

polyG2G: A Novel Machine Learning Algorithm Applied to the Generative Design of Polymer Dielectrics

Rishi Gurnani, Deepak Kamal, Huan Tran, Harikrishna Sahu, Kenny Scharm, Usman Ashraf, and Rampi Ramprasad*



Cite This: <https://doi.org/10.1021/acs.chemmater.1c02061>



Read Online

ACCESS |



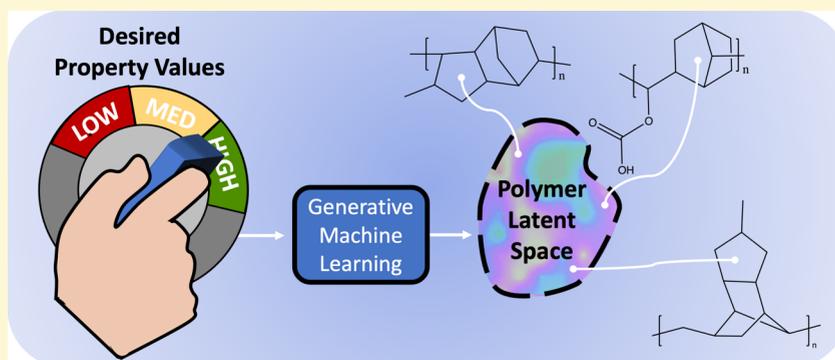
Metrics & More



Article Recommendations



Supporting Information



ABSTRACT: Polymers, due to advantages such as low-cost processing, chemical stability, low density, and tunable design, have emerged as a powerhouse class of materials for a wide range of applications, including dielectrics. However, in certain applications, the performance of dielectrics is limited by insufficient electric breakdown strength. Using this real-world application as a technology driver, we describe a novel artificial intelligence (AI)-based approach for the design of polymers. We call this approach polyG2G. The key concept underlying polyG2G is graph-to-graph translation. Graph-to-graph translation solves the inverse problem. First, the subtle chemical differences between high- and low-performing polymers are learned. Then, the learned differences are applied to known polymers, yielding large libraries of novel, high-performing, hypothetical polymers. Our approach, with respect to a host of presently adopted design methods, exhibits a favorable trade-off between generation of chemically valid materials and available chemical search space. polyG2G finds thousands of potentially high-value targets (in terms of glass-transition temperature, band gap, and electron injection barrier) from an otherwise intractable search space. Density functional theory simulations of band gap and electron injection barrier confirm that a large fraction of the polymers designed by polyG2G are indeed of high value. Finally, we find that polyG2G is able to learn established structure–property relationships.

1. INTRODUCTION

The maximum electric field that can be applied to a dielectric polymer without destroying its insulating characteristics is known as the dielectric breakdown strength. This property sets an upper limit on the maximum electrostatic energy that can be stored in a capacitor. Polymer dielectrics are favored in high power, high energy density, capacitors primarily due to their elevated dielectric breakdown strength relative to other materials.¹ A world with increasingly demanding high power electronics necessitates the development of polymers with breakdown strengths that surpass the limits of commercially available materials.^{1–3}

Owing to the complex mechanisms that lead to dielectric breakdown in polymers, direct simulations of dielectric breakdown under realistic conditions are not practical. A promising alternative approach is the estimation of dielectric breakdown strength through accessible proxies. Kamal et al. show that materials with high values of two proxy properties—

band gap and electron injection barrier with respect to an aluminum interface (simply referred to as “electron injection barrier” below)—are likely to exhibit high dielectric breakdown strength.⁴ In this approach, finding materials with high dielectric breakdown strength is reformulated into a multi-objective optimization involving the proxy properties.

Finding materials that optimize multiple properties is a nontrivial task. The main challenge is that the material chemical and configuration space is effectively infinite. Thus, a comprehensive search using physical experimentation or

Received: June 15, 2021

Revised: August 5, 2021

reliable computational methods (such as density functional theory, DFT) is practically prohibited. During the last decade, the use of machine learning (ML) models, trained on past materials data, to rapidly screen candidates and design for desired performances has become a powerful approach impacting many domains.^{5–7} This approach has already led to the discovery of a few high energy density polymer dielectric films.⁸

ML, applied to the search and design of novel materials, can be categorized into two flavors. The vast majority of previous work follows the “forward” predict-and-screen template: start with a list of candidate structures, predict properties of each, and see if any of the candidate materials meet application needs. If not, enumerate another list of candidate structures and start again. This process is constrained to exploration of designs that fall strictly within the purview of human imagination.

Materials design in the other direction involves solving the “inverse” problem by directly generating materials that meet a desired set of target properties. In recent years, due to advances in materials data collection and in the field of deep learning—namely, the advent of generative machine learning approaches—such design approaches are blossoming. These “dial-a-property” methods escape the shackles of iterative, human-directed, structure enumeration and therefore have the potential to accelerate materials design. Yet, care must be taken to maximize the fraction of generated candidates that are chemically feasible.

Perhaps the most popular of the generative algorithms are the genetic algorithm (GA) and the variational autoencoder (VAE). Notably, Kim et al.⁹ have used a GA, while Batra et al.¹⁰ have used a syntax-directed VAE (SD-VAE) to design polymers with large band gaps and high glass-transition temperatures. Other algorithms for the design of materials besides polymers have also been explored.^{11,12} Yet, gaps and concerns related to the design of polymers using existing generative algorithms remain. First, the GA is limited by its reliance on the Breaking of Retrosynthetically Interesting Chemical Substructures (BRICS) algorithm¹³ for generating chemical fragments. Kim et al. showed that the GA can learn to successfully join these fragments into high-performing polymer designs. Based on past data, the BRICS algorithm predefines the number of connections on a fragment and the locations of those connections. Consequently, the BRICS algorithm does not account for all realistic connection points between fragments. This constrains the space of chemically valid polymers that the GA can construct. In some cases, this constraint could be beneficial, as all generated polymers would only exhibit substructures (i.e., BRICS fragments) that are known to exist in previously-synthesized polymers. But, in other applications, relevant polymer designs may be hidden in regions not accessible to the GA. Second, of all polymers generated by the SD-VAE, 27% are chemically valid. This figure, while a notable improvement with respect to previous generative algorithms, leaves ample room for improvement.

To address these points, we introduce the polymer graph-to-graph translation algorithm polyG2G. An overarching challenge (composed of several subproblems) for generative algorithms is the design of chemically valid materials. One key subproblem for the SD-VAE is learning that chemical rings must be closed.^{10,14} polyG2G alleviates this problem by representing polymers first as junction trees and then as graphs.¹¹ In the junction tree phase, each node is assigned

either an element type or a ring type (e.g., benzene, thiophene, etc.). Since all atoms in a given ring are specified simultaneously, the need to close any rings is circumvented. In the graph step, the types of bonds connecting each node are defined. Additionally, for nodes that represent rings, the positions of bonds (e.g., ortho, meta, para, etc.) are defined.

Another unique feature of polyG2G, relative to the work of Kim et al. and Batra et al., is that the problem of polymer design is cast as a “translation” problem over polymer graphs (i.e., graph-to-graph translation). In graph-to-graph translation, rather than generating high-performing polymers from scratch, we start with a template polymer and learn how to chemically translate (or convert) it to a better polymer.¹¹

As a tangible application, here, we use polyG2G to search for new high-performing polymers for capacitive energy storage, i.e., polymers that possess a large band gap and a large electron injection barrier (properties that are strongly correlated with high electric breakdown strength) as well as a high T_g . High T_g is desirable so that a material can retain its structure, and therefore its function, stably over a wide range of temperatures. After training a polyG2G model on a data set of ~13 000 synthesized polymers—only 8 of which meet our triproperty objective—we designed 3556 novel, chemically valid candidates predicted to surpass our property objectives. To test these candidates, we selected a small subset and used DFT to compute their band gaps and electron injection barriers. We found that 50% of the polymers in this subset do indeed match our objectives. Finally, we mined rules for simultaneous maximization of T_g , band gap, and electron injection barrier directly from our polyG2G-designed candidates. Several of these rules have also been reported in the past, thus validating the efficacy of our design workflow (data sets, numerical representation of polymers, polyG2G, and property prediction) in learning practical chemical guidelines and incorporating them into new polymers.

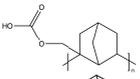
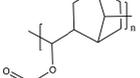
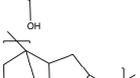
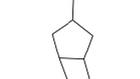
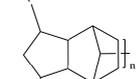
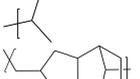
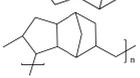
The novel contributions of this work include guidelines for designing polymers with high dielectric breakdown strength, the extension of generic graph-to-graph translation to polymer design (i.e., polyG2G), the first generative design work on simultaneous optimization of three or more polymer properties, quantitative comparisons between polyG2G, the SD-VAE, and the GA, and a set of 10, DFT-validated, polymers predicted to exhibit remarkable dielectric breakdown strength (Table 1). We propose these ten candidates for further study.

Although we optimize three properties in this work, polyG2G can, in principle, be extended to the simultaneous optimization of any number of properties. Furthermore, all properties that can be accurately predicted from the polymer repeat unit can be optimized by polyG2G. Creating polymers in this manner allows us to systematically probe the polymer chemical universe and reveal suitable novel candidates for myriads of other applications.

2. METHODS

A high-level overview of the protocol followed in this study is presented in Figure 1. We start with a sparse (in terms of the number of property entries per polymer) data set (see Section 2.1) of material properties for 13 014 synthesized polymers. The composition and chemical structure of each polymer was parsed and converted to a machine-readable numerical vector, known as a fingerprint (see Section 2.2). These fingerprints were used to train property predictors (as detailed in Section 2.5). These predictors, which instantaneously estimate polymer properties with good accuracy, were used to produce property entries for each polymer in the data set. This new,

Table 1. Selected Novel, High-Value, Targets Discovered by polyG2G^a

Structure	Properties		
	E_g , DFT (eV)	ϕ_e , DFT (eV)	T_g , ML (K)
	6.04	3.13	549
	6.52	3.05	520
	6.21	3.27	480
	6.07	3.12	511
	6.14	3.08	495
	6.28	3.24	502
	6.32	3.24	474
	6.07	3.08	524
	6.26	3.19	478
	6.29	3.17	499

^aBand gaps (E_g) and electron injection barriers (ϕ_e) are DFT estimates. Glass-transition temperatures (T_g) are ML estimates.

dense, data set, along with our desired property objectives, was fed into our implementation of graph-to-graph translation for polymers,

polyG2G, which then designed new polymers meeting the target property objectives. The performance of these new polymer designs was estimated by our trained property predictors. The 20 highest-performing (according to eq 2) polymers were down-selected and tested via direct computation of their target properties with DFT.

2.1. Data Sets. A major prerequisite for training an ML predictor is data. Data was collected for the properties of interest in this work: T_g , band gap (E_g), and charge injection barrier (ϕ_e). The experimentally measured T_g data set was accumulated from printed handbooks, including “Polymer Handbook”,¹⁵ “Handbook of Polymers”,¹⁶ “Properties of Polymers”,¹⁷ and “Polymer Data Handbook”.¹⁸ The other two data sets, i.e., those involving E_g and ϕ_e , were created using DFT^{19,20} (see Section 2.6 for more details).

2.2. Fingerprinting. The maps learned by ML algorithms require an input space that is machine readable. Thus, fingerprinting of polymers was required. Two fingerprint schemes—the Morgan Fingerprint²¹ and the Polymer Genome (PG) fingerprint^{22,23}—were used in this work. The 2048-bit Morgan Fingerprint was used to compute similarity, from a chemical structure point of view, between polymers. We opted for this fingerprint, for this task, because the Tanimoto similarity metric—which we found to be most capable of discriminating between similar and dissimilar polymers (see the SI, Section S2, for comparison)—requires bit-wise features.

For all other tasks, the PG fingerprint was used. This fingerprint has shown success in the representation of materials over a wide chemical and property space.^{23–25} The PG fingerprint operates on the simplified molecular-input line-entry system (SMILES) string²⁶ of a true polymer with infinite repeat units. In contrast, the Morgan fingerprint operates on the SMILES string of a polymer’s corresponding “pseudopolymer”. A pseudopolymer is a molecular representation of a polymer where the repeat unit is doubled and the dangling bonds are passivated by hydrogen atoms (see step A of Figure 2).

2.3. Graph-to-Graph Translation. The underlying hypothesis of graph-to-graph (G2G) translation is that the distribution of subtle chemical differences (which we call translations) between materials that lead to drastic property differences can be learned and then sampled to design new materials with desired properties from “source” (i.e., template) materials. Further, by choosing synthesized materials as our source, the newly designed (or translated) materials should be biased toward laboratory synthesis via already-published techniques.

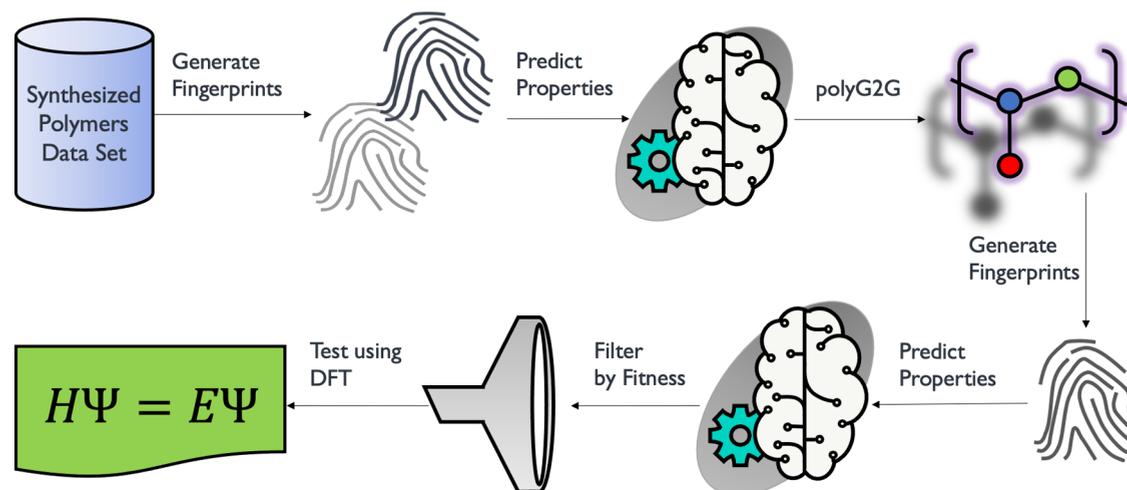


Figure 1. Flowchart describing our computational workflow: start with a set of synthesized polymers, fingerprint them, and predict their properties with ML property predictors. Then, the SMILES strings of synthesized polymers are input into polyG2G where they are translated to new polymers. The new polymers are fingerprinted to enable estimation of properties. Finally, a handful of new polymers are selected based on their estimated properties (i.e., fitness) and tested using DFT.

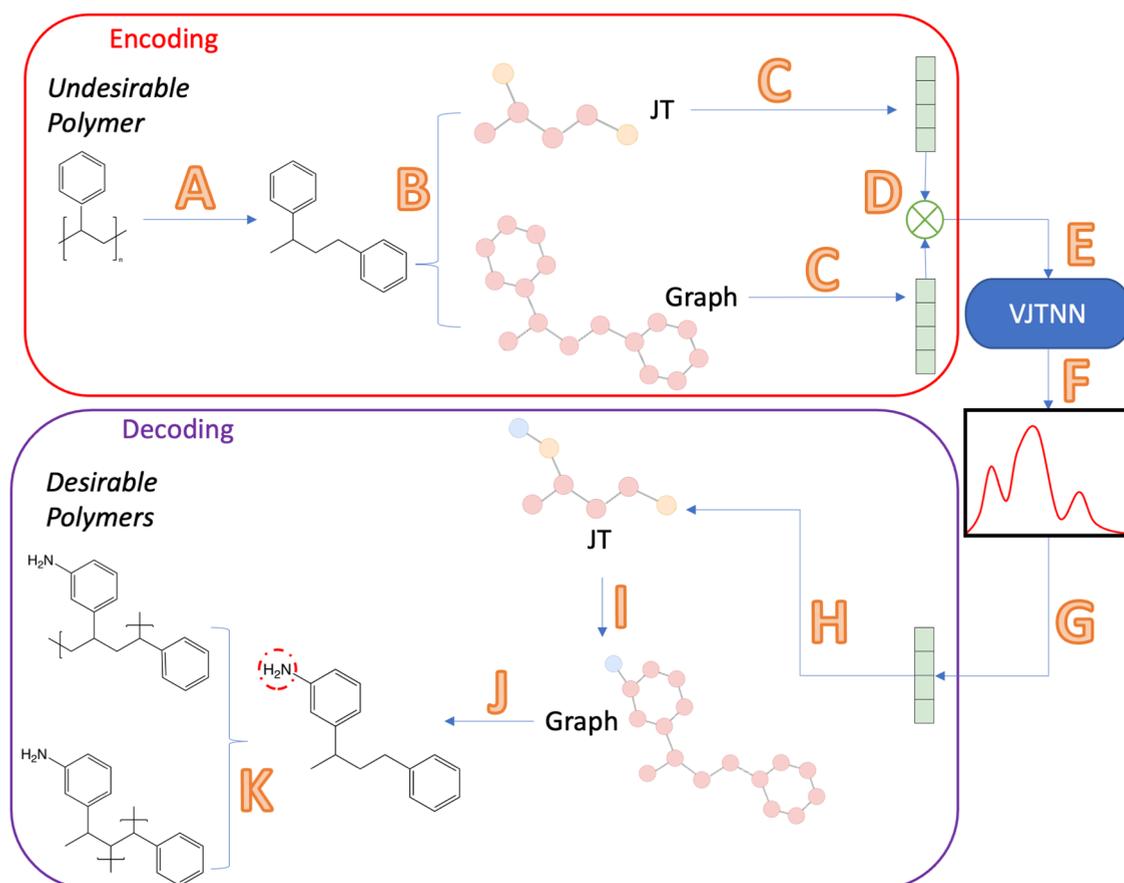


Figure 2. Diagram of the polyG2G workflow for the case of polystyrene with n_{pair} equal to two and $n_{\text{translate}}$ equal to one. The result of the latent translation is circled by the dashed red line. Pink circles are carbon atoms, yellow circles are benzene rings, and blue circles are NH_2 groups. All hydrogen atoms are implicit. (A) Create pseudopolymer from repeat unit. (B) Map pseudopolymer to a JT and a graph. (C) JT and graph are converted to a short numerical vector, the embedding, using ML. (D) Concatenate embeddings. (E) Embedding is input to VJTNN. (F) Candidate translation distribution, in latent space, is generated. (G) Embedding of a translation is sampled and added to the pseudopolystyrene embedding. This sum yields the embedding of a new pseudopolymer. (H) New pseudopolymer embedding is mapped to JT. (I) JT is mapped to graph. (J) Graph is mapped to pseudopolymer. (K) Pseudopolymer is mapped to n_{pair} polymers.

A successful implementation of graph-to-graph translation, aimed at the goal of designing materials with desirable properties, relies on four principles:

Principle 1: A way to quantify and apply translations to materials.

Principle 2: A definition of “desirable” (and “undesirable”) based upon target properties.

Principle 3: A binary similarity function, $S(\cdot)$, which evaluates an input pair of materials as either similar or dissimilar based upon chemical structure and/or composition.

Principle 4: A training set $\{(X_i^{\text{train}}, Y_i^{\text{train}})\}$ of material pairs where each X_i^{train} is an undesirable (as defined in Principle 2) material and each Y_i^{train} is a desirable material. Each pair in the set contains materials that are chemically similar (as defined in Principle 3) to one another. These pairs contain the set of successful graph-to-graph translations—one translation per pair—that machines (or humans) can learn from.

The purpose of Principle 1 is self-explanatory. polyG2G is our implementation of this principle for linear polymers (henceforth simply referred to as “polymers” unless otherwise stated). The details of our implementation are discussed in Section 2.4. Principle 2 is necessary to produce $\{(X_i^{\text{train}}, Y_i^{\text{train}})\}$. In this work, we define a desirable polymer as one that satisfies each of the following thresholds: $T_g > 450$ K, band gap > 6 eV, and electron injection barrier > 3 eV. On the other hand, an undesirable polymer is one that satisfies at least one of the following property thresholds: $T_g < 400$ K, band gap < 5.5 eV, or electron injection barrier < 2.75 eV. Principle 3 is also necessary to produce $\{(X_i^{\text{train}}, Y_i^{\text{train}})\}$. Depending on how $S(\cdot)$ is

defined, different pairs of polymers may be classified as chemically similar or dissimilar. An example of a similar pair of polymers might be polyethylene and polyacetylene, since both polymers are hydrocarbons and contain no side chains. In this work, we use a $S(\cdot)$ —formally defined in eq 1—that leverages the Tanimoto similarity

$$S(m_i, m_j) = \begin{cases} \text{True}, & T(f(m_i), f(m_j)) \geq t \\ \text{False}, & T(f(m_i), f(m_j)) < t \end{cases} \quad (1)$$

where T is the Tanimoto Similarity, $f(m)$ returns the Morgan Fingerprint²¹ of the pseudopolymer m , and t is a user-defined similarity threshold value. In this study, we set $t = 0.1$.

Our training set $\{(X_i^{\text{train}}, Y_i^{\text{train}})\}$, as prescribed by Principle 4, is constructed from a parent data set (see Section 2.1) of 13 014 synthesized polymers and corresponding property predictions of T_g , band gap, and electron injection barrier.²⁷ All undesirable parent polymers X_i are assigned to the source class X . All desirable parent polymers Y_i are assigned to the target class Y . The training set is the subset of all possible pairs (X_i, Y_j) for which $S(X_i, Y_j)$ evaluates to True. In other words, the training set contains pairs of polymers from the parent data set that are sufficiently “similar” to one another from a chemical structure point of view but drastically different than one another from a “desirability” point of view. It is important to note that polymer translation is a many-to-many learning problem. That is, for any X_i there may be multiple similar Y_j (and vice-versa). In other words, for any X_i there may be multiple successful translations. Therefore, the

map learned by a polyG2G model must output a distribution of translations for any input X_i .

The set of source polymers not present in the training set constitutes a new subset X^{infer} of source polymers. X^{infer} represents all undesirable parent polymers for which no valid translations are yet known to us. Using polymers from X^{infer} as input to a polyG2G model (trained on $\{(X_i^{\text{train}}, Y_j^{\text{train}})\}$), new and (we hope) desirable polymers are inferred via graph-to-graph translation.

2.4. polyG2G. In this work, we quantified translations between source and target polymers by assigning each translation a numerical vector in a learned latent space. This latent space was learned by training a Variational Junction Tree Encoder-Decoder (VJTNN) model²⁸ on $\{(X_i^{\text{train}}, Y_j^{\text{train}})\}$. To extend VJTNN to the polymer domain, we represent each polymer as a “pseudopolymer” (Figure 2, step A). A schematic of the polyG2G training process is shown in Figure 2, steps A–J. In step A, an undesirable polymer (from X^{train}) is converted to a corresponding pseudopolymer. In step B, the pseudopolymer is mapped to both a junction tree (JT) and a graph. The JT and graph are mapped to latent space vectors in step C using ML. The vectors are concatenated in step D. In steps E and F, the concatenated latent space vector is mapped to a latent space distribution using VJTNN. In step G, this distribution is sampled $n_{\text{translation}}$ times. In step H, the samples are converted to JTs and then to graphs using ML. The former step specifies the atoms and rings that constitute the translated pseudopolymer, while the latter step specifies how the atoms and rings are chemically bonded to one another. In step J, the graph is converted to a pseudopolymer.

During training, the objective is to translate some X_i^{train} into a sufficiently similar pseudopolymer (Figure 2, step J) that belongs to Y^{train} . The loss is computed by how frequently this objective is met. During inference (i.e., when using a trained polyG2G model to translate some polymers not in the training set), steps A–J in Figure 2 are repeated. However, this time, a polymer from X^{infer} (as opposed to X^{train}) is the input. Further, during inference, we do not compute loss. Instead, we use the new (i.e., translated) pseudopolymer to generate new polymers (step K, Figure 2). In Step K, randomly selected pairs of hydrogen atoms within the pseudopolymer are replaced by a periodic boundary to indicate ends of a polymer repeat unit. For most pseudopolymers, the number of hydrogen pairs that could be replaced is quite large, so we introduce a hyperparameter n_{pair} . n_{pair} is set to 20 in our experiments and denotes the maximum number of candidate polymers that are generated from a particular translated pseudopolymer. A candidate polymer is defined as valuable if it meets the following three criteria:

1. Validity—the candidate polymer has all atoms with a correct valence
2. High Performance—the property values, as predicted by machine-learned Gaussian Process Regression (see Section 2.5) property predictors, of the candidate meet the objective
3. Unrepeated—the candidate is distinct, in PG fingerprint²³ space (see Section 2.2), from all polymers in the parent data set and from all other candidates that have been translated and decoded up until that point

2.5. Machine Learning for Property Prediction. The principal goal of polyG2G is the generation of valuable candidate polymers. To determine value, as defined in Section 2.4, an ability to evaluate the performance of the generated candidates is required. We evaluate the performance via three property predictors used to model three properties (T_g , E_g , and ϕ_e).²³ We learn the property predictors on the aforementioned training data (see Section 2.1) using Gaussian Process Regression (GPR) with the radial basis function kernel, as discussed elsewhere.²⁹

A fourth property predictor, a neural network, was trained to predict the probability that an input polymer would be desirable (as defined in Section 2.3). We used, as training data, the polymers generated for this study by polyG2G.

2.6. Density Functional Theory. Density functional theory (DFT)^{19,20} is a first-principles-based computational method, offering a good balance between accuracy and computational cost. In the area

of polymer science, DFT has been used to develop some relatively large polymer data sets³⁰—two of which were used to train the polymer band gap and electron injection barrier property predictors. Therefore, we used exactly the DFT scheme used to generate that data to test the polymers designed in this work. In particular, our calculations were performed using Vienna Ab initio Simulation Package (VASP) code,³¹ with a plane wave cutoff of 400 eV and the PAW data sets of version 5.2. The van der Waals dispersion interactions, known to be important in stabilizing soft materials dominated by nonbonding interactions like polymers, were estimated with the nonlocal density functional vdW-DF2. The generalized gradient approximation (GGA) functional associated with vdW-DF2, i.e., refitted Perdew-Wang 86 (rPW86), was used for the exchange–correlation (XC) energies.

The initial structures for these calculations were created in a two-step procedure from the SMILES strings of polymers.²³ First, RDKit software was used to convert each polymer SMILES into a three-dimensional molecular configuration. Then, a polymer repeat unit containing several molecules was placed in a simulation box so that the periodicity along the z-axis yields an infinite chain, while a vacuum layer of at least 12 Å along the x and y axes effectively separates the chain from its periodic images.

During the optimization process, atomic and the z-lattice degrees of freedom are relaxed, while the x and y lattice parameters are fixed. The optimization is stopped when atomic forces are smaller than 10^{-2} eV/Å. We include a post-processing step, after structure optimization, to ensure that the atomic connectivity of polymer chains is not changed. Further details of the workflows for computing E_g and ϕ_e can be found in refs 4, 30.

2.7. Design Rules. In this work, we use Shapley Additive Explanations (SHAP),³² a game-theoretic approach to model interpretation, to compare and interpret the predictions of our property predictors. SHAP treats the features input to an ML model as players, and the model itself as a game in which reward is maximized by maximizing the target property. The raw outputs of SHAP are importance values to a given model, known as “Shapley values”, of each feature of each data point. The absolute Shapley values of a single feature, averaged over all data points, yield the mean importance of the feature. SHAP is useful because it can approximate Shapley values for any property predictor, GPR, deep learning, or otherwise. Thus, the importance of features of all models, which precede or succeed ours, can be directly compared with the results presented in this contribution.

3. RESULTS AND DISCUSSION

3.1. Design of Polymeric Dielectric Materials with polyG2G. In this study, we designed 3556 unique, novel polymers that meet our target property objectives. These candidates were generated from 21 trained polyG2G models. Polymers in X^{infer} were translated three times by each model. This process, over all models, resulted in the design of 58 023 pseudopolymers. Of these pseudopolymers, 93% (53,775 in total) were chemically valid. This figure surpasses that of the SD-VAE proposed by Batra et al. by 66%.

Each of the 53 775 valid pseudopolymers was converted to repeat units, yielding 784 631 valid polymers. Each polymer was fingerprinted to enable estimation of band gap, T_g , and electron injection barrier by our property predictors. Using these predictions, we found 3556 (0.45% of total) unique polymers meeting our objectives. In contrast, just 8 out of 13 014 (0.061% of total) polymers in the parent data set met our objective. This proves that polyG2G is not only able to generate a large raw number of high-performing designs but also that these designs “hit” an order of magnitude more frequently than do the space of synthesized polymers. In other words, polyG2G is able to learn targeted design rules specifically aimed at high dielectric breakdown strength

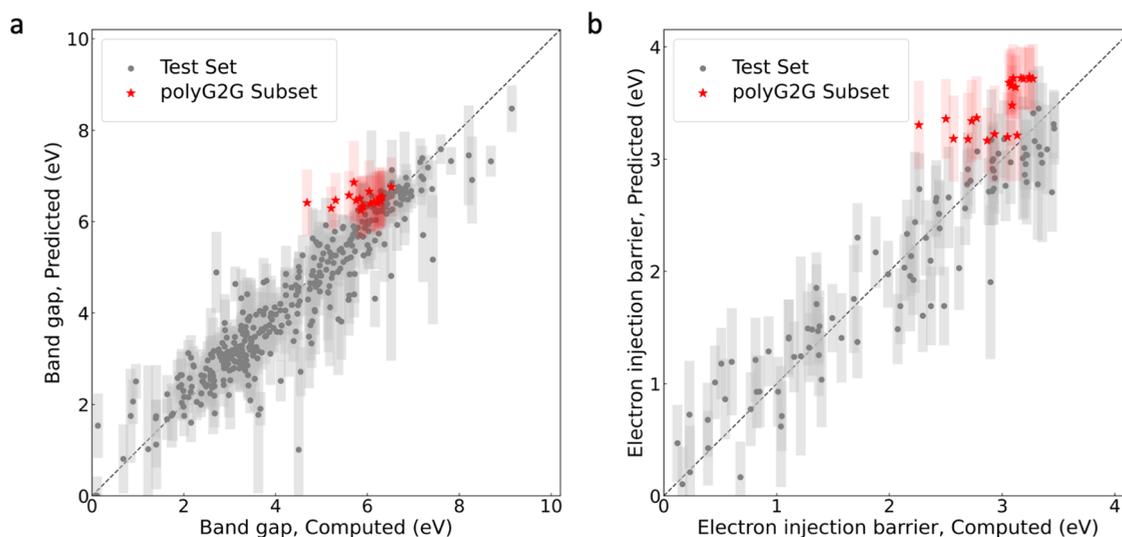


Figure 3. ML predictions of the unseen test set polymers (gray circles) and 20 polyG2G-generated polymers (red stars) for the following properties: (a) band gap and (b) electron injection barrier. Prediction uncertainties are plotted as shaded bars.

polymers. Notably, these figures are competitive with, and in some cases surpass, the GA proposed by Kim et al. (see the SI, Section S1).

3.2. DFT Validation. To down-select from our massive set of candidate materials, we subjected 20 of the 3556 polymers to DFT computations of band gap and electron injection barrier. These 20 cases were specifically chosen from the larger list as they exhibited the highest fitness, \mathcal{F} , defined in eq 2

$$\mathcal{F} = U \times T_g \times E_g \times \phi_e \times R_{mt} \quad (2)$$

where T_g , E_g , and ϕ_e are the ML property predictions of a given polymer, R_{mt} is the ratio between the number of atoms in the main chain per repeat unit and the total number of atoms per repeat unit, and U is the uniqueness of the polymer (see the SI, Section S3 for details pertaining to the calculation of U). Polymers with large U are favored for wider coverage of the chemical space. Polymers with large R_{mt} are also favored. In large R_{mt} polymers, the initial 3D atomic configurations required for DFT simulations are relatively straightforward to construct. To build the configuration, single bonds of a monomer are rotated to get a suitable conformer that can be subjected to periodic boundary conditions. As polymers with fewer and smaller side chains have a lower probability of intrachain conflict, we can autonomously construct configurations using an in-house script without any human intervention.

A comparison of the DFT-computed properties with the ML-predicted values is shown in Figure 3. We find that the agreement between DFT and ML for these new cases is comparable with the test set errors of the property predictors. These results show that, in general, ML predictions of E_g and ϕ_e are comparable with their DFT-computed counterparts. These results also give confidence that our T_g property predictor—trained using the same methods as our E_g and ϕ_e predictors—produces reasonable property estimates for the polymers designed in this work. However, explicit tests of our T_g predictions will require synthesis and testing of our polymer designs.

3.3. High-Value Polymer Designs. Two key observations can be made from Figure 3. First, polyG2G-designed new polymers (red stars in Figure 3) exhibit properties near the

upper reaches of our parent data set (gray circles in Figure 3). In fact, 10 of the 20 candidate polymers exhibit DFT-computed properties, which surpass our objectives and are therefore resistant to large electric fields. These candidates are shown in Table 1. We emphasize that the ten polymers shown in Table 1 are not hand-picked but rather are systematically derived from the multistep protocol illustrated in Figure 1.

Second, through visual inspection, it is clear that the suggested polymers are structurally similar to poly(2-norbornene) and poly(5-butylborn-2-ene) and therefore potentially synthesizable. Interestingly, each suggested polymer repeat unit contains at least one nonaromatic ring. This finding is corroborated by the work of Wu et al. who found similar polymers (derived from norbornenes using fluorination) that exhibit high breakdown resistance even at high temperatures.³³ These findings, though drawn from a limited set of data, suggest that polyG2G is able to efficiently generate feasible, high-performing polymers, meeting complex objectives.

3.4. Mining Design Rules. Now we attempt to elucidate design principles from our models. A neural network classifier was trained on the PG fingerprint of 80% of the polymers generated by polyG2G to predict whether or not a polymer will meet our triproperty objective. The optimized classifier had an accuracy of 98.0% on the remaining 20% of the data not seen during training (model architecture and hyperparameter optimization discussed in the SI). We utilized SHAP to deduce which features the classifier weighs most heavily when making predictions.³² Figure 4 shows the 11 most important features proposed by SHAP. Several rules stand out from this analysis.

First, it can be seen that "len. largest side chain" (the length of the largest, by the number of atoms, side chain) is a highly important feature to the model in classifying a polymer as desirable or undesirable. Indeed, it is already known that long, rotatable, side chains lead to (1) a decrease in the energy barrier between conformations and (2) an increase in steric repulsion between adjacent chain segments—properties that lead to a low T_g . Likewise both "3-vertex carbon: side" (the number of three-vertex nonring carbon atoms in the largest side chain) and "Ratio: main/side" (the number of atoms in the main chain divided by the number of atoms in the side chain) were also

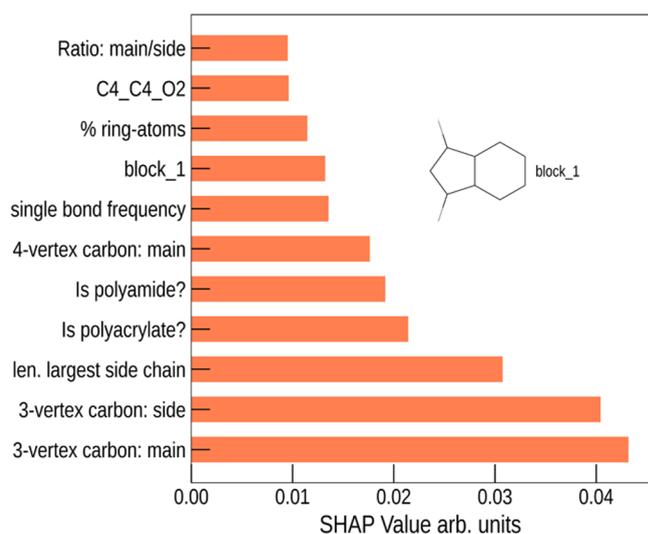


Figure 4. Importance values assigned to chemical features based on SHAP. The names of the features are shown on the y-axis, while the importance is shown on the x-axis. An explanation of each feature is given in the SI, Section S5.

found to be highly important to the classifier. This is not surprising given that these two features are naturally correlated to "len. largest side chain".

Second, the frequency of carbon–carbon single bonds per total number of atoms (denoted by the feature "single bond frequency") is an important quantity to our classifier. Indeed, the absence of π electrons in C–C bonds is known to drive up band gap. Therefore, from a theoretical perspective, having a high value of "single bond frequency" is likely to increase the probability of being a desirable polymer so long as the number of such bonds in alkyl chains is simultaneously minimized (i.e., "% ring atoms", see Figure 4, is large). These two attributes are simultaneously met by several desirable polymers in Table 1 and by "block_1" in Figure 4. If, on the other hand, "single bond frequency" was high, while "% ring atoms" were low, the band gap would tend to increase but at the expense of lower T_g .

Third, the binary feature, "Is polyacrylate?", which denotes whether or not a polymer is a polyacrylate, is also heavily weighted by our classifier. Likewise, to our knowledge, no synthesized polyacrylates exhibit a T_g above 400 K, let alone our objective of 450 K.^{34–43} Fourth, "C4_C4_O2" (i.e., the frequency of the three-atom fragment containing two four-fold coordinated carbon atoms and a two-fold coordinated oxygen atom²²) is an important feature. Indeed, a negative impact of such groups can be attributed to the lower electron injection barrier values at the polymer electrode (assumed to be aluminum) interface. This is due to the generation of large dipoles by interacting oxygen and aluminum species. The above are all examples of scientifically corroborated design rules learned by our property predictors and transferred to polyG2G. These results give confidence that the workflow proposed here can be reliably used to generate real-world materials that meet or surpass the current state of the art.

4. SUMMARY

In this contribution, we introduced polyG2G, a novel, translation-based, pipeline for the generative design of

polymers. We illustrated the potential of polyG2G by applying it to the design of promising dielectric polymers with superior resistance to high electric fields at high temperatures. polyG2G found thousands of promising designs, an exciting feat considering that only eight such polymers were previously known to us. For validation purposes, a small subset of the designed polymers was studied using DFT computations. Ultimately, we recommend ten validated polymer designs that meet our target property objectives and are worthy of further investigation and synthetic validation. We showed that polyG2G surpasses past generative design efforts in terms of chemical validity and the rate of high-value candidate generation.

The chemical and structural space available to polyG2G covers a meaningful portion of the polymer chemical universe. However, since chemical rings are not generated atom by atom, polyG2G is not able to produce rings apart from those present in the training data. Previous methods, such as the SD-VAE, do not have this constraint and can thus, in principle, access any conceivable ring structure. Further, like past generative algorithms, polyG2G is limited to the design of linear homopolymers and alternating co-polymers. The linear constraint, by definition, excludes branched and network polymers as well as other macromolecules such as dendrimers. Meanwhile, the homopolymer constraint excludes all non-alternating co-polymers, polymer blends, and polymeric systems with additives. The technological importance of these systems is reflected by their frequent use in industrial applications. As such, future work enabling design beyond linear homopolymers is critical.⁴⁴

Here, as a step toward addressing the development of novel polymer dielectrics that are resistant to dielectric breakdown, we have reformulated the design objective, focusing on creating new polymers with superior band gaps, charge injection barriers, and glass-transition temperatures. While these three properties are desirable for potential high breakdown polymer candidates, it is important to consider other properties that are known to be correlated to the breakdown strength. These properties include the cohesive energy density, electron mobility (due to hopping and band transport), yield strength, fractional free volume, polarization, and loss factor of a given polymer.⁴⁵ Given suitable data sets for these additional properties, the polyG2G generative approach described in this work can easily be extended to search for better, breakdown-resistant, polymers as well as polymers for other applications. Nonetheless, the ten selected materials identified in this work serve as starting points for closer examination. We present these candidates to the community and leave a detailed study of their synthetic feasibility to future works.

■ ASSOCIATED CONTENT

SI Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.chemmater.1c02061>.

Predictive accuracy comparison between polyG2G and GA; computing polymer similarity; method for computing polymer uniqueness; optimization strategy and values for neural network hyperparameters; explanations of feature labels (PDF)

AUTHOR INFORMATION

Corresponding Author

Rampi Ramprasad – School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States; orcid.org/0000-0003-4630-1565; Email: rampi.ramprasad@mse.gatech.edu

Authors

Rishi Gurnani – School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States; orcid.org/0000-0002-2744-2234

Deepak Kamal – School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States; orcid.org/0000-0003-1943-7774

Huan Tran – School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States; orcid.org/0000-0002-8093-9426

Harikrishna Sahu – School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States; orcid.org/0000-0001-5458-9488

Kenny Scharm – College of Computing, Georgia Institute of Technology, Atlanta, Georgia 30332, United States

Usman Ashraf – School of Chemical and Biomolecular Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States

Complete contact information is available at:

<https://pubs.acs.org/10.1021/acs.chemmater.1c02061>

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

R.G. would like to acknowledge Aubrey Toland and Mark Weber for insightful discussions. This work was supported as part of the Center for Understanding and Control of Acid Gas-Induced Evolution of Materials for Energy (UNCAGE ME), an Energy Frontier Research Center funded by the U.S. Department of Energy (DOE), Office of Science, Basic Energy Sciences (BES), under Award # DE-SC0012577, and by the Office of Naval Research through a Multi-University Research Initiative (MURI) grant (Grant No. N00014-17-1-2656).

REFERENCES

- Huan, T. D.; Boggs, S.; Teyssedre, G.; Laurent, C.; Cakmak, M.; Kumar, S.; Ramprasad, R. Advanced polymeric dielectrics for high energy density applications. *Prog. Mater. Sci.* **2016**, *83*, 236–269.
- Sharma, V.; Wang, C.; Lorenzini, R. G.; Ma, R.; Zhu, Q.; Sinkovits, D. W.; Pilania, G.; Oganov, A. R.; Kumar, S.; Sotzing, G. A.; et al. Rational design of all organic polymer dielectrics. *Nat. Commun.* **2014**, *5*, No. 4845.
- Mannodi-Kanakkithodi, A.; Treich, G. M.; Huan, T. D.; Ma, R.; Tefferi, M.; Cao, Y.; Sotzing, G. A.; Ramprasad, R. Rational Co-Design of Polymer Dielectrics for Energy Storage. *Adv. Mater.* **2016**, *28*, 6277–6291.
- Kamal, D.; Wang, Y.; Tran, H. D.; Chen, L.; Li, Z.; Wu, C.; Nasreen, S.; Cao, Y.; Ramprasad, R. Computable Bulk and Interfacial Electronic Structure Features as Proxies for Dielectric Breakdown of Polymers. *ACS Appl. Mater. Interfaces* **2020**, *12*, 37182–37187.
- Batra, R.; Song, L.; Ramprasad, R. Emerging materials intelligence ecosystems propelled by machine learning. *Nat. Rev. Mater.* **2021**, 655–678.
- Gómez-Bombarelli, R.; et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **2016**, *15*, 1120–1127.
- Barnett, J. W.; Bilchak, C. R.; Wang, Y.; Benicewicz, B. C.; Murdock, L. A.; Bereau, T.; Kumar, S. K. Designing exceptional gas-separation polymer membranes using machine learning. *Sci. Adv.* **2020**, *6*, No. eaaz4301.
- Mannodi-Kanakkithodi, A.; Chandrasekaran, A.; Kim, C.; Huan, T. D.; Pilania, G.; Botu, V.; Ramprasad, R. Scoping the polymer genome: A roadmap for rational polymer dielectrics design and beyond. *Mater. Today* **2018**, *21*, 785–796.
- Kim, C.; Batra, R.; Chen, L.; Tran, H.; Ramprasad, R. Polymer design using genetic algorithm and machine learning. *Comput. Mater. Sci.* **2021**, *186*, No. 110067.
- Batra, R.; Dai, H.; Huan, T. D.; Chen, L.; Kim, C.; Gutekunst, W. R.; Song, L.; Ramprasad, R. Polymers for Extreme Conditions Designed Using Syntax-Directed Variational Autoencoders. *Chem. Mater.* **2020**, *32*, 10489–10500.
- Jin, W.; Yang, K.; Barzilay, R.; Jaakkola, T. In *Learning Multimodal Graph-to-Graph Translation for Molecular Optimization*, 7th International Conference on Learning Representations; ICLR, 2019; pp 1–13.
- Yao, Z.; Sánchez-Lengeling, B.; Bobbitt, N. S.; Bucior, B. J.; Kumar, S. G. H.; Collins, S. P.; Burns, T.; Woo, T. K.; Farha, O. K.; Snurr, R. Q.; Aspuru-Guzik, A. Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nat. Mach. Intell.* **2021**, *3*, 76–86.
- Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; Rarey, M. On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem* **2008**, *3*, 1503–1507.
- Dai, H.; Tian, Y.; Dai, B.; Skiena, S.; Song, L. In *Syntax-Directed Variational Autoencoder for Structured Data*, 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings; ICLR, 2018.
- Polymer Handbook*, 4th ed.; Brandup, J.; Immergut, E. H.; Grulke, E. A., Eds.; John Wiley & Sons: New York, 1999.
- Handbook of Polymers*, 2nd ed.; Wypych, G., Ed.; ChemTec Publishing: Toronto, Canada, 2016.
- Van Krevelen, D. W.; Te Nijenhuis, K. *Properties of Polymers: Their Correlation with Chemical Structure; Their Numerical Estimation and Prediction from Additive Group Contributions*; Elsevier, 2009.
- Polymer Data Handbook*, 2nd ed.; Mark, J. E., Ed.; Oxford University Press: New York, 2009.
- Hohenberg, P.; Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **1964**, *136*, B864–B871.
- Kohn, W.; Sham, L. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **1965**, *140*, A1133–A1138.
- Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
- Huan, T. D.; Mannodi-Kanakkithodi, A.; Ramprasad, R. Accelerated Materials Property Predictions and Design Using Motif-Based Fingerprints. *Phys. Rev. B* **2015**, *92*, No. 014106.
- Doan Tran, H.; Kim, C.; Chen, L.; Chandrasekaran, A.; Batra, R.; Venkatram, S.; Kamal, D.; Lightstone, J. P.; Gurnani, R.; Shetty, P.; Ramprasad, M.; Laws, J.; Shelton, M.; Ramprasad, R. Machine-learning predictions of polymer properties with Polymer Genome. *J. Appl. Phys.* **2020**, *128*, No. 171104.
- Gurnani, R.; Yu, Z.; Kim, C.; Sholl, D. S.; Ramprasad, R. Interpretable Machine Learning-Based Predictions of Methane Uptake Isotherms in Metal-Organic Frameworks. *Chem. Mater.* **2021**, *33*, 3543–3552.
- Ma, R.; Luo, T. PIIM: A benchmark database for polymer informatics. *J. Chem. Inf. Model.* **2020**, *60*, 4684–4690.
- Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **1988**, *28*, 31–36.
- Doan Tran, H.; Kim, C.; Chen, L.; Chandrasekaran, A.; Batra, R.; Venkatram, S.; Kamal, D.; Lightstone, J. P.; Gurnani, R.; Shetty, P.; Ramprasad, M.; Laws, J.; Shelton, M.; Ramprasad, R. Machine-

learning predictions of polymer properties with Polymer Genome. *J. Appl. Phys.* **2020**, *128*, No. 171104.

(28) Jin, W.; Barzilay, R.; Jaakkola, T. In *Junction Tree Variational Autoencoder for Molecular Graph Generation*, 35th International Conference on Machine Learning; ICML, 2018; Vol. 5, pp 3632–3648.

(29) Lightstone, J. P.; Chen, L.; Kim, C.; Batra, R.; Ramprasad, R. Refractive index prediction models for polymers using machine learning. *J. Appl. Phys.* **2020**, *127*, No. 215105.

(30) Huan, T. D.; Mannodi-Kanakkithodi, A.; Kim, C.; Sharma, V.; Pilania, G.; Ramprasad, R. A polymer dataset for accelerated property prediction and design. *Sci. Data* **2016**, *3*, No. 160012.

(31) Kresse, G.; Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **1996**, *6*, 15–50.

(32) Lundberg, S. M.; Lee, S. I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, 4766–4775.

(33) Wu, C.; Deshmukh, A. A.; Li, Z.; Chen, L.; Alamri, A.; Wang, Y.; Ramprasad, R.; Sotzing, G. A.; Cao, Y. Flexible Temperature-Invariant Polymer Dielectrics with Large Bandgap. *Adv. Mater.* **2020**, *32*, No. 2000499.

(34) Yuki, H.; Hatada, K.; Niinomi, T.; Hashimoto, M.; Ohshima, J. Stereospecific Polymerization of Methyl α -Phenylacrylate. *Polym. J.* **1971**, *2*, 629–639.

(35) Mashita, K.; Hirooka, M. Alternating copolymers of isobutylene and acrylic ester by complexed copolymerization. *Polymer* **1995**, *36*, 2983–2988.

(36) Bertolucci, P. R. H.; Harmon, J. P. *Photonic and Optoelectronic Polymers*; ACS Symposium Series; American Chemical Society, 1997; Vol. 672, pp 7–79.

(37) Jakubowski, W.; Juhari, A.; Best, A.; Koynov, K.; Pakula, T.; Matyjaszewski, K. Comparison of thermomechanical properties of statistical, gradient and block copolymers of isobornyl acrylate and n-butyl acrylate with various acrylate homopolymers. *Polymer* **2008**, *49*, 1567–1578.

(38) Mendil, H.; Baroni, P.; Noirez, L. Solid-like rheological response of non-entangled polymers in the molten state. *Eur. Phys. J. E* **2006**, *19*, 77–85.

(39) Shetter, J. A. Effect of stereoregularity on the glass temperatures of a series of polyacrylates and polymethacrylates. *J. Polym. Sci., Part B: Polym. Lett.* **1963**, *1*, 209–213.

(40) Liu, W.; Nakano, T.; Okamoto, Y. Polymerization of t-butyl acrylate using organoaluminum complexes and correlation between main-chain tacticity and glass transition temperature of the obtained polymers. *Polymer* **2000**, *41*, 4467–4472.

(41) Cypcar, C. C.; Camelio, P.; Lazzeri, V.; Mathias, L. J.; Waegell, B. Prediction of the Glass Transition Temperature of Multicyclic and Bulky Substituted Acrylate and Methacrylate Polymers Using the Energy, Volume, Mass (EVM) QSPR Model. *Macromolecules* **1996**, *29*, 8954–8959.

(42) Mark, H. F.; Kroschwitz, J. I. *Encyclopedia of Polymer Science and Engineering*, 2nd ed.; Wiley: New York, 1985.

(43) Brandrup, J.; Immergut, E. H. *Polymer Handbook*, 3rd ed.; Wiley: New York, 1989.

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

(45) Fothergill, J. C.; Eccles, A.; Houlgreave, J. A.; Dissado, L. A. Water tree inception and its dependence upon electric field, voltage and frequency. *IEE Proc. A: Sci., Meas. Technol.* **1993**, *140*, 397–403.