The Journal of Chemical Physics

Novel high voltage polymer insulators using computational and data-driven techniques

Cite as: J. Chem. Phys. **154**, 174906 (2021); https://doi.org/10.1063/5.0044306
Submitted: 15 January 2021 . Accepted: 19 April 2021 . Published Online: 04 May 2021

Deepak Kamal, Huan Tran, Chiho Kim, Yifei Wang, Lihua Chen, Yang Cao, V. Roshan Joseph, and Rampi Ramprasad

COLLECTIONS

Paper published as part of the special topic on Computational Materials Discovery

AIP Publishing

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

Novel high voltage polymer insulators using computational and data-driven techniques

Cite as: J. Chem. Phys. 154, 174906 (2021); doi: 10.1063/5.0044306
Submitted: 15 January 2021 • Accepted: 19 April 2021 •
Published Online: 4 May 2021

Deepak Kamal,[1] Huan Tran,[1] Chiho Kim,[1] Yifei Wang,[2] Lihua Chen,[1] Yang Cao,[2] V. Roshan Joseph,[3] and Rampi Ramprasad[1,a)]

AFFILIATIONS

[1]School of Materials Science and Engineering, Georgia Institute of Technology, 771 Ferst Drive NW, Atlanta, Georgia 30332, USA
[2]Department of Materials Science and Engineering, University of Connecticut, Storrs, Connecticut 06269, USA
[3]H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, 755 Ferst Dr. NW, Atlanta, Georgia 30332, USA

Note: This paper is part of the JCP Special Topic on Computational Materials Discovery.
[a)]**Author to whom correspondence should be addressed:** rampi.ramprasad@mse.gatech.edu

ABSTRACT

One of the key bottlenecks in the development of high voltage electrical systems is the identification of suitable insulating materials capable of supporting high voltages. Under high voltage scenarios, conventional polymer based insulators, which are one of the popular choices of insulators, suffer from the drawback of space charge accumulation, which leads to degradation in desirable electronic properties and facilitates dielectric breakdown. In this work, we aid the development of novel polymers for high voltage insulation applications by enabling the rapid prediction of properties that are correlated with dielectric breakdown, i.e.,the bandgap (E_{gap}) of the polymer and electron injection barrier (Φ_e) at the electrode–insulator interface. To accomplish this, density functional theory based methods are used to develop large, chemically diverse datasets of Φ_e and E_{gap}. The deviation of the computed properties from experimental observations is addressed using a statistical technique called Bayesian calibration. Furthermore, to enable rapid estimation of these properties for a large set of polymers, machine learning models are developed using the created dataset. These models are further used to predict E_{gap} and Φ_e for a set of 13k previously known polymers. Polymers with high values of these properties are selected as potential high voltage insulators and are recommended for synthesis. Finally, the models developed here are deployed at www.polymergenome.org to enable the community use.

I. INTRODUCTION

Fueled by rising environmental concerns and economic factors, there is an increased drive to improve the efficiency of electrical energy transfer and utilization. This has brought about technologies such as the High Voltage Direct Current (HDVC) system[1] and more electrical aircraft (MEA) systems,[2] which transmit electricity and operate machinery at high voltages. One of the key bottlenecks in the development of such high power electrical systems is the identification of suitable insulating materials.[3,4] Polymers have long been used as insulators in electrical applications owing to their low-cost, flexibility, attractive insulation properties, attractive chemical and thermal stability, and ease of processability.[5,6] However, under high voltage scenarios, conventional choices of polymers such as polyethylene (PE) and polypropylene (PP) and rubber-like polymers suffer from the drawback of forming internal space charge, which culminates in dielectric breakdown of the insulation.[7,8] Therefore, to facilitate the development of insulation polymers in order to design more resilient polymeric insulating materials for high power applications, it is important to find new polymer dielectrics that are resistant to space charge accumulation and dielectric breakdown.

The origin of space charge accumulation in insulating polymers can be attributed to the presence of excess mobile charges in them.[9] Primarily, these excess charges include electrons/holes that are injected from the metal electrode in contact with the insulator (Schottky injection)[9] or the charges that are excited from the valance band of the polymer to its conduction band.[10] These excess charges increase the local field inside the insulator, resulting in the creation and accumulation of local electronic defects,[11] which over a period of time leads to large electrical stresses and finally dielectric breakdown.[7] In a recent study,[12] we demonstrated that dielectric breakdown in polymers (E_{bd}) is indeed correlated with the electron injection barriers in a conventional capacitor assembly and established ϕ_e and E_{gap} as proxies of breakdown. The computational accessibility of these proxies then opens up a pathway for finding new high voltage insulators. However, computational methods such as Density Functional Theory (DFT), which can be used to reliably estimate these properties, are too time intensive to explore the vast chemical space that makes up the polymer universe. One way of tackling this problem is to use data-driven approaches to build a surrogate model that can, to some accuracy, emulate a physics-based theoretical simulator. Recently, within the materials science and chemistry community, several such Machine Learning (ML) based surrogate models have been contributed toward accelerating the discovery of new materials for several applications.[13–15] A few such works have also been published in the domain of polymer properties.[16,17] If such ML based surrogate models could be developed for E_{gap} and Φ_e, which can reliably predict these properties for a large chemical space, we could then use these models to screen for potential high breakdown polymers. The key difficulty in developing such models is the creation of a chemically diverse dataset of E_{gap} and Φ_e, which well represents the polymer chemical space.

In this paper, we attempt to facilitate the discovery of novel high voltage insulators by developing reliable ML models for E_{gap} and Φ_e. To achieve this, we start by creating a chemically diverse dataset of E_{gap} and Φ_e using DFT based high-throughput computations. While examining the parity between experimental values and DFT-calculated values of these properties, we found that while experimental properties are correlated with calculations for both E_{gap} and Φ_e, there is a lack of numerical parity in the case of Φ_e. Taking advantage of the correlation between the computed and observed values of Φ_e, the observed difference in numerical parity was addressed using a statistical tool called Bayesian calibration (BC). Furthermore, after representing the polymers in terms of machine-readable Polymer Genome fingerprints,[16,18,19] we use Gaussian process regression (GPR) to learn the relationship between the numerical representation of the polymer and the properties. These models are further used to screen for potential high breakdown polymers from a list of 13 000 known polymers. The models developed here are deployed in `Polymer Genome` to enable the community to use it.

II. TECHNICAL DETAILS

A. Data creation

The overall machine learning model development workflow is depicted in Fig. 1. The first step in the process is to develop reliable datasets for ϕ_e and E_{gap} using DFT computations. This section provides the technical details of the data creation.

1. Polymer models

Modeling polymers is notably challenging. In fact, polymer morphology is highly complex, typically containing crystalline, semi-crystalline, and amorphous domains coexisting with the degree of crystallinity falling somewhere between 0 and ≃80%.[20] Since E_{gap} and ϕ_e can only be evaluated using electronic structure-based methods such as DFT, polymers must be modeled as perfect (infinite) crystals or chains, whose periodical unit cell contains no more than a few hundred atoms.[21,22] This approximation turns out to be pretty good, reasonably capturing the true values of E_{gap}, ϕ_e, and their respective trends.[23–26] Therefore, the datasets of E_{gap} and ϕ_e required for this work were developed using DFT computations on polymer chain and crystal models.

For a vast majority of known polymers, the only available structure-related information is their atomic connectivity—in fact, there are just a few dozens of polymers whose crystal structures have been resolved experimentally.[24] Therefore, the computationally predicted polymer structure is usually required for polymer

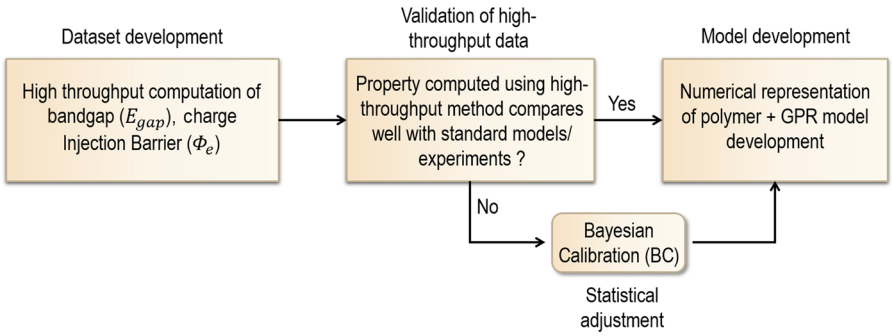

FIG. 1. A high level overview of the model development workflow used in this work. DFT based high-throughput computations are used to obtain E_{gap} and Φ_e. BC is used to correct the high-throughput data if the computed property value deviated from experimental observations, considering a handful of known cases where we have experimental observations. The electron injection barrier results from DFT were subjected to BC here. The corrected property is then used to develop the ML model.

modeling. Herein, we used the polymer structure predictor (PSP),[27] an efficient method in order to create the required polymer chains and crystals. Starting from their atomic connectivity, the PSP has been used to successfully predict the chain and crystal structures of many linear polymers.[27] By all accounts, predicting a polymer crystal structure is at least two orders more expensive than predicting a polymer chain structure. Therefore, the polymer bandgap was evaluated at two levels of polymer models, i.e., chains and crystals, leading to two datasets of E_{gap}^{chain} and $E_{gap}^{crystal}$, where the former is substantially bigger than the latter. Moreover, because setting up a crystal model required for computing charge injection barriers is even more cumbersome,[12,23] only polymer chain models were used for our high-throughput computations of ϕ_e.

2. Computational methods

Our DFT computations were performed using the version as implemented in Vienna *Ab initio* Simulation Package (VASP).[28–30] Within this scheme, a plane-wave cutoff of 600 eV and a **k**-point density of 0.1 $Å^{-1}$ were used. The van der Waals dispersion interactions, which are important for describing the polymer bulk and metal–polymer interfacial interactions, were estimated with the non-local density functional vdW-DF2.[31] Refitted Perdew–Wang 86,[32] the exchange–correlation (XC) functional associated with vdW-DF2, was also used. During the geometry optimization, convergence was assumed when the atomic forces become less than 0.01 eV/Å. Finally, to accurately compute the electronic properties such as the conduction band minimum (CBM), the valence band maximum (VBM), and the bandgap, the Heyd–Scuseria–Ernzerhof (HSE06) XC functional[33] was used on top of the optimized structures.

3. Calculations of bandgap and electron injection barrier

Representation of polymers as periodically repeating monomer units provides a simple way to incorporate the long range interaction between atoms in the polymer, which are critical in modeling the electronic properties of polymers. Plane-wave DFT was chosen for the computations here as it is better at handling periodicity. Within the plane-wave DFT framework, bandgap calculations are straightforward, i.e., E_{gap} is the energy difference between the VBM and the CBM, and both of them are obtained while solving the Kohn–Sham equations. By definitions, the estimation of charge injection barriers involves computing the electronic properties of the interface between the metal electrode and the insulating polymer. Here, to keep the study consistent, we considered only one electrode material, i.e., aluminum (Al). Conventionally, to compute ϕ_e using DFT, three properties are required. They are the maximum energy at which electrons reside in the metal or the metal fermi level E_F; the energy of the first vacant energy level in the polymer, which is the CBM; and the interaction between Al and polymer, which introduces an interface dipole moment D and shifts the polymer CBM with respect to Al E_F by $\Delta\Phi = -eD/(2a)$.[23,34,35] Finally, the electron injection barrier is then determined using $\phi_e = E_F - E_{CBM} + \Delta\Phi$. Recently, we established[12] that this approach can be effectively simplified by ignoring $\Delta\Phi$ and considering the polymer single chain model. This simplification allows us to decouple the computation into two parts where we independently determine the CBM of the polymer and the E_F of Al. Hence, the computation of ϕ_e finally boils down to the computation of the CBM of the polymer. A comparison between this scheme and the standard model is reported by Kamal *et al.*[12] This scheme is employed to perform high throughout ϕ_e computations within this work.

4. Statistical adjustment

One of the drawbacks created by the assumptions—ignoring the metal–polymer interfacial interaction and bulk interactions—made to simplify DFT calculations is that it underestimates the computed value of ϕ_e when compared to the experimental observations. This difference in numerical parity can be observed in Fig. 2(e). However, despite the numerical difference between ϕ_e^{exp} and ϕ_e^{DFT}, it can be seen that both values are correlated. Hence, to create a model that can predict ϕ_e^{exp}, we utilize this correlation to bridge the gap between computed results and experimental observations. Here, we do this using a statistical adjustment technique called the Bayesian Calibration (BC) method.[36,37]

Within this method, the experimental observation of ϕ_e, ϕ_e^{exp}, for a given polymer x is thought of to be random due to the presence of noise (uncontrollable) factors and measurement errors, ε. Then,

$$\phi_e^{exp} = \mu(x) + \varepsilon, \tag{1}$$

where $\mu(x)$ is the mean of ϕ_e^{exp} and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. The objective of this method then is to find the unknown function $\mu(x)$, given ϕ_e^{DFT} for a large number of polymers and the experimental value of these properties for a limited number of cases. This is achieved by assuming that we have a DFT model $f^{DFT}(x)$ that gives us access to ϕ_e^{DFT} for a given polymer x and using it to postulate a prior distribution for $\mu(x)$ within a Bayesian framework. Specifically, the value from $f^{DFT}(x)$ is taken as the mean of the prior distribution of $\mu(x)$. The posterior distribution is then computed using the Bayes theorem using the experimental observations for a few polymers. The posterior distribution thus derived incorporates the information from the DFT model as well as the experimental observations. The BC result used here is just the mean of this posterior distribution. In our case, we use a location-scale model[37] to postulate the prior for $\mu(x)$ because of the limited data. Thus, $\mu(x)$ is estimated by

$$\hat{\mu}(x) = f^{DFT}(x) + \hat{\beta}_0 + \hat{\beta}_1(f^{DFT}(x) - \bar{f}), \tag{2}$$

where $\hat{\beta}_0 = \max(1 - 1/z_0^2, 0)\tilde{\beta}_0$, $\hat{\beta}_1 = \max(1 - 1/z_1^2, 0)\tilde{\beta}_1$, $\bar{f} = \sum_{i=1}^{n} f^{DFT}(x_i)/n$, $z_0 = \sqrt{n}\tilde{\beta}_0/\sigma$, $z_1 = \sqrt{S}\tilde{\beta}_1/\sigma$, $S = \sum_{i=1}^{n}\{f^{DFT}(x_i) - \bar{f}\}^2$, $\tilde{\beta}_0 = \sum_{i=1}^{n}\phi_{e,i}^{exp}/n - \bar{f}$, $\tilde{\beta}_1 = \sum_{i=1}^{n}\{\phi_{e,i}^{exp} - f^{DFT}(x_i)\}f^{DFT}(x_i)/S$, and n is the number of observations.

B. Fingerprinting and machine learning

The process of developing a ML based model contains several steps. First, we represent the polymers in a numerical machine-readable manner, referred to as fingerprint. In this work, inspired by the success of our past property prediction models,[16,38] we use a four-level hierarchical fingerprint first introduced by Tran *et al.*[19] This leads to a fingerprint of about 800 components for each property. Then, recursive feature elimination[39] (RFE) is used to reduce

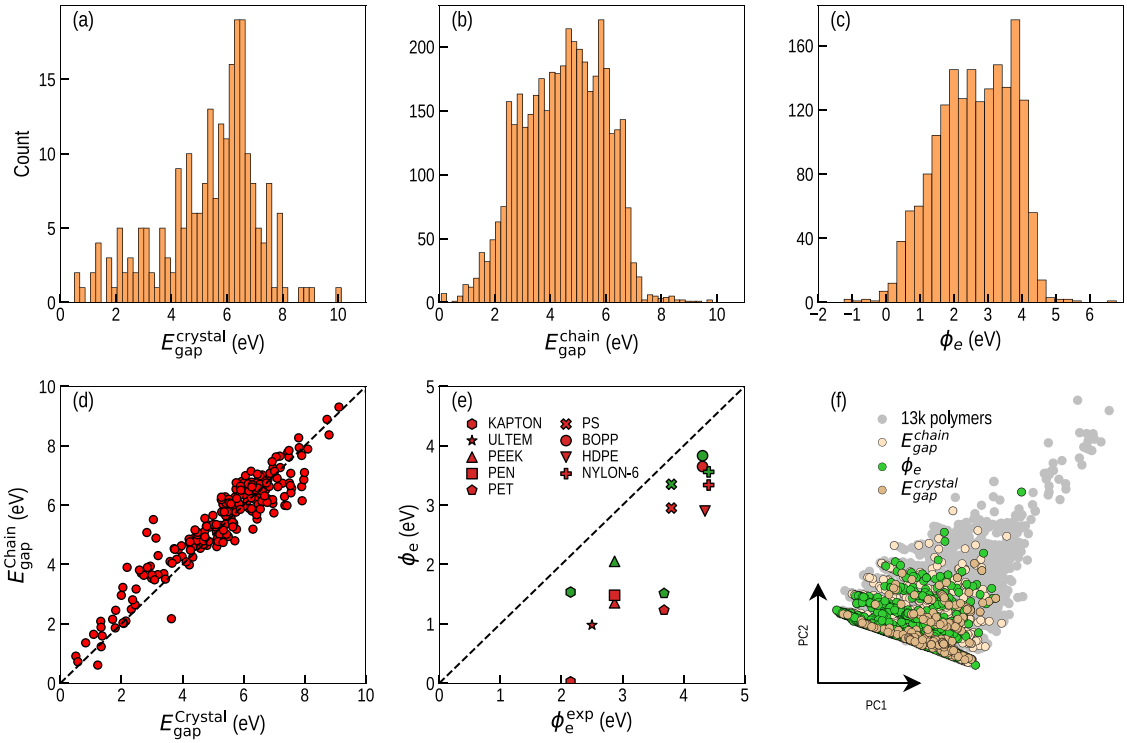

FIG. 2. (Top row) The distribution of computed values of $E_{gap}^{crystal}$ (a), E_{gap}^{chain} (b), and ϕ_e (c). (Bottom row) A comparison between $E_{gap}^{crystal}$ and E_{gap}^{chain} for a selected set of polymers (d), a comparison between ϕ_e computed using a single chain polymer model (red) and crystalline polymer model (green) with experimental observations for a selected set of polymers (e), and the distribution of polymers used for computing $E_{gap}^{crystal}$, E_{gap}^{chain}, and ϕ_e in the chemical and conformational space defined by Polymer Genome fingerprints, overlaid against that of 13k known polymers (f).

the size (the number of components) of the fingerprint. The final fingerprint size and details are provided in Sec. III A. Here, Gaussian process regression (GPR) was chosen over other ML methods such as graph neural networks (NN) and convolutional neural networks,[40–42] which indeed show great promise, mainly because the dataset that we are using is comparatively small and because GPR provides a more robust estimate of prediction uncertainty compared to that provided by NN based methods. Within GPR, we employ a squared exponential kernel of the following form to learn the relationship between the fingerprint of the polymer and the properties:

$$k(x_i, x_j) = \sigma_f \exp\left(-\frac{1}{2\sigma_l^2}\|x_i - x_j\|^2\right) + \delta_{ij}\sigma_n^2, \quad (3)$$

where σ_f, σ_l, and σ_n are the hyper-parameters controlling the characteristics of the prior. During this (model training) step, the posterior distribution is obtained by maximizing the log marginal likelihood of the observed data. Having the trained model, predictions were made by maximizing the conditional likelihood, given the fingerprint of the polymer in consideration. We used the scikit-learn package[43] for training the models and predicting the polymer properties.

III. RESULTS AND DISCUSSIONS

A. Dataset

To enable the development of ML models, which can reliably predict E_{gap} and Φ_e for new polymers, we used DFT calculations to create large datasets containing 4100 data points of the polymer chain bandgap E_{gap}^{chain} and 1800 data points of Φ_e. A dataset of $E_{gap}^{crystal}$ containing 200 polymers was also created to assess the reliability of the single chain based E_{gap}^{chain} calculations.

To enhance the generalizability of the models developed in this work, the diversity in terms of the property range and chemistry of the polymers was ensured while creating the dataset. Figures 2(a)–2(c) and 2(f) depict the diversity of the dataset. Figures 2(a)–2(c) show the number of polymers in each dataset and their distribution over property space. The dataset of $E_{gap}^{cyrstal}$ and E_{gap}^{chain} contains values ranging from 0 to 10 eV, covering the range of bandgap expected for polymeric materials. The dataset of Φ_e contains values ranging from −2 to 6 eV, which also covers the typical range of electron injection barriers reported for aluminum polymer interfaces.[12,44] The negative value of electron injection barriers here is an artifact caused due to the method used to compute them and is addressed in Sec. III A 2. To ensure chemical diversity, polymers containing nine common elements including C, H, B, O, N, S, F, Cl,

and Br and various polymer classes, including polycarbonates, polyimide, polyamide, polyolefins, polyvinyl, polyethers, and polyesters, were chosen from a the list of 13 000 known polymers and included in the datasets. Figure 2(f) shows the distribution of the polymers contained in each of the datasets in terms of the first two principal components obtained from a Principal Component Analysis (PCA) of the Polymer Genome fingerprint[19] of the polymers. It can be observed that the distribution of the data points in each dataset (represented by color) chemically and morphologically encompasses the diversity of the known polymer space (represented in gray) to a good extend.

1. Bandgap

For semiconductors and insulating materials (including polymers), the bandgap computed for their crystal models at the HSE level of DFT was established[10,45,46] to be a good estimation of the experimentally measured property. Here, we argue that the polymer chain model is also good for estimating the bandgap using DFT. Indeed, Fig. 2(d) clearly shows that two levels of polymer models, i.e., polymer chains and crystals, could lead to the comparable DFT-computed values of the bandgap. This suggests that in most linear polymers, the inter-chain interactions are weak and do not contribute significantly to the electronic structure. Therefore, using the single chain model to compute E_{gap} of a polymer could be a reasonable approximation. We also note that deviations of ~1 eV from this general behavior can be observed throughout the plot, hinting at the importance of considering bulk effects to accurately determine polymer E_{gap}.

2. Electron injection barrier

In the case of ϕ_e, a strong correlation between experimental values and DFT computations was established by Kamal *et al.*[12] and shown in Fig. 2(e). However, there seem to be some "systematic" deviations between the results of the two methods. To develop an ML model that can predict the experimental values of ϕ_e, this gap in numerical parity was removed using the BC method. To assess the robustness of the method, a leave-one-out cross-validation was performed on a subset of the high-throughput dataset for which the experimental observations of interfacial electron injection were available. The blue points in Fig. 3(a) show the results of this exercise. It can be seen that numerical parity was greatly improved using this technique. Having validated the BC method, it was employed to adapt the DFT-computed electron injection barriers of all polymers in the dataset to estimate their experimental values. Figure 3(b) shows the values of ϕ_e before and after BC.

B. Machine learning results

The next important step toward building accurate and reliable ML models is to generate relevant features that uniquely represent each polymer. To capture the polymer chemistry, a total of about 800 chemical features were used to numerically fingerprint polymers in each dataset. After a dimensionality reduction step using RFE, the polymers in $E_{gap}^{crystal}$, E_{gap}^{chain}, and ϕ_e datasets were represented by 126 features, 600 features, and 400 features, respectively.

Using the curated and pre-processed data described in Sec. III A, we developed three ML models to predict E_{gap} and Φ_e. As described previously, GPR was employed to learn the relationship between the numerical representation of the polymer and the properties. Figures 4(a)–4(c) show the learning curves for models trained on the $E_{gap}^{crystal}$ dataset, the E_{gap}^{chain} dataset, and the Φ_e dataset, respectively. To perform model training, the datasets are split into two: training set, containing 80% of the total data, and test set, containing 20% of the data. The x-axis of these figures shows the percentage of the training dataset used for training the model. The performance of the GPR models when evaluated on the training and test sets is represented along the y axis in terms of Root Mean Square Errors (RMSEs) in blue and red, respectively. The error bars on each of the points denote the 1σ deviation in the reported RMSE values over 50 runs. From Figs. 4(a)–4(c), we can see that the ML models in all three cases tend to saturate when at least ≃80% of the training data are used for model training, indicating that the inherent data in the dataset are sufficiently representative of polymers in the chemical space described.

Looking more closely at the performance of the models, Figs. 4(d)–4(f) represent parity plots that depict the prediction performance of the models trained on 80% of the data. To evaluate the

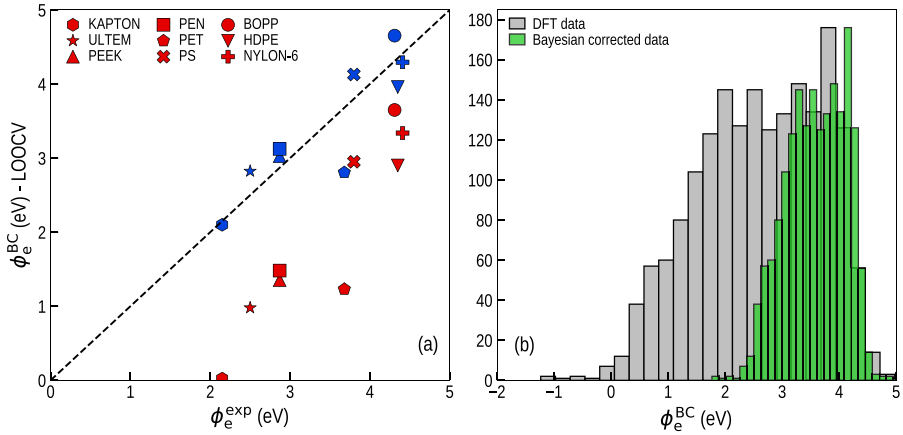

FIG. 3. (a) A comparison between the experimental value of the electron injection barrier ϕ_e^{exp} and the leave-one-out cross-validation (blue) result of BC applied to the computed value (red) and (b) the distribution of magnitudes of computed electron injection barriers before and after correction.

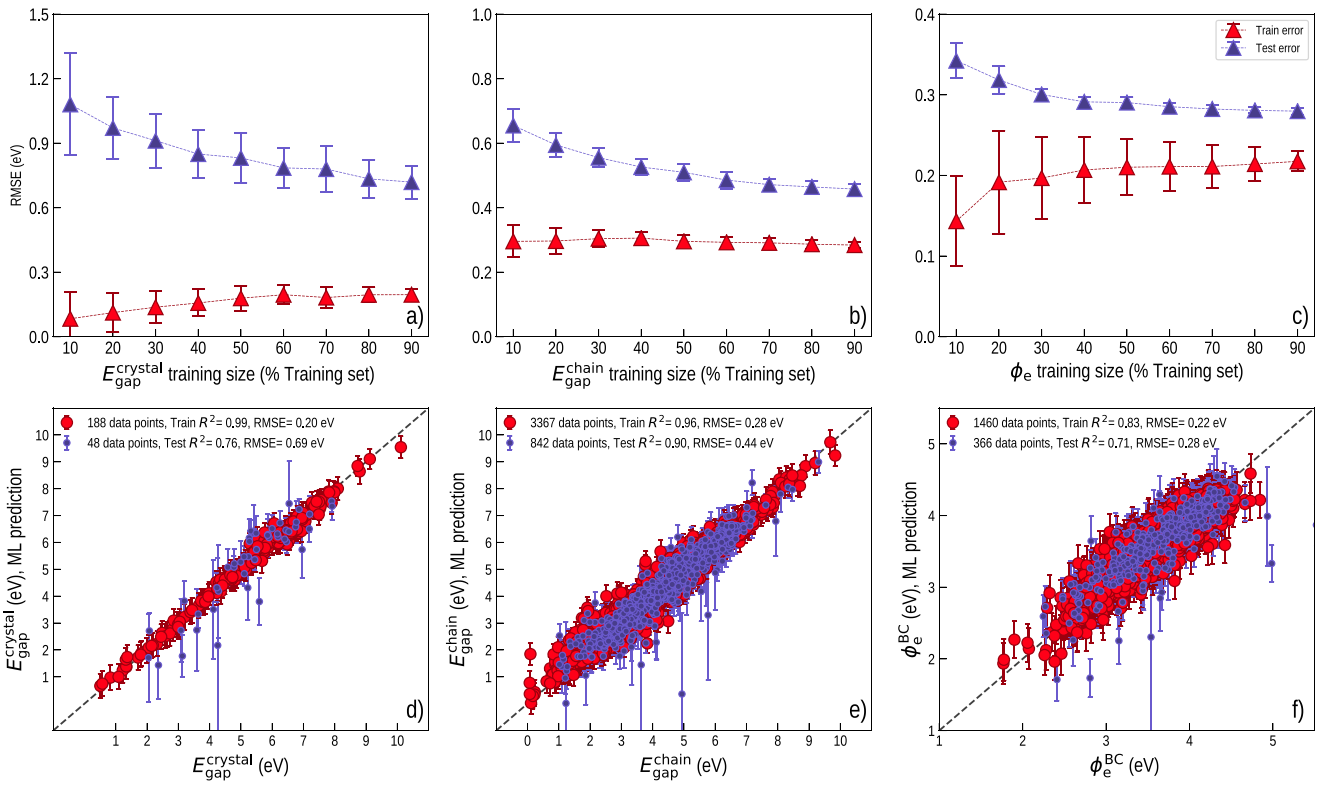

FIG. 4. Learning curves characterizing the performance of the ML models for the polymer crystal bandgap, chain bandgap, and charge injection barrier, respectively, are shown in (a)–(c). Note that each point in these plots was obtained by averaging the RMSE of 50 models trained on a given training set size. The best performing ML models of these properties are shown in (d)–(f). Each parity plot was obtained from a model trained on 80% of the data and tested on the remaining 20% of the data.

overall performance of the model at a high level, the RMSE and coefficient of determination (R^2) of the model is evaluated on the training and test sets. For the $E_{gap}^{crystal}$ models, RMSE on the test set is 0.5 eV and R^2 is 0.94, which is similar to the bandgap model performances reported by Patra *et al.*[47] For the ϕ_e model, the RMSE on the test set is 0.3 eV and R^2 is 0.78. A more local perspective can be obtained by looking more closely at the parity plots shown in Figs. 4(d)–4(f). It can be seen in all cases that most of the training and test points are on the parity line. A few points can be seen to be quite far from the parity line, especially in Figs. 4(e) and 4(f). However, interestingly, the model has a high uncertainty at the point, meaning that points such as the outlier are underrepresented in the dataset. The uncertainty of the predictions can hence be used as a tool to evaluate our confidence in the model. On visual inspection of Fig. 4(f), we can see that the low value of R^2 for ϕ_e is due to only a few outliers. From Figs. 4(d) and 4(e), it can hence be seen that all three models are robust and dependable when their uncertainties are low. Finally, to create the final models, 100% of data were utilized. Out of abundance of caution to prevent overfitting, fivefold-cross-validation was employed to make 50 sets of models, and the best model among them was chosen based on the magnitude of validation errors.

C. Candidates for high voltage insulation applications

Now, to find polymers that may be useful for high voltage applications, we attempt a high-throughput screening exercise. This is very similar to the DFT data based screening for high breakdown polymers attempted by Kamal *et al.*,[12] for ~1000 polymers. The major difference here is that since we are equipped with ML models, we screen from a much larger list set of ≃13 000 previously known polymers. Figure 5(a) shows the predictions for the list of ≃13 000 polymers. To reduce the space charge accumulation and hence facilitate the longer life span of the insulation, we choose polymers with high E_{gap} and Φ_e values. Here, a criterion of $E_{gap} \geq 5$ eV and $\phi_e \geq 3$ eV is employed to achieve this. To also account for the thermal stability of the polymer, we further add an extra screening criterion based on the glass transition temperature (T_g) of the polymer. To do this, T_g models developed by Jha *et al.*[48] were used. A criterion of $T_g \geq 400$ K was used for the purposes of downselection. Using these criteria [indicated by the shaded region on the top right of Fig. 5(a)], we identified 81 polymers that potentially could be suitable for high potential insulation applications. Because of the past synthesis evidence of all the selected polymers, we hope that they will be re-synthesized and tested for high potential applications. A

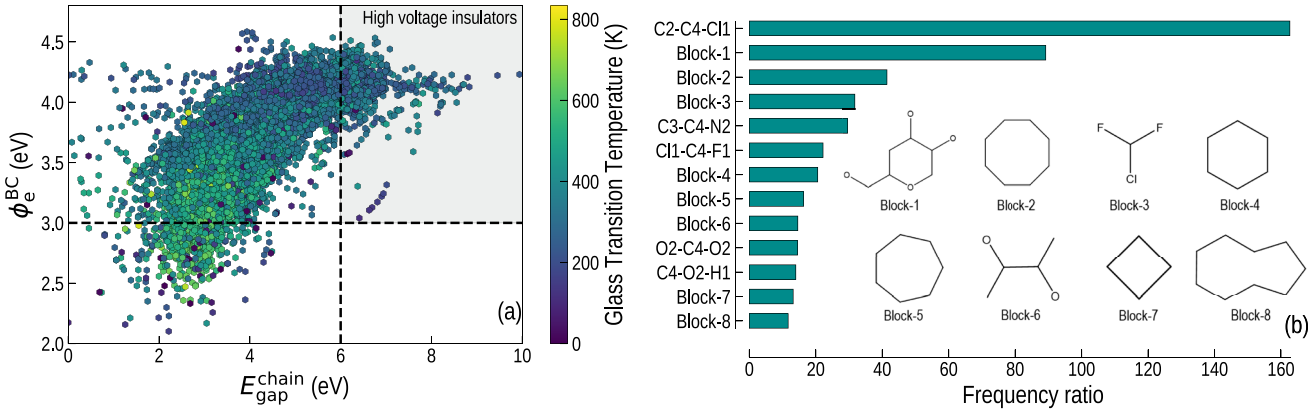

FIG. 5. (a) E_{gap}, ϕ_e, and T_g predicted for the dataset of 13k polymers. Polymers in the shaded area with $T_g \geq 400$ K are predicted to be suitable for high voltage insulators. (b) Chemical features that are more frequent in the proposed list of potential high breakdown polymers compared to those of the list of 13 000 known polymers. Note that the numbers beside the atomic triplets shown in the plot depict the number of atoms the group is bonded to (e.g., C4 means carbon bonded to four groups).

full list of these candidate polymers with details is given in Sec. III E of the supplementary material.

To reveal what chemical features stand out in the proposed list of high breakdown polymers, a frequency analysis of different chemical groups was performed. The analysis considered the frequency of occurrence of different atomic triplets and common chemical blocks proposed by Tran *et al.*[19] [going forward, referred to as "feature(s)," short for chemical feature(s)] in the proposed polymers and compared it with that of their occurrence in the list of ~13k known polymers. Finally, the ratio of the frequency of occurrence of each feature in the list of proposed and known polymers was used as the metric to ascertain feature importance. The frequency ratio (f_{ratio}) used here is given by the following formula:

$$f_{ratio} = \frac{N_{feature}^{proposed}/N_{polymers}^{proposed}}{N_{feature}^{known}/N_{polymers}^{known}}, \tag{4}$$

where $N_{feature}^{proposed}$ is the number of occurrences of a given feature in the list of proposed polymers, $N_{polymers}^{proposed}$ is the total number of polymers in the list of proposed polymers, $N_{feature}^{known}$ is the number of occurrences of a given feature in the list of known polymers, and $N_{polymers}^{known}$ is the total number of polymers in the list of proposed polymers. Figure 5(b) shows the features that are more than ten times more frequent in the list of proposed polymers than in the list of known polymers.

Several features stand out from this analysis. It can be seen that non-aromatic hydrocarbon rings are prolific in the proposed polymers. `Block-3`, `block-4`, `block-5`, `block-7`, and `block-8` are examples of such ring structures. This observation is fairly intuitive as the presence of rings in the monomer structure results in the restriction of chain mobility, which in turn results in the increase in T_g. In addition, the non-aromatic nature of the ring will ensure a high bandgap as the rings are composed of sp3-hybridized carbon atoms. Terminal halogen groups are also profuse in the proposed list. `Block-3` and `Cl1-C4-F1` are examples of terminal halides that are prolific in the proposed list. The features `C2-C4-C1`, though shown to have a high value of frequency ratio, only occur two times in the proposed list. However, the fact that it only occurs two times in the list of known polymers makes it worth further consideration. The presence of a large number of terminal halogens could be explained by the fact that the presence of halogens in the polymer tends to shift the conduction band minima of the polymer with respect to those of the Fermi level of Al[12] increasing the electron injection barrier at the metal-polymer interface and by the fact that they tend to increase T_g due to the increase in inter-chain interactions. Cellulose-like groups shown in block-1 are also seen to be prolific. This is also intuitive when considering high T_g in cellulose-like polymers[49] and their high bandgap with their composition of sp3-hybridized carbon and oxygen atoms. The high frequency of these features suggests that a polymer that contains one or more of these features has a greater chance of being a high breakdown polymer compared to a polymer picked at random. The presence of these features can then be considered as design principles to design the next generation of high breakdown dielectrics.

D. Outlook

In an attempt to develop novel polymer dielectrics, in a previous work,[12] we established that E_{gap} and ϕ_e are both correlated to E_{bd} and laid out computational pathways to rapidly estimate them. This allowed us to reformulate the difficult problem of finding breakdown resistant polymers into an easier problem of finding polymers with a high bandgap and electron injection barrier. In this work, we take a step further toward discovering novel polymer dielectrics by developing large datasets of the bandgap and electron injection barrier by drawing from a set of 13 000 known polymers. Furthermore, ML models were then built using these datasets to enable rapid reliable prediction of E_{gap} and ϕ_e. These ML models in conjunction with that of T_g were then used to screen for high breakdown dielectrics from the larger list of 13 000 known polymers.

While the approach taken here is a good starting point toward the discovery of high breakdown polymers, further refinement in multiple aspects considered here can make the adopted strategy more effective. One such avenue is the properties considered here to look for high breakdown polymers, E_{gap}, ϕ_e, and T_g. While these properties are desirable for high breakdown polymers, several other properties can be used to further refine the choice of potential high breakdown polymers. These include polymer features such as the cohesive energy density,[50] dielectric constant of the polymer, charge mobility in the polymer, and thermal conductivity of the polymer. The bottleneck in using these properties currently is the lack of a reliable dataset for them. Hence, development of a reliable dataset of these properties would help further refine the strategy. Another possible avenue is the use of advanced learning strategies such as multi-task learning[51] to learn multiple polymer properties using the same NN based ML model. This method has the added advantage that these models learn the correlation between the properties in addition to the relationship between the fingerprint and the properties, which enables them to reliably predict properties that may have fewer training data but correlated with other properties. Nonetheless, the ML models developed herein and the polymers selected using them form a starting point in the exploration of high breakdown polymers using the proposed approach. Furthermore, the ML models developed here can also be used in conjunction with generative methods such as the genetic algorithm,[52] variational autoencoders,[53] and graph to graph translations[54] to systematically create new polymers that satisfy set objectives.

E. Conclusions

In summary, in an attempt to facilitate the discovery of novel high voltage insulators, we have developed and deployed ML models to predict the essential electronic properties of polymers, i.e., bandgap E_{gap} and electron injection barrier Φ_e. To facilitate the development of the ML models, we have created reliable datasets of E_{gap} and Φ_e using DFT computations and utilized statistical techniques to address systematic deviations of DFT results from experiments wherever necessary. Furthermore, Polymer Genome fingerprints in conjunction with GPR were used to develop accurate ML models of E_{gap} and Φ_e, which are faster than modeling and experiments. These models were deployed at www.polymergenome.org to facilitate easy access for the community. These models were also used to find novel high voltage insulators by screening from a large list of known polymers. We hope that these polymers will go through (re)synthesis and tests for high voltage electrical insulation applications.

SUPPLEMENTARY MATERIAL

The full list of 81 potential high breakdown polymers that were found using the methods described in this paper is presented in the supplementary material. Details listed for each polymer include their IUPAC name, a 2D representation of the polymer repeat unit, and the mean prediction value (using ML models) of E_{gap}, 2, and T_g. References to synthesis method(s) for most polymers are also added for the benefit of the community.

AUTHORS' CONTRIBUTIONS

D.K. and H.T. contributed equally to this paper.

ACKNOWLEDGMENTS

This work was financially supported by the Office of Naval Research through a Multi-University Research Initiative (MURI) grant (Grant No. N00014-17-1-2656) and the Toyota Research Institute through the Accelerated Materials Design and Discovery program. The computational resource provided by XSEDE through Project No. "DMR080058N" is also acknowledged.

The authors declare no competing financial interests.

DATA AVAILABILITY

The data that support the findings of this study are openly available in Khazana at https://khazana.gatech.edu/.[55]

REFERENCES

[1] D. Jovcic, *High Voltage Direct Current Transmission: Converters, Systems and DC Grids* (John Wiley & Sons, 2019).
[2] J. A. Rosero, J. A. Ortega, E. Aldabas, and L. Romeral, "Moving towards a more electric aircraft," IEEE Aerosp. Electron. Syst. Mag. **22**, 3–9 (2007).
[3] G. C. Montanari, P. Seri, X. Lei, H. Ye, Q. Zhuang, P. Morshuis, G. Stevens, and A. Vaughan, "Next generation polymeric high voltage direct current cables—A quantum leap needed?," IEEE Electr. Insul. Mag. **34**, 24–31 (2018).
[4] G. Pietrini, D. Barater, G. Franceschini, P. Mancinelli, and A. Cavallini, "An open problem for more electrical aircraft (MEA): How insulation systems of actuators can be qualified?," in *2016 IEEE Energy Conversion Congress and Exposition (ECCE)* (IEEE, 2016), pp. 1–8.
[5] J. F. Hall, "History and bibliography of polymeric insulators for outdoor applications," IEEE Trans. Power Delivery **8**, 376–385 (1993).
[6] B. Knapp and P. A. Kohl, "Polymers for microelectronics," J. Appl. Polym. Sci. **131**, 41233 (2014).
[7] Y. Zhang, J. Lewiner, C. Alquie, and N. Hampton, "Evidence of strong correlation between space-charge buildup and breakdown in cable insulation," IEEE Trans. Dielectr. Electr. Insul. **3**, 778–783 (1996).
[8] T. Tanaka, "Space charge injected via interfaces and tree initiation in polymers," in *2001 Annual Report Conference on Electrical Insulation and Dielectric Phenomena (Cat. No. 01CH37225)* (IEEE, 2001), pp. 1–15.
[9] M. Ieda, "Carrier injection, space charge and electrical breakdown in insulating polymers," IEEE Trans. Electr. Insul. **EI-22**, 261–267 (1987).
[10] T. D. Huan, S. Boggs, G. Teyssedre, C. Laurent, M. Cakmak, S. Kumar, and R. Ramprasad, "Advanced polymeric dielectrics for high energy density applications," Prog. Mater. Sci. **83**, 236 (2016).
[11] A. Huzayyin, S. Boggs, and R. Ramprasad, "Density functional analysis of chemical impurities in dielectric polyethylene," IEEE Trans. Dielectr. Electr. Insul. **17**, 926–930 (2010).
[12] D. Kamal, Y. Wang, H. D. Tran, L. Chen, Z. Li, C. Wu, S. Nasreen, Y. Cao, and R. Ramprasad, "Computable bulk and interfacial electronic structure features as proxies for dielectric breakdown of polymers," ACS Appl. Mater. Interfaces **12**, 37182 (2020).
[13] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder *et al.*, "Commentary: The materials project: A materials genome approach to accelerating materials innovation," APL Mater. **1**, 011002 (2013).
[14] A. Jain, K. A. Persson, and G. Ceder, "Research update: The materials genome initiative: Data sharing and the impact of collaborative *ab initio* databases," APL Mater. **4**, 053102 (2016).
[15] R. Batra, L. Song, and R. Ramprasad, "Emerging materials intelligence ecosystems propelled by machine learning," Nat. Rev. Mater. **2020**, 1–24 (2020).
[16] C. Kim, A. Chandrasekaran, T. D. Huan, D. Das, and R. Ramprasad, "Polymer genome: A data-powered polymer informatics platform for property predictions," J. Phys. Chem. C **122**, 17575–17585 (2018).

[17]L. Chen, G. Pilania, R. Batra, T. D. Huan, C. Kim, C. Kuenneth, and R. Ramprasad, "Polymer informatics: Current status and critical next steps," Mater. Sci. Eng., R **144**, 100595 (2021).
[18]T. D. Huan, A. Mannodi-Kanakkithodi, and R. Ramprasad, "Accelerated materials property predictions and design using motif-based fingerprints," Phys. Rev. B **92**, 014106 (2015).
[19]H. D. Tran, C. Kim, L. Chen, A. Chandrasekaran, R. Batra, S. Venkatram, D. Kamal, J. P. Lightstone, R. Gurnani, P. Shetty *et al.*, "Machine-learning predictions of polymer properties with polymer genome," J. Appl. Phys. **128**, 171104 (2020).
[20]G. W. Ehrenstein, *Polymeric Materials: Structure, Properties, Applications* (Carl Hanser Verlag GmbH & Co. KG, 2012).
[21]A. Mannodi-Kanakkithodi, G. Pilania, T. D. Huan, T. Lookman, and R. Ramprasad, "Machine learning strategy for the accelerated design of polymer dielectrics," Sci. Rep. **6**, 20952 (2016).
[22]V. Sharma, C. C. Wang, R. G. Lorenzini, R. Ma, Q. Zhu, D. W. Sinkovits, G. Pilania, A. R. Oganov, S. Kumar, G. A. Sotzing, S. A. Boggs, and R. Ramprasad, "Rational design of all organic polymer dielectrics," Nat. Commun. **5**, 4845 (2014).
[23]L. Chen, T. D. Huan, Y. C. Quintero, and R. Ramprasad, "Charge injection barriers at metal/polyethylene interfaces," J. Mater. Sci. **51**, 506–512 (2016).
[24]T. D. Huan, A. Mannodi-Kanakkithodi, C. Kim, V. Sharma, G. Pilania, and R. Ramprasad, "A polymer dataset for accelerated property prediction and design," Sci. Data **3**, 160012 (2016).
[25]A. Mannodi-Kanakkithodi, G. M. Treich, T. D. Huan, R. Ma, M. Tefferi, Y. Cao, G. A. Sotzing, and R. Ramprasad, "Rational co-design of polymer dielectrics for energy storage," Adv. Mater. **28**, 6277–6291 (2016).
[26]R. Ma, A. F. Baldwin, C. Wang, I. Offenbach, M. Cakmak, R. Ramprasad, and G. A. Sotzing, "Rationally designed polyimides for high-energy density capacitor applications," ACS Appl. Mater. Interfaces **6**, 10445 (2014).
[27]T. D. Huan and R. Ramprasad, "Polymer structure prediction from first principles," J. Phys. Chem. Lett. **11**, 5823–5829 (2020).
[28]G. Kresse and J. Hafner, "*Ab initio* molecular dynamics for liquid metals," Phys. Rev. B **47**, 558 (1993).
[29]G. Kresse and J. Furthmüller, "Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set," Comput. Mater. Sci. **6**, 15–50 (1996).
[30]G. Kresse and J. Furthmüller, "Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set," Phys. Rev. B **54**, 11169 (1996).
[31]K. Lee, É. D. Murray, L. Kong, B. I. Lundqvist, and D. C. Langreth, "Higher-accuracy van der Waals density functional," Phys. Rev. B **82**, 081101 (2010).
[32]É. D. Murray, K. Lee, and D. C. Langreth, "Investigation of exchange energy density functional accuracy for interacting molecules," J. Chem. Theory Comput. **5**, 2754–2762 (2009).
[33]J. Heyd, G. E. Scuseria, and M. Ernzerhof, "Hybrid functionals based on a screened Coulomb potential," J. Chem. Phys. **118**, 8207–8215 (2003).
[34]H. Zhu and R. Ramprasad, "Effective work function of metals interfaced with dielectrics: A first-principles study of the Pt-Hfo_2 interface," Phys. Rev. B **83**, 081416 (2011).
[35]Y. Cardona Quintero, H. Zhu, and R. Ramprasad, "Adsorption of CH_3S and CF_3S on Pt(111) surface: A density functional theory study," J. Mater. Sci. **48**, 2277–2283 (2013).
[36]M. C. Kennedy and A. O'Hagan, "Bayesian calibration of computer models," J. R. Stat. Soc.: Ser. B **63**, 425–464 (2001).
[37]V. R. Joseph and S. N. Melkote, "Statistical adjustments to engineering models," J. Qual. Technol. **41**, 362–375 (2009).
[38]C. Kim, A. Chandrasekaran, A. Jha, and R. Ramprasad, "Active-learning and materials design: The example of high glass transition temperature polymers," MRS Commun. **9**, 860 (2019).
[39]I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," Mach. Learn. **46**, 389–422 (2002).
[40]C. W. Coley, R. Barzilay, W. H. Green, T. S. Jaakkola, and K. F. Jensen, "Convolutional embedding of attributed molecular graphs for physical property prediction," J. Chem. Inf. Model. **57**, 1757–1772 (2017).
[41]C. Chen, W. Ye, Y. Zuo, C. Zheng, and S. P. Ong, "Graph networks as a universal machine learning framework for molecules and crystals," Chem. Mater. **31**, 3564–3572 (2019).
[42]S. Wang, Z. Li, S. Zhang, M. Jiang, X. Wang, and Z. Wei, "Molecular property prediction based on a multichannel substructure graph," IEEE Access **8**, 18601–18614 (2020).
[43]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res. **12**, 2825–2830 (2011).
[44]G. Chen, T. Y. G. Tay, A. E. Davies, Y. Tanaka, and T. Takada, "Electrodes and charge injection in low-density polyethylene using the pulsed electroacoustic technique," IEEE Trans. Dielectr. Electr. Insul. **8**, 867–873 (2001).
[45]J. P. Perdew, W. Yang, K. Burke, Z. Yang, E. K. U. Gross, M. Scheffler, G. E. Scuseria, T. M. Henderson, I. Y. Zhang, A. Ruzsinszky *et al.*, "Understanding band gaps of solids in generalized Kohn–Sham theory," Proc. Natl. Acad. Sci. U. S. A. **114**, 2801–2806 (2017).
[46]H. Xiao, J. Tahir-Kheli, and W. A. Goddard III, "Accurate band gaps for semiconductors from density functional theory," J. Phys. Chem. Lett. **2**, 212–217 (2011).
[47]A. Patra, R. Batra, A. Chandrasekaran, C. Kim, T. D. Huan, and R. Ramprasad, "A multi-fidelity information-fusion approach to machine learn and predict polymer bandgap," Comput. Mater. Sci. **172**, 109286 (2020).
[48]A. Jha, A. Chandrasekaran, C. Kim, and R. Ramprasad, "Impact of dataset uncertainties on machine learning model predictions: The example of polymer glass transition temperatures," Modell. Simul. Mater. Sci. Eng. **27**, 024002 (2019).
[49]T. T. Kararli, J. B. Hurlbut, and T. E. Needham, "Glass–rubber transitions of cellulosic polymers by dynamic mechanical analysis," J. Pharm. Sci. **79**, 845–848 (1990).
[50]C. Wu, L. Chen, A. Deshmukh, D. Kamal, Z. Li, P. Shetty, J. Zhou, H. Sahu, H. Tran, G. Sotzing, R. Ramprasad, and Y. Cao, "Targeted co-designs of dielectric polymers tolerant to enormous electric field and temperature via material-informatic discovery acceleration," Chem. Mater. (submitted).
[51]C. Künneth, A. C. Rajan, H. Tran, L. Chen, C. Kim, and R. Ramprasad, "Polymer informatics with multi-task learning," Patterns **2**, 100238 (2021).
[52]C. Kim, R. Batra, L. Chen, H. Tran, and R. Ramprasad, "Polymer design using genetic algorithm and machine learning," Comput. Mater. Sci. **186**, 110067 (2020).
[53]R. Batra, H. Dai, T. D. Huan, L. Chen, C. Kim, W. R. Gutekunst, L. Song, and R. Ramprasad, "Polymers for extreme conditions designed using syntax-directed variational autoencoders," Chem. Mater. **32**(24), 10489–10500 (2020).
[54]W. Jin, K. Yang, R. Barzilay, and T. Jaakkola, "Learning multimodal graph-to-graph translation for molecular optimization," arXiv:1812.01070 (2018).
[55]D. Kamal, H. D. Tran, and R. Ramprasad, "Novel high-voltage polymer insulators using computational and data-driven techniques," https://khazana.gatech.edu/ (2021).