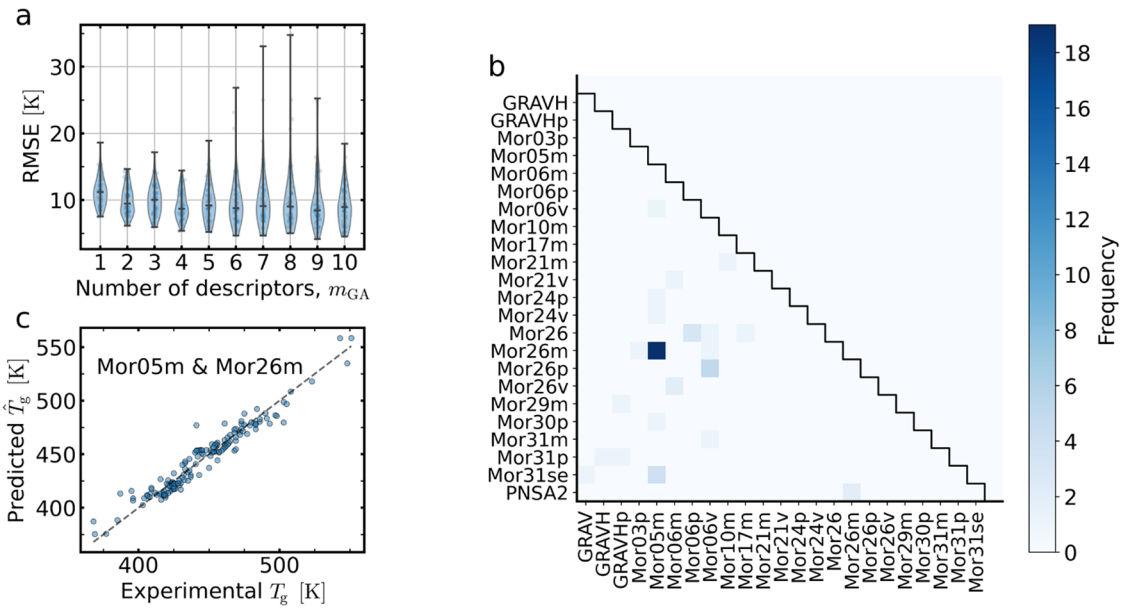


Figure 3. Analysis of descriptors. (a) Distributions (kernel density estimation) of the root mean square error (RMSE) for ten genetic algorithm (GA) models, consisting of a GA-based selection of m_{GA} = 1, m_{GA} = 2, ... m_{GA} = 10 optimal descriptors (from a pool of 213), followed by OLS regression. The RMSE is determined from a fivefold cross validation repeated ten times (while randomly shuffling each time), resulting in 50 distinct training−test splits. (b) GA results for m_{GA} = 2: from the complete set of two-descriptor combinations, we show all pairs selected at least once. As shown, the descriptor pair `Mor05m` and `Mor26m` was selected 19 times for the 50 training−test splits. (c) Predicted (in-sample) vs actual T_g values from an OLS regression on the full data set based only on the two descriptors `Mor05m` and `Mor26m`.

GAP model for the three outlier polymers. The root mean square error (RMSE) from the full external validation, i.e., the results from the full 50 training−test splits, is presented in Figure 2d for all investigated QSPR, QSPR−GAP, and GAP models. For the GAP model, in addition to the *L*−*Ar*−*L* fragment definition, we also investigated the definitions of *L*−*Ar* and *Ar*−*L*−*Ar*. From 50 splits, out-of-sample fragments are found 30 times for *L*−*Ar*, 34 times for *L*−*Ar*−*L*, and 40 times for the *Ar*−*L*−*Ar* fragment choice. The increase in the number of out-of-sample fragments grows with the number of available combinations of *L* and *Ar* groups (Figure 1A).

As shown in Figure 2d, even though all investigated QSPR−GAP models perform similarly (apart from OLS, as expected), the lasso model is the most accurate, with a RMSE range of ≃5−12K (depending on the partitioning of training/test data) and a median RMSE of ≃8 K. The QSPR models generally show a weaker predictive performance compared to QSPR−GAP, with median RMSE values ranging between 11 and 13 K (excluding OLS), compared to 8−9 K for QSPR−GAP. Thus, the QSPR−GAP models (PCR, ridge, PLS, and lasso) are more robust against the outlier (out-of-sample fragment-containing) polymers than the GAP models, and improve the predictive performance compared to the QSPR models, as shown in Figure 2.

Since the predictive ability (characterized by the RMSE) for the GAP models is significantly affected by the outlier polymers caused by out-of-sample fragments, we also compared models for which all outliers were removed (Figure S15, SI). We find that the predictions of the QSPR−GAP models are slightly improved, as exemplified by an RMSE ≃5−9 K for the lasso method, whereas the GAP models demonstrate a highly improved RMSE of ≃5−8 K. Interestingly, when comparing the predictive performance with out-of-sample fragment occurrences removed, the QSPR−GAP OLS results are identical to those of the GAP *L*−*Ar*−*L* model (Figure S15, SI).

Overall, QSPR−GAP leverages the strengths of QSPR with respect to its robustness against out-of-sample fragment-containing polymers while achieving comparable accuracy to GAP for polymers comprising only in-sample fragments. The GAP model shows excellent predictive performance for polymers containing fragments that are well represented by the data, with the downfall that polymers comprising out-of-sample fragments cannot conventionally be predicted. Although the method of averaging in-sample fragment contributions offers a solution, as demonstrated for GAP, a more accurate approach is to use descriptors to inform these contributions, which is what QSPR−GAP does.

When the QSPR method is applied (to an entire monomer), the descriptors implicitly encode conformational information, which appears to decrease predictive accuracy for T_g. In contrast, when the descriptor calculations are focused on submonomer fragments (characterized by minimal conformational degrees of freedom), as in QSPR−GAP, the predictive accuracy improves. Thus, for QSPR, 3D descriptor encoding of conformational information for the isolated monomeric unit may introduce more error than insight, whereas the most critical information for T_g resides at the local intrafragment scale (with little or no conformational degrees of freedom).

DESCRIPTOR ANALYSIS

Identifying the Most Important Descriptors. The key feature of the genetic algorithm models (QSPR−GAP GAm_{GA}) is that they explicitly select a subset of (m_{GA} = 1, ..., 10) descriptors that best predict T_g. This allows analysis of the direct impact of a discrete set of optimized descriptors on T_g, which may help develop an understanding of their physical significance.

https://doi.org/10.1021/acs.macromol.5c00178
Macromolecules XXXX, XXX, XXX−XXX

In testing the performance of the GA models, the same external validation consisting of a five-fold CV (repeated 10 times) is used, resulting in a total of 50 independent out-of-sample predictions, i.e., 50 different training−test splits (SI, Sec. S−III). The RMSE results for the 10 investigated QSPR−GAP GAm_{GA} models are shown in Figure 3a. Remarkably, only a few descriptors are needed for a good prediction. The most significant improvement in the predictive accuracy occurs between m_{GA} = 1 and 2, while for m_{GA} > 2, there is no significant improvement. Hence, only two descriptors are necessary (for this data sample) to predict T_g with an RMSE of ≃6−15 K. Thus, in the following, we restrict our attention to the two-descriptor model, QSPR−GAP GA2.

Figure 3b illustrates the frequency by which the GA selected different pairs of descriptors. It is clear that one pair stands out: `Mor05m` and `Mor26m`. From the 22,578 possible pairs arising from 213 descriptors, this pair was selected 19 out of 50 times (Figure 3b) from the 50 random data partitions. Descriptors `Mor05m` and `Mor26m` belong to the 3D-MoRSE family (Molecular Representation of Structures based on Electron diffraction),[48−50] which describes the 3D structure of a given fragment (or molecule) by a "form factor" based on atom-to-atom pair distances

$$I(q) = \sum_{l=1}^{N-1} \sum_{k=l+1}^{N} A_k A_l \frac{\sin(q r_{kl})}{q r_{kl}} \tag{4}$$

where k and l label specific atoms, N is the number of atoms in the fragment, A_k and A_l are weighting factors for atoms k and l, q is the "scattering" wave vector, and r_{kl} is the Euclidean distance between atoms k and l.

Our descriptor set (m = 213 descriptors calculated using `Mordred`) contains 160 3D-MoRSE descriptors, characterized by 32 different q values: 0 (`Mor01`), 1 Å^{-1} (`Mor02`), 2 Å^{-1} (`Mor03`), ..., 31 Å^{-1} (`Mor32`) and five different weighting schemes for A_k and A_l: unweighted ($A_k = A_l = 1$), atomic mass (`MorXXm`), van der Waals atomic volume (`MorXXv`), Sanderson electronegativity (`MorXXse`),[55] and polarizability (`MorXXp`); all weighting schemes are scaled by their value for carbon. MoRSE descriptors `Mor05m` and `Mor26m` correspond to q = 4 Å^{-1} and q = 25 Å^{-1}, where A_i is the ratio of the mass of atom i to the mass of carbon.

The T_g contribution $\hat{\beta}_i$ for fragment i can thus be accurately estimated by

$$\hat{\beta}_i = \hat{\gamma}_0 + \hat{\gamma}_1 I_i(4) + \hat{\gamma}_2 I_i(25) \tag{5}$$

where $I_i(4)$ and $I_i(25)$ are calculated from the set of atoms in fragment i and $\hat{\gamma}_{\mu=0,1,2}$ are the regression coefficients, estimated with an OLS regression applied to the full data sample of 146 polymers (no training−test splits); these estimates are provided in Table 1. The predicted against experimental T_g results are shown in Figure 3c, with an in-sample RMSE of ≃8 K.

Atomic-Level T_g Contributions. Using eqs 4 and 5, we can express the estimated T_g contribution of fragment i as a sum over atomic pair contributions

$$\hat{\beta}_i = \hat{\gamma}_0 + \sum_{l=1}^{N_i-1} \sum_{k=l+1}^{N_i} \hat{\pi}_{kl} \tag{6}$$

where $\hat{\pi}_{kl}$ denotes the (estimated) T_g contribution given by the pair of atoms k and l, expressed in terms of the two descriptors `Mor05m` and `Mor26m`,

Table 1. Regression Coefficients from the Best Two-Descriptor Genetic Algorithm Model, as Estimated from the Full Data Set of 146 Polymers (OLS)[a]

μ	descriptor	$\hat{\gamma}_\mu$ [K]	CI (95%) L/U [K]
0		298	286/310
1	`Mor05m`	−58	−67/−50
2	`Mor26m`	−198	−239/−157

[a]Upper and lower 95% confidence intervals (CIs) are presented for the estimated coefficients γ_0 and γ_μ. For assumptions and diagnostics of distributions, see the SI, Sec. S−IV.

$$\hat{\pi}_{kl} = \hat{\gamma}_1 M_k M_l \frac{\sin(4 r_{kl})}{4 r_{kl}} + \hat{\gamma}_2 M_k M_l \frac{\sin(25 r_{kl})}{25 r_{kl}} \tag{7}$$

M_k is the mass of atom k (divided by the mass of carbon), and r_{kl} is the distance between atom pair k,l. To determine the fragment contribution $\hat{\beta}_i$, the $\hat{\pi}_{kl}$ contributions are summed over the total number of atoms N_i in the ith fragment according to eq 6.

Function $\hat{\pi}_{kl}$ (r_{kl}, $M_k M_l$) is shown in Figure 4a, and at constant r_{kl}, $\hat{\pi}_{kl}$ is a linear function of the product $M_k M_l$ of atomic masses, as illustrated by the color gradient, and at constant $M_k M_l$, $\hat{\pi}_{kl}$ is an oscillating function of the pair distance. As r_{kl} increases, the pair contributions $\hat{\pi}_{kl}$ become less relevant to the overall $\hat{\beta}_i$, with negative and positive contributions canceling out in the summation of eq 6. Thus, the most important contributions lie within the range r_{kl} ≃ 1.2−1.5Å.

In Figure 4b, we show four L_1−Ar−L_2 fragments for which only the linker L_2 differs. As discussed above, the structure of each fragment is energy-minimized (using MMFF) and is represented by a unique set of r_{kl}, M_k, and M_l values. For each of the four fragments, the function $\hat{\pi}_{kl}(r_{kl}, M_k M_l)$ is evaluated for all atomic pairs, and the results are provided in the tiles (also illustrated by the corresponding heat map); each tile represents the pairwise atomic T_g contribution from atoms k and l. The overall fragment T_g contribution $\hat{\beta}_i$ results from a sum over all atomic pair contributions $\hat{\pi}_{kl}$ and the constant $\hat{\gamma}_0$.

Since hydrogen-containing pairs show very small T_g contributions (as shown in Figure 4a), they are omitted for clarity in Figure 4b. The contributions from atoms 1−7 are very similar because these atoms represent the same molecular structure motif (ether-linked phenyl). Hence, the sum over $\hat{\pi}_{kl}$ for atoms k,l = 1, ..., 7 gives contributions of $\Delta\hat{\beta}_i$ = 82.2, 86.5, 85.1, and 83.5 K, respectively, for the four fragments, as shown in Figure 4b. The slight differences between these values arise when minimizing the energy because the interatomic distances r_{kl} are influenced by the atoms in linker L_2. The differences in linker L_2 (atoms 8+) lead to significant differences in T_g, with contributions of $\Delta\hat{\beta}_i$ = 3.0, 7.8, 31.2, and 103.2 K for the four structures. The total T_g contributions for the four fragments are $\hat{\beta}_i$ = 383, 393, 415, and 485 K, respectively, determined using eq 6 while excluding hydrogens from the sum.

We conclude that for PAEK polymers the T_g contribution of each fragment (β_i) is very well approximated as a constant plus a sum over all atomic pair contributions ($\hat{\pi}_{kl}$), where the main contributions correspond to short atomic pair distances of 1.2−1.5 Å. The genetic algorithm identified two important 3D-MoRSE descriptors corresponding to length scales ($\sim 2\pi/q$) of 1.6 and 0.3 Å (see the scale bars in Figure 4b). The former is approximately the size of the average single C−C bond in the data set (Figure 4a), whereas the latter corresponds to a length

https://doi.org/10.1021/acs.macromol.5c00178
Macromolecules XXXX, XXX, XXX−XXX

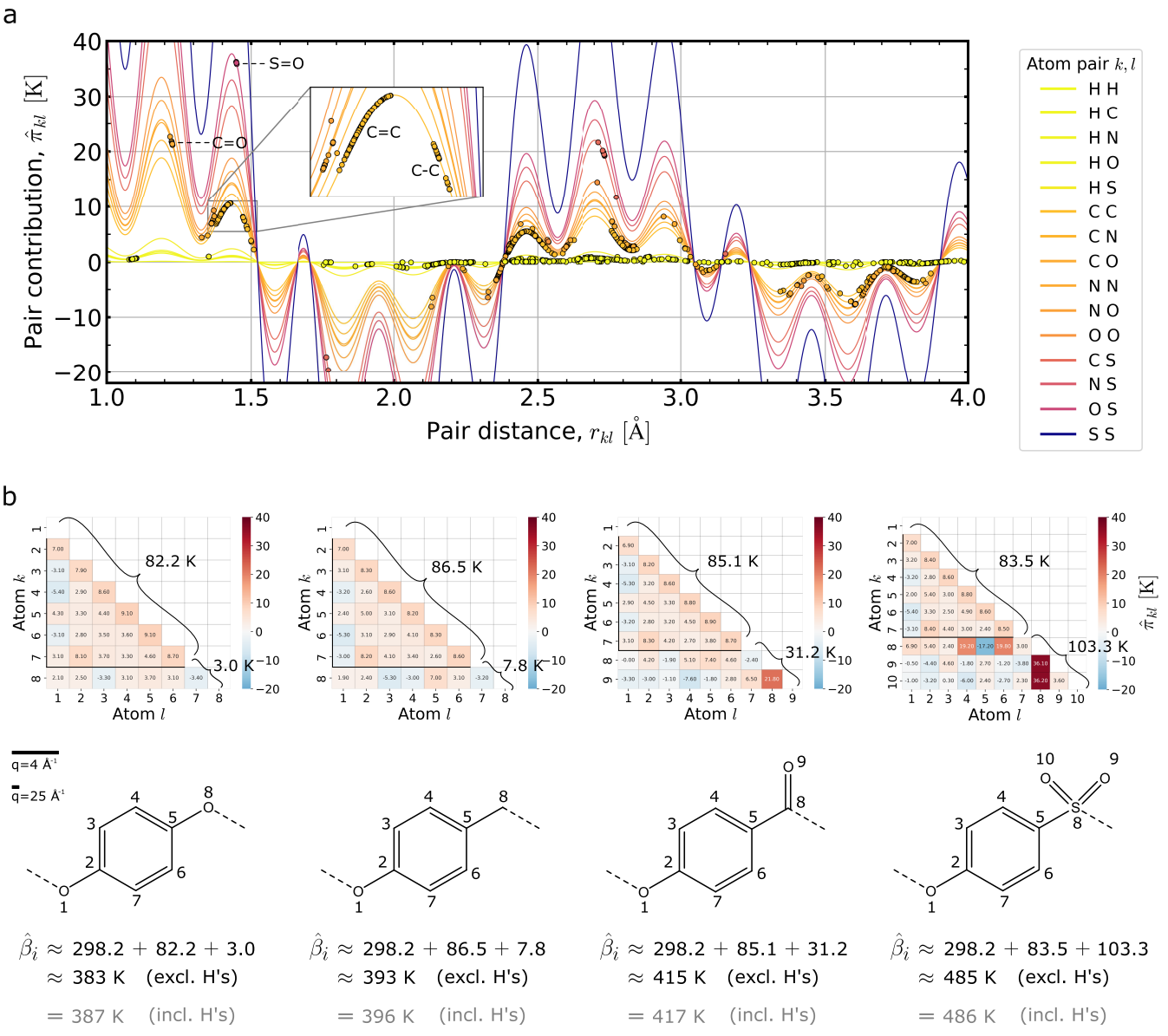


Figure 4. Estimated interatomic pair T_g contributions. (a) Estimated function $\hat{\pi}_{kl}$ of pairwise atomic T_g contributions, as expressed in eq 7; $\hat{\pi}_{kl}$ is a function of the pair distance r_{kl} between k and l atoms and the product M_kM_l , where the latter is illustrated by the color gradient. The lines represent the estimated function $\hat{\pi}_{kl}(r_{kl}$, $M_kM_l)$, and the solid points represent the function evaluated for the specific atom pairs within every polymer in the data set. The inset is a magnification of the range containing contributions from different carbon−carbon single and double bonds, whose bond lengths (when energy-minimized) vary slightly depending on the specific fragment. (b) Four example fragment IDs that share the same structure (in atoms 1−7) but vary by a single functional group (in atoms 8+). The pairwise contributions $\hat{\pi}_{kl}$ are shown as colored tiles with their values shown inside each tile. The scale bars show the length scales (1.6 and 0.3Å) corresponding to q = 4 Å^{-1}, 25 Å^{-1}, relative to the size of the (planar) structures. In the sum over pair contributions (eq 6) atoms l = 1, ..., 6 and k = 2, ..., 7, where $l < k$ (indicated by the partition) correspond to the same structure and yield nearly the same contribution $\Delta\hat{\beta}_i \simeq 82.2{-}85.5$ K for all four fragments. The remaining atom pairs set the fragments apart in their summed contribution to $\hat{\beta}_i$, which varies from $\Delta\hat{\beta}_i \simeq 3.0$ to 103.3 K. Atomic pair contributions including hydrogen atoms have been ignored from the plot since they show only small contributions to $\hat{\beta}_i$ (according to eq 7).

scale much smaller than an interatomic bond. Physically, the reliability of MMFF would be lost at a length scale of 0.3 Å; however, based on the figure, it is evident that it still encodes the length differences between the types of bonds. For example, in the zoomed inset of Figure 4a, the higher-frequency q component in the function induces a modulation that distinguishes between the C−C and C = C bonds. Without the length scale at q = 25Å^{-1}, the fit would not have enough flexibility to appropriately capture the differences among the S = O, C = O, C = C, C−C, ... (as labeled in the figure). We note that regularized sparse linear regression methods (such as lasso) applied to all 32 mass-weighted 3D-MoRSE descriptors likely yield a model with improved predictive accuracy and finer differentiation between bond types.

The example shown in Figure 4b indicates that the linker properties, such as their bulkiness, have the greatest impact on T_g; see, for example, the high contribution from S = O (36 K) and C = O (22 K) in Figure 4a,b. Based on these findings, we speculate that these particular groups strongly restrict torsional rotations and increase the chain's stiffness, thus raising T_g. For the glass transition in general, the interplay between packing

https://doi.org/10.1021/acs.macromol.5c00178
Macromolecules XXXX, XXX, XXX−XXX

and chain flexibility[5,16] suggests that three-body features, such as intramolecular angles, play an important role. In this study, such features are likely implicitly incorporated due to the chosen $L-Ar-L$ motif.

CONCLUSIONS

We present a new method for predicting T_g from the monomer structure in polymers. The method combines group additive properties (GAP) with a quantitative structure−property relationship (QSPR) approach. The GAP method assumes that T_g can be expressed by a molar mass-weighted average over T_g contributions from submonomer motifs (fragments), and our QSPR−GAP model uses molecular descriptors to relate the physical properties of these fragments to their GAP-like T_g contributions. We apply this model to a data set of 146 linear poly(aryl ether ketone) (PAEK) homo- and copolymers, resulting in a median root mean square error of 8 K (out-of-sample).

Compared to the standalone GAP and QSPR methods, the QSPR−GAP method improves robustness and accuracy in out-of-sample T_g predictions. Furthermore, 3D descriptor calculations for submonomer fragments are significantly faster than in traditional QSPR approaches, which are based on monomers (or oligomers), due to the reduction in conformational degrees of freedom.

Using a genetic algorithm, we show that only two molecular descriptors (from a pool of 213) are necessary to predict T_g with an RMSE of ≃6−15 K. Moreover, we identify a direct mapping between T_g and the monomer structure through pairwise atomic contributions.

This work offers an accurate, accessible, and broadly applicable predictive model suitable for small data sets and deployment on a standard laptop. The QSPR−GAP method is transferable to other classes of polymers, both synthetic and natural (e.g., conjugated or biopolymers), and to physical behavior beyond the glass transition, such as mechanical, optical, or transport properties.

METHODS

The number of occurrences of fragment i in polymer a is

$$(\mathbf{X})_{ai} \equiv X_{ai} \tag{8}$$

where $\mathbf{X}$ is an $n \times p$ dimensional count matrix, with n rows representing the full set of polymer IDs and p columns representing the full set of unique fragment IDs. We normalize $\mathbf{X}$ by the molar mass of the repeating unit, resulting in the mass-weighted composition matrix

$$\overline{\mathbf{X}} = (\mathrm{diag}[\mathbf{X}\boldsymbol{M}])^{-1}\mathbf{X}(\mathrm{diag}[\boldsymbol{M}]) \tag{9}$$

where $\boldsymbol{M} \in \mathbb{R}^p$ is a p-vector that enumerates the fragment molar masses. Note that the molar mass M_i of an $L-Ar-L$ fragment is the molar mass of half of each L group and the full Ar group: $M_i = M_{L_{i1}}/2 + M_{Ar_i} + M_{L_{i2}}/2$. Since the same L groups are counted twice when building a repeat unit structure from a given set of fragment IDs, the product $\mathbf{X}\boldsymbol{M}$ encompasses the molar mass of the repeating unit for all polymers in the data set (or correspondingly the molar mass of a copolymer's repeating unit, averaged over its comonomer mass fractions).

GAP Model. We used ordinary least-squares (OLS) to estimate the p coefficients $\hat{\beta}_i$, i.e., the p-vector $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$, by minimizing the residual sum of squares

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{a=1}^{n} (T_g^a - f((\overline{\mathbf{X}})_a))^2 \right\} \tag{10}$$

where the linear fitting function f that approximates T_g^a (of the ath polymer in the training data) is given by $f((\overline{\mathbf{X}})_a) = \sum_{i=1}^{p} \overline{X}_{ai}\beta_i \equiv (\overline{\mathbf{X}}\boldsymbol{\beta})_a$. The solution, i.e., the least-squares estimate of β, is given by

$$\hat{\boldsymbol{\beta}} = (\overline{\mathbf{X}}^{\mathrm{T}}\overline{\mathbf{X}})^{-1}\overline{\mathbf{X}}^{\mathrm{T}}\boldsymbol{T}_g \tag{11}$$

where the n-vector $\boldsymbol{T}_g \in \mathbb{R}^n$ contains all n T_g^a values in the training sample. Note that each estimated coefficient $\hat{\beta}_i$ corresponds to the predicted glass transition temperature of a polymer solely comprising fragment i as its repeating monomer.

An out-of-sample prediction of the glass transition temperature for a polymer b with fragment composition $\overline{X}_{bi}$ (given all $i = 1, ..., p$) can now be determined as

$$\hat{T}_g^b = \sum_{i=1}^{p} \overline{X}_{bi}\hat{\beta}_i \tag{12}$$

Predictions of T_g^b are restricted to polymers consisting of fragments whose contributions β_i have already been estimated from eq 11, meaning that GAP predictions are chemically constrained to polymers consisting of fragments within the set $\{1, ..., p\}$.

QSPR−GAP Model. The QSPR−GAP model contains a *descriptor matrix* $D_{i\mu}$, which encodes the chemical and physical properties of the ith fragment ID in terms of $\mu = 1, ..., m$ descriptor values

$$(\mathbf{D})_{i\mu} \equiv D_{i\mu} \tag{13}$$

We then express each fragment T_g contribution β_i as a linear combination of the m descriptors

$$\beta_i = \gamma_0 + \sum_{\mu=1}^{m} D_{i\mu}\gamma_\mu \tag{14}$$

where the $(m+1)$-vector $\boldsymbol{\gamma} \in \mathbb{R}^{(m+1)}$ contains the regression coefficients. The zeroth column index is included in the matrix $\mathbf{D}$ as $D_{i0} = 1$ for all fragments $i = 1, ..., p$ to accommodate the constant term γ_0; therefore, $\mathbf{D} \in \mathbb{R}^{p\times(m+1)}$.

The methods used to estimate $\boldsymbol{\gamma}$ include (1) principal component regression, (2) ridge regression, (3) lasso regression, and (4) partial least-squares regression and are discussed further in the SI, Sec. S−II. However, to illustrate the application of the QSPR−GAP model in its simplest form, we discuss the genetic algorithm (GA) model here.

The GA uses concepts analogous to evolution to select an optimal subset ($m_{GA} \leq 10$) from a total of 213 potential descriptors. The descriptors chosen will have the greatest influence on T_g and, once selected, are included in the function

$$f((\overline{\mathbf{X}})_a, \mathbf{D}) = \sum_{i=1}^{p} \overline{X}_{ai}\beta_i \tag{15a}$$

$$= \gamma_0 + \sum_{i=1}^{p}\sum_{\mu=1}^{m_{GA}} \overline{X}_{ai}D_{i\mu}\gamma_\mu \tag{15b}$$

(since $\sum_{i=1}^{p}\overline{X}_{ai} = 1$), followed by a least-squares minimization

$$\hat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}} \left\{ \sum_{a=1}^{n} (T_g^a - f((\overline{\mathbf{X}})_a, \mathbf{D}))^2 \right\} \tag{16}$$

yielding the solution

$$\hat{\boldsymbol{\gamma}} = (\mathbf{D}^{\mathrm{T}}\overline{\mathbf{X}}^{\mathrm{T}}\overline{\mathbf{X}}\mathbf{D})^{-1}\mathbf{D}^{\mathrm{T}}\overline{\mathbf{X}}^{\mathrm{T}}\mathbf{T}_g \tag{17}$$

Once the coefficients $\boldsymbol{\gamma}$ are estimated from the training data, the estimated T_g contribution of any new fragment j is given by

$$\hat{\beta}_j = \hat{\gamma}_0 + \sum_{\mu=1}^{m_{GA}} D_{j\mu}\hat{\gamma}_\mu \tag{18}$$

https://doi.org/10.1021/acs.macromol.5c00178
Macromolecules XXXX, XXX, XXX−XXX

which may include fragments that are not in the original data sample.

An out-of-sample predicted glass transition temperature for a new polymer b given fragment compositions $\bar{X}_{bj}$ (for all $j = 1, ..., q$) and molecular descriptors $D_{j\mu}$ (for all $\mu = 1, ..., m_{GA}$) is

$$\hat{T}_g^b = \hat{\gamma}_0 + \sum_{j=1}^{q} \sum_{\mu=1}^{m_{GA}} \bar{X}_{bj} D_{j\mu} \hat{\gamma}_\mu \tag{19}$$

noting that if $\{1, ..., p\}$ is the set of in-sample fragments, then $\{1, ..., q\}$ is the set of in-sample and out-of-sample fragments, where $\{1, ..., p\} \subseteq \{1, ..., q\}$.

■ ASSOCIATED CONTENT

Data Availability Statement

Source data files are available at the University of Leeds Data Repository at 10.5518/1596.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.macromol.5c00178.

Description of the data set; QSPR feature generation including the conformer generation and descriptor calculations; background information on the machine learning models and algorithms used; description of the genetic algorithm implementation, model evaluation strategy, and cross validation; all fragment-level T_g contributions; and a few examples of atomic-level T_g contributions (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Johan Mattsson − *School of Physics and Astronomy, University of Leeds, Leeds LS2 9JT, U.K.*; orcid.org/0000-0001-7057-2192; Email: k.j.l.mattsson@leeds.ac.uk

Authors

Sebastian Brierley-Croft − *School of Physics and Astronomy, University of Leeds, Leeds LS2 9JT, U.K.*

Peter D. Olmsted − *Department of Physics and Institute for Soft Matter Synthesis and Metrology, Georgetown University, Washington, D.C. 20057, United States*; orcid.org/0000-0001-6538-9385

Peter J. Hine − *School of Physics and Astronomy, University of Leeds, Leeds LS2 9JT, U.K.*

Richard J. Mandle − *School of Physics and Astronomy, University of Leeds, Leeds LS2 9JT, U.K.; School of Chemistry, University of Leeds, Leeds LS2 9JT, U.K.*

Adam Chaplin − *Victrex PLC, Hillhouse International, Thornton Cleveleys, Lancashire FY5 4 QD, U.K.*

John Grasmeder − *Victrex PLC, Hillhouse International, Thornton Cleveleys, Lancashire FY5 4 QD, U.K.*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.macromol.5c00178

Author Contributions

S.B.C., P.D.O. and J.M. cowrote the manuscript. S.B.C. performed all of the data analysis. J.M. and P.D.O. supervised the project. All the authors reviewed and edited the manuscript and contributed to useful discussions.

Notes

The authors declare no competing financial interest.

Code Availability Statement: The codes that support the findings of this study are available upon reasonable request from the corresponding author.

■ ACKNOWLEDGMENTS

The authors acknowledge financial support from the Engineering and Physical Sciences Research Council (EPSRC) funded Centre for Doctoral Training in Soft Matter and Functional Interfaces (grant EP/L015536/1) and Victrex PLC. The authors thank Victrex PLC for the data they made available to conduct this study. P.D.O. thanks the Ives foundation and Georgetown University for support. R.J.M. thanks UKRI for funding via Future Leaders Fellowship (grant MR/W006391/1) and the University of Leeds for funding via a University Academic Fellowship.

■ REFERENCES

(1) Angell, C. A.; Ngai, K. L.; McKenna, G. B.; McMillan, P. F.; Martin, S. W. Relaxation in glassforming liquids and amorphous solids. *J. Appl. Phys.* **2000**, *88*, 3113−3157.

(2) Fox, T. G., Jr; Flory, P. J. Second-order transition temperatures and related properties of polystyrene. I. Influence of molecular weight. *J. Appl. Phys.* **1950**, *21*, 581−591.

(3) Cown, J.M.G. Some general features of Tg-M relations for oligomers and amorphous polymers. *Eur. Polym. J.* **1975**, *11*, 297−300.

(4) Novikov, V. N.; Sokolov, A. Universality of the dynamic crossover in glass-forming liquids: A magic relaxation time. *Phys. Rev. E* **2003**, *67*, No. 031507.

(5) Baker, D. L.; Reynolds, M.; Masurel, R.; Olmsted, P. D.; Mattsson, J. Cooperative Intramolecular Dynamics Control the Chain-Length-Dependent Glass Transition in Polymers. *Phys. Rev. X* **2022**, *12*, No. 021047.

(6) Bernabei, M.; Moreno, A. J.; Colmenero, J. Dynamic Arrest in Polymer Melts: Competition between Packing and Intramolecular Barriers. *Phys. Rev. Lett.* **2008**, *101*, No. 255701.

(7) Colmenero, J. Are Polymers Standard Glass-Forming Systems? The Role of Intramolecular Barriers on the Glass-Transition Phenomena of Glass-Forming Polymers. *J. Phys.:Condens. Matter* **2015**, *27*, No. 103101.

(8) Boyer, R. F. The Relation of Transition Temperatures to Chemical Structure in High Polymers. *Rubber Chem. Technol.* **1963**, *36*, 1303.

(9) Bershtein, V. A.; Egorov, V.; Podolsky, A.; Stepanov, V. Interrelationship and Common Nature of the β Relaxation and the Glass Transition in Polymers. *J. Polym. Sci., Polym. Lett. Ed.* **1985**, *23*, 371.

(10) Boyer, R. F. Mechanical Motions in Amorphous and Semi-Crystalline Polymers. *Polymer* **1976**, *17*, 996.

(11) Boyd, R. H. Relaxation processes in crystalline polymers: molecular interpretation - a review. *Polymer* **1985**, *26*, 1123.

(12) Boyd, R. H. Relaxation Processes in Crystalline Polymers: Experimental Behaviour - A review. *Polymer* **1985**, *26*, 323.

(13) Ngai, K. L.; Plazek, D. Relation of internal rotational isomerism barriers to the flow activation energy of entangled polymer melts in the high-temperature Arrhenius region. *J. Polym. Sci. and Polym. Phys. Ed.* **1985**, *23*, 2159.

(14) Roland, C. M.; Hensel-Bielowka, S.; Paluch, M.; Caslini, R. Supercooled dynamics of glass-forming liquids and polymers under hydrostatic pressure. *Rep. Prog. Phys.* **2005**, *68*, 1405−1478.

(15) Schneider, H. A.; Di Marzio, E. A. The glass temperature of polymer blends: comparison of both the free volume and the entropy predictions with data. *Polymer* **1992**, *33*, 3453−3461.

(16) Matsuoka, S. Entropy, Free Volume, and Cooperative Relaxation. *J. Res. Natl. Inst. Stand. Technol.* **1997**, *102*, 213.

(17) Schneider, H. A. Polymer class specificity of the glass temperature. *Polymer* **2005**, *46*, 2230−2237.

(18) Schut, J.; Bolikal, D.; Khan, I.; Pesnell, A.; Rege, A.; Rojas, R.; Sheihet, L.; Murthy, N.; Kohn, J. Glass transition temperature prediction of polymers through the mass-per-flexible-bond principle. *Polymer* **2007**, *48*, 6115−6124.

(19) Xie, R.; Weisen, A. R.; Lee, Y.; Aplan, M. A.; Fenton, A. M.; Masucci, A. E.; Kempe, F.; Sommer, M.; Pester, C. W.; Colby, R. H.; Gomez, E. D. Glass transition temperature from the chemical structure of conjugated polymers. *Nat. Commun.* **2020**, *11*, No. 893.
(20) Alesadi, A.; Cao, Z.; Li, Z.; Zhang, S.; Zhao, H.; Gu, X.; Xia, W. Machine learning prediction of glass transition temperature of conjugated polymers from chemical structure. *Cell Rep. Phys. Sci.* **2022**, *3*, No. 100911.
(21) Pilania, G.; Iverson, C. N.; Lookman, T.; Marrone, B. L. Machine-Learning-Based Predictive Modeling of Glass Transition Temperatures: A Case of Polyhydroxyalkanoate Homopolymers and Copolymers. *J. Chem. Inf. Model.* **2019**, *59*, 5013−5025.
(22) Boyer, R. F. The Relation of Transition Temperatures to Chemical Structure in High Polymers. *Rubber Chem. Technol.* **1963**, *36*, 1303−1421.
(23) Weyland, H.; Hoftyzer, P.; Van Krevelen, D. Prediction of the glass transition temperature of polymers. *Polymer* **1970**, *11*, 79−87.
(24) van Krevelen, D. W. *Properties of Polymers*; Elsevier: Amsterdam, 1990; Vol. 3, p xi.
(25) Katritzky, A. R.; Sild, S.; Lobanov, V.; Karelson, M. Quantitative structure- property relationship (QSPR) correlation of glass transition temperatures of high molecular weight polymers. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 300−304.
(26) Hopfinger, A. J.; Koehler, M. G.; Pearlstein, R. A.; Tripathy, S. K. Molecular modeling of polymers. IV. Estimation of glass transition temperatures. *J. Polym. Sci., Part B:Polym. Phys.* **1988**, *26*, 2007−2028.
(27) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley, 2000.
(28) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; Wiley, 2009.
(29) Bejagam, K. K.; Lalonde, J.; Iverson, C. N.; Marrone, B. L.; Pilania, G. Machine Learning for Melting Temperature Predictions and Design in Polyhydroxyalkanoate-Based Biopolymers. *J. Phys. Chem. B* **2022**, *126*, 934−945.
(30) Le, T.; Epa, V. C.; Burden, F. R.; Winkler, D. A. Quantitative Structure-Property Relationship Modeling of Diverse Materials Properties. *Chem. Rev.* **2012**, *112*, 2889−2919.
(31) Sanchez-Lengeling, B.; Roch, L. M.; Perea, J. D.; Langner, S.; Brabec, C. J.; Aspuru-Guzik, A. A Bayesian Approach to Predict Solubility Parameters. *Adv. Theory Simul.* **2019**, *2*, No. 1800069, DOI: 10.1002/adts.201800069.
(32) Parandekar, P. V.; Browning, A. R.; Prakash, O. Modeling the flammability characteristics of polymers using quantitative structure-property relationships (QSPR). *Polym. Eng. Sci.* **2015**, *55*, 1553−1559.
(33) Duchowicz, P. R.; Fioressi, S. E.; Bacelo, D. E.; Saavedra, L. M.; Toropova, A. P.; Toropov, A. A. QSPR studies on refractive indices of structurally heterogeneous polymers. *Chemom. Intell. Lab. Syst.* **2015**, *140*, 86−91.
(34) Gasteiger, J. *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley, 2003.
(35) Katritzky, A. R.; Kuanar, M.; Slavov, S.; Hall, C. D.; Karelson, M.; Kahn, I.; Dobchev, D. A. Quantitative correlation of physical and chemical properties with chemical structure: utility for prediction. *Chem. Rev.* **2010**, *110*, 5714−5789.
(36) Kim, C.; Chandrasekaran, A.; Huan, T. D.; Das, D.; Ramprasad, R. Polymer Genome: A Data-Powered Polymer Informatics Platform for Property Predictions. *J. Phys. Chem. C* **2018**, *122*, 17575−17585.
(37) Jha, A.; Chandrasekaran, A.; Kim, C.; Ramprasad, R. Impact of dataset uncertainties on machine learning model predictions: the example of polymer glass transition temperatures. *Modell. Simul. Mater. Sci. Eng.* **2019**, *27*, No. 024002.
(38) Volgin, I. V.; Batyr, P. A.; Matseevich, A. V.; Dobrovskiy, A. Y.; Andreeva, M. V.; Nazarychev, V. M.; Larin, S. V.; Goikhman, M. Y.; Vizilter, Y. V.; Askadskii, A. A.Machine learning with enormous synthetic data sets: predicting glass transition temperature of polyimides using graph convolutional neural networks *ACS Omega* 2022; Vol. 7, pp 43678−43691.
(39) Huan, T. D.; Mannodi-Kanakkithodi, A.; Kim, C.; Sharma, V.; Pilania, G.; Ramprasad, R. A polymer dataset for accelerated property prediction and design. *Sci. Data* **2016**, *3*, No. 160012.
(40) Otsuka, S.; Kuwajima, I.; Hosoya, J.; Xu, Y.; Yamazaki, M.*PoLyInfo: Polymer Database for Polymeric Materials Design*, 2011 International Conference on Emerging Intelligent Data and Web Technologies, 2011; pp 22−29.
(41) Kuenneth, C.; Ramprasad, R. polyBERT: a chemical language model to enable fully machine-driven ultrafast polymer informatics. *Nat. Commun.* **2023**, *14*, No. 4099.
(42) Kemmish, D.Update on the technology and applications of polyaryletherketones. In *ISmithers* 2010; p 142.
(43) Song, C.; Gu, H.; Zhu, L.; Jiang, W.; Weng, Z.; Zong, L.; Liu, C.; Hu, F.; Pan, Y.; Jian, X. A polymer genome approach for rational design of poly(aryl ether)s with high glass transition temperature. *J. Mater. Chem. A* **2023**, *11*, 16985−16994.
(44) Fougnies, C.; Dosière, M.; Koch, M. H. J.; Roovers, J. Morphological Study and Melting Behavior of Narrow Molecular Weight Fractions of Poly(aryl ether ether ketone) (PEEK) Annealed from the Glassy State. *Macromolecules* **1998**, *31*, 6266−6274.
(45) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.
(46) *RDKit: Open-Source Cheminformatics*, 2023 https://www.rdkit.org DOI: 10.5281/zenodo.10893044.
(47) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminf.* **2018**, *10*, No. 4.
(48) Schuur, J. H.; Selzer, P.; Gasteiger, J. The Coding of the Three-Dimensional Structure of Molecules by Molecular Transforms and Its Application to Structure-Spectra Correlations and Studies of Biological Activity. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 334−344.
(49) Schuur, J.; Gasteiger, J. Infrared Spectra Simulation of Substituted Benzene Derivatives on the Basis of a 3D Structure Representation. *Anal. Chem.* **1997**, *69*, 2398−2405.
(50) Devinyak, O.; Havrylyuk, D.; Lesyk, R. 3D-MoRSE descriptors explained. *J. Mol. Graphics Modell.* **2014**, *54*, 194−203.
(51) Firth, N. C.; Brown, N.; Blagg, J. Plane of Best Fit: A Novel Method to Characterize the Three-Dimensionality of Molecules. *J. Chem. Inf. Model.* **2012**, *52*, 2516−2525.
(52) Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc., Ser. B* **1996**, *58*, 267−288.
(53) Wegelin, J. A. *A Survey of Partial Least Squares (PLS) Methods, with Emphasis on the Two-Block Case*, 2000https://stat.uw.edu/research/tech-reports/survey-partial-least-squares-pls-methods-emphasis-two-block-case.
(54) Mattioni, B. E.; Jurs, P. C. Prediction of Glass Transition Temperatures from Monomer and Repeat Unit Structure Using Computational Neural Networks. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 232−240.
(55) Sanderson, R. T. Electronegativity and bond energy. *J. Am. Chem. Soc.* **1983**, *105*, 2259−2261.

https://doi.org/10.1021/acs.macromol.5c00178
Macromolecules XXXX, XXX, XXX−XXX