THE JOURNAL OF PHYSICAL CHEMISTRY B

pubs.acs.org/JPCB Article

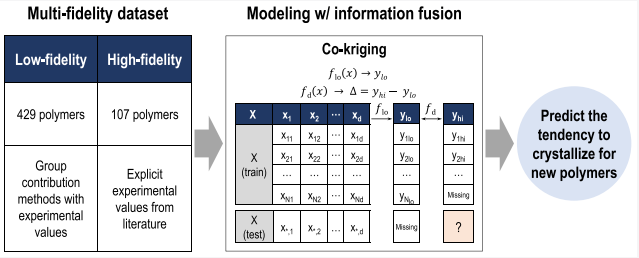
Predicting Crystallization Tendency of Polymers Using Multifidelity Information Fusion and Machine Learning

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

Shruti Venkatram, Rohit Batra, Lihua Chen, Chiho Kim, Madeline Shelton, and Rampi Ramprasad*

Cite This: https://dx.doi.org/10.1021/acs.jpcb.0c01865 Read Online

ABSTRACT: The degree of crystallinity of a polymer is a critical parameter that controls a variety of polymer properties. A high degree of crystallinity is associated with excellent mechanical properties crucial for high-performing applications like composites. Low crystallinity promotes ion and gas mobility critical for battery and membrane applications. Experimental determination of the crystallinity for new polymers is time and cost intensive. A data-driven machine learning-based method capable of rapidly predicting the crystallinity could counter these disadvantages and be used to screen polymers for a myriad of applications in a fast, inexpensive fashion. In this work, we developed the first-of-its-kind, data-driven machine learning model to predict the most-likely polymer crystallinity trained on experimental data and theoretical group contribution methods. Since polymer data under consistent processing conditions are unavailable, we tackled process variability by using the "most-likely" polymer values which we refer to as the polymer's tendency to crystallize. Experimental data for polymers' tendency to crystallize is limited by number and diversity, and to tackle this, we augmented experimentation-based data with data using group contribution methods. Therefore, this work utilized two data sets, viz., a high-fidelity, experimental data set for 107 polymers and a more diverse, less accurate low-fidelity data set for 429 polymers which used group contribution methods. We used a multifidelity information fusion strategy to utilize all the information captured in the low-fidelity data set while still predicting at the high-fidelity accuracy. Although this model inherently assumed "typical" processing conditions and estimated the "most-likely" percent crystallinity value, it can help in the estimation of a polymer's tendency to crystallize in a far more cost-effective and efficient manner.



INTRODUCTION

In the 1940s, a seminal development in polymer science established that polymers are semicrystalline, in that they have well-ordered crystalline domains and distinctly less ordered amorphous domains.[1] This initiated the concept of the degree of crystallinity of semicrystalline polymers, which is the percent fraction of the polymer that exists as an ordered state. The degree of crystallinity is a critical kinetics-driven parameter reflective of a polymer's diverse structure−property relationships. It is strongly correlated to crucial mechanical properties which are important for high-performance structural applications.[2−4] Lower polymer crystallinity is associated with increased ion mobility and gas diffusivity; these are useful for battery and membrane applications.[5−9]

The percent crystallinity of a polymer has associated process-dependent variability and measurement-related uncertainties. The process variability mainly arises from intrinsic factors like the molecular weight, number of side chains, polydispersity index, extrinsic processing conditions like the extrusion technique, form of the polymer (film, fiber, etc.), temperature, and pressure as well as postprocessing techniques like annealing and drawing. Since the percent crystallinity impacts numerous critical polymer properties, several experimental and theoretical techniques with varying fidelity levels have been developed to arrive at a percent crystallinity value for a polymer. This, therefore, leads to different percent crystallinity values for a single polymer sample (whence process variability is not an issue) leading to uncertainties in the crystallinity values.

Some experimental methods like X-ray diffraction (XRD) and nuclear magnetic resonance (NMR) directly capture the degree of crystallinity in 3-dimensional order. However, most experimental methods calculate the degree of crystallinity via

Received: March 2, 2020
Revised: June 12, 2020
Published: June 15, 2020

https://dx.doi.org/10.1021/acs.jpcb.0c01865
J. Phys. Chem. B XXXX, XXX, XXX−XXX

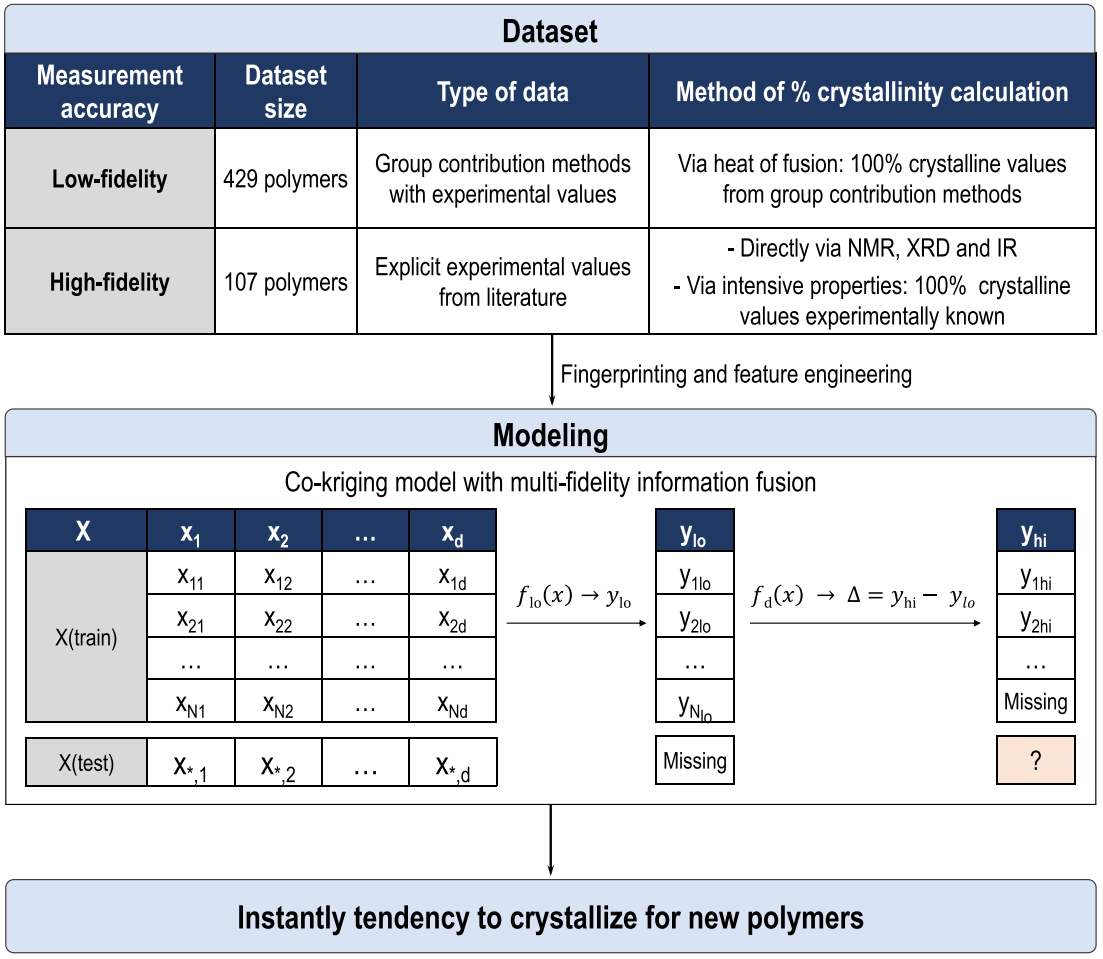


Figure 1. Workflow schematic used to instantly predict the crystallization tendencies for new polymers.

related intensive properties namely specific volume, specific heat, specific enthalpy, specific enthalpy of fusion and density. For a given intensive property P, these methods define the degree of crystallinity = $(P_a - P)/(P_a - P_c)$, where P_a and P_c are the intensive properties for the purely amorphous and purely crystalline components of the polymer.[10,11]

A common complication in methods which use intensive properties to measure its degree of crystallinity is the unavailability of the pure-component properties (P_a and P_c).[12] A common resolution to this complication is using group contribution methods. Group contribution methods have been widely used to estimate pure-component properties for a variety of intensive polymer properties. They assume that any property is a sum of contributions made by the constituent functional groups. This assumption makes them necessarily approximate since the contribution made by one group is assumed to be independent of another. Despite these drawbacks, group contribution methods can be used to estimate pure component properties for a large number of polymers, albeit with varying accuracy.[13]

Despite several methods for determining the polymer crystallinity, estimating the crystallinity values for a new polymer is not trivial. Direct experimental methods require the necessary infrastructure and tedious reproducible sample preparation to estimate the polymer crystallinity making these methods expensive and time-consuming. Group contribution methods, although less time-consuming, are limited by the constituent functional groups for which reliable component parameters exist.

Recently, data-driven efforts in materials science have been fairly successful and have paved a path to instantly predicting pertinent material properties.[14−17] However, previous work to predict polymer crystallinity using data-driven methods have been limited and no data-driven methods have been used to predict the percent crystallinity for a broad polymer chemical space.[18,19] As a result of the aforementioned process variability and uncertainties in the percent crystallinity values, a consistent and uniform data set which is unavailable for a number of polymers. Therefore, data-driven predictions of the tendency of a given polymer to crystallize to certain degree is presently unavailable.

In this work, we established a workflow to predict the most-likely degree of polymer crystallinity which we refer to as the polymer's tendency to crystallize using a diverse data set from multiple sources. We also successfully demonstrated a methodology for tackling the measurement-related uncertainties commonly encountered in percent crystallinity calculations. However, as mentioned above, there are several sources for process-related variability of the percent crystallinity and therefore, percent crystallinity values are unavailable for consistent process-related factors for several polymers. To circumvent this, we analyze process-related variability on percent crystallinity for 30 different polymers. We established that one may define a polymer's tendency to crystallize

https://dx.doi.org/10.1021/acs.jpcb.0c01865
J. Phys. Chem. B XXXX, XXX, XXX−XXX

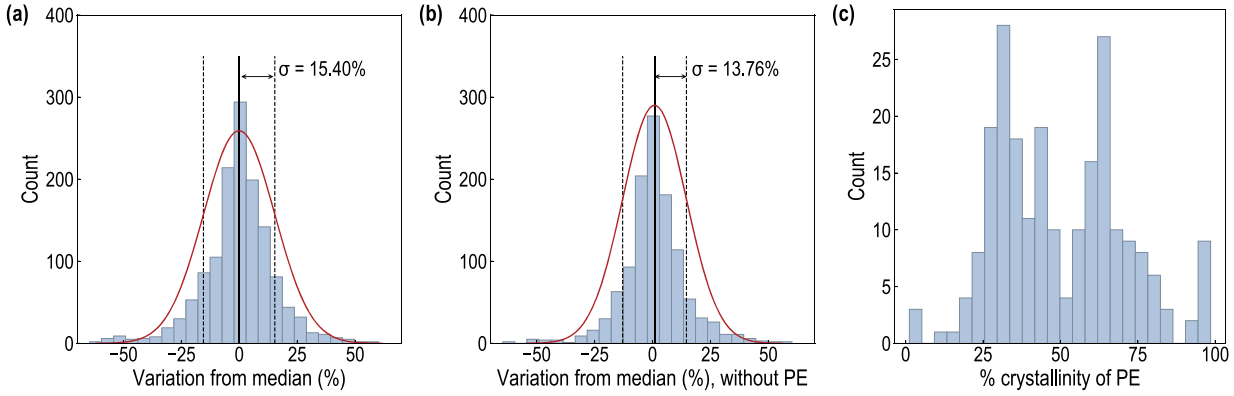


Figure 2. Capturing the variability of the high-fidelity data set due to polymer processing. (a) Normal distribution fit for variation from median for 30 polymers with multiple reported values of percent crystallinity. (b) Normal distribution fit for variation from median for 29 polymers, without polyethylene with multiple reported values of percent crystallinity. (c) Percent crystallinity values of polyethylene for 229 distinct samples.

corresponding to the median of the spectrum of polymer crystallinity values.

To effectively tackle and utilize measurement-related uncertainties, we categorized our data set into two types—high-fidelity and low-fidelity (see Figure 1). In the high-fidelity data set, the crystallization tendencies of the polymers are highly accurate, comprising of either values obtained directly via experiments or obtained via intensive properties like the heat of fusion and density measurements for polymers where pure-component (100% crystalline) values have been experimentally established. However, the high-fidelity data set comprised of just 107 polymers with limited chemical diversity. The low-fidelity data set was developed via approximate group contribution methods to overcome the diversity limitations of the high-fidelity data set. This data set comprised of polymers whose crystallization tendencies were calculated using heat of fusion values. The observed heat of fusion was obtained from experimental measurements and the 100% crystalline heat of fusion values were calculated using group contribution methods.[13] This data set comprised of 429 polymers and was far more diverse than the high-fidelity data set. However, the use of group contribution methods to develop this data set makes its accuracy lower than the high-fidelity data set.

To estimate the tendency to crystallize for new polymers, we have developed multifidelity cokriging model based on a fusion of the high and low-fidelity data sets. Details of the data sets and a schematic of the workflow for the models are depicted in Figure 1. Our results suggest that fusing group contribution methods with a sparse high-fidelity experimentally obtained data for polymers can be used successfully to make predictive models. Also, this methodology can be extended to a variety of polymer properties like the glass transition temperature, the entropy, and the density. The ability of the cokriging model to make predictions for a new polymer's tendency to crystallize based entirely on its features and without the corresponding low-fidelity estimate makes it an extremely useful and versatile method. Although this model inherently assumes "typical" processing conditions and estimates the "most-likely" percent crystallinity value occurs, it can help in the estimation of a polymer's tendency to crystallize in a far more cost-effective and efficient manner.

DATA SETS

Data Set Variability: Polymer Processing. As mentioned previously, the process-related variability is a factor while estimating the percent crystallinity for a polymer. While a uniform and consistent data set which encapsulates a subset of process parameters is unavailable, we can analyze the process-related variability for a selected number of polymers in our data set. Here, we evaluated the spread in the high-fidelity polymer crystallinity values for 30 different polymers over 1375 distinct samples.[13,20−24] For each polymer, the percent crystallinity is evaluated using experimental heat of fusion ΔH_{fus} values as

$$\%\ \text{crystallinity} = \frac{\Delta H_{fus,observed}}{\Delta H_{fus,100\%crystalline}} \times 100 \quad (1)$$

where the 100% crystalline heat of fusion values for all 30 polymers have been accurately experimentally determined by Wunderlich et al.[25] and are, therefore, high-fidelity values.

For these 30 polymers, a histogram for their deviations from their respective median measurements is shown in Figure 2a. From this plot, it is evident that the processing-related variability of the data set follows a normal distribution with a standard deviation of 15.40% with a peak at 0% which represents the median value. Polyethylene constituted around 16.5% of the sample set size. Therefore, we performed the analysis again, without polyethylene, and found that the noise still follows a normal distribution but the standard deviation drops to 13.76% (Figure 2b). The large variation in the percent crystallinity values of polyethylene can be attributed to the fact that low-density and high-density polyethylene are widely studied in literature and have distinct structural differences. The bimodal nature of the percent crystallinity values of polyethylene is can be attributed to this (Figure 2c), and it is unlike that for the other polymers (Figure S1) whose variance followed a unimodal normal distribution.

This analysis results in two important results. First, the normal distribution of parts a and b of Figure 2 implies that there is a "most-likely" value of the percent crystallinity (equal to its median) for polymers which exhibit multiple values. Therefore, choosing the median percent crystallinity value to represent each polymer in the data set is justified. We will refer to this value as the polymer's tendency to crystallize for the remainder of this work. Furthermore, this analysis quantified

https://dx.doi.org/10.1021/acs.jpcb.0c01865
J. Phys. Chem. B XXXX, XXX, XXX−XXX

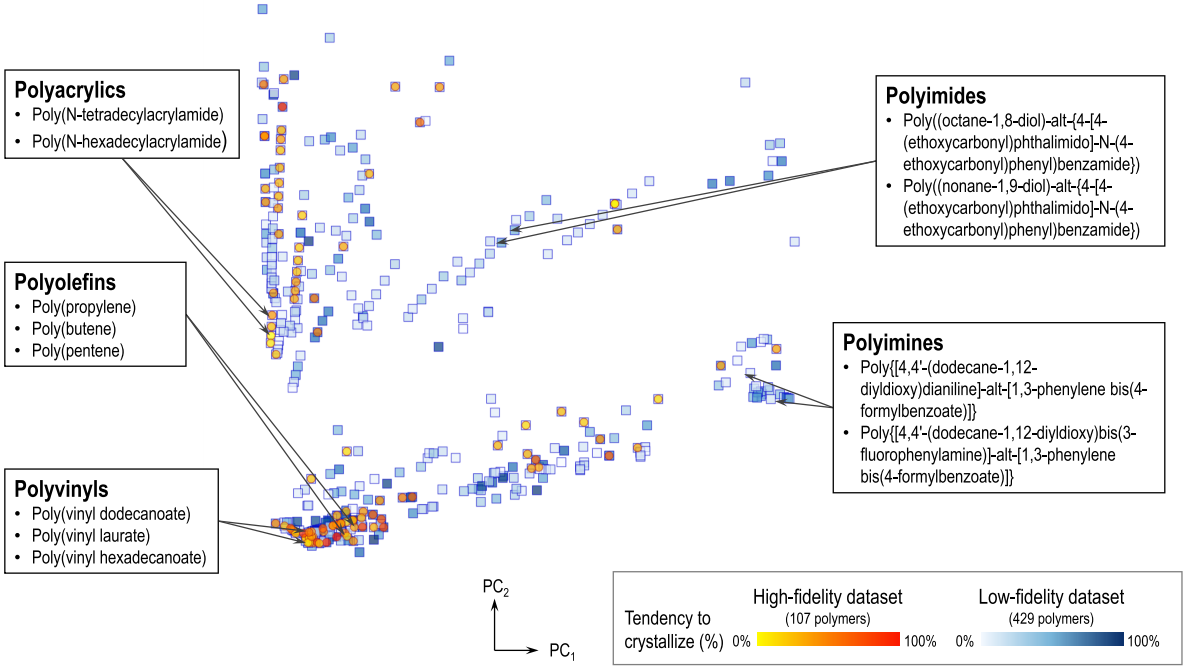


Figure 3. Graphical summary of the high and low fidelity data sets against the chemical space of polymers. Two leading components, PC1 and PC2, are produced by principal component analysis, and assigned to axes of the plot.

the processing-related variability for polymer crystallinity to about 13−15%. This also gives us a best-case estimate of the expected error from machine learning models that use only high-fidelity data.

High- and Low-Fidelity Data Sets. The data sets used in this study belong to two categories—high- and low-fidelity. The high-fidelity data set comprises of accurate and explicit experimental values where the variance in the percent crystallinity values for a polymer is only due to its processing and experimental measurement heterogeneity. The high-fidelity data set comprises of 107 polymers with associated percent crystallinity values. All values are curated from existing sources of experimental measurements like handbooks, published papers and online sources.[13,20−24] The percent crystallinity values in this data set comprise of either explicit percent crystallinity values via X-ray diffraction (XRD) and nuclear magnetic resonance (NMR) or are obtained via intensive properties like the heat of fusion and density measurements for polymers where pure-component (100% crystalline) values have been experimentally established. For cases where multiple values were reported, we used the median percent crystallinity to train machine learning models. This value corresponds to the polymer's tendency to crystallize at a high-fidelity level.

The low-fidelity data set comprises of percent crystallinity values estimated using a combination of experimental and group-contribution methods. Therefore, the variance in this data set arises from not just processing and experimental heterogeneity but also the inherent uncertainty of the group-contribution method. This makes the accuracy of this data set lower than the high-fidelity data set. The low-fidelity data set includes 429 polymers. For this data set, the percent crystallinity value for each polymer is calculated using eq 1 described above. Here, the heat of fusion $\Delta H_{fus,\ observed}$ is curated from differential scanning calorimetry (DSC) experiments from a variety of existing sources like polymer handbooks and prior published works. For polymers where multiple experimental heat of fusion values were reported, we used the median value to calculate the percent crystallinity and train machine learning models. This value is the polymer's tendency to crystallize at the low-fidelity level. However, the $\Delta H_{fus,\ 100\%\ crystalline}$ is not available for all polymers. Therefore, we used group contribution methods using values established by Van Krevelen[13] to calculate the $\Delta H_{fus,\ 100\%\ crystalline}$ values for this data set. Additionally, we believe the spread of the calculated low-fidelity values to be similar to the distribution shown in Figure 2 since the heat of fusion directly correlates to the polymer's tendency to crystallize. For all polymers whose high-fidelity value is known, the low-fidelity value is also available.

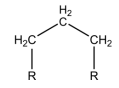
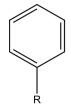
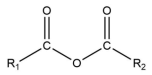
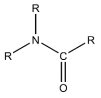
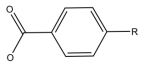
To demonstrate the diversity of the data sets, we compared both data sets using principal component analysis (PCA) (Figure 3). Since the high-fidelity data set is a subset of the low-fidelity data set, we performed a principal component analysis (PCA) of the low-fidelity fingerprint vector (explained in the Methods section below). The horizontal and vertical axes are the first two principal components, PC_1 and PC_2. The high-fidelity data set is less diverse than the low-fidelity data set—the high diversity data set includes polymers that belong to the class of polyolefins, polyvinyls, and polyacrylics. However, the low-fidelity data set also includes polymers in the polyimines and polyimides chemical space where the high-fidelity data set is absent.

METHODS

Feature Set and Dimensionality Reduction. A hierarchical fingerprinting method was used to capture descriptors that control the tendency to crystallize in polymers and was described in a previous work.[14] The fingerprinting scheme comprises of four hierarchical levels of descriptors. The first level is at the atomic scale comprising of the count of atomic triples (e.g., O2−C3−C4, representing of a 2-fold

https://dx.doi.org/10.1021/acs.jpcb.0c01865
J. Phys. Chem. B XXXX, XXX, XXX−XXX

Table 1. Representative Features Strongly Correlated to the Crystallization Tendency of Polymers[a]

Correlation with tendency to crystallize	Representative features		
Positive correlation	Atomic and block level	 (Linear CH_2 groups)	 (Phenyl)  (Carbonyl)  (Anhydride)
	Chain level	Relative number of atoms in the main chain	
	Special features	Melting temperature (T_m)	
Negative correlation	Atomic and block level (in side chains)	 (Amide)	 (Sulfone)  (Bulky pendant groups)  (Terminal CH_3 group)
	Chain level	Length of longest side chain	

[a]R represents an arbitrary chemical group of C, O, N or H elements.[28]

coordinated oxygen, a 3-fold coordinated carbon, and a 4-fold coordinated carbon). The second set of fingerprint components captures a population of predefined chemical building blocks (e.g., $-C_6H_4-$, $-CH_2-$, $-C(=O)-$). The third hierarchical level comprises of quantitative structure−property relationship (QSPR) descriptors, such as van der Waals surface area, topological surface area and the fraction of rotatable bonds, implemented in the RDKit cheminformatics library. The fourth and last fingerprinting level includes morphological features such as the topological distance between rings, fraction of atoms that are part of side chains and length of the largest side chain. Additionally, the melting temperature of each polymer was added as an additional fingerprint given its positive correlation with the heat of fusion (ΔH_{fus}).[26] For new polymers whose melting temperature is unknown, they can be predicted using the machine-learning based model implemented at www.polymergenome.org. Consequently, this feature space includes 256 features.

To retain only relevant features, LASSO (least absolute shrinkage and selection operator) was performed on the initial 256-dimensional feature vector and both, the high and low-fidelity data sets with 5-fold cross-validation.[27] LASSO is a shrinkage and selection method for linear regression which minimizes the usual sum of squared errors, with a bound on the sum of the absolute values of the coefficients. Since a single-fidelity GPR model is trained on the high-fidelity data set as a baseline and a multifidelity cokriging model is trained using data sets where the low-fidelity data set is larger (with 429 polymers), feature engineering was performed on both data sets. For the high-fidelity data set, LASSO retains 14 pertinent features (including the melting temperature), which are then used to train the single-fidelity GPR model. For the low-fidelity data set, LASSO retained 97 important features (including the melting temperature) which were used to train the multifidelity cokriging model. LASSO retains more features on the low-fidelity data set since it is larger and more chemically diverse than the high-fidelity data set.

Factors Affecting the Crystallization Tendency of Polymers. In addition to performing feature elimination to retain only relevant features, it is also valuable to analyze the retained features and its correlation to the tendency to crystallize. In this analysis, we study the 97 features retained after performing feature elimination on the low-fidelity data set since the data set is more diverse and representative of the polymer chemical space. The representative features are tabulated in Table 1 and their corresponding coefficients which determine their correlations are summarized in Table S1. As expected, there are positive correlations to atomic and block-level features like phenyl, carbonyl, and anhydride groups that are known to increase the chain stiffness and, therefore, its crystallization tendencies.[28] However, an interesting observation was that chain stiffening groups promote crystallization only when they are present in the main chain. In our data set, chain stiffening groups like amide and sulfones were mostly present in the side chain (in polymers like acrylamides and certain sulfur-containing polyoxides). Since the crystallization is a function of the packing of the polymer, these functional groups were negatively correlated since the presence of chain stiffening functional groups reduce the packing including the aforementioned amides, sulfones as well as bulky ring-containing side groups. Since the majority of terminal CH_3 groups are present in side chains, this group is also negatively correlated to the tendency to crystallize. Further, there are correlations to block and chain-level features which propagated greater packing. For instance, it is positively correlated to the linear CH_2 chains, the length of the main chain (relative to the side chain) and negatively correlated to the length of the side chains (since they enhance branching, reducing the crystallinity). Additionally, the melting temperature is known to increase with the increase in the crystallinity,[26] and we see a strong positive correlation between them. These chemical guidelines can help in rationally designing polymers with desirable crystallization tendencies.

Machine Learning Models. The single-fidelity machine learning model utilizes Gaussian process regression (GPR) with a rational quadratic kernel to map the high-fidelity data set to its tendency to crystallize.[29] GPR uses a Bayesian

https://dx.doi.org/10.1021/acs.jpcb.0c01865
J. Phys. Chem. B XXXX, XXX, XXX−XXX

framework, wherein a Gaussian process is used to obtain the mapping from the polymer to its associated tendency to crystallize based on the available training set and the Bayesian prior, incorporated using the kernel function. In this case, the kernel function between two materials with features x and x' is given by

$$k(x, x') = \sigma^2\left(1 + \frac{(x - x')^2}{2\alpha l^2}\right)^{-\alpha}$$

Here, the three hyperparameters σ, l, and α are the variance, the length-scale parameter, and the expected noise in the data, respectively. These hyper parameters were determined during the training of the models by maximizing the log-likelihood. Further, 5-fold cross validation was adopted to avoid overfitting.

Multifidelity analysis is closely related to machine learning methods like multitask learning, which relies on learning correlations among different material properties.[30−33] The multifidelity information fusion model uses a cokriging method to effectively utilize the high and low fidelity data sets simultaneously.[16,34] The flexibility of the cokriging approach allows it to have a variable number of low and high-fidelity data points. For all cases whose high-fidelity value is known, the respective low-fidelity value is also available. In analogy to GPR, the CK model assumes the high-fidelity data to be a realization of the Gaussian process Z_{hi}, which is further defined as the sum of a low-fidelity process Z_{lo} scaled by a factor ρ plus another independent Gaussian process Z_d which captures the difference between the available low- and high-fidelity data points. Therefore,

$$Z_{hi}(x) = \rho Z_{lo}(x) + Z_d(x) \quad (2)$$

The root-mean-square error (RMSE and R^2) values were used to evaluate and compare the performances of the GPR and the cokriging models. To estimate the prediction errors on unseen data, learning curves were generated by varying the size of the training and the test sets. For all ML models, i.e., GPR and cokriging, the prediction accuracy was computed on a completely unseen and randomly chosen test set consisting of 22 data points (20% of the high-fidelity data set). Additionally, for each case, statistically meaningful results were obtained by averaging RMSE results over 50 runs with varying training and test splits.[14,16]

RESULTS AND DISCUSSION

Correlations between the Low- and High-Fidelity Data Sets. Group contribution methods have been commonly used to calculate a variety of intensive properties;[13] however, there has been no reported work on its prediction capability for percent crystallinity. For this work, all polymers in the high-fidelity data set have associated low-fidelity values calculated using 100% crystalline heat of fusion values using group contribution methods. Figure 4 depicts a parity plot of 107 polymers and its associated high-fidelity and the low-fidelity values. The associated RMSE and R^2 are calculated as shown in Figure 4.

This analysis yields two results. First, this analysis shows that for this limited data set, the high- and low-fidelity data sets are quite closely correlated. The group contribution methods work especially well for polyolefins and polyvinyls. However, for polyesters and polyamides, the heat of fusion values waver from the experimental values as the complexity of the polymer

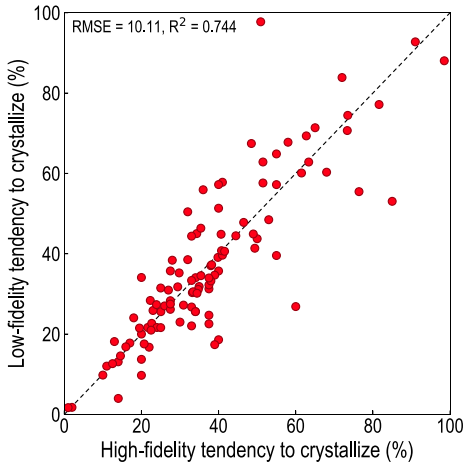


Figure 4. Prediction capability of the group contribution method for crystallization tendencies as a function of the high- and low-fidelity data sets.

(and thus, the number of groups) increases. Second, the good agreement between the high- and low-fidelity data sets make them excellent candidates for multifidelity based machine learning using cokriging. The cokriging method utilizes the difference between the high- and low-fidelity values and a good correlation between the two values is beneficial to this model. Additionally, this quantifies the uncertainty error from the different methods of measurements.

Machine Learning Models: Performance and Comparison. The high-fidelity data set and a combination of the high- and low-fidelity data sets were used to train the machine learning models. The performance of machine learning models is evaluated using learning curves with varying test-train splits. The learning curve represents the learning trend of the model upon adding more data. For all the machine learning models in this work, the test set comprises of 22 points, which is around 20% of the high-fidelity data set. Using only the high-fidelity data set, a Gaussian process regression (GPR) model was trained using 14 pertinent features selected using LASSO. Using a combination of the high- and the low-fidelity data sets, cokriging multifidelity machine learning models were trained using 97 important features selected using LASSO. The first multifidelity model assumes the unavailability of the low-fidelity data for the test set, and therefore comprises of 407 low-fidelity points. The second multifidelity model accounts for the case where the low-fidelity values of the test set is available, and therefore uses all the 429 low-fidelity data points. The performance of the machine learning models can be evaluated from the learning curves presented in Figure 5a, wherein average RMSE on the training and the test sets as a function of training set size are included. The error bars denote the 1σ deviation in the reported RMSE values over 50 runs. Parts b and c of Figures 5 are parity plots that depict the prediction performance on individual cases included in the training and test sets for the single-fidelity GPR model and multifidelity cokriging model with 407 multifidelity points which uses 85 training points and 22 test points. Along with learning curves, parity plots are a valuable comparison tool, in this case between the single and multifidelity model.

From Figure 5a, we can make several pertinent conclusions. First, for all train-test combinations (Figure 5a−c), the cokriging multifidelity models outperforms the GPR single-fidelity model. The average RMSE of the GPR single-fidelity

https://dx.doi.org/10.1021/acs.jpcb.0c01865
J. Phys. Chem. B XXXX, XXX, XXX−XXX

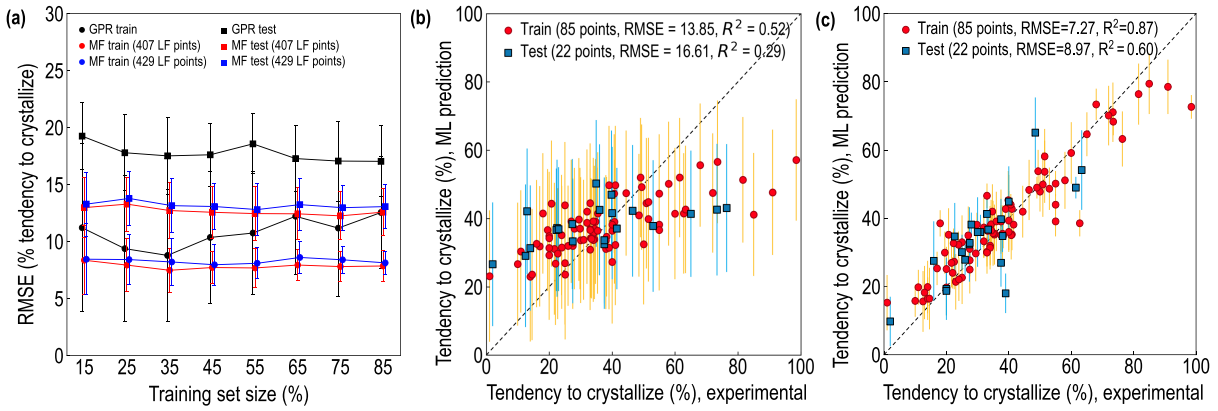


Figure 5. Prediction accuracy for machine learning models trained using single-fidelity GPR and multifidelity cokriging methods. Part a comprises learning curves for the single-fidelity GPR model and the multifidelity models with 407 and 429 low-fidelity points where the test sets consists of 22 randomly selected points and averaged over 50 runs, with error bars illustrating 1σ deviation. Parts b and c illustrate example parity plots with 107 high-fidelity train points and 22 test points for the GPR and cokriging models, respectively. The cokriging model uses 429 low-fidelity points.

model which uses 85 training points is 17.04%. In comparison, the multifidelity models with 407 and 429 low-fidelity points has an average RMSE of 12.58% and 13.06% respectively. The performance of both the multifidelity models are quite similar, but as expected, the standard deviation for the model with 429 low-fidelity points is lower than that of the model with 407 low-fidelity points. The fairly consistent performance of the models can be attributed to the limited chemical diversity of the high-fidelity data set where the learning is extremely rapid. This is also suggested from the similar performance trend displayed by both the single-fidelity and the multifidelity models. Additionally, the multifidelity models also utilize a large number of low-fidelity points in its training. With this, we have established that group contribution methods which were used in the low-fidelity data set captures meaningful information that can be successfully used with cokriging machine learning in this manner. Second, both multifidelity models perform similarly. This implies that, in order to make predictions for tendency of new polymers to crystallize under conventional synthesis and processing conditions, the low-fidelity crystallization tendency is not required. Therefore, this model does not require any further experimentation or information to make new predictions at the high-fidelity level of accuracy. Next, the RMSE of the multifidelity model can be discussed in comparison to the group contribution method's prediction capability. The variance in the high- and low-fidelity data sets arises from two sources—the variance from different polymer processing conditions which was previously discussed and due to the group contribution method used to curate the low-fidelity data set. The group contribution method is devoid of processing variability which is higher in value (13−15%) than its prediction error (10.11%). Additionally, the low-fidelity data set is significantly more diverse and is not included in the group contribution analysis. This explains the higher RMSE of the multifidelity model compared to the group contribution method's prediction error. Finally, we also believe that the multifidelity model captures the polymer processing variance; the processing variability established previously is 13−15% which is similar to the RMSE of the multifidelity models.

Multifidelity Model Verification. For the verification of the multifidelity cokriging model, we utilized the multifidelity model trained on all data points in both, the high- and low-fidelity data sets. We used five polymers which were not utilized in this work previously for the verification. These polymers were chosen specifically for their wide range in the percent crystallinity values as well as their varying representation in the training data sets. The polymers are poly(vinylphenol), nylon-5,6, poly(methacrylonitrile), poly(*p*-styrene sulfonic acid), and kevlar, and their crystallinity values were gathered from the literature (monomer structures in Table S2)[21,35−40] As demonstrated in Figure 6, the polymers have crystallinities ranging from noncrystalline/amorphous to 70%. Additionally, we also selected them on the basis of their representation in the data set to assess the predicted uncertainty values. Polymers like poly(vinylphenol), poly(*p*-styrene sulfonic acid), and nylon-5,6 are well-represented in the data set whereas poly(methacrylonitrile) and kevlar have few similar training examples. We predicted their crystallinities using the methodology described in this work, and the results are presented in Figure 6. We find that there is excellent agreement for all polymers except for poly(vinylphenol) which

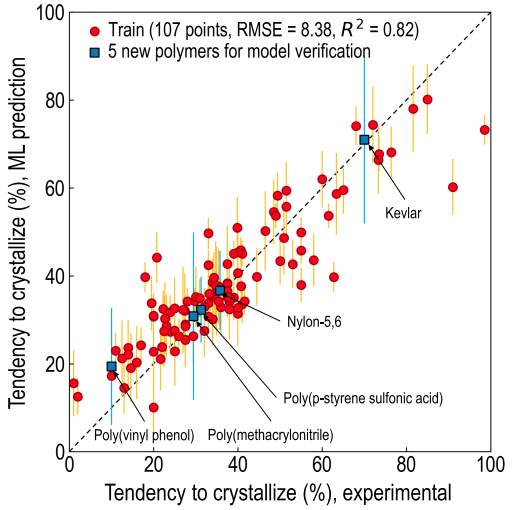


Figure 6. Validation of the multifidelity cokriging model using five new polymers whose experimental percent crystallinity values are compared with its predicted values.

https://dx.doi.org/10.1021/acs.jpcb.0c01865
J. Phys. Chem. B XXXX, XXX, XXX−XXX

is amorphous and therefore, we chose to use 10% as its experimental value for representation in Figure 6. The predicted uncertainties are also in agreement with the presence of the polymers in the data set. Since poly(methacrylonitrile) and kevlar do not have similar training examples in the data set, they present high uncertainty values. This analysis provides valuable insight about realistic scenarios where the tendency to crystallize of an new polymer candidate is predicted.

OUTLOOK

Predicting polymer properties which are extremely process-dependent like the polymer crystallinity is nontrivial for several reasons. A conspicuous assumption that was made in this work was to assign singular tendency to crystallize values (therefore, not accounting for process variability) to all the polymers in the data sets. The multiparameter nature of this data set makes it extremely difficult to curate a uniform data set that is tailored to predict a percent crystallinity value for a specific case. Therefore, we decided to use the most-likely values for polymers. However, this work can be still used to predict the ability of a polymer to crystallize, i.e., polyethylene generally exhibits greater crystallinity than polystyrene.[13,20−24] While the exact value of the percent crystallinity is dependent on the processing conditions, the inherent ability to crystallize depends on chemical factors examined in the Features section. Therefore, for a new polymer, this tool can be useful to determine its tendency to crystallize and we have implemented it at www.polymergenome.org. Additionally, the lack of uniform, high-fidelity data is commonly encountered in materials science. The methodology used in this work can be used to overcome this issue for a variety of properties for organic and inorganic materials. Several polymer properties like the glass transition temperature, the melting temperature, entropy, solubility have established group contribution values by Van Krevelen[13] for commonly occurring substructures. Furthermore, inorganic materials properties like the lattice thermal conductivity and the formation enthalpy have established theoretical models which can be utilized in conjunction with limited experimental data. This work demonstrates the viability of this methodology for data sets which have experimental and theoretical values. Going forward, a possible solution to this could be a natural language processing (NLP) based machine learning model which can actively collect and curate highly complex data sets to overcome these issues.[41]

CONCLUSIONS

In conclusion, we have developed a multifidelity cokriging model to predict crystallization tendencies for polymers based on a fusion of the high and low-fidelity data sets. This work successfully develops two differing data sets to predict crystallization tendencies and also quantified the polymer processing-related variations for percent crystallinity. Additionally, we also identify pertinent features and corresponding chemical guidelines which affect the polymer's tendency to crystallize. We find that the position of functional groups (like amides, sulfones, carbonyl, phenyls), viz, in the main chain or the side chain was critical to its crystallization tendency. Functional groups that promote chain stiffening are positively correlated to the crystallization if they are present in the main chain but negatively affect the tendency to crystallize if present in the side chains since they reduce the overall packing density. These guidelines can be utilized to rationally design polymers with desired crystallization tendencies. Further, we find that we found the multifidelity information fusion framework using cokriging worked successfully and better than a single-fidelity GPR model by 23%. Although this model inherently assumes "typical" processing conditions and estimates the "most-likely" percent crystallinity values, it can help in the estimation of a polymer's tendency to crystallize in a far more cost-effective and efficient manner. The ability of the cokriging model to make predictions for a new polymer entirely based on its features and without the corresponding low-fidelity estimate makes it an extremely useful and versatile method and can be extended to other polymer properties.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcb.0c01865.

Process-related variation for poly(propylene), poly(ethylene oxide), poly(caprolactone), and poly(butylene terephthalate), LASSO coefficients for key features which affects the percent crystallinity, monomers of the polymers used for model verification (PDF)

AUTHOR INFORMATION

Corresponding Author

Rampi Ramprasad − *School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States;* orcid.org/0000-0003-4630-1565; Email: rampi.ramprasad@mse.gatech.edu

Authors

Shruti Venkatram − *School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States;* orcid.org/0000-0003-0306-8222

Rohit Batra − *School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States;* orcid.org/0000-0002-1098-7035

Lihua Chen − *School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States;* orcid.org/0000-0002-9852-8211

Chiho Kim − *School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States;* orcid.org/0000-0002-1814-4980

Madeline Shelton − *School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.0c01865

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work is supported through a grant by the Kolon Center for Lifestyle Innovation.

REFERENCES

(1) Bunn, C.; Alcock, T. The texture of polythene. *Trans. Faraday Soc.* **1945**, *41*, 317−325.

(2) Krimm, S.; Tobolsky, A. V. Quantitative x-ray studies of order in amorphous and crystalline polymers. Quantitative x-ray determination of crystallinity in polyethylene. *J. Polym. Sci.* **1951**, *7*, 57−76.
(3) Kong, Y.; Hay, J. The enthalpy of fusion and degree of crystallinity of polymers as measured by DSC. *Eur. Polym. J.* **2003**, *39*, 1721−1727.
(4) Seymour, R. B.; Kirshenbaum, G. S. *High performance polymers, their origin and development*; Springer: 1986.
(5) Meyer, W. H. Polymer electrolytes for lithium-ion batteries. *Adv. Mater.* **1998**, *10*, 439−448.
(6) Lin, H.; Freeman, B. D. Gas solubility, diffusivity and permeability in poly (ethylene oxide). *J. Membr. Sci.* **2004**, *239*, 105−117.
(7) Armand, M. B. Polymer electrolytes. *Annu. Rev. Mater. Sci.* **1986**, *16*, 245−261.
(8) Chen, L.; Venkatram, S.; Kim, C.; Batra, R.; Chandrasekaran, A.; Ramprasad, R. Electrochemical Stability Window of Polymeric Electrolytes. *Chem. Mater.* **2019**, *31*, 4598.
(9) Das, D.; Chandrasekaran, A.; Venkatram, S.; Ramprasad, R. Effect of crystallinity on Li adsorption in polyethylene oxide. *Chem. Mater.* **2018**, *30*, 8804−8810.
(10) Kavesh, S.; Schultz, J. Meaning and measurement of crystallinity in polymers: A review. *Polym. Eng. Sci.* **1969**, *9*, 331−338.
(11) Murthy, N.; Minor, H. General procedure for evaluating amorphous scattering and crystallinity from X-ray diffraction scans of semicrystalline polymers. *Polymer* **1990**, *31*, 996−1002.
(12) Blaine, R. L. Thermal Applications Note. *Polymer Heats of Fusion*; TA Instruments: 2002.
(13) Van Krevelen, D. W.; Te Nijenhuis, K. *Properties of polymers: their correlation with chemical structure; their numerical estimation and prediction from additive group contributions*; Elsevier: 2009.
(14) 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.
(15) Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Computational Materials* **2017**, *3*, 54.
(16) Batra, R.; Pilania, G.; Uberuaga, B. P.; Ramprasad, R. Multifidelity Information Fusion with Machine Learning: A Case Study of Dopant Formation Energies in Hafnia. *ACS Appl. Mater. Interfaces* **2019**, *11*, 24906−24918.
(17) Patra, A.; Batra, R.; Chandrasekaran, A.; Kim, C.; Huan, T. D.; Ramprasad, R. A multi-fidelity information-fusion approach to machine learn and predict polymer bandgap. *Comput. Mater. Sci.* **2020**, *172*, 109286.
(18) Yamamoto, T. Computer modeling of polymer crystallization−Toward computer-assisted materials' design. *Polymer* **2009**, *50*, 1975−1985.
(19) Welch, P. Examining the role of fluctuations in the early stages of homogenous polymer crystallization with simulation and statistical learning. *J. Chem. Phys.* **2017**, *146*, 044901.
(20) Mark, J. E. *Physical properties of polymers handbook*; Springer: 2007; Vol. *1076*.
(21) Polymer Database. https://polymerdatabase.com/.
(22) 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 (EIDWT)* **2011**, 22−29.
(23) Brandrup, J.; Immergut, E. H.; Grulke, E. A.; Abe, A.; Bloch, D. R. *Polymer handbook*; Wiley: New York, 1999; Vol. *89*.
(24) Wypych, G. *Handbook of polymers*; Elsevier, 2016.
(25) Gaur, U.; Wunderlich, B. Advanced thermal analysis system (ATHAS) polymer heat capacity data bank. *Computer Applications in Applied Polymer Science* **1982**, *197*, 355.
(26) Dole, M.; Wunderlich, B. Melting points and heats of fusion of polymers and copolymers. *Makromol. Chem.* **1959**, *34*, 29−49.
(27) Gauraha, N. Introduction to the LASSO. *Resonance* **2018**, *23*, 439−464.
(28) Balani, K.; Verma, V.; Agarwal, A.; Narayan, R. Physical, thermal, and mechanical properties of polymers. *Biosurfaces: A Materials Science and Engineering Perspective* **2015**, 329.
(29) Rasmussen, C. E. Gaussian processes in machine learning. *Advanced Lectures on Machine Learning* **2004**, *3176*, 63−71.
(30) Pilania, G.; Gubernatis, J. E.; Lookman, T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Comput. Mater. Sci.* **2017**, *129*, 156−163.
(31) Yamada, H.; Liu, C.; Wu, S.; Koyama, Y.; Ju, S.; Shiomi, J.; Morikawa, J.; Yoshida, R. Predicting Materials Properties with Little Data Using Shotgun Transfer Learning. *ACS Cent. Sci.* **2019**, *5*, 1717−1730.
(32) Kaya, M.; Hajimirza, S. Using a Novel Transfer Learning Method for Designing Thin Film Solar Cells with Enhanced Quantum Efficiencies. *Sci. Rep.* **2019**, *9*, 5034.
(33) Li, X.; Zhang, Y.; Zhao, H.; Burkhart, C.; Brinson, L. C.; Chen, W. A transfer learning approach for microstructure reconstruction and structure-property predictions. *Sci. Rep.* **2018**, *8*, 1−13.
(34) Le Gratiet, L. *Multi-fidelity Gaussian process regression for computer experiments*. Ph.D. thesis, 2013.
(35) Hietala, S.; Paronen, M.; Holmberg, S.; Näsman, J.; Juhanoja, J.; Karjalainen, M.; Serimaa, R.; Toivola, M.; Lehtinen, T.; Parovuori, K.; et al. Phase separation and crystallinity in proton conducting membranes of styrene grafted and sulfonated poly (vinylidene fluoride). *J. Polym. Sci., Part A: Polym. Chem.* **1999**, *37*, 1741−1753.
(36) Joh, Y.; Kotake, Y.; Yoshihara, T.; Ide, F.; Nakatsuka, K. Stereospecific polymerization of methacrylonitrile. I. Characterization of crystalline polymethacrylonitrile. *J. Polym. Sci., Part A-1: Polym. Chem.* **1967**, *5*, 593−603.
(37) Chen, H.-L.; Liu, H.-H.; Lin, J. Microstructure of semicrystalline poly (L-lactide)/poly (4-vinylphenol) blends evaluated from SAXS absolute intensity measurement. *Macromolecules* **2000**, *33*, 4856−4860.
(38) Yi, J.; Goh, S. Miscibility and interactions in poly (n-propyl methacrylate)/poly (vinyl alcohol) blends. *Polymer* **2005**, *46*, 9170−9175.
(39) Puiggalí, J.; Franco, L.; Alemán, C.; Subirana, J. Crystal structures of nylon 5, 6. A model with two hydrogen bond directions for nylons derived from odd diamines. *Macromolecules* **1998**, *31*, 8540−8548.
(40) Hultin, D. P. *Quantifying Crystallinity in Kevlar Using Raman Spectroscopy*. https://www.nnci.net/sites/default/files/inline-files/D-04-Drennan.pdf, accessed on 29 February, 2020.
(41) Kim, E.; Huang, K.; Saunders, A.; McCallum, A.; Ceder, G.; Olivetti, E. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem. Mater.* **2017**, *29*, 9436−9444.

https://dx.doi.org/10.1021/acs.jpcb.0c01865
J. Phys. Chem. B XXXX, XXX, XXX−XXX