**Supporting Information:**

**Design of Polymers for Energy Storage Capacitors Using Machine Learning and Evolutionary Algorithms**

Joseph Kern, Lihua Chen, Chiho Kim, Rampi Ramprasad #

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

\*rampi.ramprasad@mse.gatech.edu

ML models of 5 target properties:

For each machine learning (ML) model of the five targeted properties, 80%-20% randomly selected test-train splits, 5-fold cross validation was used with the average root mean square error (RMSE) reported for each model. After, 100% of data was used to create the final model. The charge injection barrier (ecoh), glass transition temperature (Tg), and bandgap (Eg) models [1], charge injection barrier () model [2], and dielectric constant (e) [3] model are explained further in their associated references. The number of training datapoints and average five-fold RMSE for each model is summarized in Table S1.

**Table S1** Summary of Gaussian Process Regression models. Each model was trained on the associated number of training datapoints with the five-fold cross validation average root mean square error (RMSE) shown

|  |  |  |
| --- | --- | --- |
| Model | Training datapoints | 5-fold cross validation average test RMSE |
| Tg | 5083 | 35.7 K |
| e | 1200 | 0.35 |
| Eg | 3883 | 0.51 eV |
| ecoh | 235 | 25.21 cal cm^-3 |
|  | 1448 | 0.25 eV |

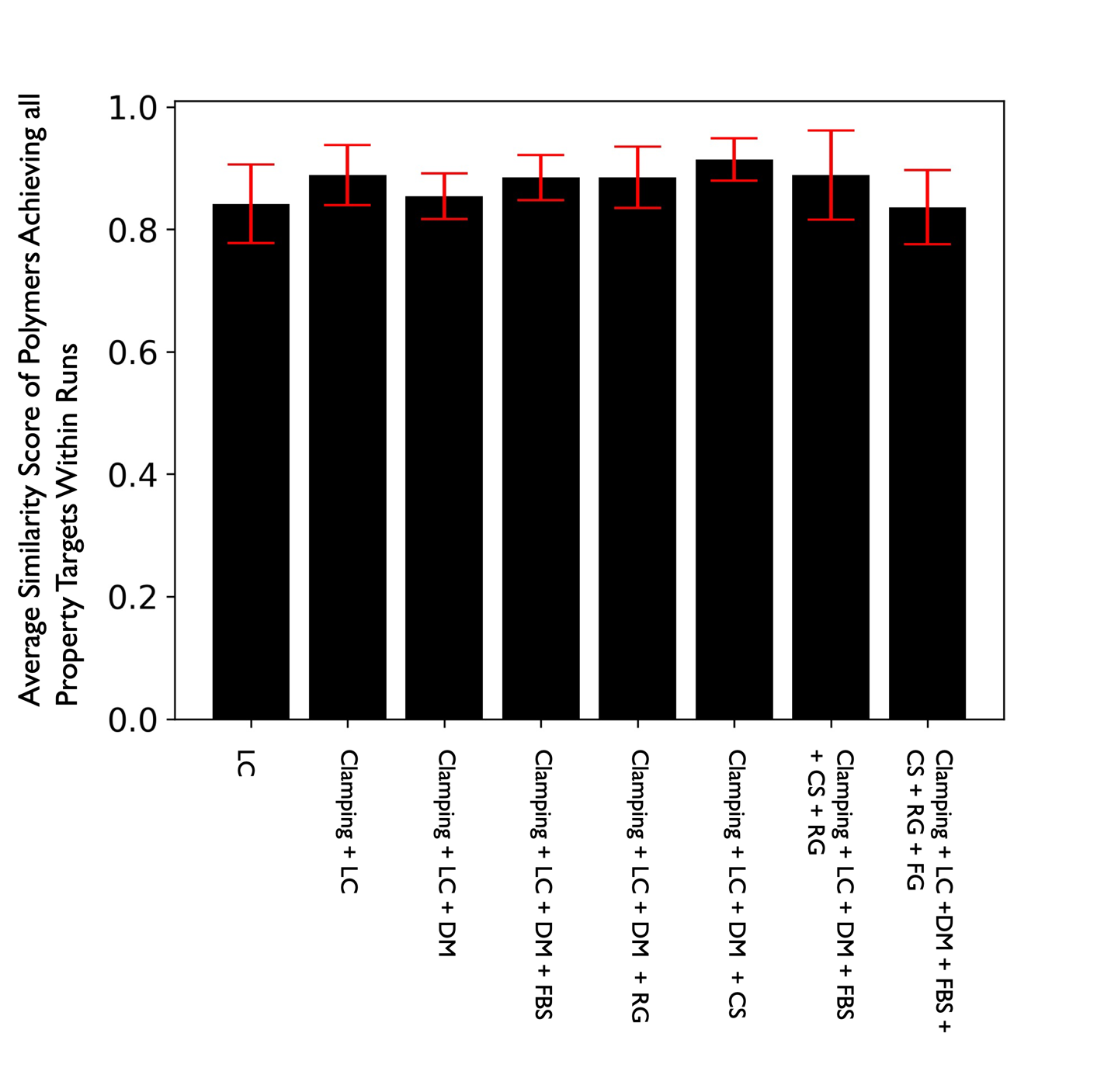
Hypothetical Polymer Diversity:

Diversity of polymers achieving all target properties remained low throughout each experiment of the GA. Adding duplication mutation (DM) failed at improving these low diversity scores, as seen in Figure S1. Diversity of all polymers within a run of an experiment was not assessed due to the large number of polymers generated for each run and because the goal was to increase the diversity of polymers achieving target properties, not the general space explored.

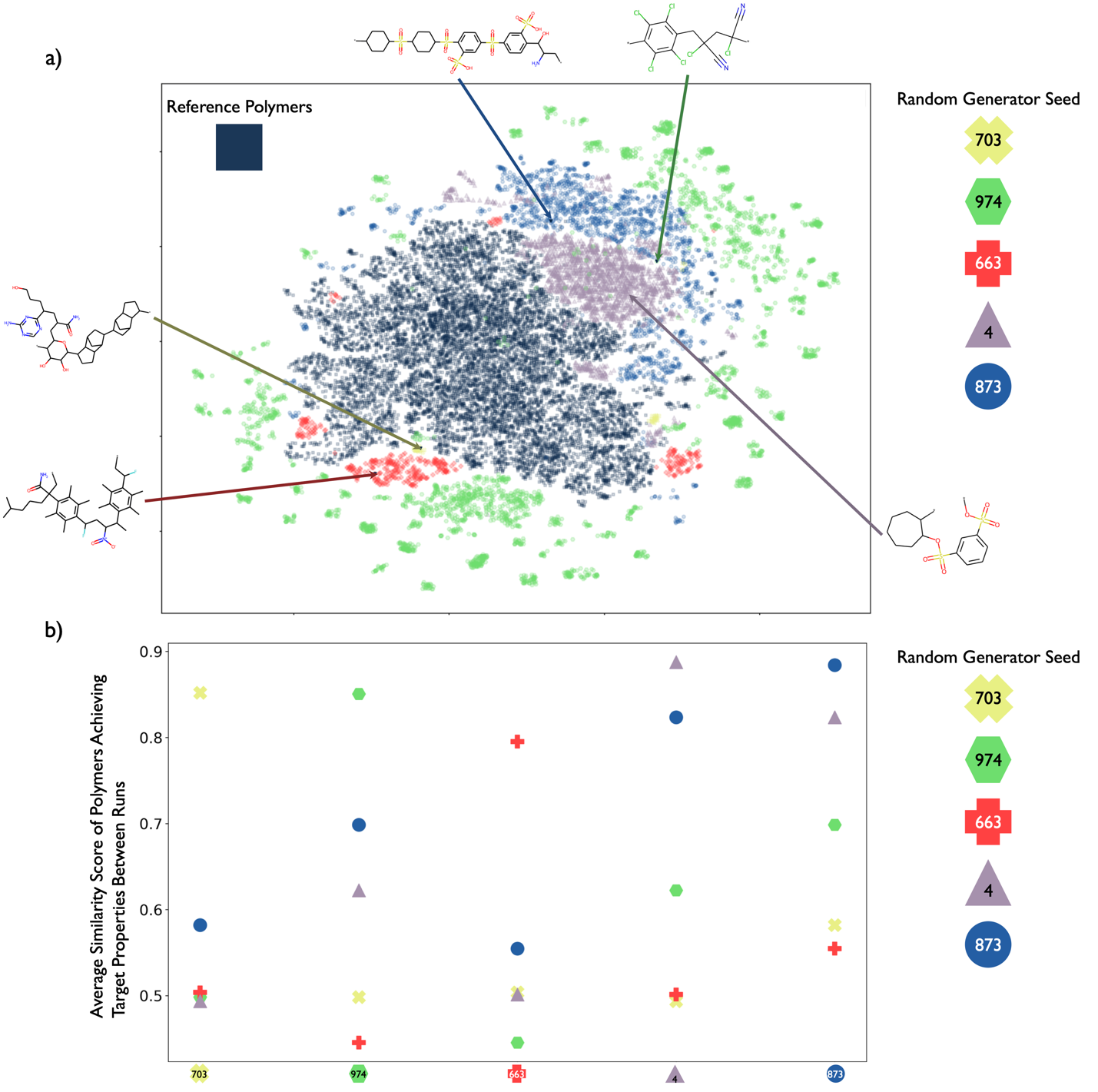
While diversity was low within runs of a genetic algorithm, it was higher between runs, as shown in Figure S2. The Uniform Manifold Approximation and Projection (UMAP) shown in Figure S2 (a) is a 2-d representation of a high-dimension space [4]. Each datapoint corresponds to a polymer. Polymers that are closer together are, generally, more similar. The GA generated polymers that achieve target properties typically fall outside the region of known polymer space (reference polymers in the image), and each seed within the run typically finds a unique chemical space that achieve target properties. Note that polymers in seed 974 are still chemically similar despite appearing to be very spaced on the map. This is because a complex, high-dimensional chemical space is being mapped to two dimensions.

Figure S2 (b) quantitatively corroborates the suggestion that, between runs of the genetic algorithm, unique chemical spaces are explored. For each polymer achieving target properties in a run, its Tanimoto similarity score was compared with every polymer achieving target properties in a separate run. The average similarity score for all polymers compared to the other polymers was then plotted. The x-tick corresponds to the run being compared with all polymers of the associated y-tick. The plot shows that when compared with itself, polymers in a run are very similar (~0.8 to 0.9), but when compared to other runs, they are more dissimilar (~0.4 to 0.7). This is not always the case, however, as seen with random generator seeds 4 and 873.

This inconsistency is because any change in the way polymers are generated or propagate to the next generation can cause a massive change in the chemical space explored, as seen in Figure S3. This UMAP of all polymers between a run of two experiments shows a major divergence in the explored space, despite starting with the same first generation of polymers, with the only difference being frequency-based selection (FBS)



**Figure S1** Average Tanimoto similarity score of all polymers achieving target properties within a run of the genetic algorithm. Each experiment was performed five times. If a run of the experiment did not have any polymers achieving target properties, it was excluded from averaging. LC stands for linear combination, DM for duplicate mutatiton, FBS for frequency-based selection, RG for remove adverse functional groups, CS for chemical screening, and FG for manually selected first generation



**Figure S2** Uniform Manifold Approximation and Projection (UMAP) (a) and average tanimoto similarity score (b) of polymers achieving target properties (Tg > 500 K, > 4, Eg > 5 eV, ecoh < 80 cal cm-3, and Φe > 3 eV) between five runs of the genetic algorithm with a linear combination + Clamping fitness function and duplicate mutation of repeat polymers. (a) UMAP projects the chemical fingerprints of the polymers down to two-dimensions with a cosine metric, 200 nearest neighbors, and 0.1 minimum distance. Reference polymers are the polymers breaking of retrosynthetically interesting chemical substructures was used on. UMAP shows that the GA typically predicts polymers within a run that are chemically similar, but between runs, the polymers form distinct groups. Predicted polymers are typically distinct from the reference polymers blocks were taken from. (b) average tanimoto similarity scores between polymers achieving target properties of runs shows that polymers are very similar within a run but are much less similar to polymers of other runs. Seeds are colored the same as they are in the (a)

A picture containing chart

Description automatically generated

**Figure S3** UMAP of all polymers in two separate runs of the GA that used a random generator seed of 4. First generations of both are the same

Hypothetical Polymer Synthesizability:

Unfortunately, most polymers achieving target properties are not retrosynthetically feasible. As seen in Figure S3, only a few runs within an experiment can find any polymers that are retrosynthetically feasible. Runs that find a large number that are feasible typically generate very similar polymers. Efforts to increase synthetic feasibility by adding chemical screening (CS) and frequency-based selection (FBS) did not consistently work.

Chart

Description automatically generated

**Figure S4** Number of polymers achieving target properties (Tg > 500 K, > 4, Eg > 5 eV, ecoh < 80 cal cm-3, and Φe  > 3 eV) that are either medium or high retrosynthetically scoring

Reference:

1. Doan Tran H, Kim C, Chen L, et al (2020) Machine-learning predictions of polymer properties with Polymer Genome. Journal of Applied Physics 128:171104. https://doi.org/10.1063/5.0023759

2. 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

3. 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

4. McInnes L, Healy J, Melville J (2020) UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:180203426 [cs, stat]