



Design of polymers for energy storage capacitors using machine learning and evolutionary algorithms

Joseph Kern¹ , Lihua Chen¹ , Chiho Kim¹ , and Rampi Ramprasad^{1,*}

¹ School of Materials Science and Engineering, Georgia Institute of Technology, 771 Ferst Drive NW, Atlanta, GA 30332, USA

Received: 2 August 2021

Accepted: 8 September 2021

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

ABSTRACT

To meet the demands of emerging electrification technologies, polymers that are capable of withstanding high electric fields at high temperatures are needed. Given the staggeringly large search space of polymers, traditional, intuition- and experience-based Edisonian approaches are too slow at discovering new polymers that can meet these demands. In this work, a genetic algorithm was combined with five machine learning-based property predictors to design over 50,000 hypothetical polymers that achieve target properties. Additionally, a polymer synthesizability-based criterion was used to narrow these polymers down to 23 candidates likely to be synthesizable and 3665 that may be synthesizable. A version of the genetic algorithm code is also made available for public use on GitHub.

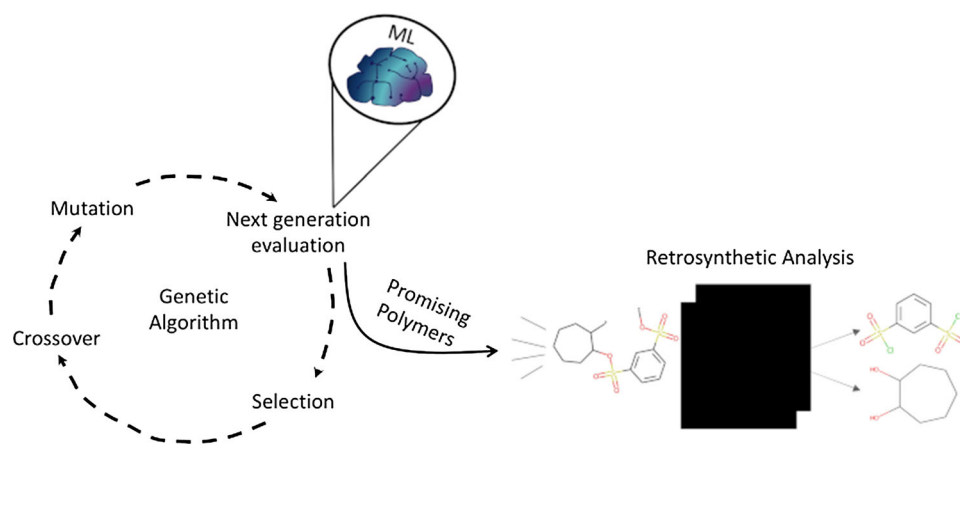
Handling Editor: Maude Jimenez.

Address correspondence to E-mail: rampi.ramprasad@mse.gatech.edu

<https://doi.org/10.1007/s10853-021-06520-x>

Published online: 18 September 2021

GRAPHICAL ABSTRACT



Introduction

Polymers such as polypropylene have, historically, been used as the dielectric materials of choice in high energy density capacitors because of their graceful failure due to self-clearing and low production costs [1–3]. As the demand for electrification under extreme conditions becomes more prevalent, these capacitors may experience high temperatures ranging from 150 °C in electric vehicles to 250 °C in aircraft [4, 5]. Presently available polymer dielectrics are not suitable for these applications due to low thermal and electrical stability at high temperatures [6]. Consequently, there have been efforts to develop new polymer dielectrics that can withstand these conditions [6–8].

Like with many applications, polymer dielectrics for high-temperature capacitors need to meet multiple property criteria, including a high energy density to reduce the size of capacitors, high thermal stability to survive high operating temperatures, and high breakdown field strength to withstand high electric fields. It is time- and cost-intensive to search for polymers satisfying these properties using traditional, Edisonian experimental methods. However, their discovery can be aided and accelerated by computation- and data-driven design algorithms [9–12]. Indeed, past work has revealed the power of such informatics-guided pursuits for the design of

high energy density capacitor dielectrics [13–16], especially when combined synergistically with experimental validation [17–19].

In this paper, we utilize a genetic algorithm (GA), along with machine learning (ML) property prediction models for the associated target properties, to design hypothetical polymer dielectrics for high temperatures and electric field applications. In contrast to a previous study that also utilized GA to design polymers for this application [14], the present work involves several significant advancements as detailed below.

- (1) First, a set of five property-based target screening criteria was adopted in the present study to guarantee superior performance as opposed to the two criteria used in the previous study. These five property criteria include high band-gap (E_g), high charge injection barrier (Φ_e), low cohesive energy density (e_{coh}), high dielectric constant (ϵ), and high glass transition temperature (T_g). The first three criteria ensure high electrical breakdown strength [20–24], the fourth criterion allows for high energy density [6, 20, 25, 26], and the last criterion is necessary for good thermal stability. The present implementation, through a “clamping” fitness function, can handle a varying number of target criteria in a robust manner.

- (2) Second, several algorithmic improvements were made to our GA to explore the search space more exhaustively than before, thus increasing the diversity of the designs. This has led to an increase in the number of polymers achieving target properties and the discovery of over 50,000 promising, hypothetical polymers for high-temperature dielectric applications.
- (3) Third, a polymer synthesizability-based criterion was also included to ensure that the polymers designed using our GA have a high probability of synthetic success, drawing on a recent machine-guided synthesis planning development [27].
- (4) Finally, flexibility has been imparted to the GA design framework so that a chemist's preference (or lack thereof) for certain blocks or functional groups may be handled with ease.

Below, we describe the enhancements implemented in our GA workflow and present the most promising polymer designs that have emerged from this effort. The present GA implementation may be utilized in a general-purpose manner for designing polymers for any application, so long as the target design criteria may be set up and models (ML or otherwise) are available for predicting the relevant properties.

Methods

Target screening criteria and machine learning models

As mentioned above, five screening criteria were set up as the goals for our GA to achieve. These criteria are listed in Table 1. As stated previously, for high-temperature polymer dielectric applications, polymers should have large ϵ and T_g , as these, respectively, correlate with a larger energy density and

higher operating temperature. They should also have a high dielectric breakdown strength, which is the maximum electric field an insulator can operate at before becoming a conductor due to the creation of free electrons. Since the breakdown strength is related to the number of free electrons in the material, it is correlated with properties related to electron concentration, such as E_g , Φ_e and e_{coh} . Polymers should have a high E_g , as the larger the E_g the more energy is required to transfer electrons from the valence band to the conduction band. Φ_e should also be large, as metal electrodes can inject free electrons into the material if the barrier is too low. For e_{coh} , there is an inverse relationship with E_g . Higher e_{coh} is attributed to strong inter-chain interactions (hydrogen-bonding and van der Waals forces), which impact anti-bonding energy level overlaps and thus decreases E_g . Thus, e_{coh} should be small.

For each of these five target properties, machine learning models based on Gaussian process regression (GPR) are available for instantaneous property predictions. Details pertaining to these models, including training data used, algorithmic details and accuracy have been discussed elsewhere [19, 28, 29]. We have provided a short reference to these values in Table S1.

Genetic algorithm

The GA is a simulated evolution-based search algorithm that has been used since the 1990s to tackle the design of molecules and polymers [13–17]. It has also been used to design or improve other materials and applications, such as optical glasses, induction motors, Li–S cathodes, and image processing software [30–33]. Each version of the GA has a different design, but the fundamental concepts underlying the algorithm are the same: first, crossover and mutation operations are performed on a parent dataset to create a new child dataset, then each child is scored according to some fitness function, and finally, the

Table 1 Design goals for high temperature polymer dielectrics, proxy properties correlated with the goals, and their targets during the genetic algorithm design process

Property	Target	Goal
Bandgap (E_g)	> 5 eV	Increase breakdown strength
Electron injection barrier (Φ_e)	> 3 eV	Increase breakdown strength
Cohesive energy density (e_{coh})	< 80 cal cm ⁻³	Increase breakdown strength
Dielectric constant @ 100 Hz (ϵ)	> 4	Increase energy density
Glass transition temperature (T_g)	> 500 K	Increase thermal stability

parents of the next generation are chosen based on their fitness scores [34]. This cycle continues until a set of prescribed target criteria is achieved.

In our previous work, we explained how we followed this paradigm with small modifications to make it suitable for the polymer domain. First, 3045 building blocks were extracted from ~ 12,000 reference polymers via the “breaking of retrosynthetically interesting chemical substructures” (BRICS) scheme [35]. These building blocks were used as the chromosomes of our polymers. Then, we followed conventional GA protocols and utilized ML-based surrogate models for the rapid prediction of the relevant properties (in the screening criteria) of new polymers, scoring their fitness based on their predicted property values [14]. In the present work, we have made further modifications to achieve the following goals: expand the number of property criteria, increase chemical diversity, improve synthetic accessibility, and facilitate user-friendly chemical block selection. These modifications are summarized in Table 2.

To test if the modifications achieved these goals, virtual experiments were performed. Each experiment consisted of five runs of the GA. Each run continued for 100 generations and was seeded with one of the following five random number generator seeds: 4, 663, 703, 873 or 974. The first generation of polymers was randomly generated for the first experiment and then manually selected for subsequent ones, so each experiment had the same first generation of polymers based on the seed (the first generation was different between seeds). During the segmentation phase, polymers were segmented at the

center with a standard deviation of 0.2 blocks to the left or right. This allowed predicted polymers to grow or shrink. During crossover, all combinations of 10 parent polymer segments were generated (~ 180 children were created). During the mutation phase, unless otherwise specified, $0.2 \times$ the number of blocks in a child polymer (with a standard deviation of 0.25 blocks) would mutate. There was also a 5% chance of an additional random block being appended to the end of the polymer chain. In the next section, we will describe each of the enhancements listed in Table 2 in greater detail.

Enhancements to the genetic algorithm

Expand property criteria (“clamp” fitness function)

Setting up GA to achieve a single target property is trivial. For two target properties, using a weighted summation of the properties after min-max normalization (i.e., with property values scaled to be in the 0–1 range) works effectively. However, if more than two target properties are involved and if some pair of properties are inversely related to each other, a weighted summation is not effective as the exact weighting required to improve all property values above their targets can be challenging to find. Since we have five target properties, and since our preliminary analysis indicated an inverse relationship between some pairs of properties, e.g., T_g and E_g , the GA struggled to find the ideal chemical space with a weighted summation fitness function alone.

A clamping fitness function helps alleviate this problem. It is defined as:

Table 2 Modifications added to the genetic algorithm to improve polymer design

Goal	Modification	Summary
Expand property criteria	Clamping fitness function	During fitness scoring, all values exceeding target property goal are clamped at goal
Increase chemical diversity	Duplication check	Polymers that have already been discovered are deleted or mutated prior to parent selection
Improve synthetic accessibility	Chemical rules	Polymers are screened for chemical realism, only a certain minimum or maximum number of blocks are allowed, and/or blocks are chosen according to the frequency at which they exist in a reference set of existing polymers
	Size restrictions	
	Frequency-based selection	
User-friendly chemical block selection	Functional group screening	Functional groups are automatically screened from the block list according to chemist specifications before running the GA

Each modification was added to achieve one of the four goals

$$y_{ij} = \begin{cases} -\max(x_{ij}, \text{target}_j), & \text{goal}_j = \text{less than target}_j \\ \min(x_{ij}, \text{target}_j), & \text{goal}_j = \text{greater than target}_j \end{cases} \quad (1)$$

with x_{ij} representing the j th property value of polymer i , target_j representing property j 's target value, goal_j representing if the goal is to have polymers achieve less than or greater than values of the target, and y_{ij} representing the clamped value of property j for polymer i used for min-max normalization and fitness scoring. For instance, if the target value for E_g is a value greater than 5 eV, and if predicted values for three child polymers are 4, 5.5, and 6 eV, then the fitness function values for E_g would first be clamped at 5 eV (i.e., 4 eV will remain 4 eV, but 5.5 eV and 6 eV will be clamped at 5 eV) then scaled (to 0, 1, 1). This allowed the GA to weight polymers achieving all target properties higher than those that only excel at specific properties and avoided the need to fine-tune weights in a weighted linear summation. In the results section, we will show that with this modification, the GA can quickly find polymers achieving five different property targets.

Increase chemical diversity

A common problem with GAs is that they quickly find a local optimum and then stay within that area. This can lead to duplicate polymer predictions and a lack of chemical diversity in the design outcomes. To explore a larger, more diverse chemical space, duplication checking was added.

Duplication checking compares the string equivalency of the canonicalized SMILES strings of all child polymers with the canonicalized SMILES of all previously generated polymers. A canonicalized SMILES string is one that has a standardized format [36]. For instance, *[C]OCCCCO[*] and *[C]CCO[*] are both poly(ethylene glycol), but comparing their string equivalency returns false. Canonicalization transforms *[C]OCCCCO[*] into *[C]CCO[*], and comparing string equivalency will now return true. The general procedure for canonicalization of polymers involves first ensuring the joint atoms are equivalent (e.g., *[C]OCCOC[*] \rightarrow *[C]CCOCCO[*]) and then removing repetitions (e.g., *[C]CCOCCO[*] \rightarrow *[C]CCO[*]). If a child polymer occurred in a previous generation, it was either deleted or one fragment of the repeat polymer was mutated until a previously unseen

polymer was generated. This step increased the number of unique polymers the GA predicted but did not affect diversity as hoped.

Improve synthetic feasibility

Regardless of diversity, many of the predicted polymers tend to be unrealistic from a practical (or synthetic feasibility) point of view. They may have bonds that are unlikely to occur or have many chemical fragments, increasing monomer molecular weight and synthetic complexity. To address this, specific chemical screening (CS) rules were integrated into the GA to screen child polymers. These rules (shown in Table S2) were a list of desirable and undesirable bonds. If no undesirable bonds were present and at least one desirable one was present, polymers were kept, else, they were discarded. The size (number of blocks) in a hypothetical polymer was also restricted. Additionally, it was previously hypothesized that biasing chosen blocks based on the frequency they appeared in the reference set of polymers would improve chemical realism (frequency-based selection (FBS)) [14]. To assess this, the ability to select blocks with or without frequency weighting was added.

To test for synthetic complexity, a polymer retrosynthesis algorithm [27] was used to generate a synthesizability score for the predicted polymers. This algorithm compares the Tanimoto similarity score of the predicted polymer with a database of known polymers. The Tanimoto similarity score is defined as:

$$\text{Tanimoto}(x, y) = \frac{\sum x_i y_i}{\sum x_i^2 + \sum y_i^2 - \sum x_i y_i} \quad (2)$$

where x and y are either both polymers or both monomers and i is their i th fingerprint. A Tanimoto score of 0 means x and y are completely different while a value of 1 indicates they are the same.

The synthesis pathway, whether it be condensation, addition, or ring-opening, of the closest scoring known polymer was taken, and the reverse reaction was performed on the predicted polymer. The reactants from this reverse reaction had their Tanimoto score compared against a database of known polymer reactants, with the highest scoring reactants being saved along with their score. The similarity score between the predicted polymer and known polymer and the scores between the predicted reactants and known reactants were then combined to create one

synthesizability score. The exact calculation for this score can be found in [27], but a medium score is one classified as $0.5 < \text{score} \leq 0.7$, and a high is a $0.7 < \text{score} \leq 1$.

User-friendly chemical block selection

Even if a GA-designed polymer was synthetically accessible according to our algorithm, it may still not be favored by a synthetic chemist. Synthetic-friendly polymers that meet target property criteria may include functional groups that are undesirable (due to functional, cost or toxicity reasons). For high-temperature polymer dielectrics, for instance, $-\text{OH}$ groups are beneficial towards achieving the five property goals but can participate in H-bonding, restricting the ability for dipoles to orient in an electric field. Thus, to further target polymer design to chemist intuition, the ability to automatically remove certain functional groups from the block list was added to our framework.

Results and discussion

Expand property criteria

Clamping improved the GA's ability to design polymers achieving all target properties. A comparison of the number of polymers achieving all target properties for experiments with just an evenly weighted linear combination (LC) fitness function and ones with LC + clamping is shown in Fig. 1. LC experiments, on average, predict 63 ± 120 polymers that achieve all target properties, while clamping + LC runs predict 1380 ± 528 polymers.

Average child values for each property in a GA run with LC + clamping fitness function improved until all were close to achieving their targets, as shown in Fig. 2. Φ_e started with most polymers achieving the target property of 3 eV. Within one to three generations, average values for E_g and e_{coh} achieved their targets of 5 eV and 80 cal cm^{-3} . Over 10 generations, average T_g values gradually increased and then plateaued until a sharp step-up occurred between generations 18 and 20. This step also resulted in a reduction in Φ_e . This long period of limited change followed by a sudden evolutionary step is known as a punctuated equilibrium and commonly occurs within GAs and nature [37]. ε experienced a similar jump

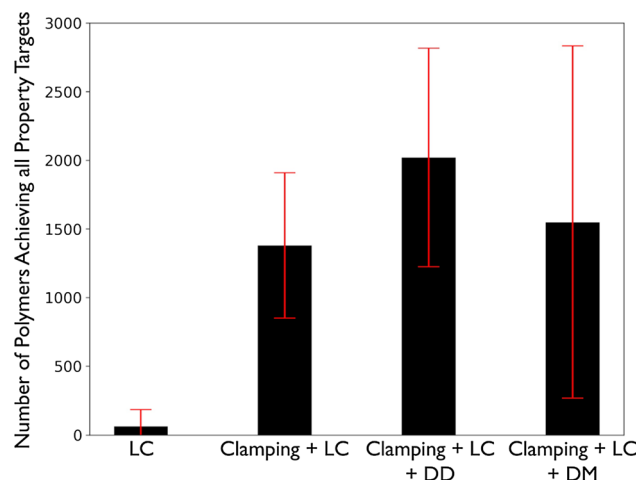


Figure 1 Average number of polymers that achieve all target properties ($T_g > 500 \text{ K}$, $\varepsilon > 4$, $E_g > 5 \text{ eV}$, $e_{\text{coh}} < 80 \text{ cal cm}^{-3}$, and $\Phi_e > 3 \text{ eV}$) per GA run. Average values and standard deviations per run are shown for runs using a linear combination (LC) fitness function, runs using a clamping + LC fitness function, runs using a clamping + LC fitness function, and either duplicate delete (DD) or duplicate mutate (DM). Five runs of 100 generations were done for each experiment.

between generations 40 and 50. Overall, 11,113 unique polymers were predicted with 1217 achieving all property targets for this run. Similar behavior was observed in the other four runs.

Clamping automatically tunes the GA chemical search space during a run. After a property target is achieved, the GA focuses on improving properties that have not yet been achieved, while maintaining the achieved property above or below its target. For instance, when, in Fig. 2, T_g increases above its goal while lowering Φ_e , because the Φ_e average is still larger than 3 eV, the GA permits this. Without clamping, this jump in T_g would not occur if the reduction in Φ_e was too severe.

Clamping target values must be realistic, however. For instance, if the T_g target was 1000 K, clamping would never take effect since no polymer would have that predicted T_g . If none of the goals were achievable, then the fitness function would behave like a typical LC run.

A non-evenly weighted LC scheme may also be suitable for finding polymers that achieve all target properties; however, nine runs were attempted with different weights being used for each property and none successfully generated large numbers of polymers achieving all targets.

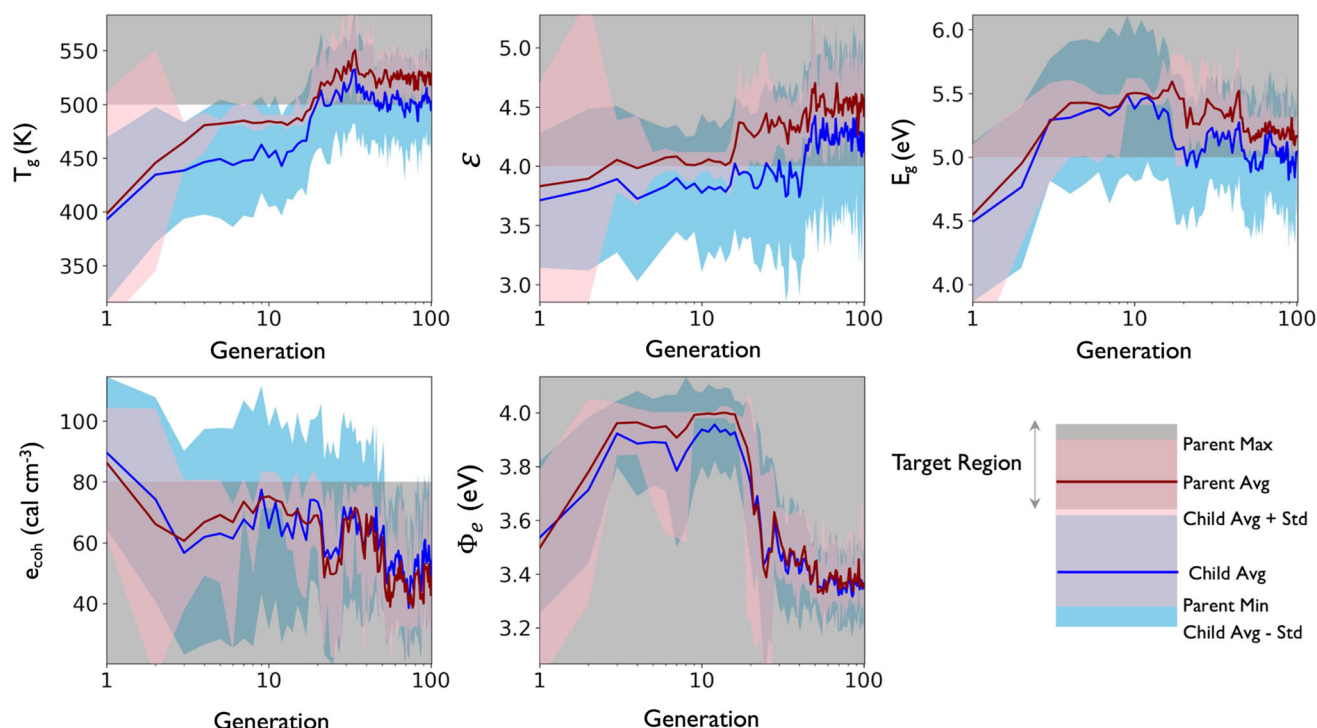


Figure 2 Average property values for parent and child polymers in each generation of the genetic algorithm run with a linear combination + clamping fitness function. The target design region for each property is shaded gray. The region between the parent

min and max for each generation is shaded red, and the region shaded between the child average \pm standard deviation is shaded blue.

While clamping helped the GA predict more polymers that achieved all property targets, many polymers were repeated throughout a GA run, and the polymers were not chemically diverse. For instance, in one run of a clamping + LC experiment, 3859 hypothetical polymers achieved all target properties, but only 1486 (39%) were unique. Of these unique polymers, the average Tanimoto similarity score, seen in Fig. S1, was ~ 0.96 .

Increase chemical diversity

By adding duplication checking, whether with mutating the polymer further (duplicate mutation (DM)) or removing it entirely (duplicate deletion (DD)), the number of polymers achieving target properties went up slightly on average, as seen in Fig. 1, with clamping + LC + DD generating 2019 ± 796 polymers and clamping + LC + DM generating 1549 ± 1283 . A larger spread occurs because one run can get unlucky and find zero polymers, whereas when another run finds an ideal space, it will explore it more thoroughly. Unfortunately, duplication

checking did not influence similarity scores, as seen in Fig. S1.

Duplication checking allows a larger chemical space to be explored as repeat polymer configurations are further mutated; however, it does not increase diversity because the microevolution caused by mutations does not dramatically alter polymer compositions [13]. Other researchers have successfully increased the diversity of runs by creating “niches” of polymers that may fail some targets but that excel at others. These niche polymers get passed along to the next generation along with the polymers best suited for the overall fitness function [15]. A similar methodology will be implemented in the future to further improve chemical diversity within runs.

Still, because each initial, randomly generated first generation is composed of different polymers, the evolutionary process finds different optimal chemical spaces between runs. Thus, although there is a lack of diversity within runs, between runs of an experiment, the diversity is greater, as shown in Fig. S2. By varying the first generation of the GA and

performing multiple runs, this diversity issue was somewhat mitigated.

Improve synthetic feasibility

Frequency-based selection

In addition to increasing the diversity of predicted polymers, the GA should generate polymers that are synthesizable. Despite the increased number of polymers achieving target properties with clamping, only 23 unique polymers out of 6898 polymers generated in all five runs of the LC + Clamping experiment had a medium or high retro-synthesis score. Adding DM increased this to 257 out of 7744. These 257 polymers had a high average Tanimoto similarity score of 0.88. This is because only two of five runs found a feasible polymer, and similarity scores are high within the runs.

Frequency-based selection (FBS), when combined with LC + clamping + DM, increased the number of unique, synthetically feasible polymers predicted from 257 to 358; however, the average similarity score increased slightly to 0.89. The number of runs that found at least one synthetically feasible polymer increased from two to four, but the majority came from only two runs, which is why there is a high average similarity.

While there is a slight increase in feasible polymers, this could be due to serendipity. Evidence to support this is shown in Figs. S3 and S4. As seen in Fig. S4, when FBS is added to an experiment with chemical screening (CS), seed 4 (which always has the same first generation) goes from finding zero medium retrosynthesis scoring polymers to 298. Yet, when FBS is added to a run without CS, it reduces the number of medium or high scoring polymers found from over 100 to close to zero. A comparison of the CS run with and without FBS is illustrated in a uniform manifold approximation and projection (UMAP) in Fig. S3. UMAP is a method to map high-dimensional fingerprints down to two dimensions, similar to principle component analysis, and can offer some insight into the chemical spaces being explored [38]. FBS immediately guides the GA search space to an area distinct from the run without FBS. Thus, it could just be a matter of luck that more are synthesizable. Even if a true improvement occurred, the space explored is biased towards chemical substructures prominent in already existing polymers. This will hinder the

exploration of chemical spaces that have yet to be fully explored by chemists, slowing the discovery of novel spaces. Thus, FBS does not seem to be effective at exploring further spaces and may be detrimental to discovering novel ones.

Chemical screening

CS also did not consistently affect realism. As seen in Fig. S4, for some runs it increased the number of realistic polymers found, but for others, it decreased it. CS after FBS improved the GA's ability to find feasible polymers in three runs but hindered it in one. With only CS, one run was improved, but two were hindered.

The reason CS is inconsistent is the same as for FBS; it guides the GA to different chemical spaces. For instance, CS could make the GA throw out predicted polymers that have mutations that are beneficial towards achieving target properties, but that have a bond unrealistic for polymers. This prevents that beneficial mutation from occurring in the next generation, and the GA explores a different chemical space. This new space may or may not have polymers that can achieve target properties while being synthesizable.

Restricting size

Restricting size had a positive effect on finding retrosynthetically feasible polymers when the mutation rate was not optimized. Figure 3 shows the results of modifying mutation rate for an LC + clamping + DM run. In Fig. 3a, no size restrictions were used. When the mutation rate was too low ($0 \pm 25\%$), the GA could not mutate to find an ideal chemical space that improved target properties. Thus, it compensated by increasing the size of the polymers. When the mutation rate was too high ($75 \pm 25\%$), the GA could not determine which mutations improved the target properties and shrank the polymer sizes. Once all polymers shrank to a length of two, the crossover could no longer alter polymer sizes. The random appending of a block to the end of a polymer did increase the size once more, but polymers shrank back down. At the ideal mutation rate ($20 \pm 25\%$) the polymer size was stable, and the GA could find an ideal chemical space that achieved all target properties.

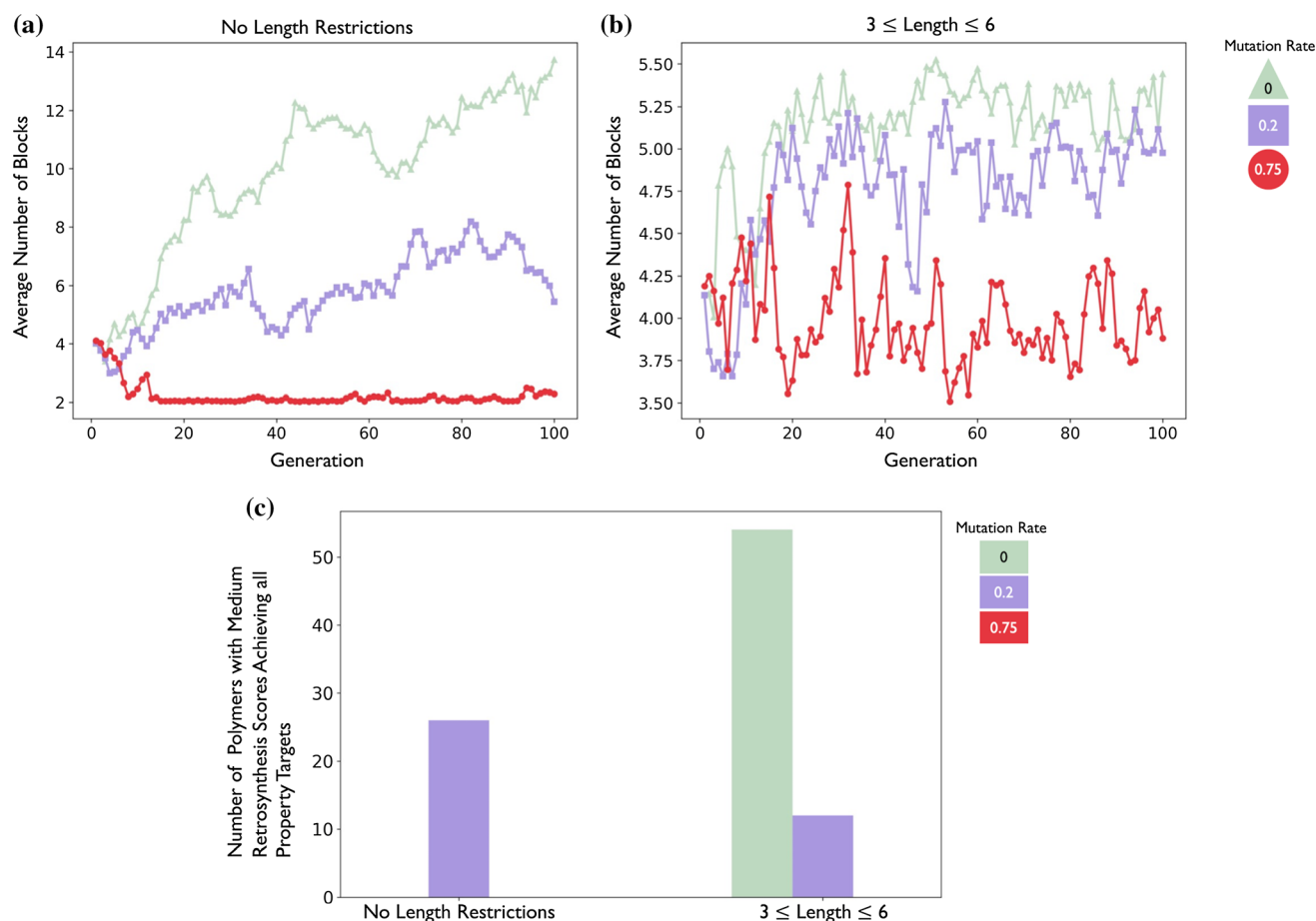


Figure 3 **a** Average number of blocks per generation for all polymers the genetic algorithm generates with a linear combination + clamping fitness function and duplication checking and mutation of repeat polymers. **b** Same as (a), but

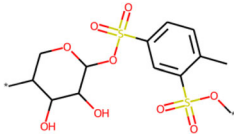
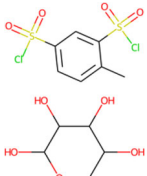
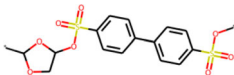
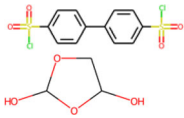
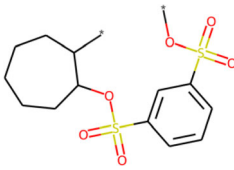
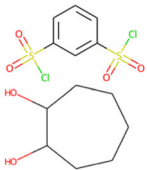
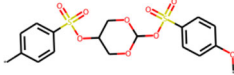
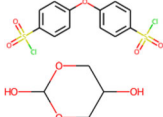
with a minimum of three blocks allowed and a max of six for each polymer repeat unit. **c** Number of retrosynthetically medium or high scoring polymer repeat units for runs described in (a) and (b).

By restricting size to lengths of three to six, as shown in Fig. 3b, retrosynthesizability was improved for low mutation rate runs, as shown in Fig. 3c. However, ensuring the polymer length never reached two did not improve the performance of high mutation rate runs. The GA still could not find the ideal chemical space. For these experiments, even a 50% mutation rate was too high, and the GA was unable to bring all averages to their target goals. Given the ineffectiveness of restricting the minimum size, it is best to only restrict the maximum size of polymers to improve realism. Even then, one should experiment to find a mutation rate that maintains a stable polymer length that is relatively low.

User-friendly chemical block selection

While the GA can find realistic polymers that achieve target properties, these polymers may still not be useful for specific design applications. Take the table of high scoring polymers predicted by the GA displayed in Table 3. Polymer 1 comes from an experiment that includes LC + clamping + DM + FBS, with a restricted max block size of 6. It achieves all target values and has a very high retrosynthesis score, but it also has $-OH$ groups. These $-OH$ groups would participate in H-bonding, restricting rotation and making it hard for dipoles to orient themselves in an electric field. Dipole orientation can improve the dielectric constant, so restricting rotation should be avoided. Thus, even though hypothetical polymers with these hydroxyl groups can achieve all target properties, because the

Table 3 Table of polymers the GA has predicted that achieve all target properties ($T_g > 500$ K, $\epsilon > 4$, $E_g > 5$ eV, $e_{\text{coh}} < 80$ cal cm $^{-3}$, and $\Phi_e > 3$ eV) and that have a high retrosynthesis score

Polymer	Structure	Properties	Predicted reactants	S_{score}
1		$E_g: 5.6 \pm 0.4$ eV $\epsilon: 4.7 \pm 1.5$ $T_g: 526 \pm 91$ K $e_{\text{coh}}: 47 \pm 90$ cal cm $^{-3}$ $\Phi_e: 3.2 \pm 0.2$ eV		0.88
2		$E_g: 5.1 \pm 0.5$ eV $\epsilon: 4.2 \pm 0.4$ $T_g: 539 \pm 88$ K $e_{\text{coh}}: 57 \pm 85$ cal cm $^{-3}$ $\Phi_e: 3.04 \pm 0.17$ eV		0.80
3		$E_g: 5.3 \pm 0.4$ eV $\epsilon: 4.5 \pm 0.5$ $T_g: 533 \pm 86$ K $e_{\text{coh}}: 51 \pm 88$ cal cm $^{-3}$ $\Phi_e: 3.1 \pm 0.19$ eV		0.77
4		$E_g: 5.2 \pm 0.5$ eV $\epsilon: 4.3 \pm 0.4$ $T_g: 530 \pm 81$ K $e_{\text{coh}}: 67 \pm 81$ cal cm $^{-3}$ $\Phi_e: 3.2 \pm 0.16$ eV		0.71

Predicted reactants are the reactants the retrosynthesis code generates when the inverse reaction known to create a similar polymer is used on the polymer. The inverse reaction was found to be a condensation reaction for each polymer, as opposed to addition or ring-opening. Each predicted reactant, except the bottom one for polymer 2, have been found for sale in Reaxys

hydroxyl groups are expected to negatively affect the target application, they should be removed from the block list so the GA does not get stuck in a local optimum in a hydroxyl-containing chemical space.

To avoid this problem, we explicitly incorporated a feature by which a user could avoid a chemical block that they think would negatively impact the material being designed. Polymer 2 in Table 3 is from an experiment with the same parameters as polymer 1, but that also included chemical screening and removed all hydroxyl groups from the block list. The GA successfully found a high scoring polymer without hydroxyl groups, whereas previous experiments did not. However, this also resulted in a fewer number of polymers achieving target properties being predicted on average. Additionally, the GA can find high scoring polymers that do not have adverse functional groups without explicitly removing them. Polymer 3 of Table 3 shows one that came from a run with LC + clamping + DM.

Polymer 4 comes from an experiment where the first generation of polymers was manually selected

from 10 medium synthesis scoring polymers taken from the experiment referenced for polymer 2. This was done to assess if having a retrosynthetically feasible starting point resulted in a larger number of retrosynthetically feasible polymers. It did not, however, that run did discover this high scoring hypothetical polymer.

Targeting polymer designs further can improve the GA's predictions for specific design criteria but can hinder the ability of the GA to achieve all target properties since the chemical space being explored is more restricted. While the GA can find a suitable space without restricting the chemical space, it can get "stuck" after finding a chemical space that achieves all target properties. Thus, it may be better to guide it towards chemical spaces that are suitable for the specific design task. On the other hand, if diversity can be improved, this may not be needed.

All polymers predicted by the GA have very high uncertainties for e_{coh} values, as shown in Table 3. They are of the same order of magnitude as the predicted values because the model was trained on

only 294 polymers. This exemplifies that the GA method of polymer prediction will be heavily dependent on model accuracy.

All predicted reactants, except for the bottom one for polymer 2, are commercially available, as per the Reaxys database [39]. Thus, three of these predicted polymers could, in principle, be synthesized, and one could be, if the monomer not in the Reaxys database could be synthesized.

Conclusion

Our primary objective in this effort was to design polymers suitable for high energy density applications, with the attributes of high temperature stability at high electric fields. This objective leads to a highly non-trivial materials search problem entailing the achievement of multiple property objectives. Our approach to achieving this objective was to impart significant enhancements to the legendary genetic algorithm (GA), specifically geared to address polymer design. A public version of our GA code for polymer design is released with this manuscript [40].

Several modifications were added to the GA to achieve four goals: allowance for the consideration of any number of target criteria, exploration of a diverse chemical space during GA exploration, designing polymers that are synthetically accessible and choosing (or avoiding) specific chemical blocks during design. Four retrosynthetically feasible example designs were shown and an additional 19 high scoring and 3665 medium scoring designs were provided as supplementary material.

High temperature dielectrics are not the only application accessible through this approach. Many other applications, such as polymer membranes for batteries and fuel cells [41], polymer membranes for the separation of complex mixtures of gases or solvents [42], polymers with metal-like conductivities for electronics [33], and recyclable plastics [43] are some other examples. Each application will require materials that have numerous property requirements and will require reliable property predictors (to accompany the GA algorithm) to circumvent the Edisonian guess-and-check experiments that are too slow to search the vast polymer chemical space.

Acknowledgments

This work is supported by the Office of Naval Research through a Multi-University Research Initiative (MURI) Grant (N00014-20-1-2586). We greatly appreciate it.

Author contributions

JK: Writing and reviewing of manuscript, modifying of genetic algorithm, experiment design, and analysis. LC: Original designer of retrosynthesis algorithm, review and editing of the manuscript. CK: Original designer of genetic algorithm. RR: Supervision, Methodology, Funding acquisition, Resources, Writing—review & editing.

Code availability

A modified version of the genetic algorithm is available at: <https://github.com/Ramprasad-Group/polyga>

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Supplementary Information: The online version contains supplementary material available at <http://doi.org/10.1007/s10853-021-06520-x>.

References

- [1] Rabuffi M, Picci G (2002) Status quo and future prospects for metallized polypropylene energy storage capacitors. *IEEE Trans Plasma Sci* 30:1939–1942. <https://doi.org/10.1109/TPS.2002.805318>
- [2] Qin S, Ma S, Boggs SA (2012) The mechanism of clearing in metalized film capacitors. In: 2012 IEEE international symposium on electrical insulation. IEEE, San Juan, PR, USA, pp 592–595
- [3] Reed CW, Cichanowskil SW (1994) The fundamentals of aging in HV polymer-film capacitors. *IEEE Trans Dielect Electr Insul* 1:904–922. <https://doi.org/10.1109/94.326658>
- [4] Zhou Y, Wang Q (2020) Advanced polymer dielectrics for high temperature capacitive energy storage. *J Appl Phys* 127:240902. <https://doi.org/10.1063/5.0009650>

- [5] Johnson RW, Evans JL, Jacobsen P et al (2004) The changing automotive environment: high-temperature electronics. *IEEE Trans Electron Packag Manufact* 27:164–176. <https://doi.org/10.1109/TEPM.2004.843109>
- [6] Ho JS, Greenbaum SG (2018) Polymer capacitor dielectrics for high temperature applications. *ACS Appl Mater Interfaces* 10:29189–29218. <https://doi.org/10.1021/acsami.8b07705>
- [7] Qiao Y, Yin X, Zhu T et al (2018) Dielectric polymers with novel chemistry, compositions and architectures. *Prog Polym Sci* 80:153–162. <https://doi.org/10.1016/j.progpolymsci.2018.01.003>
- [8] Venkat N, Dang TD, Bai Z et al (2010) High temperature polymer film dielectrics for aerospace power conditioning capacitor applications. *Mater Sci Eng B* 168:16–21. <https://doi.org/10.1016/j.mseb.2009.12.038>
- [9] Wang CC, Pilania G, Boggs SA et al (2014) Computational strategies for polymer dielectrics design. *Polymer* 55:979–988. <https://doi.org/10.1016/j.polymer.2013.12.069>
- [10] Huan TD, Boggs S, Teyssedre G et al (2016) Advanced polymeric dielectrics for high energy density applications. *Prog Mater Sci* 83:236–269. <https://doi.org/10.1016/j.pmatsci.2016.05.001>
- [11] Kim C, Chandrasekaran A, Huan TD et al (2018) Polymer genome: a data-powered polymer informatics platform for property predictions. *J Phys Chem C* 122:17575–17585. <https://doi.org/10.1021/acs.jpcc.8b02913>
- [12] Batra R, Song L, Ramprasad R (2020) Emerging materials intelligence ecosystems propelled by machine learning. *Nat Rev Mater* 6:655–678. <https://doi.org/10.1038/s41578-020-00255-y>
- [13] Venkatasubramanian V, Chan K, Caruthers JM (1995) Evolutionary design of molecules with desired properties using the genetic algorithm. *J Chem Inf Model* 35:188–195. <https://doi.org/10.1021/ci00024a003>
- [14] Kim C, Batra R, Chen L et al (2021) Polymer design using genetic algorithm and machine learning. *Comput Mater Sci* 186:110067. <https://doi.org/10.1016/j.commatsci.2020.110067>
- [15] Verhellen J, Van den Abeele J (2020) Illuminating elite patches of chemical space. *Chem Sci* 11:11485–11491. <https://doi.org/10.1039/D0SC03544K>
- [16] Berardo E, Turcani L, Miklitz M, Jelfs KE (2018) An evolutionary algorithm for the discovery of porous organic cages. *Chem Sci* 9:8513–8527. <https://doi.org/10.1039/C8SC03560A>
- [17] Sheridan RP, Kearsley SK (1995) Using a genetic algorithm to suggest combinatorial libraries. *J Chem Inf Model* 35:310–320. <https://doi.org/10.1021/ci00024a021>
- [18] Mannodi-Kanakathodi A, Chandrasekaran A, Kim C et al (2018) Scoping the polymer genome: a roadmap for rational polymer dielectrics design and beyond. *Mater Today* 21:785–796. <https://doi.org/10.1016/j.mattod.2017.11.021>
- [19] Kamal D, Tran H, Kim C et al (2021) Novel high voltage polymer insulators using computational and data-driven techniques. *J Chem Phys* 154:174906. <https://doi.org/10.1063/5.0044306>
- [20] Sharma V, Wang C, Lorenzini RG et al (2014) Rational design of all organic polymer dielectrics. *Nat Commun* 5:4845. <https://doi.org/10.1038/ncomms5845>
- [21] Zeng Q, Oganov AR, Lyakhov AO et al (2014) Evolutionary search for new high- k dielectric materials: methodology and applications to hafnia-based oxides. *Acta Crystallogr C Struct Chem* 70:76–84. <https://doi.org/10.1107/S2053229613027861>
- [22] Sun Y, Boggs SA, Ramprasad R (2012) The intrinsic electrical breakdown strength of insulators from first principles. *Appl Phys Lett* 101:132906. <https://doi.org/10.1063/1.4755841>
- [23] Hou Y, Zhang J, Zhang Z (2016) Significantly improved breakdown performances of propylene carbonate-based nano-fluids. *Micro Nano Letters* 11:490–493. <https://doi.org/10.1049/mnl.2016.0214>
- [24] Chen L, Huan TD, Quintero YC, Ramprasad R (2016) Charge injection barriers at metal/polyethylene interfaces. *J Mater Sci* 51:506–512. <https://doi.org/10.1007/s10853-015-9369-2>
- [25] Tan Q, Irwin P, Cao Y (2006) Advanced dielectrics for capacitors. *IEEE TransFM* 126:1153–1159. <https://doi.org/10.1541/ieejfms.126.1153>
- [26] Chu B (2006) A dielectric polymer with high electric energy density and fast discharge speed. *Science* 313:334–336. <https://doi.org/10.1126/science.1127798>
- [27] Chen L, Kern J, Lightstone JP, Ramprasad R (2021) Data-assisted polymer retrosynthesis planning. *Appl Phys Rev* 8:031405. <https://doi.org/10.1063/5.0052962>
- [28] Chen L, Kim C, Batra R et al (2020) Frequency-dependent dielectric constant prediction of polymers using machine learning. *npj Comput Mater* 6:61. <https://doi.org/10.1038/s41524-020-0333-6>
- [29] Doan Tran H, Kim C, Chen L et al (2020) Machine-learning predictions of polymer properties with Polymer Genome. *J Appl Phys* 128:171104. <https://doi.org/10.1063/5.0023759>
- [30] Cassar DR, Santos GG, Zanotto ED (2021) Designing optical glasses by machine learning coupled with a genetic algorithm. *Ceram Int* 47:10555–10564. <https://doi.org/10.1016/j.ceramint.2020.12.167>
- [31] Mallik S, Mallik K, Barman A et al (2017) Efficiency and cost optimized design of an induction motor using genetic

- algorithm. *IEEE Trans Ind Electron* 64:9854–9863. <https://doi.org/10.1109/TIE.2017.2703687>
- [32] Katoch S, Chauhan SS, Kumar V (2021) A review on genetic algorithm: past, present, and future. *Multimed Tools Appl* 80:8091–8126. <https://doi.org/10.1007/s11042-020-10139-6>
- [33] Gao G, Zheng F, Pan F, Wang L (2018) Theoretical investigation of 2D conductive microporous coordination polymers as Li–S battery cathode with ultrahigh energy density. *Adv Energy Mater* 8:1801823. <https://doi.org/10.1002/aenm.201801823>
- [34] Yang X-S (2014) Genetic algorithms. In: *Nature-inspired optimization algorithms*. Elsevier, pp 77–87. <https://doi.org/10.1016/B978-0-12-416743-8.00005-1>
- [35] Degen J, Wegscheid-Gerlach C, Zaliani A, Rarey M (2008) On the art of compiling and using “drug-like” chemical fragment spaces. *ChemMedChem* 3:1503–1507. <https://doi.org/10.1002/cmdc.200800178>
- [36] O’Boyle NM (2012) Towards a Universal SMILES representation—a standard method to generate canonical SMILES based on the InChI. *J Cheminform* 4:22. <https://doi.org/10.1186/1758-2946-4-22>
- [37] McCall J (2005) Genetic algorithms for modelling and optimisation. *J Comput Appl Math* 184:205–222. <https://doi.org/10.1016/j.cam.2004.07.034>
- [38] McInnes L, Healy J, Melville J (2020) UMAP: uniform manifold approximation and projection for dimension reduction. *arXiv:180203426 [cs, stat]*
- [39] Reaxys. <https://www.reaxys.com/#/search/quick>. Accessed 26 Jul 2021
- [40] Ramprasad Group (2021) polyga
- [41] Yang W-J, Wang H-Y, Lee D-H, Kim Y-B (2015) Channel geometry optimization of a polymer electrolyte membrane fuel cell using genetic algorithm. *Appl Energy* 146:1–10. <https://doi.org/10.1016/j.apenergy.2015.01.130>
- [42] Ali FAA, Alam J, Shukla AK et al (2020) A novel approach to optimize the fabrication conditions of thin film composite RO membranes using multi-objective genetic algorithm II. *Polymers* 12:494. <https://doi.org/10.3390/polym12020494>
- [43] Pilia G, Iverson CN, Lookman T, Marrone BL (2019) Machine-learning-based predictive modeling of glass transition temperatures: a case of polyhydroxyalkanoate homopolymers and copolymers. *J Chem Inf Model* 59:5013–5025. <https://doi.org/10.1021/acs.jcim.9b00807>

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