JOURNAL OF
EMERGING INVESTIGATORS

Article

Assessing and Improving Machine Learning Model Predictions of Polymer Glass Transition Temperatures

Manav Ramprasad[1], Chiho Kim[2]

[1] Joseph Wheeler High School, Marietta, Georgia

[2] School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia

SUMMARY

The success of the Materials Genome Initiative has led to opportunities for data-driven approaches for materials discovery. The recent development of Polymer Genome (PG), which is a machine learning (ML) based data-driven informatics platform for polymer property prediction, has significantly increased the efficiency of polymer design. Nevertheless, continuous expansion of the 'training data' is necessary to improve the robustness, versatility, and accuracy of the ML predictions. Accurate prediction of polymer properties, such as glass transition temperature (T_g), is advantageous for the design of polymers, particularly for high temperature applications. We hypothesized that by adding more data with increased chemical diversity to the dataset, the predictive capabilities of the PG model would improve. In order to test the performance and transferability of the predictive model for T_g (previously trained on a dataset of 450 polymers), we have carefully collected additional experimental T_g data for 871 polymers from multiple data sources. The T_g values predicted by the present PG models for the polymers in the newly collected dataset were compared directly with the experimental T_g to estimate the accuracy of the present model. Using the full dataset of 1321 polymers, a new ML model for T_g was built following past work. The root mean square error (RMSE) of prediction for the extended dataset, when compared to the earlier one, decreased to 27 K from 57 K, thereby supporting our initial hypothesis that increasing the dataset would improve the predictions. To further improve the performance of the T_g prediction model, we are continuing to accumulate new data and exploring new ML approaches.

INTRODUCTION

A polymer is a large molecular system composed of a chemical repeating unit (monomer). Polymers, displaying a dizzying diversity of physical and chemical properties, constitute an important and ubiquitous class of materials (1). Although they are made up of atomic species found from the periodic table, such as carbon, hydrogen, and oxygen, their limited chemical palette still leads to a rich and diverse spectrum of distinct polymers with a broad range of properties. Thus, it is highly non-trivial to find a suitable polymer for a particular application with desired properties in the practically infinite chemical space. As a result, selection of polymers has hitherto proceeded largely by intuition and trial-and-error efforts, which generally tend to advance the materials discovery landscape in a painstakingly slow manner.

In 2011, the White House unveiled the Materials Genome Initiative (MGI) to accelerate the discovery, manufacture, and deployment of advanced materials to a speed twice as fast as in the past, but at a fraction of the cost (2). One of the central pillars of the MGI is the use of data-driven approaches, such as machine learning (ML), to speed up materials discovery, including in polymer science and engineering. Data-driven ML approaches are complementary to traditional approaches, such as trial-and-error methods involving serendipity, used in materials science and engineering (3). ML approaches utilize prior data, information, and knowledge in an effective and efficient manner, as has been demonstrated in many other domains in the past. Classic examples of ML approaches include facial, fingerprint, or object recognition systems; machines that can play sophisticated games such as chess, Go, or poker; and automation systems such as in robotics or self-driving cars (3, 4).

Within the domain of materials science and engineering, the synthesis and testing process in the laboratory tends to be expensive and time-consuming, especially when handling the polymeric system. In order to utilize the data-driven framework, a dataset of several similar materials and their properties must be first collected. This data constitutes "prior knowledge" on this situation, i.e., the data is obtained from previously performed dedicated experiments or from the literature. Each of the materials in the dataset is then converted to a unique numerical representation, typically referred to as the "fingerprint." Finally, a mapping is established between the fingerprint and its properties using ML algorithms such as Gaussian process regression (GPR), thus leading to a predictive surrogate model (5). Subsequently, this model can be used to make instantaneous predictions of the properties of a new material, by simply following the fingerprinting and mapping procedures. The essential elements of this workflow are portrayed in **Figure 1**.

The efficacy of this method has been recently demonstrated as part of the "Polymer Genome" (PG) Project (6). In order to improve upon the predictive capabilities of the ML models implemented, increased data collection is extremely important. The present work deals with testing the capability

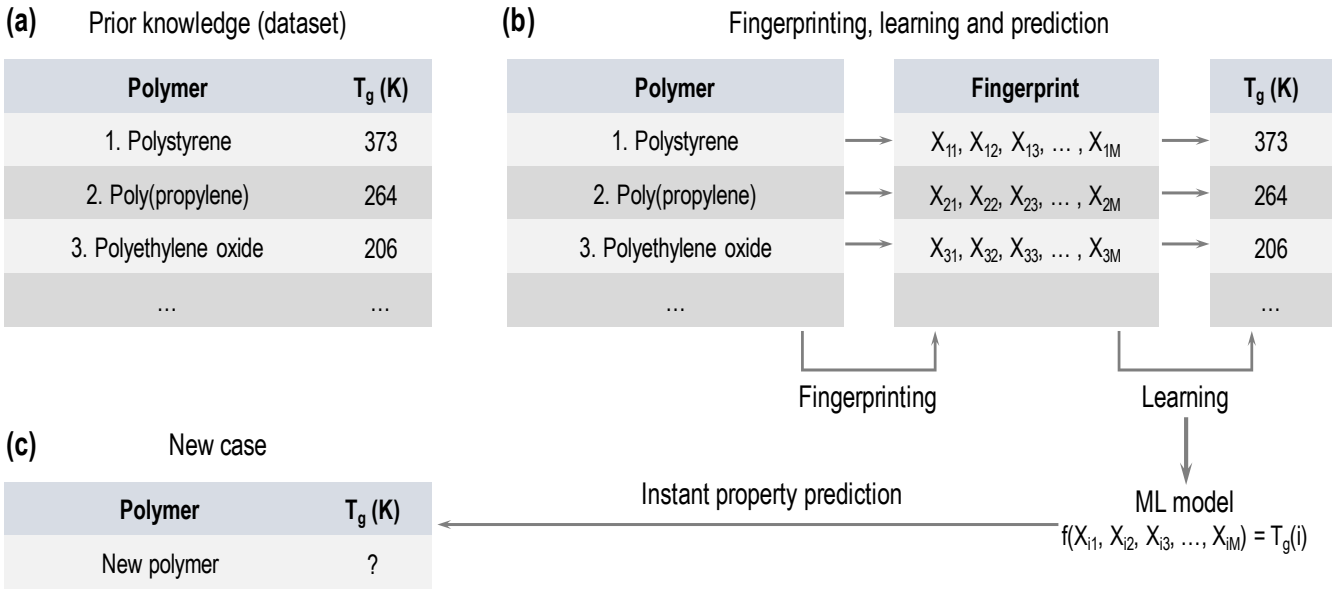


Figure 1: The key elements of machine learning in materials science. (a) Schematic view of an example data set. (b) Creation of a prediction model via the fingerprinting and learning steps. (c) Statement of the problem "What is the T_g of new polymer?"

of PG on new polymers, then using the results of this test to improve the predictive models. The property chosen for this test was the glass transition temperature (T_g) the temperature above which a polymer transitions from brittle and glass-like to viscous and rubber-like. T_g is an important property for many applications, as it determines the temperature ranges at which it is safe to use a polymer. Previously, the model hosted by PG was trained on 450 polymers. Current work demonstrates how expansion of the dataset affects the performance of the ML model. Therefore, we believe that by significantly increasing the size of the dataset, we can improve the predictive capabilities of the PG model in its performance and transferability. We have collected additional experimental T_g data for 871 polymers. The predictions of PG for these new polymers were compared directly with the collected T_g data, and conclusions have been drawn regarding the deficiencies of PG. The original training set was then augmented with this new data, and retraining was performed, ultimately leading to an improvement in the predictive capability of PG.

RESULTS

We refer to the earlier version of PG, which was trained on 450 T_g values, as PG-0. The newer version of PG in which the new T_g data for 871 additional polymers has been incorporated, is referred to as PG-1 (details of data distribution and example polymers in the dataset are shown in the section Methods). Since PG-0 was trained on the original 450 data points, the predictions for those 450 points are fairly accurate. The prediction for the new polymers, on the other hand, is inaccurate, and uncertainty of the prediction is higher. **Figure 2a** shows a parity plot of the performance of PG-0 on both the new dataset of 871 polymers and the initial 450. While many polymers fall closer to the parity line, indicating good agreement between predicted and actual values (values found from the literature), predictions for a certain portion of new polymers are off the parity line.

The poor predictive capabilities for those polymers in the range 300 K - 500 K is mainly due to the difference in fingerprint for the new data points compared to the benchmark (original) data points. In the case of very high T_g values, the PG-0 model performs poorly due to a lack of benchmark data points in the high T_g region (see also: **Figure 3**, showing the distribution of T_g values found in the original and new datasets). In all cases for which the predictions are poor, the uncertainty of the predictions, which is depicted by error bar around data points (**Figure 2**), is relatively higher than those for the original 450 polymers. High uncertainty for a particular case indicates that the polymer is 'not very similar' to the 450 training set polymers of PG-0. Had the scope of the training set been larger, more polymers would have been considered 'more similar' to the training set polymers and would have had more accurate predictions. Overall, the performance in terms of the root mean square error (RMSE) for PG-0 is greater than 50 K for the set of new 871 polymers. This RMSE is higher than desired for T_g predictions (below 30 K). Additionally, of the 871 polymers, 43% have a difference of at least 30 K between the experimental and predicted T_g. This observation indicates that more data points are necessary to improve the predictive performance of the ML model.

Next, we used the 871 new polymers and their corresponding T_g values to augment the original T_g dataset used for PG-0, then retrained to create a new PG-1 GPR predictive model for T_g (**Figure 2b**). As can be seen, a remarkable improvement in predictions emerges. The RMSE in this case is well below 30 K, which is acceptable, as the uncertainties in the actual measurement of T_g is in the

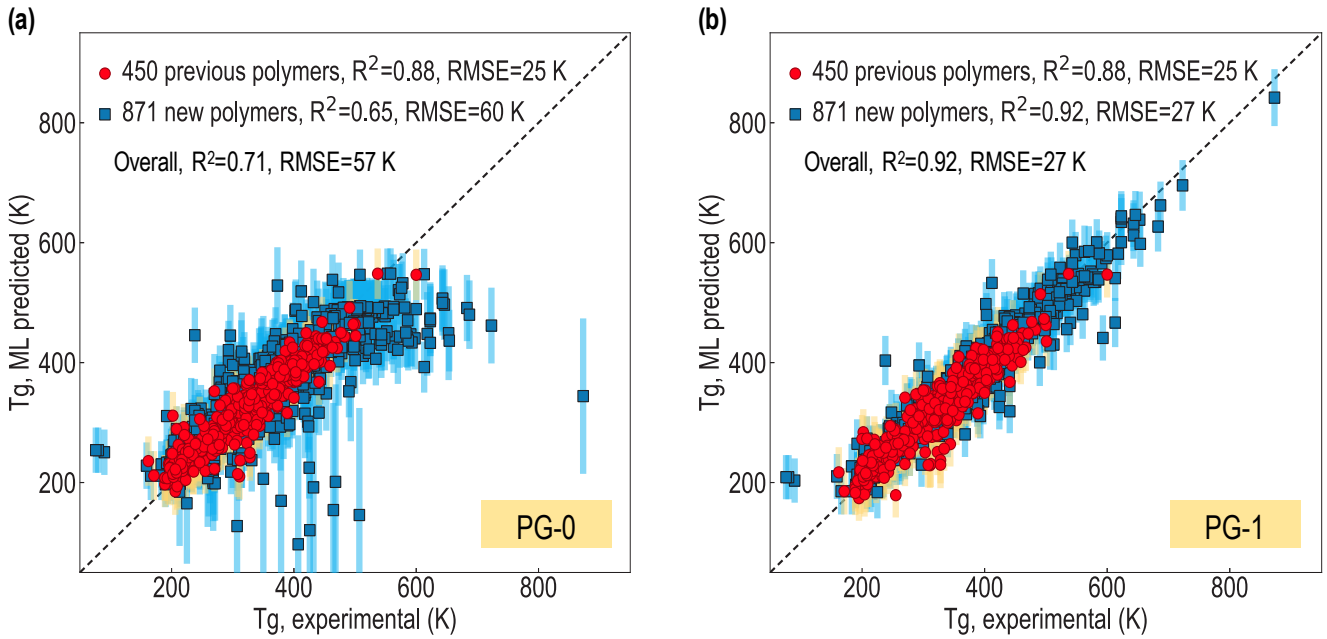


Figure 2: Performance of ML prediction model. Comparison of models trained on (a) 450 previous polymers and (b) 1321 polymers, including 871 new polymers. Error bar represents GPR uncertainty (confidence of prediction).

same range. The uncertainties calculated by GPR, shown by the error bars, have also decreased significantly, again showing an improvement in prediction capabilities. Relative to the original dataset, the new dataset has specifically and purposefully added polymers in new chemical spaces, and has added polymers with high T_g values, i.e., in the 500-700 K range. These aspects have led to a significantly better predictive capability of PG. Further progress can be achieved by systematically adding more diverse data.

DISCUSSION

Although efficient, ML models are accurate and reliable only within the domain of the dataset on which the model was trained. Predictions made for cases that fall outside the domain of the training data (i.e., the dataset originally used to create the models) are not expected to be reliable. In such cases, the new data points that fall outside the original domain of applicability have to be necessarily included in a retraining process to make the predictive model more versatile and transferable.

In summary, to improve upon an existing ML model to predict polymer T_g, a comprehensive dataset of polymer T_g was collected. Machine learning predictions for these new polymers revealed the deficiencies of the previous model. In retraining the machine learning model on the new data, the performance of the predictions dramatically improved, supporting our initial hypothesis that dataset expansion can significantly improve prediction. This work has thus led to a T_g prediction model that has been exposed to a more diverse dataset than before and is hence more versatile. The new model reduced the RMSE for not only new polymers, but also the polymers from the original dataset. The new prediction model presented for T_g, as well as the other polymer properties listed above, is available for free at the PG online platform (6).

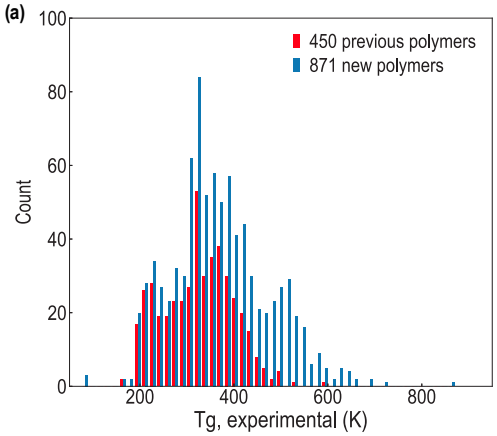
Looking into the future, it would be useful if the prediction pipeline could be inverted, such that polymers could be recommended to meet a specific set of property objectives, such as T_g between 600 K and 650 K. A variety of artificial intelligence-based algorithms (7, 8) may be utilized for such purposes. Solving this inverse problem effectively will significantly accelerate polymer discovery, as inputting requirements for particular polymer properties would result in the suggestion of possible polymers to meet the need.

Besides T_g, many other properties of polymers are important as well. In addition to the T_g prediction, PG also offers predictions of other properties, including 1) electrical properties like bandgap, ionization energy, and electron affinity, 2) dielectric and optical properties such as the dielectric constant and the refractive index, 3) physical and thermodynamic properties like density and atomization energy, 4) solubility properties like Hildebrand solubility parameter and a list of solvents and non-solvents, 5) mechanical properties like tensile strength and Young's modulus, and 6) permeability properties like gas (He, H_2, CO_2, N_2, O_2, and CH_4) permeability. Each of these predictive models within PG could potentially go through an improvement due to new data infusion.

MATERIALS AND METHODS

Data for this work were obtained from publicly available collections of experimental measurements (9, 10) and an online repository of polymer properties (11). The new polymer dataset is highly diverse, and its constituent polymers are composed of nine atomic species: carbon, hydrogen, oxygen,

(a)



(b)

Polymer name	SMILES	$T_{g,experimental}$ (K)	$T_{g,PG-0}$ (K)	$T_{g,PG-1}$ (K)
Polystyrene	CC(C1=CC=CC=C1)	373	368	353
Poly(propylene)	CC(C)CC(C)CC(C)	264	280	256
Poly(caprolactone)	CCCCCOC(=O)CCCCCOC(=O)	208	229	217
Poly(ethylene adipate)	CCCCC(=O)OCCOC(=O)	230	243	235
Poly(ethylene glycol)	CCOCCO	206	185	189
Polyglycolide	CC(=O)OCC(=O)O	312	308	312
Polyacetal	COCO	211	203	204
Poly(phenylene sulfide)	C(C=C1)=CC=C1SC(C=C2)=CC=C2S	391	378	372
Poly(p-xylene)	CC(C=C1)=CC=C1C	343	357	349
...	...	...	...	...

Figure 3: (a) Distribution of the T_g values for all polymers considered in this work. (b) Sample polymer dataset with SMILES representations and experimental T_g values.

nitrogen, sulfur, fluorine, chlorine, bromine and iodine. The T_g of the 1321 polymers (450 polymers from previous work and 871 newly collected polymers) in the dataset varied widely, ranging from 76 K to 873 K with a mean of 354 K (**Figure 3a**). The repeat units of the polymers were represented using the simplified molecular-input line-entry system (SMILES) (12). Examples of SMILES representations are shown in **Figure 3b** with the original name of polymers and T_g.

In order to capture the key features that may control T_g, we utilized the hierarchical polymer fingerprinting scheme (13). The fingerprint building process involves assessing three hierarchical levels of features. The first is at the atomic scale, wherein atomic fragments occur. This set of descriptors captures the type of atoms and atomic connectivity in the polymers. For our 1321 polymers, there are 128 such components. The next level deals with quantitative structure property relationship descriptors (14), such as the estimated surface area of the polymer repeating unit and fraction of rotatable bonds. Such descriptors, 39 in total, form the next set of components of our overall fingerprint. The third level descriptors captured morphological features, such as the topological distance between aromatic rings and the length of sidechains. We include 22 morphological features in the fingerprint.

The ML model was built by mapping the descriptors to the T_g values using GPR with a sum-kernel consisting of a radial basis function kernel and a white-noise kernel. During the model development step, we used an out-of-sample testing scheme to validate the ML model. We randomly partitioned the T_g dataset so that 80% was used for the model training and 20% was used for the validation of the trained model (5-fold cross-validation). Among five models trained on different random choices of training set, the single model delivering the most accurate predictions for the test set was selected as the best cross-validation model. Using the fixed kernel with the hyper-parameters extracted from this model, we obtained the final model which was retrained on our entire dataset. In the prediction step, data points with fingerprints very close to the new fingerprint value are weighted more than data points with fingerprints farther away. This means that if the new polymer is similar in terms of fingerprint to some polymers already in the data set, GPR will give a T_g value close to that of those similar polymers. Details of the approach used may be found in previously published work (13).

Received: August 3, 2019
Accepted: March 4, 2020
Published: March 18, 2019

REFERENCES

1. Krevelen, Dirk Willem van. *Properties of Polymers: Their Correlation with Chemical Structure*. Elsevier, 1997.
2. "Materials Genome Initiative." www.mgi.gov
3. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A., and Kim, C. "Machine Learning in Materials Informatics: Recent Applications and Prospects." *npj Computational Materials*, vol. 3, 2017, pp. 54.
4. Jordan, M. I. and Mitchell, T. M. "Machine Learning: Trends, Perspectives, and Prospects." *Science*, vol. 349, no. 6245, 2015, pp. 255-260.
5. Williams, C. K. I., and Rasmussen, C. E. "Gaussian Processes for Regression." *Advances in Neural Information Processing Systems 8*. MIT Press: Cambridge, 1996, pp. 514-520.
6. "Polymer Genome: An informatics platform for polymer property prediction and design using machine learning" www.polymergenome.org
7. Shafkat, I. "Intuitively Understanding Variational Autoencoders." *Towards Data Science*, https://towardsdatascience.com/intuitively-understanding-variational-autoencoders-1bfe67eb5daf.
8. Sadeghi, J., Sadeghi, S., and Niaki, Seyed Taghi A. "Optimizing a hybrid vendor-managed inventory and transportation problem with fuzzy demand: An improved particle swarm optimization algorithm." *Information*

Sciences, vol. 272, no. C, 2014, pp. 126-144.
9. Brandup, J., Immergut, E.H., and Grulke, E.A. *Polymer Handbook*. 4th ed., 1999, John Wiley and Sons, New York.
10. Askadskii, A. A. *Computational Materials Science of Polymers*. 2003, Cambridge: Cambridge Int.
11. "Chemical Retrieval on the Web (CROW)" http://polymerdatabase.com
12. Weininger, D. "SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules." *Journal of Chemical Information and Modeling*, vol. 28, no. 1, 1988, pp. 31-36.
13. Kim, C., Chandrasekaran, A., Huan, T.D., Das, D., and Ramprasad, R. "Polymer Genome: A Data-powered Polymer Informatics Platform for Property Predictions." *Journal of Physical Chemistry C*, vol. 122, no. 31, 2018, pp. 17575-17585.
14. Nantasenamat, C., Isarankura-Na-Ayudhya, C., Naenna, T., and Prachayasittikul, V. A. "Practical Overview of Quantitative Structure-Activity Relationship." *EXCLI Journal*, vol. 8, 2009, pp. 74-88.