

ARTICLE OPEN



Frequency-dependent dielectric constant prediction of polymers using machine learning

Lihua Chen¹, Chiho Kim¹, Rohit Batra¹, Jordan P. Lightstone¹, Chao Wu², Zongze Li², Ajinkya A. Deshmukh³, Yifei Wang², Huan D. Tran¹, Priya Vashishta⁴, Gregory A. Sotzing³, Yang Cao² and Rampi Ramprasad¹✉

The dielectric constant (ϵ) is a critical parameter utilized in the design of polymeric dielectrics for energy storage capacitors, microelectronic devices, and high-voltage insulations. However, agile discovery of polymer dielectrics with desirable ϵ remains a challenge, especially for high-energy, high-temperature applications. To aid accelerated polymer dielectrics discovery, we have developed a machine-learning (ML)-based model to instantly and accurately predict the frequency-dependent ϵ of polymers with the frequency range spanning 15 orders of magnitude. Our model is trained using a dataset of 1210 experimentally measured ϵ values at different frequencies, an advanced polymer fingerprinting scheme and the Gaussian process regression algorithm. The developed ML model is utilized to predict the ϵ of synthesizable 11,000 candidate polymers across the frequency range 60–10¹⁵ Hz, with the correct inverse ϵ vs. frequency trend recovered throughout. Furthermore, using ϵ and another previously studied key design property (glass transition temperature, T_g) as screening criteria, we propose five representative polymers with desired ϵ and T_g for capacitors and microelectronic applications. This work demonstrates the use of surrogate ML models to successfully and rapidly discover polymers satisfying single or multiple property requirements for specific applications.

npj Computational Materials (2020)6:61 | <https://doi.org/10.1038/s41524-020-0333-6>

INTRODUCTION

Polymers are important dielectric materials that are often used for a wide range of applications, including high-energy-density capacitors^{1–9}, high-voltage cables¹⁰, microelectronics¹¹, and photovoltaic devices^{12,13}. Each application requires a given range of the polymer dielectric constant ϵ , also referred to as the relative permittivity. High ϵ polymers are needed for high-energy-density capacitors and photovoltaic devices to allow facile charge extraction. On the other hand, polymers with low ϵ are needed in other applications, e.g., to reduce signal-delay time in microelectronics, and lower conduction loss in high-voltage cables. Extensive efforts are undertaken to optimize device performance by tailoring the ϵ of a given polymer. As a common example in the capacitor domain, many strategies have been proposed to increase ϵ of polymers via doping/coating high ϵ inorganic particles (e.g., BaTiO₃)^{14,15}, grafting/blending with highly polar polymers (e.g., polyvinylidene fluoride, PVDF)¹⁶ or metal-organic polymers¹⁷. However, such modifications are almost always accompanied with new challenges, e.g., reduced breakdown strength, high dielectric loss and increased film processing cost. Therefore, it is highly desirable to design pure all-organic polymers with tailored ϵ values^{4,8,18,19}, while not compromising other attractive and necessary attributes.

ϵ is related to the electric polarization of a material under an alternating electric field^{20,21}. It consists of three contributions, arising from electronic (ϵ_{elec}), ionic (ϵ_{ionic}), and dipolar ($\epsilon_{\text{dipolar}}$) polarization. Each of these polarization mechanisms have different response times, resulting in different contributions to the overall ϵ as a function of the applied frequency—above optical frequencies only ϵ_{elec} contributions are relevant, in THz regime both ϵ_{elec} + ϵ_{ionic} contribute, and at lower frequencies all of the ϵ_{elec} + ϵ_{ionic} + $\epsilon_{\text{dipolar}}$ contributions are significant. Thus, generally, ϵ decreases

with an increase in the applied frequency (ignoring certain near-singularity artifacts at the resonant frequencies). This also suggests the significance of obtaining the complete frequency-dependent ϵ behavior for polymers, rather than a particular ϵ value at a single frequency. Extensive computational efforts have been made to compute the ϵ of polymers in the higher-frequency (THz) regimes^{7,22}. For example, density functional perturbation theory (DFPT) has been used to compute the ϵ of crystalline polymers with acceptable accuracy^{7,22}. However, this method is computationally expensive and restricted to small systems (<50 atoms). As a result, the computed ϵ can only account for ϵ_{elec} and ϵ_{ionic} parts, excluding the $\epsilon_{\text{dipolar}}$ contributions arising from block- and chain-level changes in the polymers. Furthermore, the assumption of dense ordered crystalline structures commonly made in these computations (to allow for a small unit cell) leads to an overestimation of the ϵ_{ionic} part. These issues can be partly solved by using large-scale classical molecular dynamics (MD) simulations²³, but these are restricted to polymer systems with reliable classical force field.

Data-driven techniques are popular and powerful alternatives to build surrogate models for property prediction and material design, greatly accelerating the (discovery and application of new materials^{8,24–29}). In the polymer domain, group contribution methods have been developed to predict various properties of polymers, such as ϵ ²¹. However, major problems with this approach include the inability to generalize to new polymers containing functional groups outside the library of considered groups, and the disregard of sequence and connections of the constituting functional groups. A recent successful development has been to use hand-crafted features (also called descriptors or fingerprints) within the context of machine-learning (ML) models for polymer property prediction^{6,22,30–32}. Although reliable

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

ϵ -prediction models were developed in our previous work³², those are limited by the accuracy of the underlying DFPT dataset, especially due to the assumption of crystalline polymer structures (as mentioned above). More importantly, those models cannot predict the complete frequency-dependent ϵ behavior.

In this work, we develop an ML model to predict the frequency-dependent ϵ behavior of polymers, using a dataset of 1210 experimentally measured values at various frequencies (spanning 15 orders of magnitude). This is achieved using a 3-level hierarchical polymer fingerprinting scheme and the Gaussian process regression (GPR) algorithm to train the model, as shown in Fig. 1. The resulting ML model can accurately and rapidly predict ϵ of new polymer candidates across a wide range of frequencies, as validated using the performance on unseen test set. To better understand the ML models developed and derive simple chemical trends, we investigate the key chemical features that dominate the ϵ of polymers. Furthermore, to showcase the predictive power and the usefulness of the developed surrogate models, we computed the frequency-dependent ϵ of a candidate set of 11,000 unseen polymers manually accumulated from various available sources^{7,21,32–34}. Another critical design property (glass transition temperature, T_g), reflective of the thermal stability of these polymers, was predicted using our previously developed ML model³². Using these two predicted properties, five representative polymers satisfying specific ϵ and T_g requirements are proposed for capacitor and microelectronic applications.

RESULTS

Dataset and polymer fingerprints

As illustrated in Fig. 2a, 1210 experimental ϵ values belonging to 738 unique polymers were collected from the literature^{9,19,21,33,35–42} to train the ML models. These measurements were made at 9 frequency values (i.e., 60, 10^2 , 10^3 , 10^4 , 10^5 , 10^6 , 10^7 , 10^9 , and 10^{15} Hz), at room temperature and under dry conditions. Here, ϵ values at 10^{15} Hz represent the optical frequency region and were obtained by taking the square of the experimental refractive

index. Given the limitation of available experimental values, each polymer in Fig. 2a has ϵ values available at 1–8 frequency values. Furthermore, this 738-polymer dataset includes 11 elements, i.e., C, H, B, O, N, S, P, Si, F, Cl, and Br and various polymer classes, e.g., polycarbonates, polyimide, polyamide, polyolefins, polyvinyl, polyethers and polyesters. The ϵ distribution as a function of frequency (in Hz) is presented in Fig. 2a, along with the corresponding polymer count at each frequency. We note that the ϵ dataset ranges from 1.3 to 11 and is slightly unbalanced in terms of data count at different frequencies. This can be attributed to the difficulties experienced when making empirical measurements at various frequencies, but we believe that the data diversity is sufficient to build reliable regression models. The trends in ϵ values for 6 common and diverse polymers highlighted in Fig. 2a signify the importance of polymer chemistry. It is worth noting that ϵ of polar polymers like PVDF and polyvinyl alcohol (PVA) significantly decreases with an increase in frequency while for non-polar polymers, such as polypropylene (PP) and ETFE, ϵ is not sensitive to the applied frequency. Therefore, for the ML model to capture such trends accurately, it is essential that the dataset is representative and balanced in terms of polymer chemistry and count, respectively. More details on the ϵ dataset are provided in the “Methods” section.

The next important step towards building accurate and reliable ML models is to generate relevant features that uniquely represent each polymer and also capture its frequency-dependent ϵ behavior. To capture the polymer chemistry, we used features from three hierarchical levels, i.e., (1) atomic-level fragments, (2) block-level fragments, and (3) chain-level features. A total of 411 chemical features were used to numerically fingerprint 738 polymers. Additionally, the frequency in log-scale ($\log F$) was incorporated as the key feature to capture the frequency-dependent behavior, overall resulting in a 412-dimensional feature vector. Next, the least absolute shrinkage and selection operator (LASSO) method was adopted for dimensionality reduction and elimination of irrelevant features. The details on the fingerprinting scheme and the use of the LASSO method are

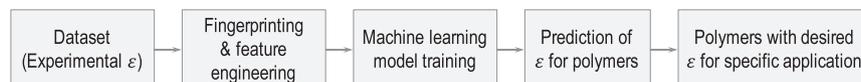


Fig. 1 Machine-learning workflow. Schematic of the workflow adopted to build general data-driven models of frequency-dependent ϵ for polymers.

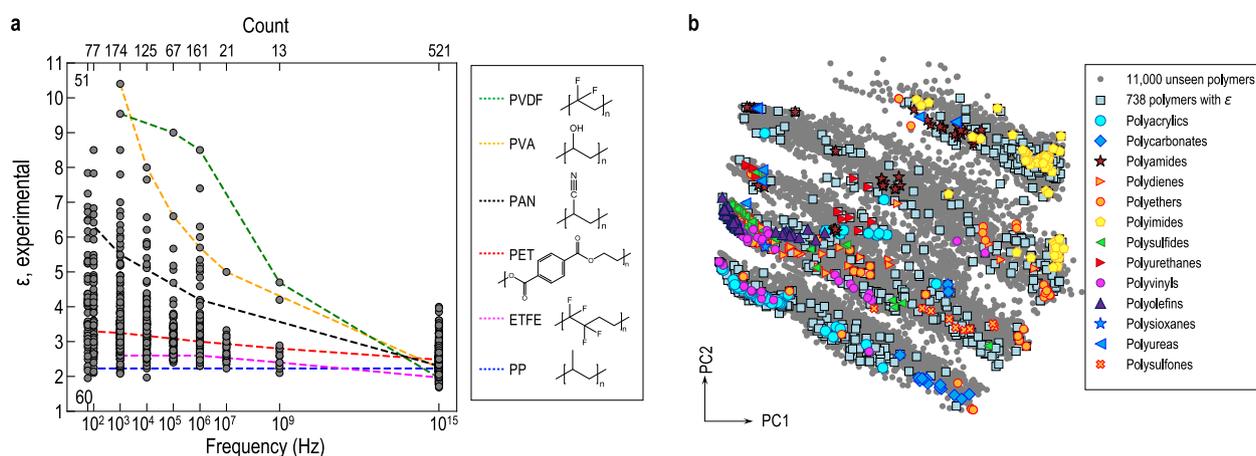


Fig. 2 Experimental dielectric constant dataset and the chemical space of training and unseen datasets. **a** Experimental ϵ as a function of the frequency (unit, Hz), along with the data count at each frequency. The trends in ϵ values of six representative polymers are also shown using dashed lines. **b** Chemical space of the training set (738 polymers) considered this work (light blue squares), with respect to a larger unseen dataset of 11,000 polymers (gray circles), illustrated using the first two principal components (PC1 and PC2). A few representative polymer classes of the training dataset are highlighted with colored symbols.

Table 1. Details of ML models. N_x is the number of features.

Models	Train-validation-test split	Feature reduction	N_x	ML
a	Polymer-types (738),	None	412	GPR-RBF
	Group-shuffle-split, fivefold	LASSO	57	
b	Data-points (1210),	None	412	GPR-RBF
	K-fold, fivefold	LASSO	53	

included in the “Methods” section, while the final number of features retained for model development are summarized in Table 1.

To validate the generality, reliability and usefulness of the ML models developed in this work, the frequency-dependent ϵ of an unseen dataset of 11,000 candidate polymers previously synthesized elsewhere (but for which no dielectric characterization has been done)^{7,21,32–34}, were predicted. This unseen dataset contains polymers distinct from the training dataset (of 738 polymers), but is made up of the same 11 elements, i.e., C, H, B, O, N, S, P, Si, F, Cl, and Br. Furthermore, the chemical diversity of this unseen dataset is quite similar to that of the training dataset (of 738 polymers), as illustrated in Fig. 2b using the first two (PC1 and PC2) components obtained from the principal component analysis (PCA) on chemical features of all polymers. The similarity of two datasets is further discussed using the agglomerative hierarchical clustering analysis in Supplementary Section 1. Note that the training dataset (light blue square) spans the chemical space well, indicating that it is representative of the unseen polymer dataset (gray circles). Several representative polymer classes of the training dataset are also labeled with colored symbols in Fig. 2b.

Frequency-dependent machine-learning models of dielectric constant

Considering that ϵ depends on both polymer-type and the applied frequency, the ML models (using the GPR algorithm) were trained in two different fashions with varying train-validation-test splits, referred to here as the (1) polymer-types-split (738 polymers) and (2) data-points-split (1210 points) approach. In the former split, the test set consists of completely different polymers than those in the training set, resulting in evaluation of ML performance on unseen polymer cases. While both random and stratified sampling methods were used in the latter to split train-validation-test sets across all polymers and all frequencies, as discussed in Supplementary Section 2.1. The random sampling method is selected in the present work due to the comparable ML performance of two sampling methods. For all models, fivefold cross-validation (CV) was used to avoid overfitting, and two error metrics, namely, root mean square error (RMSE) and the coefficient of determination (R^2), were used to evaluate their performance.

Figure 3a1, b1 show the learning curves of the ML models trained using polymer-types-split and data-points-split methods, respectively. The average training and test RMSE of ϵ prediction as a function of training set size is plotted, with the error bars denoting 1σ standard deviation in the reported RMSE values over 50 runs. Results for both the cases, i.e., with all 412 features (GPR- X_{All}) and with those retained after LASSO dimensionality reduction (GPR- X_{LASSO}) are included. As expected, the test RMSE decreases with an increase in training set size for all cases. We note that the GPR- X_{LASSO} does a better job of improving the ML performance when trained using the data-points-split approach in comparison with the polymer-types-split approach. Further, a higher test RMSE of 0.67 resulted in polymer-types-split models using 90 % training set (664 polymers), while a test RMSE of 0.35 was obtained in data-points-split models (with 1089 training points). Considering the ϵ

dataset ranges from 1.3 to 11, this amounts to an error of $\lesssim 7\%$. In addition to the LASSO feature reduction method, the recursive feature elimination (RFE) using linear support vector regression algorithm was used in the data-points-split model to backward eliminate irrelevant features. The corresponding learning curve is shown in Supplementary Fig. 4, revealing that the GPR- X_{LASSO} model provides higher prediction accuracy.

To further validate the generality and accuracy of the two ML models, all frequency-dependent information of five common polymers, namely, polyethylene terephthalate (PET), polypropylene (PP), polyacrylonitrile (PAN), polyvinyl chloride (PVC) and PDTC-HK511, was intentionally included in the 10% test set (completely unseen by the 90% train set). These five polymers were selected based on their difference in polarity, wide range of ϵ values, and larger availability of frequency-dependent data. The resulting parity plots between ML prediction vs. experimental ϵ using the GPR- X_{LASSO} models are portrayed in Fig. 3a2, b2. The error bars in these cases represent the GPR uncertainty and the size of markers denote the frequency applied. It can be seen that the R^2 for the test set of polymer-types splits and data-points splits models is 0.74 and 0.92, respectively. The corresponding frequency-dependent ϵ behavior for PP, PVC, and PAN polymers is shown in Fig. 3a3, b3. The remaining two polymers (PET and PDTC-HK511) are available in Supplementary Fig. 5. It can be observed that frequency-dependent ϵ trend for PP and PAN are predicted fairly well using the polymer-type-split models, although the GPR uncertainties are slightly high due to absence of similar polymer chemistry within the training set. This issue is, however, greatly improved in the data-points-split model, wherein more polymer types (695) are included in the training set as compared to that in the polymer-type-split method (with 664 polymers).

A major benefit of the presented ML models is their ability to predict ϵ across a wide range of frequencies (60– 10^{15} Hz). In Fig. 3a3, b3, we also show the ϵ predictions for the three unseen polymers at 10^{12} Hz, where empirical data is unavailable. The ML predictions can be seen to closely follow the available frequency-dependent ϵ trend. We also compare these models with our previous work utilizing DFPT-based computed ϵ values at THz frequency (denoted as ML-DFPT). As illustrated in Supplementary Fig. 6, the ML-DFPT predicted ϵ of PET, PP, and PVC are much higher than their corresponding experimental values at 10^9 Hz, leading to incorrect frequency-dependent ϵ trend; ϵ value should decrease with increase in frequency. The reason for this discrepancy is the overestimation 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. On the other hand, the present ML models utilize information available at different frequencies (both the lower regime and the higher optical region) to accurately predict the ϵ values at 10^{12} Hz.

Overall, Fig. 3 shows that the data-points-split-based ML models perform better than their polymer-types-split-based counterparts in terms of test RMSE, the error trends in the learning curve, and the prediction capability of five completely unseen polymers. Such observation is expected and understandable because of inclusion of fewer polymer types in the polymer-types-split training set. Moreover, in the data-points-split approach it is possible that the same polymers with different frequencies are randomly sampled in the training and the test sets, thus improving the ML performance. From a theoretical standpoint, these two ML models provide predictive capability of ϵ at two extremes: data-points-split model is appropriate for polymer cases with some known frequency-dependent ϵ values, while polymer-types-split model is applicable for completely new polymers with no ϵ information. With these systematic and careful studies, we believe that the random data-points-split approach is reliable and appropriate to

a Polymer-types-split

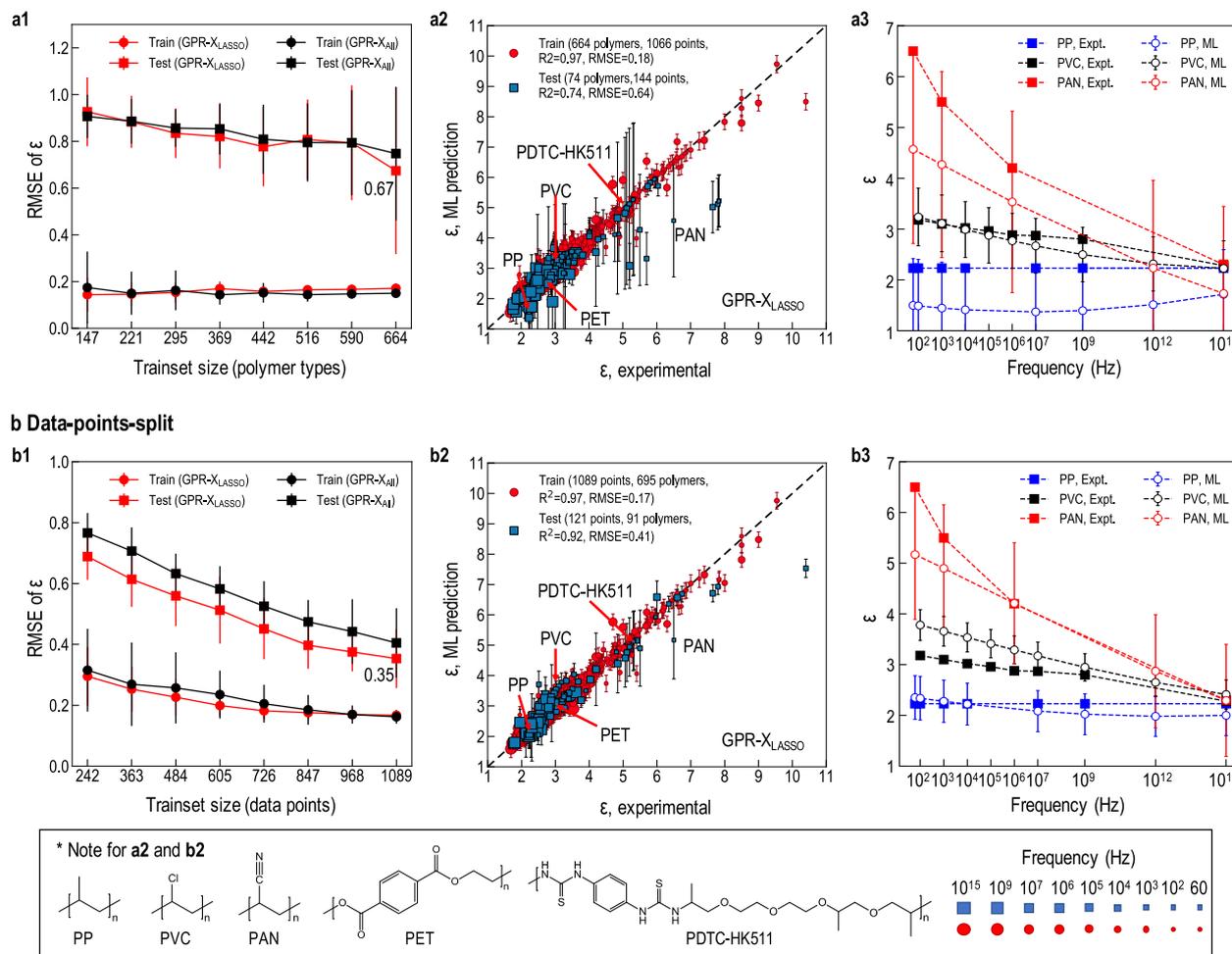


Fig. 3 Machine-learning models of dielectric constant. ML models of ϵ based on polymer-types-split **a** and data-points-split **b**. **a1** and **b1** are learning curves trained using all features (GPR- X_{all}) and LASSO (GPR- X_{LASSO}) reduced features, with the error bars denoting 1σ standard deviation in the reported RMSE values over 50 runs. **a2** and **b2** are parity plots using GPR- X_{LASSO} and the 90% train set, where all frequency-dependent information of five polymers (PP, PET, PAN, PVC, and PDTC-HK511) were intentionally included in the 10% test set. Symbol sizes represent the frequency applied. **a3** and **b3** show Expt. vs ML predicted ϵ of PP, PVC and PAN in **a2** and **b2**, respectively, with frequency = 60, 10^2 , 10^3 , 10^4 , 10^5 , 10^6 , 10^7 , 10^9 , and 10^{15} Hz. The remaining two polymers (PET and PDTC-HK511) are available in Supplementary Fig. 5. Furthermore, the additional ML predicted ϵ values at 10^{12} Hz of these three polymers are shown. Error bars in **a2**, **a3**, **b2**, and **b3** are predicted GPR uncertainties.

be used to train the final predictive model with the entire dataset and CV.

Factors affecting dielectric constant

In addition to building the ML models, it is valuable to analyze the key features that correlate highly with the measured ϵ behavior in polymers. In the data-points-split approach, 53 features were retained from the initial set of 412 after LASSO-based dimensionality reduction. Figure 4 summarizes some representative features with strong negative or positive correlation with ϵ , with the corresponding coefficients available in Supplementary Fig. 7. As expected, there is a negative correlation between $\log F$ (frequency in log-scale) and ϵ with a coefficient of -0.93 . Additionally, the presence of certain atomic- and block-level features, including CH_2CH_2 , CF_2CF_2 , benzene rings, CH_3 , CF_3 , $(\text{CH}_3)_3$, and $\text{CH}_2\text{CH}_2\text{CH}$, and chain-level features, such as the high number of 3-vertex carbon atoms, number of cyclic double bonds and presence of a purely single bond, lead to lower ϵ . The main reason being that these functional groups introduce zero or negligible net dipole moments but larger free volumes, resulting in small net dipole

density and thus lower ϵ_{elec} . In contrast, the presence of polar groups, such as $\text{CH}_2\text{CF}_2\text{CH}_2$, C-F, C-Cl, -OH, ketone, thioketones, NH, amide, pyridine, pyrrole, $\text{CH}_2\text{CH}_2\text{O}$, and various fragments including NH/amide could strongly enhance the electronic polarity (ϵ_{elec}) of polymers. Consequently, these positive (negative) correlated features can increase (decrease) the total ϵ across the entire frequency regime by controlling ϵ_{elec} . Furthermore, the structural arrangement of these functional groups strongly affects the polymer ϵ value, e.g., PVDF ($\text{CF}_2\text{CH}_2\text{CF}_2\text{CH}_2$) has an ϵ of 9.45 at 100 Hz while ETFE ($\text{CH}_2\text{CH}_2\text{CF}_2\text{CF}_2$) has an ϵ of just 2.6. Thus, it was essential to cover such special sequence-controlled block-level features in our fingerprinting scheme (e.g., $\text{CH}_2\text{CH}_2\text{CF}_2$ and $\text{CH}_2\text{CF}_2\text{CH}_2$) to distinguish polymers. Also, the chain-level features including the topological polar surfaces area of polar elements (e.g., O, N, S, F, and Cl) and the number of H-bond acceptors have a positive relationship with ϵ . These features can increase the ionic (ϵ_{ionic}) and dipolar ($\epsilon_{\text{dipolar}}$) parts by strengthening the H-bonding and dipole interactions between polymer chains, thus increasing the overall ϵ at THz and lower frequency regime. All these findings can be helpful guidelines for rational design of polymers with desired frequency-dependent ϵ values.

Correlation with ϵ		Representative features	
log F			
Negative	Atomic-block level	 (CH ₂ CH ₂) (CF ₂ CF ₂) (Benzene) (CH ₃) (CF ₃) (CH ₃) ₃ (CH ₂ CH ₂ CH)	
	Chain-level	Number of 3-vertex carbon atoms (N _{3v_C}), Number of cyclic double bonds in a repeat unit, Presence of a purely single bond	
Positive	Atomic-block level	 (CH ₂ CF ₂ CH ₂) (C-F) (C-Cl) (C-OH) (ketone) (thioketones) (NH) (amide) (pyridine) (pyrrole) (CH ₂ CH ₂ O)	
	Chain-level	Topological polar surface area (TPSA), Number of H-bond acceptor (N _{H-bond})	

Fig. 4 Representative features affecting dielectric constant. Representative features having strong negative or positive correlations with ϵ . R represents an arbitrary chemical group of C, O, H, N elements, and log F denotes the log-scale frequency value used as a feature in the ML model.

Application-specific polymers design with desired dielectric constant

Next, we move on to apply the developed ML model to discover novel polymers with desired ϵ for capacitors and microelectronic devices. As illustrated in Fig. 5a, the frequency-dependent ϵ of the 11,000 unseen candidate polymers in Fig. 2b were predicted using the GPR- X_{LASSO} model trained on the full dataset (1210 points), the data-points-split approach and fivefold CV. We note that ϵ predictions can be made across a wide range of frequencies (e.g., 60, 10², 10³, 10⁴, 10⁵, 10⁶, 10⁷, 10⁸, 10⁹, 10¹², and 10¹⁵ Hz), although no training data is available at THz frequency. The inverse relation of predicted ϵ with frequency for these new polymers can be observed in Fig. 5a and further validations are shown in Supplementary Fig. 8.

To optimize polymer candidates for capacitor and microelectronic applications, in addition to ϵ , another critical design property, T_g , is considered. Polymers with high T_g are expected to be thermally stable, which is essential for these two applications^{9,43,44}. Thus, in Fig. 5a, we also provide ML predicted T_g using our previously developed models³². Based on the past considerations appropriate for high-temperature energy density capacitors^{2,3,43,44}, $T_g \geq 450$ K was used as the first criterion to discover polymers for high-temperature applications. As mentioned earlier, polymers with high ϵ are required for capacitors, thus, 85 polymers with $\epsilon \geq 5$ (at 100 Hz) were selected from Fig. 5a expected to display high-energy density. As insulating films in microelectronic devices need polymers with low ϵ to decrease the signal-delay time, 191 polymers with ϵ in a range of 2.0–2.5 (at 100 Hz) were identified. For each application, the frequency-dependent ϵ of five representative polymers is shown in Fig. 5b. The corresponding monomer unit, and the ML-based ϵ (at 100 Hz) and T_g (in K) predictions are summarized in Fig. 6. Here, ID 1–5 represent cases with high ϵ for capacitors and ID 6–10 are polymers with low ϵ for microelectronic devices.

As shown in Fig. 5b, the frequency-dependent ϵ trend of ten polymers is correctly captured. Moreover, the monomer chemistry for the selected 5 polymer with high ϵ (ID 1–5) includes either amide, OH or C–Cl groups, agreeing with the positive correlation trend discussed above (and shown in Fig. 4). Similarly, the presence of CF₃ group and benzene rings greatly decrease the

polymer ϵ , as mentioned earlier and can be seen from the selected list of low ϵ polymer with amides groups (ID 6–8) and OH groups (ID 9) in Fig. 6. We also note that all of the selected 10 polymers contain rigid benzene rings, resulting in high T_g . Based on the prediction accuracy reached by our models on the unseen test set, the ability of the model to correctly capture inverse ϵ vs. frequency behavior, and the chemical arguments made above, we believe that these proposed ten polymers are good candidates for further experimental validations.

DISCUSSION

Using an experimental ϵ dataset of 738 polymers (or 1210 data-points) at various frequencies, unique 3-level hierarchical polymer features and the GPR algorithm, we built a single ML model to accurately predict the frequency-dependent ϵ behavior of polymers. There are several advantages of the ML models presented here: first, it can predict ϵ of polymers across a wide range of frequencies (60–10¹⁵ Hz, excluding the resonant frequency regions). The single ML model developed here more accurately capture the inverse relationship between ϵ and frequency, compared with separate ML models for ϵ at different frequency regimes, as discussed in Supplementary Section 4. As the frequency in log-scale was used as a feature in the single ML model, 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 the frequency in log-scale as a feature is that it allows us to make ϵ predictions at any arbitrary frequency value, which is not possible with separate ML models. This complete frequency-dependent picture provides comprehensive information to assist rational design of new polymers. The present ϵ -prediction model is already implemented in our Polymer Genome platform (<http://www.polymergenome.org>).

Second, the predicted GPR uncertainty acts as a useful guide to know when the ML predictions can be trusted. The present ML model is more suitable for homo-polymers containing C, H, B, O, N, S, P, Si, F, Cl, and Br atoms. Also, higher uncertainties can be expected within the frequency range of 10¹⁰–10¹⁴ Hz owing to the unavailability of training data in this regime. These uncertainties

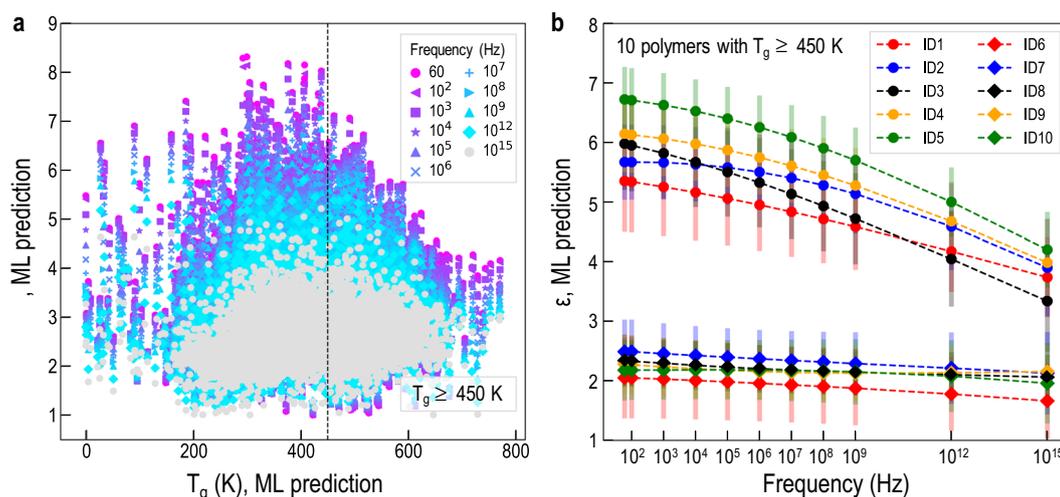


Fig. 5 Machine-learning-predicted dielectric constant of 11,000 unseen polymers. **a** ML predicted ϵ at various frequencies (i.e., 60, 10^2 , 10^3 , 10^4 , 10^5 , 10^6 , 10^7 , 10^8 , 10^9 , 10^{12} , and 10^{15} Hz) for 11,000 unseen polymers from Fig. 2b, along with their ML predicted T_g values. **b** Ten representative polymers with high T_g (≥ 450 K) selected from **a**, such that five polymers (ID 1–5) have high ϵ (≥ 5), and remaining five (ID 6–10) have low ϵ (2–2.5).

ID	Monomer	T_g (K)	ϵ at 100 Hz	ID	Monomer	T_g (K)	ϵ at 100 Hz
1		451±40	5.34±0.81	6		528±31	2.05±0.64
2		581±29	5.67±0.60	7		501±27	2.49±0.50
3		531±39	5.95±0.73	8		495±28	2.33±0.39
4		474±27	6.13±0.60	9		493±30	2.27±0.46
5		516±31	6.71±0.50	10		576±27	2.18±0.46

Fig. 6 Details of ten representative polymers. The monomer unit, and the ML predicted T_g and ϵ (at 100 Hz) of ten representative polymers shown in Fig. 5b. Polymers with ID 1–5 have high ϵ (≥ 5), while ID 6–10 are polymers with low ϵ (2–2.5). The associated ML prediction uncertainty is also provided.

can provide useful guidance for next experiments via active learning, with the newly generated data aiding model improvement⁴⁵.

Third, key features that strongly affect the polymer ϵ behavior were analyzed, forming a crude first stage criteria to find polymers with the desired ϵ . To attain high ϵ , common polar groups, including C–F, –OH, C=O and amides, and rigid groups such as

pyridine and pyrrole can be introduced into polymers. On the other hand, the introduction of non-polar groups (e.g., benzene rings and CH₃) or functional groups with low polarization density (e.g., CF₃) leads to low ϵ . However, we note that presence of some flexible polar groups may induce an unwanted high dielectric loss, which can be further eliminated by introducing additional screening criteria on other polymer properties, e.g., low dielectric loss and high breakdown strength.

Finally, ϵ and T_g of about 11,000 polymers have been predicted using the ML models developed in this and our previous work³², respectively, providing a huge pool of polymers for various applications. Using the T_g and ϵ as the screening criteria, 5 high and 5 low ϵ polymers are proposed for capacitors and microelectronic devices, respectively. While this work initiates a great opportunity to select polymers satisfying two properties, it can be easily extended to three or more properties.

Although we believe that the developed ML model is fairly accurate and universal, more efforts are envisioned in the future. First, Fig. 3 shows that a test RMSE of 0.67 and 0.35 is achieved for the polymer-types-split and data-points-split-based ML models using 90% training set and 10% test set, respectively. Therefore, it is expected that the average RMSE of predicted values for new cases ranges from 0.35 to 0.67. For polymers in applications requiring a high ϵ of 5–11, even the RMSE of 0.67 leads to an acceptable relative error of 6–13.4%. For applications require polymers with ϵ ranging from 2 to 3.5, the RMSE of 0.35 results in a relative error 10–17%, which is slightly high but acceptable. The relative error of some completely unseen polymers may reach to 19–33% with respect to the RMSE of 0.67. However, their predicted GPR uncertainties should also be high. Therefore, more data should be collected from literature either manually or using natural language processing techniques⁴⁶ to improve the model performance and dataset diversity. Second, almost no empirical data is available in the THz region. First-principles MD simulations with the reactive force fields have been recently shown to accurately estimate ϵ values at THz frequencies using amorphous phases of polymers⁴⁷. Such method can successfully overcome the problem of ϵ overestimation introduced because of the unrealistically higher densities of crystalline polymer models used in the DFPT method. There is a great opportunity to incorporate theoretical data to fill the empty THz region of our dataset. Third, new polymer features can be included at the morphological-level, e.g., molecular weights, cross-link and torsion angles, to represent more complicated polymer chemical space. Also, more advanced feature reduction methods can be developed to replace the present linear LASSO method.

METHODS

Dataset

The experimental ϵ of 738 polymers, measured at room temperature, under dry conditions and at 9 frequency values, i.e., 60, 10², 10³, 10⁴, 10⁵, 10⁶, 10⁷, 10⁹, and 10¹⁵ Hz, were considered in this work. These values were taken from refs.^{9,19,21,33,35–42}. The ϵ measurements within the frequency range of 60–10⁹ Hz is commonly made using the impedance analyzer, the precision inductance, and capacitance and resistance (LCR) meter^{18,42}. ϵ values at 10¹⁵ Hz were obtained by taking the square of the experimental refractive index measured using refractometers. Since experimental conditions significantly impact the measured ϵ , we collected the data only when the measurements were made at room temperature (295 ± 5 K) and under dry conditions (with relative humidity <1%). We note that it is almost impossible to find consistent sample qualities across the literature, with the common variations observed in sample thickness and different order of polymer crystallinity. While such uncertainties are unavoidable in experimental datasets, we believe they are acceptable to train reliable ML models. For cases where multiple data-points were available we used the average ϵ value.

Our developed ML model was used to make prediction for a completely unseen dataset of roughly 11,000 homo-polymers that have previously

been synthesized and reported (but for which no dielectric characterization has been done). This dataset is substantially diverse, containing numerous polymers classes, e.g., polyolefins, polyimides, polycarotenedamides, polyvinyls, polyethers, polyesters, polydienes, polyoxides, and polycarbonates, but not more complex polymers such as copolymers, polymer blends, as well as ladder, cross-linked, and metal-containing polymers. Because of the evidence of past synthetic work, polymer candidates identified for specific applications from this candidate list using our model are expected to have good potential to be synthesized (again) and tested. This large dataset, which contains polymer identities, names/labels, and/or monomer representations, was collected from various available sources, including published articles, handbooks, and online repositories^{7,21,32–34}.

Feature engineering

To build accurate and reliable ML models, it is important to include relevant features that numerically represent materials and collectively capture the trends in ϵ values across wide frequency range and across varying polymer chemistry. Our polymer fingerprinting scheme is based on a pre-defined list of possible components covering various length scales, including (1) atomic-level fragments, (2) block-level fragments, and (3) chain-level, i.e., extended features that capture higher level morphological information in polymers. The atomic-level fragments are specified by the generic label "A_iB_jC_k", representing an *i*-fold coordinated A atom, a *j*-fold coordinated B atom, and a *k*-fold coordinated C atom, connected in the specified order. For example, N3-C3-C4 represents a threefold coordinated N, a threefold coordinated carbon and a fourfold coordinated carbon. The block-level fingerprint components track the presence of 363 pre-defined building blocks that frequently occur in conventional polymers with some representative examples being C₆H₆, C=O, CH₂, and CF₂. More importantly, a series of triplet-blocks were defined to represent the specific structural arrangements of functional groups, e.g., CH₂CH₂CF₂ and CH₂CF₂CH₂. The occurrence of each block in the polymer repeat unit (monomer) normalized by the number of atoms (of the monomer) is used as a block-level fingerprint component. The chain-level features capture information at the highest length scale, including quantitative structure-property relationship (QSPR) and morphological features. The QSPR features, e.g., van der Waals surface area, topological polar surface area, and the fraction of rotatable bonds, were generated using the RDKit library. The morphological features, e.g., the length of the longest/shortest side chains with/without rings and the shortest topological distance between rings, were developed by us. Using this fingerprinting scheme, 155 atomic-level, 197 block-level and 59 chain-level features were generated for each of the 738 polymers, leading to a total of 411 chemical features for each polymer. Additionally, the frequency in log-scale (log *F*) was incorporated as a feature in the ML model development process, resulting in a total of 412 features. As per standard ML practices, all features were scaled from 0 to 1 during the model training.

The least absolute shrinkage and selection operator (LASSO) method was used to retain the relevant features by optimizing the regularization term to achieve the highest *R*². Subsequently, the remaining features with non-zero coefficients were used to construct the ML models. For the LASSO dimensionality reduction scheme, all 412-dimensional features and the entire ϵ dataset was used. Furthermore, the group-shuffle-split and *K*-fold libraries implemented in sklearn python package were respectively used for the polymer-types-split and the data-points-split approach. The resulting number of feature (*N_x*) is summarized in Table 1, including the frequency feature internally selected by the LASSO method.

To visualize the chemical diversity of the training (738 polymers) and the unseen (11,000 polymers) datasets adopted here, PCA was performed on the complete chemical features of these two datasets (706 features in total), excluding the frequency feature. The first two (PC1 and PC2) components are shown in Fig. 2b and used to analysis the similarity of two datasets with the agglomerative hierarchical clustering method. As illustrated in Supplementary Fig. 1, there are 90% shared chemical space of two datasets, revealing that the training dataset fairly covers the chemical space of the unseen dataset.

Gaussian process regression

We used the Gaussian process regression (GPR) with the radial basis function (RBF) kernel to train the ML models. In this case, the co-variance

function between two materials with features x and x' is given by

$$k(x, x') = \sigma_f \exp\left(-\frac{1}{2\sigma_l^2} \|x - x'\|^2\right) + \sigma_n^2. \quad (1)$$

Here, three hyperparameters σ_f , σ_l , and σ_n represent the variance, the length-scale parameter and the expected noise in the data, respectively. These were determined during the model training by maximizing the log-likelihood estimate. Further, as shown in Table 1, K -fold and group-shuffle-split methods with fivefold cross-validation were adopted in the polymer-types-split and the data-points-split models to avoid overfitting, respectively. The root mean square error (RMSE) and the coefficient of determination (R^2) were used to evaluate the performance of the ML models. Further, learning curves (Fig. 3) were generated by varying the size of the training and the test sets to estimate the prediction errors on unseen data. Model performance (RMSE) was evaluated by averaging over 50 statistical runs with random training and test splits.

DATA AVAILABILITY

The dielectric constant dataset will be made available upon reasonable request for academic use.

CODE AVAILABILITY

The codes that support the findings of this study are not publicly available as they are the Intellectual Property of Georgia Tech Research Corporation. However, they may be created using the descriptions provided in ref. ³² <Polymer Genome: A Data-Powered Polymer Informatics Platform for Property Predictions>, and the freely available RDKit and scikit-learn python modules.

Received: 28 January 2020; Accepted: 28 April 2020;

Published online: 21 May 2020

REFERENCES

- Chu, B. et al. A dielectric polymer with high electric energy density and fast discharge speed. *Science* **313**, 334–336 (2006).
- Li, Q. et al. Flexible high-temperature dielectric materials from polymer nanocomposites. *Nature* **523**, 576 (2015).
- Tan, Q., Irwin, P. & Cao, Y. Advanced dielectrics for capacitors. *IEEE Trans. FM* **126**, 1153–1159 (2006).
- Sharma, V. et al. Rational design of all organic polymer dielectrics. *Nat. Commun.* **5**, 4845 (2014).
- Huan, T. D. et al. Advanced polymeric dielectrics for high energy density applications. *Prog. Mater. Sci.* **83**, 236–269 (2016).
- Mannodi-Kanakkithodi, A. et al. Rational co-design of polymer dielectrics for energy storage. *Adv. Mater.* **28**, 6277–6291 (2016).
- Huan, T. D. et al. A polymer dataset for accelerated property prediction and design. *Sci. Data* **3**, 160012 (2016).
- Mannodi-Kanakkithodi, A. et al. Scoping the polymer genome: a roadmap for rational polymer dielectrics design and beyond. *Mater. Today* **21**, 785–796 (2018).
- Ho, J. S. & Greenbaum, S. G. Polymer capacitor dielectrics for high temperature applications. *ACS Appl. Mater. Interfaces* **10**, 29189–29218 (2018).
- Dissado, L. A. & Fothergill, J. C. *Electrical Degradation and Breakdown in Polymers*, Vol. 9 (IET, 1992).
- Maier, G. Low dielectric constant polymers for microelectronics. *Prog. Polym. Sci.* **26**, 3–65 (2001).
- Dang, M. T., Hirsch, L. & Wantz, G. P3ht: Pcbm, best seller in polymer photovoltaic research. *Adv. Mater.* **23**, 3597–3602 (2011).
- Facchetti, A. π -conjugated polymers for organic electronics and photovoltaic cell applications. *Chem. Mater.* **23**, 733–758 (2011).
- Huang, X. & Jiang, P. Core-shell structured high-k polymer nanocomposites for energy storage and dielectric applications. *Adv. Mater.* **27**, 546–554 (2015).
- Wang, Y. et al. Ultrahigh energy density and greatly enhanced discharged efficiency of sandwich-structured polymer nanocomposites with optimized spatial organization. *Nano Energy* **44**, 364–370 (2018).
- Smith, O. L. et al. Enhanced permittivity and energy density in neat poly(vinylidene fluoride-trifluoroethylene-chlorotrifluoroethylene) terpolymer films through control of morphology. *ACS Appl. Mater. Interfaces* **6**, 9584–9589 (2014).
- Nasreen, S. et al. Sn-polyester/polyimide hybrid flexible free-standing film as a tunable dielectric material. *Macromol. Rapid Commun.* **40**, 1800679 (2019).
- Wu, C. et al. Dipole-relaxation dynamics in a modified polythiourea with high dielectric constant for energy storage applications. *Appl. Phys. Lett.* **115**, 163901 (2019).
- Ma, R. et al. Rationally designed polyimides for high-energy density capacitor applications. *ACS Appl. Mater. Interfaces* **6**, 10445–10451 (2014).
- Ku, C. C. & Liepins, R. *Electrical Properties of Polymers* (Hanser Publishers, New York, 1987).
- Bicerano, J. *Prediction of polymer properties* (CRC Press, 2002).
- Wang, C. et al. Computational strategies for polymer dielectrics design. *Polymer* **55**, 979–988 (2014).
- Misra, M., Mannodi-Kanakkithodi, A., Chung, T., Ramprasad, R. & Kumar, S. K. Critical role of morphology on the dielectric constant of semicrystalline polyolefins. *J. Chem. Phys.* **144**, 234905 (2016).
- Jordan, M. I. & Mitchell, T. M. Machine learning: trends, perspectives, and prospects. *Science* **349**, 255–260 (2015).
- Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: recent applications and prospects. *NPJ Comput. Mater.* **3**, 54 (2017).
- Gaultois, M. W. et al. Data-driven review of thermoelectric materials: Performance and resource considerations. *Chem. Mater.* **25**, 2911–2920 (2013).
- Chen, L., Tran, H., Batra, R., Kim, C. & Ramprasad, R. Machine learning models for the lattice thermal conductivity prediction of inorganic materials. *Comp. Mat. Sci.* **170**, 109155 (2019).
- 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* **11**, 24906–24918 (2019).
- Chandrasekaran, A. et al. Solving the electronic structure problem with machine learning. *NPJ Comput. Mater.* **5**, 22 (2019).
- Wu, K. et al. Prediction of polymer properties using infinite chain descriptors (icd) and machine learning: toward optimized dielectric polymeric materials. *J. Polym. Sci. Pol. Phys.* **54**, 2082–2091 (2016).
- Mannodi-Kanakkithodi, A., Pilania, G., Huan, T. D., Lookman, T. & Ramprasad, R. Machine learning strategy for accelerated design of polymer dielectrics. *Sci. Rep.* **6**, 20952 (2016).
- 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* **122**, 17575–17585 (2018).
- Mark, J. *Polymer Data Handbook* (Oxford University Press, 1999).
- Otsuka, S., Kuwajima, I., Hosoya, J., Xu, Y. & Yamazaki, M. In *2011 International Conference on Emerging Intelligent Data and Web Technologies*, 22–29 (IEEE, 2011).
- Baldwin, A. F. et al. Rational design of organotin polyesters. *Macromolecules* **48**, 2422–2428 (2015).
- Baldwin, A. F. et al. Poly (dimethyltin glutarate) as a prospective material for high dielectric applications. *Adv. Mater.* **27**, 346–351 (2015).
- Ma, R. et al. Rational design and synthesis of polythioureas as capacitor dielectrics. *J. Mater. Chem. A* **3**, 14845–14852 (2015).
- Lorenzini, R., Kline, W., Wang, C., Ramprasad, R. & Sotzing, G. The rational design of polyurea & polyurethane dielectric materials. *Polymer* **54**, 3529–3533 (2013).
- Chisca, S., Sava, I., Musteata, V.-E. & Bruma, M. Dielectric and conduction properties of polyimide films. In *2011 International Semiconductor Conference (CAS)*, Vol. 2, 253–256 (IEEE, 2011).
- Mandelcorn, L. & Miller, R. L. High temperature, >200 degrees C, polymer film capacitors. In *IEEE 35th International Power Sources Symposium*, 369–372 (IEEE, 1992).
- Pan, J., Li, K., Chuayprakong, S., Hsu, T. & Wang, Q. High-temperature poly (phthalazinone ether ketone) thin films for dielectric energy storage. *ACS Appl. Mater. Interfaces* **2**, 1286–1289 (2010).
- Li, Z. et al. High energy density and high efficiency all-organic polymers with enhanced dipolar polarization. *J. Mater. Chem. A* **7**, 15026–15030 (2019).
- Tan, D., Zhang, L., Chen, Q. & Irwin, P. High-temperature capacitor polymer films. *J. Electron. Mater.* **43**, 4569–4575 (2014).
- Wu, C. et al. Flexible temperature-invariant polymer dielectrics with large band-gap. *Adv. Mater.* e2000499 (2020).
- Kim, C., Chandrasekaran, A., Jha, A. & Ramprasad, R. Active-learning and materials design: the example of high glass transition temperature polymers. *MRS. Commun.* **9**, 860–866 (2019).
- Tshitoyan, V. et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
- Fukushima, S. et al. Effects of chemical defects on anisotropic dielectric response of polyethylene. *AIP Adv.* **9**, 045022 (2019).

ACKNOWLEDGEMENTS

This work is supported by the Office of Naval Research through N0014-17-1-2656, a Multi-University Research Initiative (MURI) grant. We thank Dr. Rui Ma, Dr. Gregory M.

Treich and Dr. Shamima Nasreen for collecting, organizing, and providing experimental data from their papers.

AUTHOR CONTRIBUTIONS

L.C. and R.R. initiated this research project; L.C. developed and analyzed the ML models; C.K. contributed to the development of polymer fingerprinting codes; R.B. and R. R. contributed to the model analysis and discussions; J. L., C.W, Z.L., A.D., Y.W., H.T. contributed to the data collection; all co-authors contributed to the development of the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at <https://doi.org/10.1038/s41524-020-0333-6>.

Correspondence and requests for materials should be addressed to R.R.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020