Supplementary Information: Frequency-dependent Dielectric Constant Prediction of Polymers using Machine Learning

Lihua Chen[1], Chiho Kim[1], Rohit Batra[1], Jordan P. Lightstone[1], Chao Wu[2], Zongze Li[2], Ajinkya A. Deshmukh[3], Yifei Wang[2], Huan D. Tran[1], Priya Vashishta[4], Gregory A. Sotzing[3], Yang Cao[2], and Rampi Ramprasad[1,*]

[1]School of Materials Science and Engineering, Georgia Institute of Technology, 771 Ferst Drive NW, Atlanta, GA 30332, United States
[2]Electrical and Computer Engineering, University of Connecticut, 371 Fairfield Way, Storrs, CT 06269, United States
[3]Polymer Program, University of Connecticut, 97 North Eagleville Road, Storrs, CT 06269, United States
[4]Collaboratory for Advanced Computing and Simulations, University of Southern California, Los Angeles, California 90089-0242, United States
[*]rampi.ramprasad@mse.gatech.edu

Contents

1 Similarity analysis of 11,000 unseen polymers and 738 training polymers **2**

2 Data-points-split based ML model **2**

2.1 Stratified sampling . . . 2

2.2 Feature reduction . . . 3

2.3 Dielectric constant vs. frequency of PET and PDTC-HK511 . . . 3

2.4 Validation of ML predicted dielectric constant at THz . . . 3

2.5 Model parameters . . . 3

GPR • LASSO feature reduction

3 Frequency-dependent ε trend of 11,000 unseen polymers **5**

4 Single ML model vs Two separate ML models **5**

1 Similarity analysis of 11,000 unseen polymers and 738 training polymers

In order to estimate the similarity of 11,000 unseen polymers and 738 training polymers, we have computed the percentage of shared chemical space groups of two datasets. Agglomerative hierarchical clustering (AHC), implemented in scikit-learn, was utilized to cluster the chemical space of two datasets with the PC1 and PC2 values shown in Figure 2b of the article. The Euclidean distance is used to measure the distance between two data points and Ward linkage is used to measure the distance between two clusters. Figure 1 shows the PCA with 100 groups, including 11,000 polymers (colored circles) and 738 training polymers (white squares). The corresponding percentage of shared groups of the two datasets is 90%. Therefore, we believe the chemical space of these two datasets is quite similar.

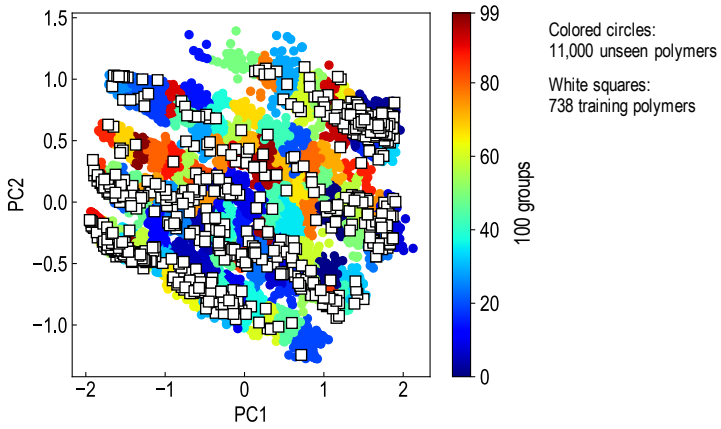


Figure 1. PCA plot with 100 groups clustered using the AHC method. Colored circles are 11,000 unseen polymers and white squares are 738 training polymers.

2 Data-points-split based ML model

2.1 Stratified sampling

We have tried to use stratified sampling of chemical space to generate train and test sets and train the ML models as follows:

(1) Data points split:

(a) Principal components analysis (PCA) was performed on the 411 chemical features of 738 training polymers to reduce the feature dimensionality.

(b) Agglomerative hierarchical clustering (AHC), implemented in scikit-learn library, was utilized to cluster the 738 training polymers using the first two principal components (PC1 and PC2) from PCA. The Euclidean distance is used to measure the distance between two data points and Ward linkage is used to measure the distance between two clusters. The resulting bottom-up dendrogram is shown in Figure 2(a). The chemical space of training polymers can be divided into different number groups based on the Euclidean distance, e.g., 5 and 13 groups (dashed lines in Figure 2(a)). The corresponding PCA plots with 5 and 13 groups are shown in Figure 2(b).

(c) We used the stratifiedshufflesplit method implemented in scikit-learn library to split train/test sets by stratified sampling different groups of chemical space (i.e., 5 or 13 groups shown in Figure 2(b)). The same procedure is also applied in all cross-validation processes.

(2) Feature reduction to train ML models:

LASSO method was used to eliminated irrelevant features from 412 features (411 chemical features + log frequency) with the stratifiedshufflesplit CV (5-fold).

(3) ML model training:

GPR algorithm is used to train the ML model using LASSO reduced features, stratifiedshufflesplit train/test sets and 5-fold stratifiedshufflesplit CV.

The learning curves of ML models trained by LASSO reduced features and two sampling methods including (random sampling used in the paper, stratified sampling-5 and 13 groups) are shown in Figure 3. The average training and test RMSE of ε prediction as a function of training set size are plotted, with the error bars denoting 1 sigma standard deviation in the reported RMSE values over 50 runs. We note that the ML performance trained by the random (in the article) and stratified sampling is comparable.

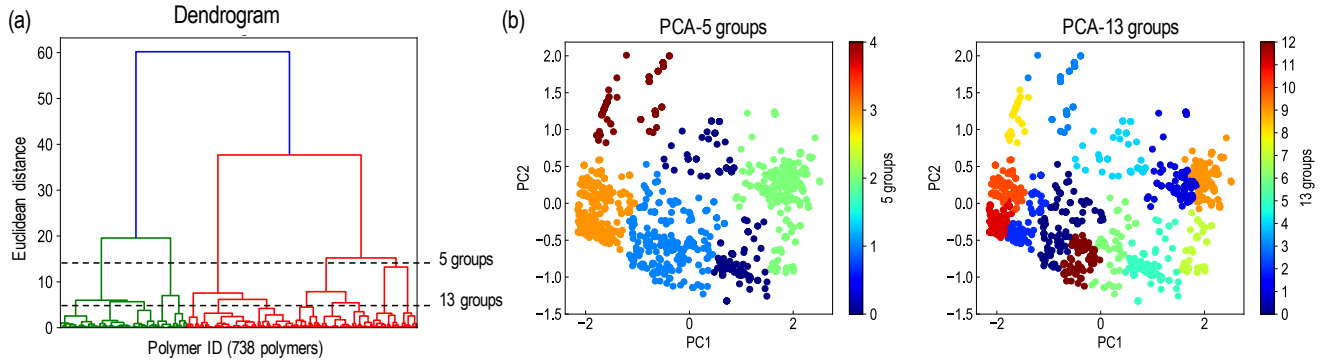


Figure 2. (a) Bottom-up dendrogram of 738 polymers, obtained from AHC. (b) PCA plots with 5 groups and 13 groups.

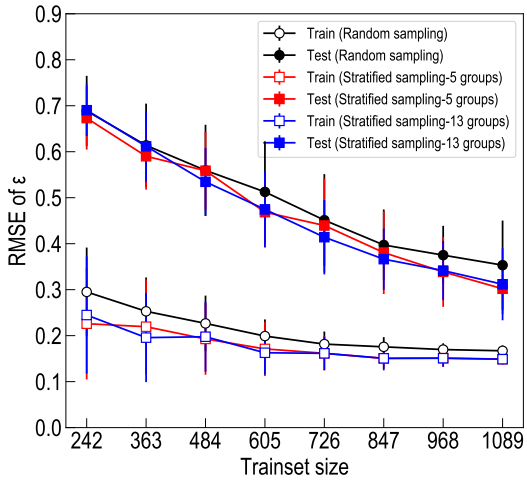


Figure 3. Learning curves of ML models trained by LASSO reduced features and two sampling methods including (random sampling used in the paper, stratified sampling-5 and 13 groups). Error bars denote 1 sigma standard deviation in the reported RMSE values over 50 runs.

2.2 Feature reduction

In addition to LASSO, we have used recursive feature elimination (RFE) using linear support vector regression algorithm (with 5-fold CV) to backward eliminate irrelevant features. Figure 4 shows learning curves of ML models trained by all features (GPR-X_{All}), LASSO reduced features (GPR-X_{LASSO}), and RFE reduced features (GPR-X_{RFE}). Error bars denote 1 sigma standard deviation in the reported RMSE values over 50 runs. We note that the test RMSE of GPR-X_{LASSO} model is generally lower than that of GPR-X_{RFE}. LASSO does a better job in this work, probably this method retains the relevant features by optimizing the regularization term to achieve the highest R2. As a result, key features in determining the sparse high-dielectric constant region (>8) are kept. However, the choice does not significantly affect the key finding of this paper.

2.3 Dielectric constant vs. frequency of PET and PDTC-HK511

Figure 5 shows Expt. vs ML predicted ε of PET and PDTC-HK511 in Figure 3a2 and b2 of the article.

2.4 Validation of ML predicted dielectric constant at THz

Figure 6 shows the ML predicted ε of PET, PP and PVC at 10^{12} Hz using the data-points-split model developed here (denoted by ML-this work) and previously trained by DFPT computed data (denoted by ML-DFPT), respectively. We also provide the experimental and ML predicted ε at 10^{9} and 10^{15} Hz. Figure 6 reveals that the present ML model provides a correct frequency vs ε trend, while the ε at 10^{12} Hz is overestimated by the ML-DFPT model. The reason for this discrepancy is the over-estimation of DFPT computed ε values, which are computed using unrealistic crystalline structures of polymers having unreasonably higher densities than realistic semi-crystalline or amorphous case.

2.5 Model parameters

2.5.1 GPR

The GPR-RBF kernel parameters for the final training model are as follows:

$3.33^{2}2$ * RBF(length-scale=1.33) + WhiteKernel(noise-level=0.0351)

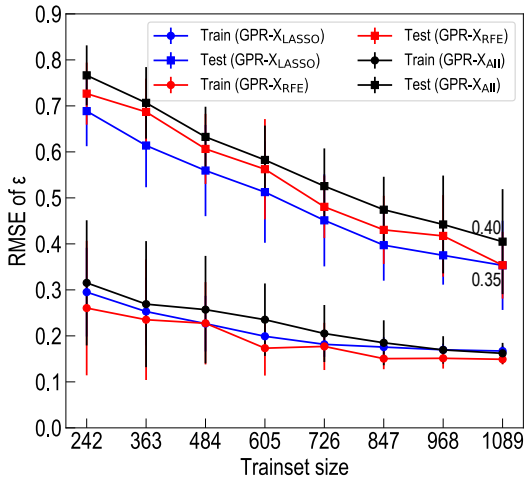


Figure 4. Learning curves of ML models trained by all features (GPR-X_{All}), LASSO reduced features (GPR-X_{LASSO}), and RFE reduced features (GPR-X_{RFE}). Error bars denote 1 sigma standard deviation in the reported RMSE values over 50 runs

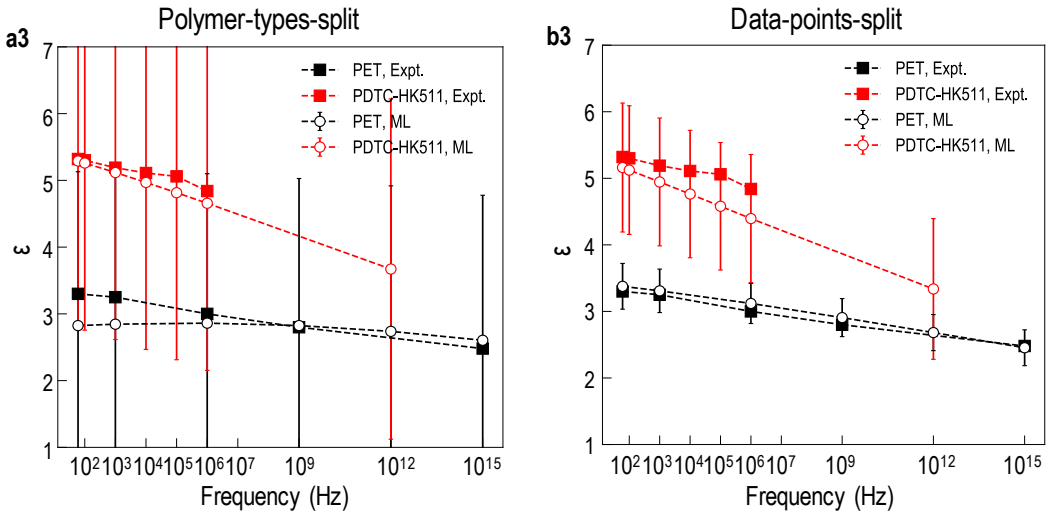


Figure 5. Expt. vs ML predicted ε of PET and PDTC-HK511 in Figure 3a2 and b2 of the article, respectively, with frequency = 60, 10^2, 10^3, 10^4, 10^5, 10^6, 10^7, 10^9 and 10^{15} Hz. Furthermore, the additional ML predicted ε values at 10^{12} Hz of these 2 polymers are shown.

2.5.2 LASSO feature reduction

In the data-points-split models, 53 features were internally selected from 412 initial features using the LASSO method with the optimal regularization term (α) of 0.003 to achieve the highest R2 of 0.685. Figure 7 summarizes the LASSO coefficients of 30 representative features.

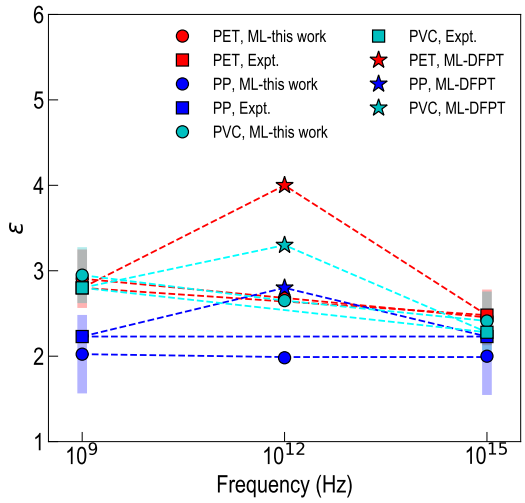


Figure 6. ML predicted ε at 10^{12} Hz of PET, PP and PVC using the models developed here (ML-this work) and previously trained by DFPT computed data (ML-DFPT), respectively. The experimental and ML predicted ε at 10^{9} and 10^{15} Hz are also provided.

3 Frequency-dependent ε trend of 11,000 unseen polymers

To quantify the prediction accuracy of the frequency-dependent ε trend shown in Figure 5a of the article, we evaluate $\Delta\varepsilon = \varepsilon_{\text{lowF}} - \varepsilon_{\text{highF}}$ of 11,000 unseen polymers, i.e., the difference of ε at the low frequency (low F) and the neighboring high frequency (high F). Figure 8 presents the ratio of polymers with $\Delta\varepsilon < 0$ and $\Delta\varepsilon < -0.01$ at 10 pairs of low and high frequencies (given in Hz), i.e., (60, 10^2), (10^2, 10^3), (10^3, 10^4), (10^4, 10^5), (10^5, 10^6), (10^6, 10^7), (10^7, 10^8), (10^8, 10^9), (10^9, 10^{12}) and (10^{12}, 10^{15}). We note that only less than 6% cases have $\Delta\varepsilon < -0.01$, suggesting our ML models are indeed reliable and accurate for predicting frequency-dependent ε of unseen polymers.

4 Single ML model vs Two separate ML models

Apart from the single ML model presented in this work, we had built two separate ML models for different frequency regimes, one for a low frequency (60-10^9 Hz, 698 data points) and other at optical frequency (521 data points), using the same workflow in the paper. The validation set accuracy of the two separate models was slightly higher than that of the single model presented in the article, since a small dataset for each frequency range was used to train the models. However, the frequency-dependent trends of the predicted dielectric constant using two separate ML models were incorrect for multiple polymers. As shown in Figure 9, 10 proposed polymers in Figure 6 of the article were predicted using the two separate ML models, together with the results obtained from the single ML model. This is because the two separate models were completely independent. When the log-frequency was used as a feature in the single ML model (presented in this work), the frequency-dependent trend was learned from the training data itself. Furthermore, we found the single ML model to be more generalizable for new cases, as it was trained using a larger polymer dataset. Additional advantages of having log-frequency as a feature is that it allows us to make predictions of dielectric constant at any arbitrary frequency value, which is not possible with separate ML models, e.g., THz values.

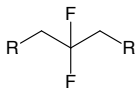
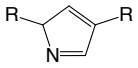
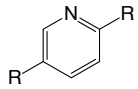
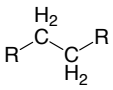
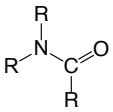
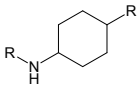
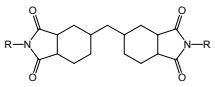
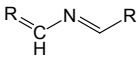
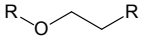
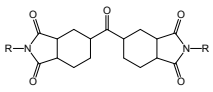
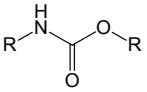
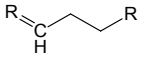
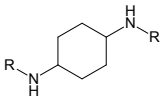
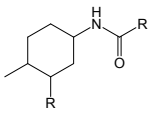
ID	Feature	Coefficient	ID	Feature	Coefficient	ID	Feature	Coefficient
1	log F	-0.95	2		3.57	3		0.43
4		-0.80	5		2.26	6		0.43
7		-0.44	8		1.31	9		0.38
10		-0.43	11		1.19	12	Number of H-bond acceptor	0.34
13	Number of 3-vertex carbon atoms	-0.24	14	Topological polar surface area	1.04	15		0.18
16		-0.13	17		0.86	18		0.08
19	Number of cyclic double bonds in a repeat unit	-0.21	20		0.51	21		0.58
22	Presence of a purely single bond	-0.44	23		0.43	24		0.35
25		-0.08	26		0.16	27		0.045
28		-0.04	29		0.038	30		0.023

Figure 7. LASSO coefficients of 30 representative features. log F denotes the log-scale frequency value used as a feature in the ML model.

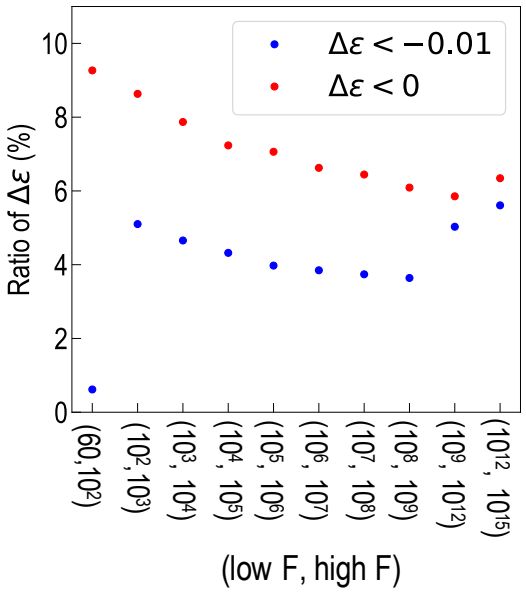


Figure 8. $\Delta\varepsilon = \varepsilon_{\text{lowF}} - \varepsilon_{\text{highF}}$ of 11,000 unseen polymers shown in Figure 5a of the article. Low and high F denote low and high frequencies (given in Hz), respectively.

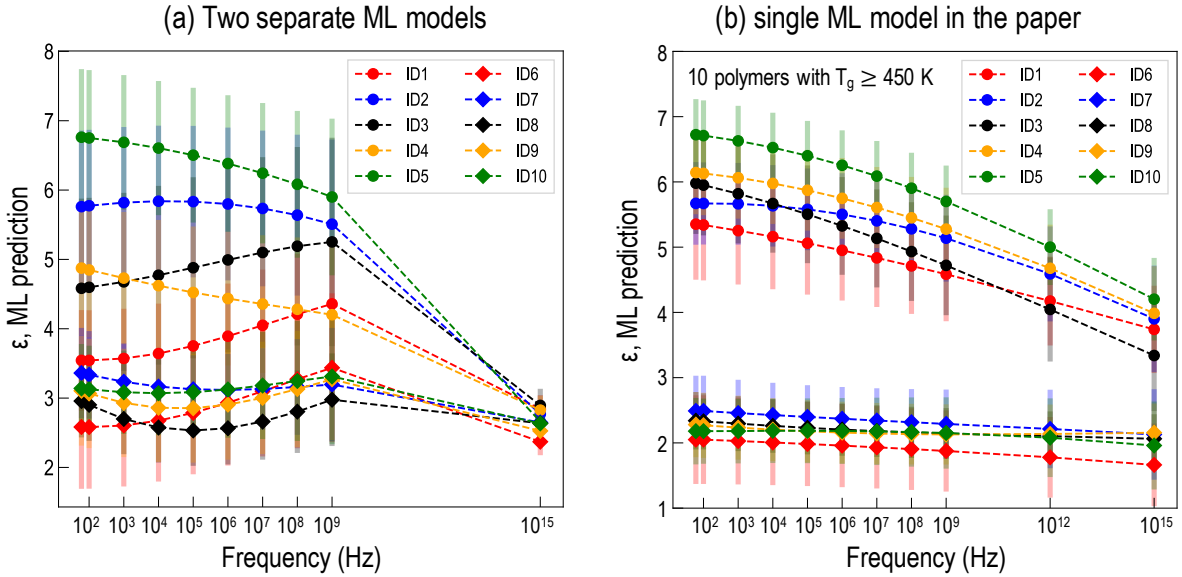


Figure 9. ML predicted frequency-dependent ε of 10 polymers proposed in Figure 6 of the article using (a) Two separate ML models: one for a low frequency (60-10^9 Hz, 689 data points) and other at optical frequency (521 data points) (b) single ML model presented in the work.