Computational Materials Science 186 (2021) 110067

Contents lists available at ScienceDirect

Computational Materials Science

journal homepage: www.elsevier.com/locate/commatsci

Polymer design using genetic algorithm and machine learning

Chiho Kim, Rohit Batra, Lihua Chen, Huan Tran, Rampi Ramprasad*

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

ARTICLE INFO

Keywords:
Genetic algorithm
Machine learning
Glass transition temperature
Bandgap
Polymer

ABSTRACT

Data driven or machine learning (ML) based methods have been recently used in materials science to provide quick material property predictions. Although powerful and robust, these predictive models are still limited in terms of their applicability towards the design of materials with target property or performance objectives. Here, we employ a nature-mimicking optimization method, the genetic algorithm, in tandem with ML-based predictive models to design polymers that meet practically useful, but extreme, property criteria (*i.e.*, glass transition temperature, $T_g > 500$ K and bandgap, $E_g > 6$ eV). Analogous to nature, the characteristic properties of a polymer are assumed to be determined by the constituting types and sequence of chemical building blocks (or fragments) in the monomer unit. Evolution of polymers by natural operations of crossover, mutation, and selection over 100 generations leads to creation of 132 new (as compared to 4 already known cases) and chemically unique polymers with high T_g and E_g. Chemical guidelines on what fragments make up polymers with extreme thermal and electrical performance metrics have been selected and revealed by the algorithm. The approach presented here is general and can be extended to design polymers with different property objectives.

1. Introduction

Polymers have found enormous use in numerous applications, due to their versatility and the richness of their chemical diversity [1]. The latter aspect also poses a challenge. The near-infinite chemical space spanned by polymers leads to a daunting search problem. *Edisonian* trial-and-error and intuition-based strategies may not be efficient, and run the risk of missing good solutions. Moreover, if such strategies use traditional experimental or computational routes, they may be time- and resource-intensive. Machine learning (ML) based surrogate models, trained on available polymer-property datasets, can make instantaneous property predictions for a new polymer, and may alleviate the burden on time and resources [2–13]. But such accelerated prediction options still leave open the challenge of accumulating a large and diverse candidate set of polymers for which predictions need to be made. It is completely unclear how one would make such a candidate set “complete” enough so as to not miss suitable and important candidates.

A more general and appropriate approach would be to solve the “inverse problem”, *i.e.*, given the desired property objectives, directly generate polymers that satisfy those objectives, as opposed to screening from a pre-defined candidate set. There have been attempts to perform such designs in the past [14,15], but they have been limited in terms of the explored chemical space, as they are constrained by the available choices of the building units. Recently, machine learning based generative models, such as variational autoencoders (VAE) and generative adversarial networks (GAN), have also been utilized to solve the inverse problem [16–23]. They learn a mapping from a continuous latent space to the materials space, using which new materials with desired properties are generated after solving the optimization problem in the latent space. While this approach remains attractive for drug discovery, its application to periodic systems such as polymers is in a state of infancy.

In this contribution, we set our goal as the design of polymers with two extreme properties: high glass transition temperature (T_g) and high bandgap (E_g). The former is desirable to find polymers that have high thermal stability at high temperatures. The latter is useful for polymers that can withstand high electric fields, and display high dielectric strength. Collectively, these two properties are essential for several applications, including high-temperature high-energy density dielectrics [24]. The difficulty in achieving these desired property objectives becomes apparent when we check the literature of known polymers: only four out of ~12,000 reference polymers collected from literature [25–28] meet the target properties ($T_g > 500$ K and $E_g > 6$ eV) as illustrated in Fig. 1. In this figure, the T_g and E_g estimates for the ~12,000 known polymers were made using our past ML models.[3] As can be seen from Fig. 1, the inverse relation between the E_g and the T_g makes it difficult to find polymers that meet both property criteria

* Corresponding author.
E-mail address: rampi.ramprasad@mse.gatech.edu (R. Ramprasad).

https://doi.org/10.1016/j.commatsci.2020.110067
Received 11 June 2020; Received in revised form 24 August 2020; Accepted 12 September 2020
Available online 28 September 2020

simultaneously. Indeed, the criteria are met only for 4 known polymers.

2. Methods

The genetic algorithm (GA), a simulated evolution-based search algorithm, is a powerful method to tackle this inverse problem of polymer design using the principle of natural selection that drives biological evolution [29]. In analogy with how nature uses the basic steps of *crossover*, *mutation* and *selection* for evolution of species, in this work we use GA to evolve a generation of polymer candidates to 'survive' user-defined property objectives. The inherent structure of a polymer makes its treatment using GA straightforward—a polymer can be thought of as a sequence of chemical building blocks connected to each other by covalent bonds (analogous to DNA base pairs), and the properties of a polymer are functions of the constituent chemical building blocks and their relative ordering (analogous to DNA). Thus, starting from a generation of candidate polymers, crossover and mutations operations may be performed to alter their sequence of chemical building blocks, and obtain new polymer offsprings. A user-defined fitness function (here, based on desired T_g and E_g property objectives) has to be then evaluated to aid in the retention of only the top performing offsprings, which then become the parent generation for the next iteration. This GA cycle may be repeated until sufficient number of candidates with desired properties are obtained.

The evaluation of polymer candidates to see if they meet the desired property objectives, *i.e.*, computation of the fitness function, is a crucial component of GA for polymer design. This step has been a major bottleneck since polymer property estimation through experiments or computations is very expensive and time-consuming. However, with the recent development of cheap and reliable ML models for polymer property estimation, the fitness function can now be computed in a fraction of a second. For this work, we developed ML-based predictive T_g and E_g models using the framework described in our previous works [3,30]. The ML models were based on a hierarchical polymer fingerprinting scheme and Gaussian process regression [3,31]. While the T_g model was trained on an experimental dataset of 5,072 polymers, the E_g model was learned using density functional theory (DFT) computed E_g values of 562 polymers. These ML models were found to be accurate with a root mean square error (RMSE) of 19 K and 0.26 eV for T_g and E_g predictions, respectively (Fig. S1). Stitching all the different pieces together, the GA can be used to search the polymer space as follows:

1. Starting from a randomly selected generation of polymer candidates, crossover and mutation operations are performed to produce new polymer candidates by altering the chemical building blocks and their sequence.
2. ML models are then be used to make quick property estimates for the newly generated polymer candidates and evaluate their fitness.
3. Only the top candidates with best fitness evaluation are retained as parent polymers, and the above steps are iterated until sufficient number of polymer candidates with desired properties are found. Here, we defined the fitness function as (normalized T_g) × (normalized E_g), where the normalization was performed to negate the effect of T_g being usually two orders of magnitude higher than E_g.

Overall workflow of the GA process is illustrated in Fig. 2a. A monomer repeat unit of polymers is represented using its constituent chemical building blocks, *e.g.*, polyvinyl chloride `CC[Cl]` can be written using chemical blocks `*C*` and `*C[Cl]*`, with "*" representing an open end of a building block. A total of 3,045 building blocks were extracted from ~12,000 reference polymers using the "breaking of retrosynthetically interesting chemical substructures" (BRICS) algorithm as implemented in the RDKit Python package [32,33]. We note that each of the chemical building blocks has 1–4 end points (represented by the symbol "*") that can act as a connection point with other chemical building blocks. To initiate the GA process, 100 polymers consisting of 8 building blocks in their monomer unit were created in the first generation. The building blocks were chosen randomly while respecting their frequency of occurrence in the reference polymer dataset. In each GA iteration, the top 10 polymer with highest fitness evaluation, *i.e.*, (normalized T_g) × (normalized E_g), were retained as parents to create the next generation offsprings through crossover and mutation operations.

During crossover, two parent polymers generate an offspring by combining one random segment of a parent polymer with another random segment of the other parent. The segmentation point of a parent polymer was chosen using a Gaussian function with a mean pointing to the center of the monomer unit. For example, for polymers with 8 building blocks in the monomer unit, the segmentation mostly occurs at their middle, with each resulting polymer segment containing 4 blocks. We allowed a small variation of choosing the segmentation position so that the monomer unit can be separated into the fragments with more or less number of blocks. Further, for each parent polymers the segmentation process was performed 4 times, resulting in a total of $^{10}C_2$ times 4 = 180 offspring polymers from 10 parent polymers in each iteration. Like in Nature, mutation operations were also incorporated to diversify

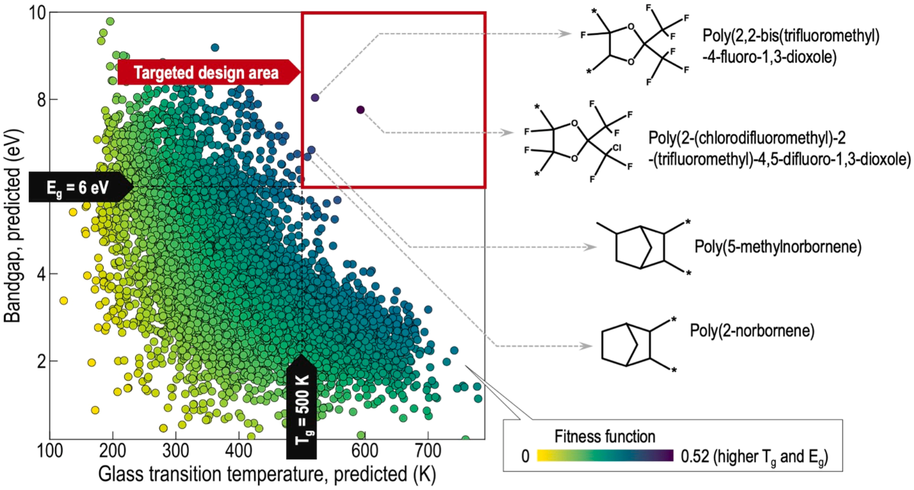

Fig. 1. Property map of glass transition temperature vs bandgap predicted by ML models. Among 12,721 known polymers, only four polymers meet the desired property objectives ($T_g > 500$ K and $E_g > 6$ eV). The T_g and E_g values are based on ML predictions. The fitness function used for color code was defined as (normalized T_g) × (normalized E_g).

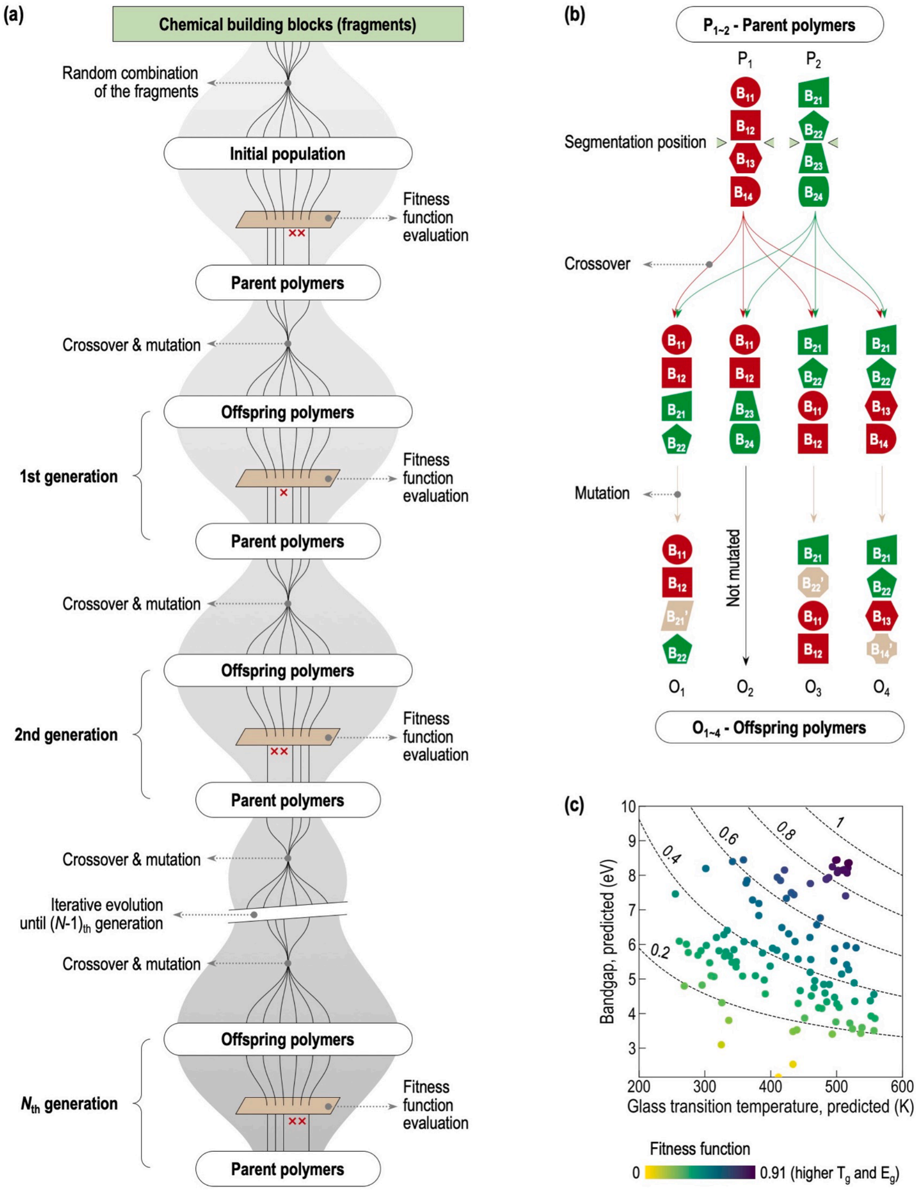

Fig. 2. Process to design polymers using genetic algorithm framework. (a) Overall workflow of iterative evolution of polymer generations. (b) Crossover and mutation to create offspring polymers from a pair of parent polymers. Polymers with four chemical building blocks (fragments) are shown for demonstration. (c) Offspring polymers mapped on to the property space of T_g vs bandgap E_g. 10 offspring polymers with highest fitness function are selected as parents in each iteration.

the "gene pool". A randomly selected polymer building block in the monomer unit was replaced with a new building block randomly chosen from the list of 3,045 blocks. We assigned a significant chance (40%) of mutation for each offspring polymer in order to promote the chemical diversification. During the evolution, offspring polymers that disobey some known chemical rules, e.g., having unstable motifs such as `*OOOO*`, or polymer assembling rules, e.g., having repeat unit with more than 3 end points such as `*CC(*)CC*`, were discarded from the candidates list. Crossover and mutation steps are schematically demonstrated in Fig. 2b using two exemplary polymers of 4 fragments each. The fitness function evaluations were made using T_g and E_g predictions from a pair of pretrained ML models. The evaluated fitness function for a few example polymers are depicted in Fig. 2c.

3. Results and discussion

Using the GA process we were able to design 132 new polymers that meet the target property objectives. This was achieved in 100 generations of evolution and from a cumulative of 12,675 offspring polymers. Fig. 3 displays the change in the chemical diversity of all polymers generated, and the corresponding evolution in the T_g and E_g property predictions, throughout the 100 generations. During the early generations (*e.g.*, 1–10), the fragments search space is still very diverse and arbitrary, with a low probability to generate target offspring polymers with both high T_g and E_g. Thus, average fitness function values can been seen to be relatively low in the earlier generations in Fig. 3a. However, with the progression of the evolutionary process, within each generation of polymer candidates, 10 offspring polymers with high fitness function are retained as parents and suitable building blocks that may contribute towards the desired properties are transferred to the next generation. This results in an overall increase in the fitness function value with every iteration. The later in the evolutionary process, higher the chance to incorporate "better" fragments, and thereby, higher the likelihood of discovering offspring polymers with desired properties.

The explored polymer chemical space is illustrated in Fig. 3b using the first two principal components (PC_1 and PC_2) obtained through

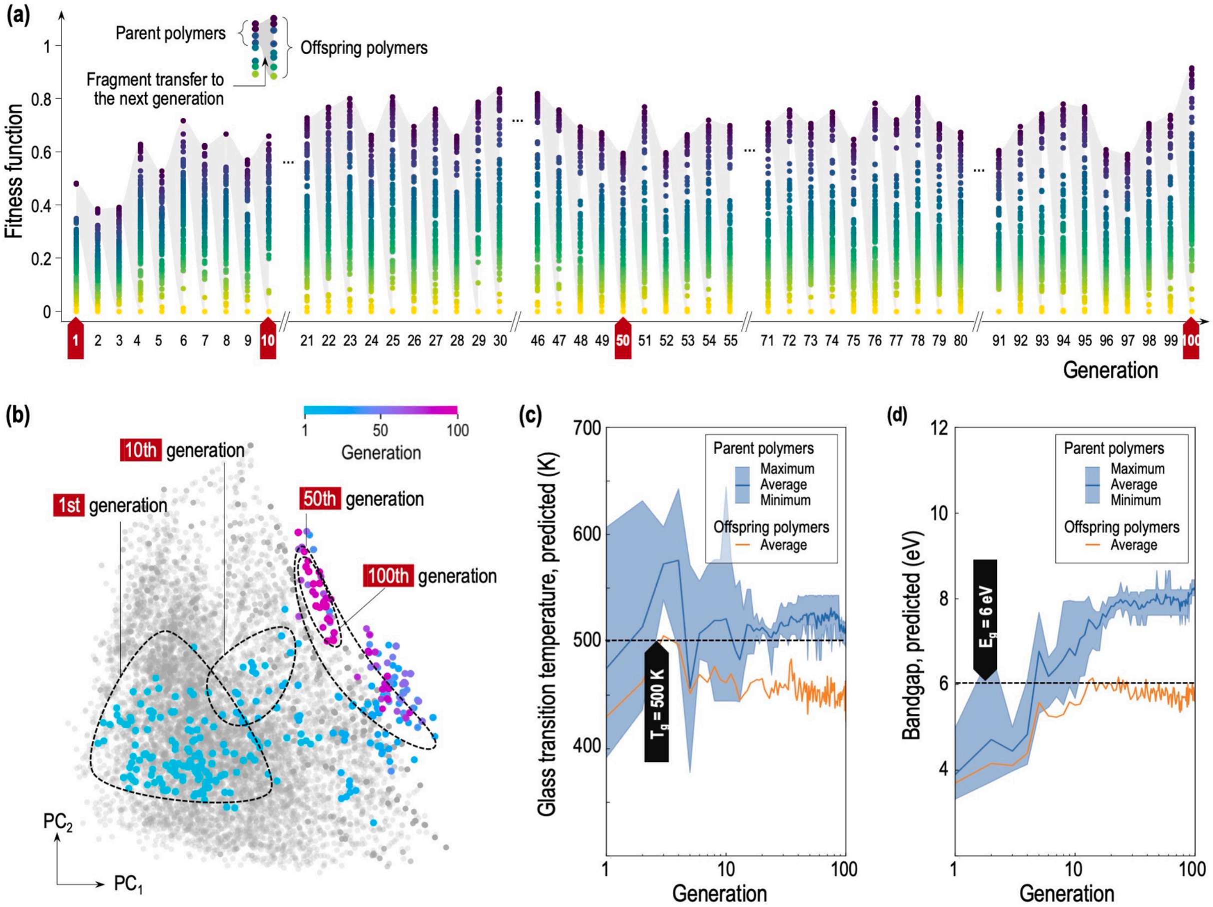

Fig. 3. Evolution of the fitness function, chemical diversity, and T_g and E_g predictions of polymers across the 100 GA iterations. (a) Fitness function evaluations for all the polymers generated in every GA iteration. From each generation, 10 offspring polymers with highest fitness function values are selected as parent polymers. 'Good' fragments from these parent polymers are transferred to the next generation, resulting in discovery of polymers with desired properties in the later generations. (b) 12,675 polymers projected on 2D PC space (PC generated using their polymer fingerprints). All polymers created during 100 generations are represented by gray points. Selected parents are color-coded by their generation number. Area of polymers created at the generation # 1, 10, 50, and 100 are selected to visualize the convergence in chemical diversity with evolution. Change in (c) T_g and (d) E_g predictions of polymers with every generation.

principal component analysis (PCA) on the fingerprint of all polymers generated during the 100 GA iterations. In the earlier generations, parent polymers are sparse and occupy a wide region of the chemical space, demonstrating their chemical diversity and distinctiveness. However, with the progression of the evolutionary process, the area occupied by each generation of parent polymers becomes narrower. Nonetheless, the mutation operation allow introduction of fragments that were originally not part of the parent polymers, and prevents a generation of polymer candidates to converge to a chemically equivalent point. Fig. 3c and d respectively capture the overall change in the T_g and E_g predictions for all the offspring and the parent polymers with every iteration. The E_g and T_g predictions for the parent polymers can be seen to increase with number of iterations, finally, converging at ~9 eV and ~520 K, respectively, which are higher than the desired property objectives.

Next, we perform analysis on the polymers discovered using the GA process. After the 100th generation, 132 new polymers pass the desired property objectives. In terms of chemical fragments, we found every discovered polymer candidate to contain at least one ring in the main-chain and/or in the pendant group. Further, more than 50% of the polymers had terminal group of difluorocarbon, *C(*)(F)F, and/or trifluoromethyl, *C(F)(F)F. Other dominant fragments that appear consistently throughout the evolutionary process and, therefore, could be responsible for high T_g and E_g property values in polymers, are shown in Fig. 4a. The color in the gene strip represents the cumulative number of occurrences of a fragment over 100 generations, with the frequently occurring fragments depicted by the darker bars on the strip. The ordering of the fragments is based on their chemical similarity (determined using the first PC of PCA performed on fingerprint vectors of 3,045 fragments). Thus, similar fragments like *C(*)(*)CC (Fragment 225) and *C(*)C* (Fragment 276) are positioned close to each other on the strip. Based on the spread of dominant fragments in Fig. 4a, it can be concluded that a diverse set of fragments can lead to high T_g and E_g property values.

Six example polymers designed in this work are presented in Fig. 4b and c, along with their ML predicted T_g and E_g values, and their constituent fragments depicted on the gene strips. The GPR based uncertainty in the T_g and E_g predictions are also included, which can be seen to be not very high; the discovered polymers had an overall GPR uncertainty in T_g and E_g predictions of 50 K and 0.7 eV, respectively. Not only the discovered polymers appears reasonable, a few common chemical blocks found during the GA process agree well with known chemical intuition. For example, fluorine atoms in *C(*)(CF$_3$)(CF$_3$) (Fragment 965) are known to contribute towards higher E_g by introducing lower (higher) C–F σ bonding (anti-bonding) orbitals. The presence of saturated rings not only induce high T_g because of their rigidity, but also provide high E_g owing to their low C–C and C–H σ bonding energy levels. Additionally, the polar groups present in the example polymers, including the *N(H)*, *OH and *C(F)* groups, can further enhance the dipole–dipole and H-bonding interactions, leading to high

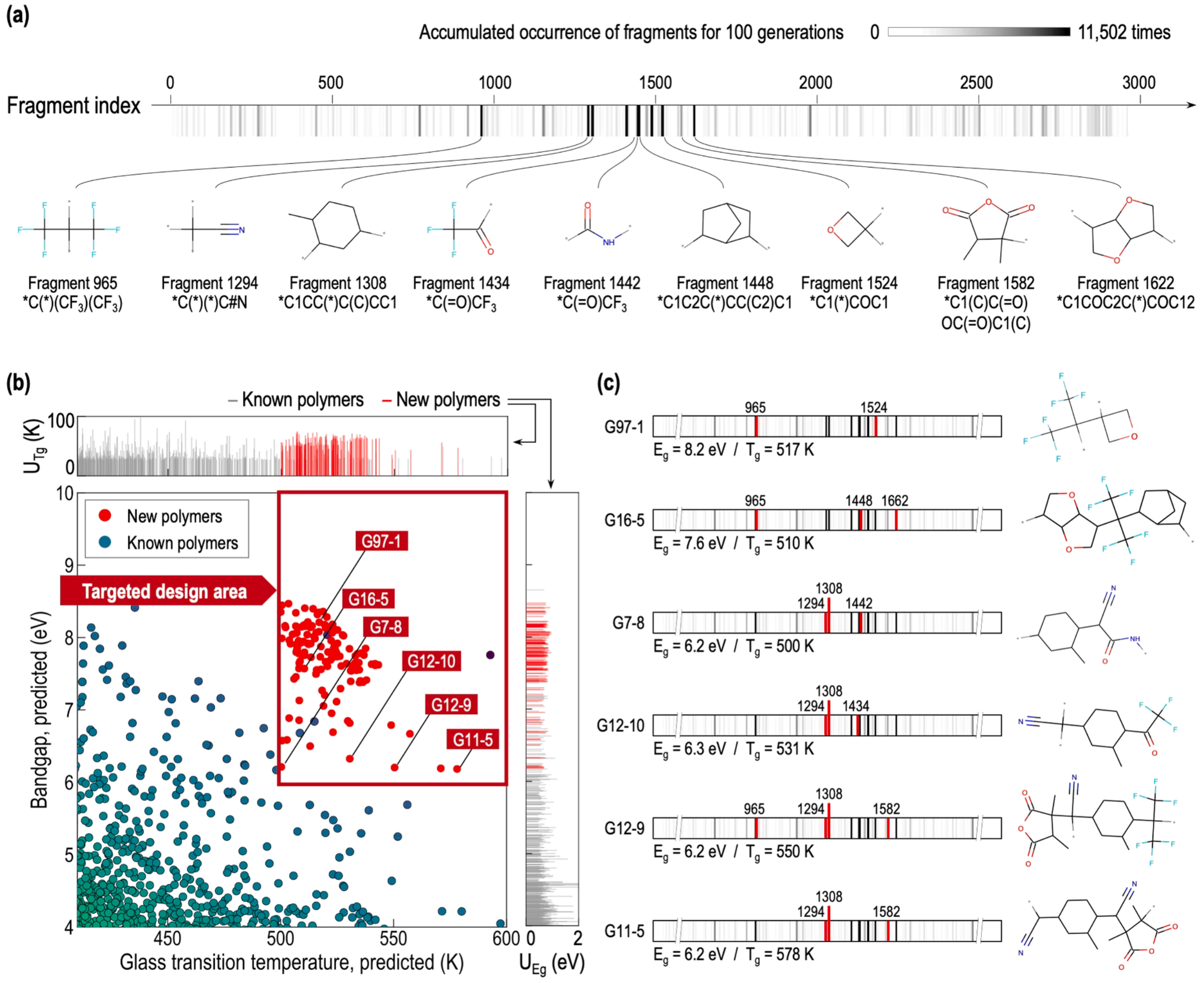

Fig. 4. Virtual gene strip of polymers and example of new polymer designs. (a) Gene strip shows cumulative occurrence of all fragments (chemical building blocks) over 100 generations of evolution. Nine fragments obtained from six hand-picked example polymer designs are indicated using their SMILES representation. (b) Position of 132 polymer designs generated during the 100 GA iterations on the map of T_g vs E_g. Uncertainty estimates for the predicted T_g (U_{Tg}) and predicted E_g (U_{Eg}) are shown together. Six hand-picked example polymers are highlighted with tags 'G#-#' representing the generation number G# and parent index #. (c) Gene strip and structure of the example polymers. A symbol '*' marks an open position in polymer chain or chemical building block. Polymers that meet the design criteria consist of 2–6 building blocks, although the GA process was initiated with polymers containing 8 blocks. This is owing to the available flexibility in the segmentation position during crossover.

T_g.

4. Conclusions

In general, we believe that the GA algorithm can rapidly and reliably assist polymer design for specific applications, especially high-temperature energy capacitors. Over a hundred new polymer candidates, six of them are shown in Fig. 4, are proposed for further experimental validation. This approach has several clear advantages. First, the GA process developed here was demonstrated with a randomly generated initial population, assuming no prior knowledge on the desirable polymers. Such information, if available, could be particularly useful. In particular, by biasing the initial population and/or the mutation operation towards favorable building blocks (or fragments), the GA search for target polymers may be significantly accelerated. Alternatively, we can further narrow down the searching in GA with desired chemical subspaces and specific structural arrangements, *e.g.*, the novel polyoxafluoronorbornene polymer dielectric for high temperature, high energy capacitors [24].

Second, because this GA scheme involves less number of tuning parameters than other generative models, *e.g.*, VAE and GAN, it provides a more viable, cheaper, and efficient option for the scenario in which training data are small. However, the different GA parameter choices, *e.g.*, the quality and the size of the initial generation, chance of mutation, and possible building blocks, which significantly impact the results obtained, must be optimized using a laborious trial-and-error approach.

Finally, we note that this polymer design approach is generalizable to other property objectives, provided corresponding reliable property prediction models (ML or otherwise) can be developed. The efficiency of this approach is dependent on the accuracy and the computational cost of the underlying property prediction models. With more efforts being devoted towards polymer database development and construction of associated ML property prediction models, we expect this approach to become more general, accurate and efficient.

Moving forward, we expect to generalize this scheme in order to handle more complex polymer systems, including ladder and cross-linked polymers. Furthermore, the scoring function can also be improved to incorporate other important design aspects, such as polymer thermodynamic and chemical stability, and/or its ease of synthesizability.

5. Data availability

The DFT computed bandgap used in this work is available at our online repository `https://khazana.gatech.edu`.

CRediT authorship contribution statement

Chiho Kim: Conceptualization, Methodology, Software, Investigation, Visualization, Writing - original draft, Writing - review & editing. **Rohit Batra:** Writing - original draft, Writing - review & editing, Software. **Lihua Chen:** Writing - review & editing, Methodology. **Huan Tran:** Data curation, Writing - review & editing. **Rampi Ramprasad:** Supervision, Methodology, Funding acquisition, Resources, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported financially by the National Science Foundation (Grant No. 1743418) and a Multidisciplinary University Research Initiative grant from the Office of Naval Research (Grant No. N00014-17-1-2656).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.commatsci.2020.110067.

References

[1] D. van Krevelen, Properties of Polymers: Their Correlation with Chemical Structure, Elsevier, 1997. URL:https://books.google.com/books?id=RQp9vgEACAAJ.

[2] A. Mannodi-Kanakkithodi, A. Chandrasekaran, C. Kim, T.D. Huan, G. Pilania, V. Botu, R. Ramprasad, Scoping the polymer genome: a roadmap for rational polymer dielectrics design and beyond, Mater. Today 21 (7) (2018) 785–796, https://doi.org/10.1016/j.mattod.2017.11.021. URL:https://www.sciencedirect.com/science/article/pii/S1369702117307344.

[3] C. Kim, A. Chandrasekaran, T.D. Huan, D. Das, R. Ramprasad, Polymer genome: A data-powered polymer informatics platform for property predictions, J. Phys. Chem. C 122 (31) (2018) 17575–17585. arXiv:https://doi.org/10.1021/acs.jpcc.8b02913, doi:10.1021/acs.jpcc.8b02913. URL: doi: 10.1021/acs.jpcc.8b02913.

[4] A. Mannodi-Kanakkithodi, G. Treich, T.D. Huan, R. Ma, M. Tefferi, Y. Cao, G. Sotzing, R. Ramprasad, Rational co-design of polymer dielectrics for energy storage, Adv. Mater. 28 (2016) 6277–6291.

[5] T.D. Huan, S. Boggs, G. Teyssedre, C. Laurent, M. Cakmak, S. Kumar, R. Ramprasad, Advanced polymeric dielectrics for high energy density applications, Prog. Mater. Sci. 83 (2016) 236.

[6] L. Chen, C. Kim, R. Batra, J.P. Lightstone, C. Wu, Z. Li, A. Deshmukh, Y. Wang, H. Tran, P. Vashishta, G. Sotzing, Y. Cao, R. Ramprasad, Frequency-dependent dielectric constant prediction of polymers using machine learning, Npj Comput. Mater. 6 (2020) 61.

[7] J.P. Lightstone, L. Chen, C. Kim, R. Batra, R. Ramprasad, Refractive index prediction models for polymers using machine learning, J. Appl. Phys. 127 (21) (2020), 215105.

[8] G. Zhu, C. Kim, A. Chandrasekarn, J.D. Everett, R. Ramprasad, R.P. Lively, Polymer genome-based prediction of gas permeabilities in polymers, J. Polym. Eng. 40 (6) (2020) 451–457, https://doi.org/10.1515/polyeng-2019-0329.

[9] L. Chen, T.D. Huan, Y.C. Quintero, R. Ramprasad, Charge injection barriers at metal/polyethylene interfaces, J. Mater. Sci. 51 (1) (2016) 506–512, https://doi.org/10.1007/s10853-015-9369-2.

[10] C. Kim, A. Chandrasekaran, A. Jha, R. Ramprasad, Active-learning and materials design: the example of high glass transition temperature polymers, MRS Commun. 9 (2019) 860–866, https://doi.org/10.1557/mrc.2019.78.

[11] A. Mannodi-Kanakkithodi, T.D. Huan, R. Ramprasad, Mining materials design rules from data: the example of polymer dielectrics, Chem. Mater. 29 (21) (2017) 9001–9010, https://doi.org/10.1021/acs.chemmater.7b02027.

[12] S. Venkatram, C. Kim, A. Chandrasekaran, R. Ramprasad, Critical assessment of the hildebrand and hansen solubility parameters for polymers, J. Chem. Inf. Model. 59 (10) (2019) 4188–4194, pMID: 31545900. arXiv:https://doi.org/10.1021/acs.jcim.9b00656, doi:10.1021/acs.jcim.9b00656. URL: doi: 10.1021/acs.jcim.9b00656.

[13] S. Nasreen, G.M. Treich, M.L. Baczkowski, A.K. Mannodi-Kanakkithodi, A. Baldwin, S.K. Scheirey, Y. Cao, R. Ramprasad, G.A. Sotzing, A material genome approach towards exploration of zn and cd coordination complex polyester as dielectrics: design, synthesis and characterization, Polymer 159 (2018) 95–105, https://doi.org/10.1016/j.polymer.2018.10.017. URL:http://www.sciencedirect.com/science/article/pii/S0032386118309339.

[14] A. Mannodi-Kanakkithodi, G. Pilania, T.D. Huan, T. Lookman, R. Ramprasad, Machine learning strategy for the accelerated design of polymer dielectrics, Sci. Rep. 6 (2016) 20952, https://doi.org/10.1038/srep20952.

[15] T.D. Huan, A. Mannodi-Kanakkithodi, R. Ramprasad, Accelerated materials property predictions and design using motif-based fingerprints, Phys. Rev. B 92 (2015), 014106.

[16] R. Gómez-Bombarelli, J.N. Wei, D. Duvenaud, J.M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T.D. Hirzel, R.P. Adams, A. Aspuru-Guzik, Automatic chemical design using a data-driven continuous representation of molecules, ACS Cent. Sci. 4 (2) (2018) 268–276, pMID: 29532027. arXiv:https://doi.org/10.1021/acscentsci.7b00572, doi:10.1021/acscentsci.7b00572. URL: doi: 10.1021/acscentsci.7b00572.

[17] M.J. Kusner, B. Paige, J.M. Hernández-Lobato, Grammar variational autoencoder, in: D. Precup, Y.W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning, Vol. 70 of Proceedings of Machine Learning Research, PMLR, International Convention Centre, Sydney, Australia, 2017, pp. 1945–1954. URL: http://proceedings.mlr.press/v70/kusner17a.html.

[18] H. Dai, Y. Tian, B. Dai, S. Skiena, L. Song, Syntax-directed variational autoencoder for structured data, arXiv:1802.08786. URL: https://arxiv.org/abs/1802.08786.

[19] W. Jin, R. Barzilay, T. Jaakkola, Syntax-directed variational autoencoder for structured data, arXiv:1802.04364. URL: https://arxiv.org/abs/1802.04364.

[20] Q. Liu, M. Allamanis, M. Brockschmidt, A. Gaunt, Constrained graph variational autoencoders for molecule design, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems 31, Curran Associates Inc, 2018, pp. 7795–7804. URL: http://papers.nips.cc/paper/8005-constrained-graph-variational-autoencoders-for-molecule-design.pdf.

[21] G. Lima Guimaraes, B. Sanchez-Lengeling, C. Outeiral, P. Luis Cunha Farias, A. Aspuru-Guzik, Objective-reinforced generative adversarial networks (organ) for sequence generation models, arXiv:1705.10843. URL: https://arxiv.org/abs/1705.10843.

[22] J. You, B. Liu, R. Ying, V. Pande, J. Leskovec, Graph convolutional policy network for goal-directed molecular graph generation, in: Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18, Curran Associates Inc., Red Hook, NY, USA, 2018, p. 6412–6422.

[23] N. De Cao, T. Kipf, Molgan: An implicit generative model for small molecular graphs, arxiv.org/abs/1805.11973. URL: https://arxiv.org/abs/1805.11973.

[24] C. Wu, A.A. Deshmukh, Z. Li, L. Chen, A. Alamri, Y. Wang, R. Ramprasad, G.A. Sotzing, Y. Cao, Flexible temperature-invariant polymer dielectrics with large bandgap, Adv. Mater. 2000499 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/adma.202000499, doi:10.1002/adma.202000499. URL:https://onlinelibrary.wiley.com/doi/abs/10.1002/adma.202000499.

[25] T.D. Huan, A. Mannodi-Kanakkithodi, C. Kim, V. Sharma, G. Pilania, R. Ramprasad, A polymer dataset for accelerated property prediction and design, Sci. Data 3 (2016), 160012, https://doi.org/10.1038/sdata.2016.12. URL:http://www.nature.com/articles/sdata201612.

[26] J. Bicerano, Prediction of Polymer Properties, Marcel Dekker Inc., New York, USA, 2002.

[27] J.E. Mark (Ed.), Polymer Data Handbook, second ed., Oxford University Press, New York, 2009.

[28] S. Otsuka, I. Kuwajima, J. Hosoya, Y. Xu, M. Yamazaki, Polyinfo: Polymer database for polymeric materials design, in: 2011 International Conference on Emerging Intelligent Data and Web Technologies (EIDWT), IEEE, Tirana, 2011, pp. 22–29. doi:10.1109/EIDWT.2011.13.

[29] J. Holland, Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence, University of Michigan Press, 1975. URL:https://books.google.com/books?id=JE5RAAAAMAAJ.

[30] A. Jha, A. Chandrasekaran, C. Kim, R. Ramprasad, Model. Simul. Mater. Sci. Eng. 27 (2) (2019) 024002, https://doi.org/10.1088/1361-651X/aaf8ca.

[31] C.E. Rasmussen, C.K.I. Williams (Eds.), Gaussian Processes for Machine Learning, The MIT Press, Cambridge, MA, 2006.

[32] RDKit, open source toolkit for cheminformatics. URL:http://www.rdkit.org/.

[33] J. Degen, C. Wegscheid-Gerlach, A. Zaliani, M. Rarey, On the art of compiling and using 'drug-like' chemical fragment spaces, ChemMedChem 3 (10) (2008) 1503–1507. arXiv:https://chemistry-europe.onlinelibrary.wiley.com/doi/pdf/10.1002/cmdc.200800178, doi:10.1002/cmdc.200800178. URL:https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/cmdc.200800178.