Macromolecules

pubs.acs.org/Macromolecules Article

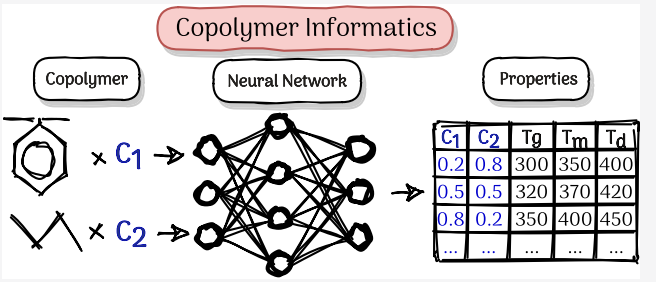
Copolymer Informatics with Multitask Deep Neural Networks

Christopher Kuenneth, William Schertzer, and Rampi Ramprasad*

Cite This: https://doi.org/10.1021/acs.macromol.1c00728 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Polymer informatics tools have been recently gaining ground to efficiently and effectively develop, design, and discover new polymers that meet specific application needs. So far, however, these data-driven efforts have largely focused on homopolymers. Here, we address the property prediction challenge for copolymers, extending the polymer informatics framework beyond homopolymers. Advanced polymer fingerprinting and deep-learning schemes that incorporate multitask learning and meta learning are proposed. A large data set containing over 18 000 data points of glass transition, melting, and degradation temperature of homopolymers and copolymers of up to two monomers is used to demonstrate the copolymer prediction efficacy. The developed models are accurate, fast, flexible, and scalable to more copolymer properties when suitable data become available.



INTRODUCTION

In less than a century, polymer consumption has become significant in everyday life and high-technology.[1,2] Extensive efforts are underway more vigorously than ever before to shape and design polymers to meet specific application needs. Given the vastness and richness of the polymer chemical and structural spaces, new capabilities are required to effectively and efficiently search this space to identify optimal, application-specific solutions. The burgeoning field of polymer informatics[3−12] attempts to address such critical search problems by utilizing modern data-driven machine learning (ML) approaches.[13−19] Such efforts have already seen significant successes in terms of the realization and deployment of on-demand polymer property predictors[13−15] and solving inverse problems by which polymers meeting specific property requirements are either identified from a candidate set or freshly designed using genetic[19] or generative algorithms.[20] Data that fuel such approaches may be efficiently and autonomously extracted from the literature using ML approaches.[21,22]

In the present contribution, we direct our efforts toward building ML models that can instantaneously predict three important temperatures—the glass transition (T_g), melting (T_m), and degradation (T_d) temperatures—of copolymers. T_g and T_m determine the mechanical properties of copolymers, while T_d indicates the overall temperature stability of copolymers. The focus on copolymers is opportune and very timely. Past informatics efforts by us and others are dominated by investigations involving homopolymers, but several application problems may require the usage of copolymers, owing to the flexibility copolymers offer in tuning physical properties.[23−25] Since typical copolymer search spaces are very large, several authors[26−28] have showcased feasible pathways to efficiently explore this space or study possible synthesis protocols[29] using ML models.

The present work has several critical ingredients. The first ingredient is the data set itself. We have curated a data set of three temperatures for copolymers containing two distinct monomer units and the corresponding end-member homopolymers. The data set utilized in this study includes a total of 18 445 data points, as detailed in Table 1. A "data point" is defined as a tuple composed of the homopolymer or copolymer specifications and one of the three temperature values. The second ingredient of our work is the method used to numerically represent each polymer using a modification of

Table 1. Number of Homopolymer and Copolymer Data Points[a]

property	symbol	homopolymer	copolymer	total
glass transition temperature	T_g	5072	4426	9498
melting temperature	T_m	2079	1988	4067
degradation temperature	T_d	3520	1360	4880
total		10 671	7774	18 445

[a]The 7774 copolymer data points encompass 1569 distinct copolymer chemistries, ignoring composition information.

Received: April 5, 2021
Revised: June 13, 2021

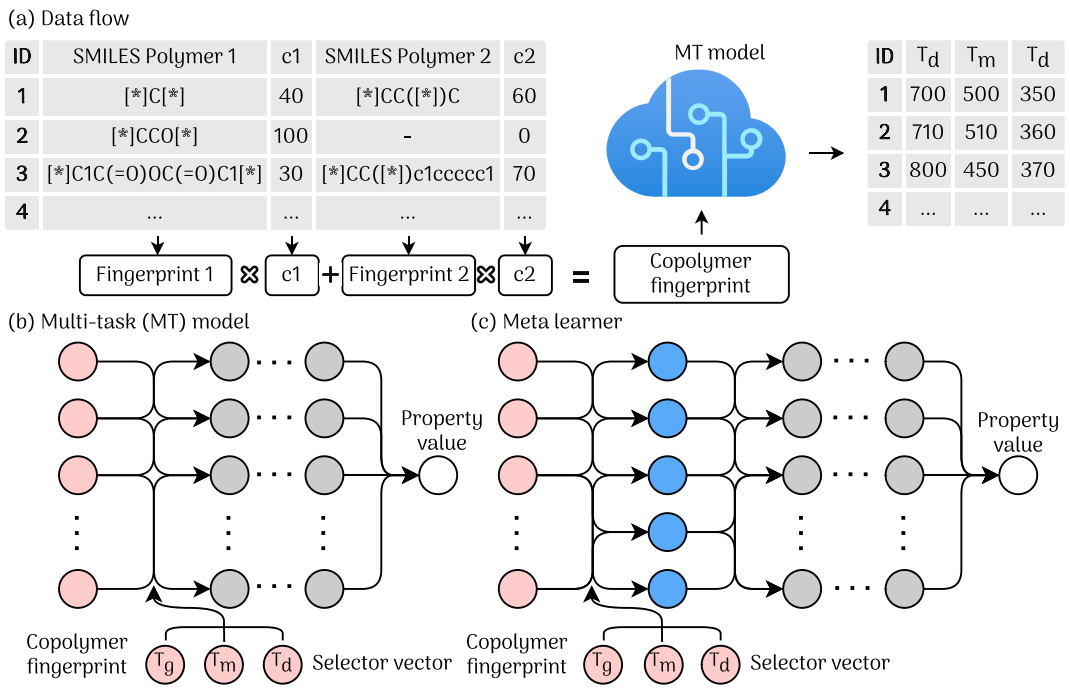


Figure 1. Data flow and machine learning model. (a) Shows the data flow through the machine learning model and sketches the copolymer fingerprint generation. `ID2` indicates a homopolymer, and `ID1` and `ID3` indicate copolymers. The two dangling bonds of the polymer repeat units are denoted using "`[*]`" in the SMILES strings. (b) Shows a concatenation-based conditioned multitask neural network. It takes in the copolymer fingerprint and a binary selector vector (e.g., `(100)` if the property is T_g, `(010)` if T_m, and `(001)` if T_g) as inputs (light red nodes) and outputs the data processed through an optimized number of dense layers (gray nodes). The `1` in the selector vector indicates the selected output property of the final layer (white node): glass transition (T_g), melting (T_m), and degradation (T_d) temperature. (c) Shows the meta learner that is composed of the five cross-validation models (blue nodes) and a neural network.

our past fingerprinting methodology. The final vital ingredient is the multitask neural network[13] that ingests the entire data set of homopolymer and copolymer fingerprints and their corresponding T_g, T_m, and T_d values. Training is performed using state-of-the-art practices involving cross-validation (CV) and meta-ensemble-learning that embeds and leverages the cross-validation models, as detailed below. The adopted workflow is portrayed in Figure 1, and the final models have been deployed at https://polymergenome.org.

RESULTS AND DISCUSSION

Data. As mentioned above, and summarized in Table 1, our data set includes T_g, T_m, and T_d values for homopolymers and copolymers involving two distinct monomers at various compositions. Of the entire data set of 18 445 data points, 10 671 (≈60%) data points correspond to homopolymers (collected from the previous studies[6,13,15,16]) and 7774 (≈40%) data points pertain to copolymers (collected from the PolyInfo repository[30]) that encompass 1569 distinct copolymer chemistries, ignoring composition information. For the sake of uniformity and consistency, only T_d data points measured via thermogravimetric analysis (TGA) and T_g and T_m data points measured via differential scanning calorimetry (DSC) were utilized in this work. All copolymers are furthermore assumed to be random copolymers because information about the copolymer types was not uniformly available. For similar reasons, the degree of polymerization and molecular weights of the polymers are neglected in the data set. Before training our multitask neural networks, the property values (namely, T_g, T_m, and T_d) were scaled to the range of [0, 1].

Cross-Validation. Up to this point, we have described our copolymer data set and how to numerically represent copolymers using fingerprints. The next step concerns the actual ML model building process. For this, the data is split such that 80% is used to develop five cross-validation models and 20% is used by the meta learner (see below). The five cross-validation models are concatenation-based conditioned multitask deep neural networks (see Figure 1b) and are implemented using Tensorflow.[31] They take in the copolymer fingerprints as well as the three-component selector vector, which indicates whether the data point corresponds to T_g, T_m, or T_d and outputs the property chosen by the selector vector. We used the Adam optimizer combined with the Stochastic Weight Averaging method and an initial learning rate of 10^{-3} to optimize the mean-square error (MSE) of the property values. Early stopping, combined with a learning rate scheduler, was deployed during the optimization. All hyperparameters, such as the initial learning rate, number of layers, neurons, dropout rates, and layer after which the selector vector is concatenated, are optimized with respect to the generalization error using the Hyperband method, as implemented in the Python package Keras-Tuner.[32] The optimized hyperparameters are summarized in Table S1 of the Supporting Information.

The low root-mean-square errors (RMSEs) and small confidence intervals of the five cross-validation models in Figure 2 attest to the strength of our copolymer fingerprints and multitask approach. Multitask deep neural networks have recently shown advantages in efficiency, scalability, and accuracy over Gaussian processes in an extensive comparison and benchmark for many polymer properties.[13] The 5-fold

https://doi.org/10.1021/acs.macromol.1c00728
Macromolecules XXXX, XXX, XXX−XXX

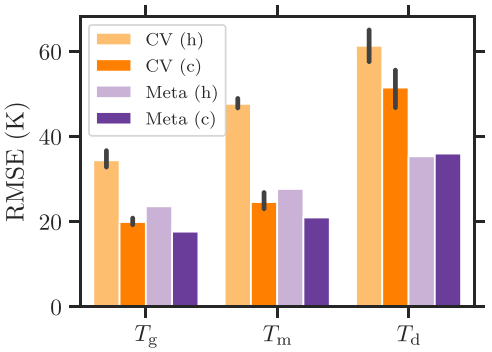


Figure 2. Five-fold cross-validation (CV) and meta learner (Meta) root-mean-square errors (RMSEs) of homopolymers (h) and copolymers (c). The orange bars indicate the average of the 5-fold cross-validation RMSEs for the validation data set, and the error bars are the 68% confidence intervals of these RMSE averages.

averaged RMSEs (red bars) of T_g, T_m, and T_d are 29, 38, and 59 K, respectively, which are similar to the values reported in other studies[13,15] that focus on homopolymers and of the order of experimentally expected uncertainties. Additionally, the RMSE for T_g is the lowest, and that for T_d is the highest (with the T_m RMSE being intermediate). These excellent results suggest that the proposed copolymer fingerprints create a well-conditioned learning problem for multitask models.

Meta Learner. The next and last element of this study is a meta learner—essentially, an ensemble learner—that makes the final property forecast based upon the predictions of the ensemble of cross-validation models, as illustrated in Figure 1c. It may be useful to think of it as consisting of two levels: at the first level, predictions are made using the five cross-validation models, and at the second level, these predictions are utilized in a neural network to predict the final value. The meta learner is trained on the 20% of the data points that were set aside before cross-validation and implemented using a neural network composed of the five cross-validation models (with fixed weights) as the first layer and two fully connected, dense layers as second and third layers. For the meta learner, just as for the cross-validation models, we use the Hyperband method to optimize all hyperparameters (documented in Table S1 of the Supporting Information). The 95% confidence intervals of the meta learner's predictions are estimated using the Monte Carlo dropout method.[33] Such error estimates are particularly of interest for high-throughput predictions of copolymer screening. In the following, we will first assess the performance of the meta learner using parity plots (Figure 3) and second examine the meta learner's prediction performance on the basis of four copolymer examples (Figure 4).

With RMSE (R^2) values as low as 21 (0.96), 24 (0.94), and 36K (0.90) when predicting T_g, T_m, and T_d, respectively, the parity plots in Figure 3 show the exceptional fitness of the meta learner. Because the meta learner's predictions are based on the cross-validation models, it can infer from all five models, effectively rendering its RMSEs lower than the average RMSEs of the cross-validation models, as illustrated in Figure 2.

Figure 4 shows the predictions (--) of the meta learner along with experimental data points (•) for T_g (blue), T_m (red), and T_d (green) of four selected copolymers across the entire composition range. The 95% confidence intervals of the predictions are shown as shaded bands. In all cases, there is a high level of agreement between predictions and experimental data points. Interestingly, the meta learner predicts averaged trends through the experimental data points. For example, in the case of copolymer (b), the predicted trend of T_d takes an averaged pathway through the scattered experimental data points across the range of the copolymer compositions. Also, the predicted trends display an appropriate level of smoothness, which indicates that the meta learner was regulated properly during training (we are using dropouts), thus avoiding both overfitting or underfitting. Another interesting finding is that the meta learner is capable of distilling key knowledge from the data set, as shown for T_m of polymer (b) in Figure 4: although no experimental data point is present at weight fraction 0, the meta learner predicts an upward trend for T_m. This T_m trend is inferred from the T_g trend for the same polymer. Apparently, the data on which the model is based condition the meta learner to predict similar trends for T_g and T_m. Clearly, the use of inherent knowledge and correlations in data sets allows for accurate ML models using fewer data points and makes our method the preferred one for small data sets.[13]

CONCLUSIONS

This work is a first step toward creating general property-predictive ML models for copolymers. Using a copolymer data

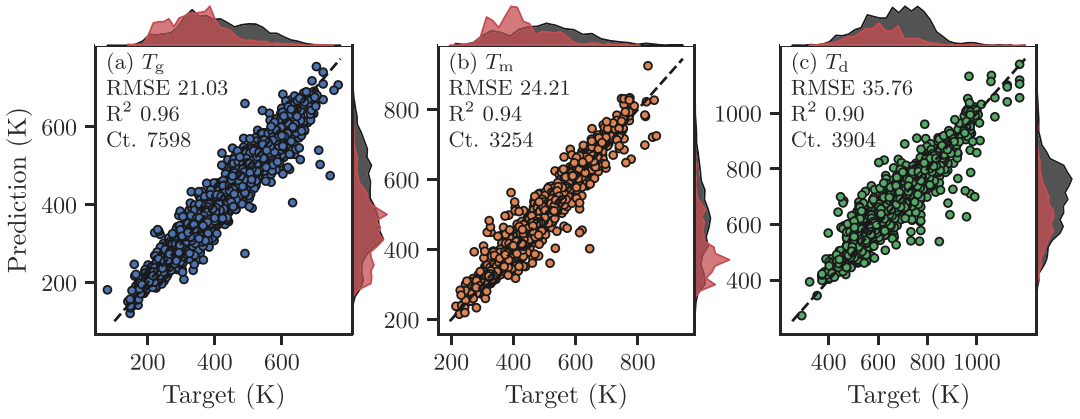
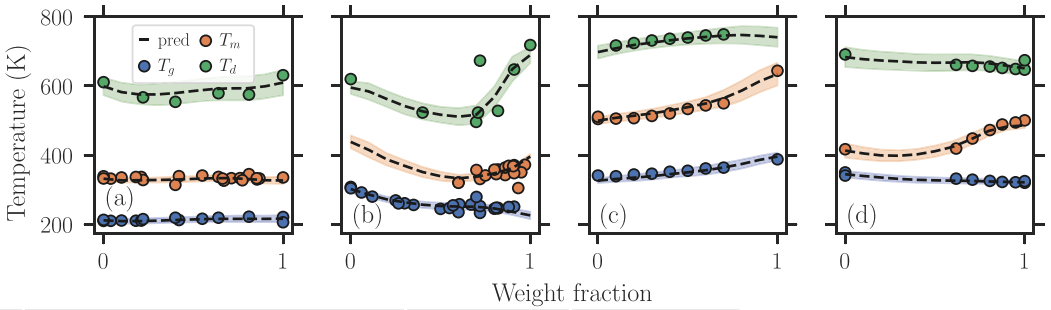


Figure 3. Meta learner parity plots. The predictions are displayed for 80% of the data set that was used to train the cross-validation models. The parity plots in (a), (b), and (c) display the glass transition (T_g), melting (T_m), and degradation temperature (T_d), respectively. The root-mean-square errors (RMSEs), coefficient of determination (R^2), and data point count (Ct.) are indicated in each subplot. The frequencies of the homopolymer and copolymer data points are indicated in black and red, respectively, in the margins of the parity plots.

https://doi.org/10.1021/acs.macromol.1c00728
Macromolecules XXXX, XXX, XXX−XXX



	Name	SMILES 1	SMILES 2
(a)	poly[oxyethylene/poly(oxy(1-oxohexane-1,6-diyl))]	[*]CCCCCC(=O)O[*]	[*]CCO[*]
(b)	poly[ethylene/1-(acetyloxy)ethylene]	[*]CC([*])OC(C)=O	[*]C[*]
(c)	poly[iminoterephthaloyliminononane-1,9-diyl/iminohexane dioyliminononane-1,9-diyl]	[*]CCCCCCCCCNC(=O)CCCCC(=O)N[*]	[*]CCCCCCCCCNC(=O)c1ccc(C(=O)N[*])cc1
(d)	poly[oxy(2,2-dimethylpropane-1,3-diyl)oxy terephthaloyl/oxypropane-1,3-diyloxyterephthaloyl]	[*]CC(C)(C)COC(=O)c1ccc(C(=O)O[*])cc1	[*]CCCOC(=O)c1ccc(C(=O)O[*])cc1

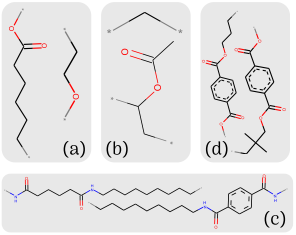


Figure 4. Sample predictions of T_g, T_m, and T_d of four different copolymers. Filled circles (●) indicate the training data points and dashed lines (--) the meta learner predictions (pred). T_g, T_m, and T_d stand for glass transition, melting, and degradation temperatures. The shaded bands indicate the 95% confidence intervals of the predictions.

set for the glass transition (T_g), melting (T_m), and degradation (T_d) temperatures captured in 18 445 data points, we first developed a scheme to numerically represent and fingerprint copolymers. These fingerprints were used as the inputs to five cross-validation multitask neural networks. Based on the trained cross-validation models, a meta learner was built for production deployment that—as expected—surpasses the performance of the cross-validation models. The meta learner leads to final models with unprecedented accuracies (overall R^2 of 0.94) and small prediction times for homopolymers and copolymers alike. The entire workflow proposed here is generalizable to copolymers with more than two monomers and for a broader range of properties. The implications of this work are far-reaching as they lay the ground work for future advancements of polymer informatics beyond homopolymers.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.macromol.1c00728.

Optimized hyperparameters of the cross-validation (CV) and meta learner models (PDF)

Accession Codes

The code is available at https://github.com/Ramprasad-Group/copolymer_informatics. The meta learner is openly available for use at https://polymergenome.org.

■ 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

Christopher Kuenneth − *School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States;* orcid.org/0000-0002-6958-4679

William Schertzer − *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.macromol.1c00728

Author Contributions

C.K. designed, trained, and evaluated the ML models. W.S. and C.K. collectively collected and curated the data points used in this study. The work was conceived and guided by R.R. All authors discussed the results and commented on the manuscript. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

C.K. thanks the Alexander von Humboldt Foundation for financial support. This work is financially supported by the Office of Naval Research through a Multidisciplinary University Research Initiative (MURI) grant (N00014-17-1-2656).

■ REFERENCES

(1) Peacock, A. J.; Calhoun, A. *Polymer Chemistry: Properties and Application*; Carl Hanser Verlag GmbH Co KG, 2012.

(2) Sequeira, C.; Santos, D. *Polymer Electrolytes: Fundamentals and Applications*; Woodhead Publishing in Materials; Elsevier Science, 2010.

(3) Chen, L.; Pilania, G.; Batra, R.; Huan, T. D.; Kim, C.; Kuenneth, C.; Ramprasad, R. Polymer informatics: Current status and critical next steps. *Mater. Sci. Eng., R* **2021**, *144*, No. 100595.

(4) Batra, R.; Song, L.; Ramprasad, R. Emerging materials intelligence ecosystems propelled by machine learning. *Nat. Rev. Mater.* **2020**, *521*, No. 436.

(5) Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Comput. Mater.* **2017**, *3*, No. 54.

(6) 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.
(7) Audus, D. J.; de Pablo, J. J. Polymer Informatics: Opportunities and Challenges. *ACS Macro Lett.* **2017**, *6*, 1078−1082.
(8) Peerless, J. S.; Milliken, N. J. B.; Oweida, T. J.; Manning, M. D.; Yingling, Y. G. Soft Matter Informatics: Current Progress and Challenges. *Adv. Theory Simul.* **2019**, *2*, No. 1800129.
(9) Adams, N.; Murray-Rust, P. Engineering Polymer Informatics: Towards the Computer-Aided Design of Polymers. *Macromol. Rapid Commun.* **2008**, *29*, 615−632.
(10) Jabeen, F.; Chen, M.; Rasulev, B.; Ossowski, M.; Boudjouk, P. Refractive indices of diverse data set of polymers: A computational QSPR based study. *Comput. Mater. Sci.* **2017**, *137*, 215−224.
(11) Wu, Y.; Guo, J.; Sun, R.; Min, J. Machine learning for accelerating the discovery of high-performance donor/acceptor pairs in non-fullerene organic solar cells. *npj Comput. Mater.* **2020**, *6*, No. 120.
(12) Afzal, M. A. F.; Haghighatlari, M.; Ganesh, S. P.; Cheng, C.; Hachmann, J. Accelerated Discovery of High-Refractive-Index Polyimides via First-Principles Molecular Modeling, Virtual High-Throughput Screening, and Data Mining. *J. Phys. Chem. C* **2019**, *123*, 14610−14618.
(13) Kuenneth, C.; Rajan, A. C.; Tran, H.; Chen, L.; Kim, C.; Ramprasad, R. Polymer informatics with multi-task learning. *Patterns* **2021**, *2*, No. 100238.
(14) Doan Tran, H.; Kim, C.; Chen, L.; Chandrasekaran, A.; Batra, R.; Venkatram, S.; Kamal, D.; Lightstone, J. P.; Gurnani, R.; Shetty, P.; Ramprasad, M.; Laws, J.; Shelton, M.; Ramprasad, R. Machine-learning predictions of polymer properties with Polymer Genome. *J. Appl. Phys.* **2020**, *128*, No. 171104.
(15) 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. 24002.
(16) Kim, C.; Chandrasekaran, A.; Jha, A.; Ramprasad, R. Active-learning and materials design: The example of high glass transition temperature polymers. *MRS Commun.* **2019**, *9*, 860−866.
(17) Chen, L.; Kim, C.; Batra, R.; Lightstone, J. P.; Wu, C.; Li, Z.; Deshmukh, A. A.; Wang, Y.; Tran, H. D.; Vashishta, P.; Sotzing, G. A.; Cao, Y.; Ramprasad, R. Frequency-dependent dielectric constant prediction of polymers using machine learning. *npj Comput. Mater.* **2020**, *6*, No. 61.
(18) 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*, No. 109286.
(19) Kim, C.; Batra, R.; Chen, L.; Tran, H.; Ramprasad, R. Polymer design using genetic algorithm and machine learning. *Comput. Mater. Sci.* **2021**, *186*, No. 110067.
(20) Batra, R.; Dai, H.; Huan, T. D.; Chen, L.; Kim, C.; Gutekunst, W. R.; Song, L.; Ramprasad, R. Polymers for Extreme Conditions Designed Using Syntax-Directed Variational Autoencoders. *Chem. Mater.* **2020**, *32*, 10489−10500.
(21) Shetty, P.; Ramprasad, R. Automated knowledge extraction from polymer literature using natural language processing. *iScience* **2021**, *24*, No. 101922.
(22) Kononova, O.; Huo, H.; He, T.; Rong, Z.; Botari, T.; Sun, W.; Tshitoyan, V.; Ceder, G. Text-mined dataset of inorganic materials synthesis recipes. *Sci. Data* **2019**, *6*, No. 203.
(23) Raval, N.; Kalyane, D.; Maheshwari, R.; Tekade, R. K. *Basic Fundamentals of Drug Delivery*; Elsevier, 2019; pp 173−201.
(24) Hadjichristidis, N.; Pispas, S.; Floudas, G. *Block Copolymers*; John Wiley & Sons, Inc.: Hoboken, USA, 2002; Chapter 21, pp 383−408.
(25) Arcos-Hernández, M. V.; Laycock, B.; Donose, B. C.; Pratt, S.; Halley, P.; Al-Luaibi, S.; Werker, A.; Lant, P. A. Physicochemical and mechanical properties of mixed culture polyhydroxyalkanoate (PHBV). *Eur. Polym. J.* **2013**, *49*, 904−913.
(26) Webb, M. A.; Jackson, N. E.; Gil, P. S.; de Pablo, J. J. Targeted sequence design within the coarse-grained polymer genome. *Sci. Adv.* **2020**, *6*, No. eabc6216.
(27) Jablonka, K. M.; Jothiappan, G. M.; Wang, S.; Smit, B.; Yoo, B. Bias free multiobjective active learning for materials design and discovery. *Nat. Commun.* **2021**, *12*, No. 2312.
(28) Shmilovich, K.; Mansbach, R. A.; Sidky, H.; Dunne, O. E.; Panda, S. S.; Tovar, J. D.; Ferguson, A. L. Discovery of Self-Assembling π-Conjugated Peptides by Active Learning-Directed Coarse-Grained Molecular Simulation. *J. Phys. Chem. B* **2020**, *124*, 3873−3891.
(29) Mohapatra, S.; Hartrampf, N.; Poskus, M.; Loas, A.; Gómez-Bombarelli, R.; Pentelute, B. L. Deep Learning for Prediction and Optimization of Fast-Flow Peptide Synthesis. *ACS Cent. Sci.* **2020**, *6*, 2277−2286.
(30) Otsuka, S.; Kuwajima, I.; Hosoya, J.; Xu, Y.; Yamazaki, M. In *PoLyInfo: Polymer Database for Polymeric Materials Design*, 2011 International Conference on Emerging Intelligent Data and Web Technologies, 2011; pp 22−29.
(31) Martin, A. et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015. https://www.tensorflow.org/.
(32) O'Malley, T.; Bursztein, E.; Long, J.; Chollet, F.; Jin, H.; Invernizzi, L. Keras Tuner, 2019. https://github.com/keras-team/keras-tuner.
(33) Gal, Y.; Ghahramani, Z. In *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning*, 33rd International Conference on Machine Learning, ICML, 2016; pp 1651−1660.

https://doi.org/10.1021/acs.macromol.1c00728
Macromolecules XXXX, XXX, XXX−XXX