# Supporting Information

## 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[*,†]

[†]*School of Materials Science and Engineering, Georgia Institute of Technology, 30332 Atlanta, Georgia, United States*

[‡]*College of Computing, Georgia Institute of Technology, 30332 Atlanta, Georgia, United States*

[¶]*School of Chemical and Biomolecular Engineering, Georgia Institute of Technology, 30332 Atlanta, Georgia, United States*

E-mail: rampi.ramprasad@mse.gatech.edu

## S1 Comparison between genetic algorithm and polyG2G

An experiment was conducted to compare the genetic algorithm (GA) of Kim et al.[1] and *polyG2G*. In their paper, Kim et al. used the GA to find polymers with a band gap greater than 6 eV and a glass-transition temperature greater than 500 K. Out of 12,675 polymers generated by their GA, 132 met these twin objectives. This is equivalent a success rate of 1.04%.

We used *polyG2G* to find polymers meeting the same objectives considered by Kim et al.—a predicted band gap greater than 6 eV and a glass-transition temperature greater than 500 K. Out of 222,464 polymers generated by *polyG2G*, 6,570 polymers met the twin

objectives. This is equivalent to a success rate of 2.95%—nearly 3 times the success rate of the GA.

## S2 Selecting a fingerprint for computing polymer similarity

Two polymer similarity metrics were considered—cosine similarity and Tanimoto similarity. Since the cosine similarity can accept continuous inputs, it was computed with the Polymer Genome (PG) fingerprint. We call this approach the "PG-cosine" approach. Since the Tanimoto similarity can only accept bit-wise inputs, it was computed with the Morgan fingerprint.[2] We call this approach the "Morgan-Tanimoto" approach. We computed the top 6 most similar pairs of polymers from our data set (see Section 2.1) using both metrics. The results are shown below in Figure S1. It is visually clear that the "Morgan-Tanimoto" approach does a better job of quantifying similarity between polymers. For this reason we chose to use the "Morgan-Tanimoto" approach in our work to compute similarity.

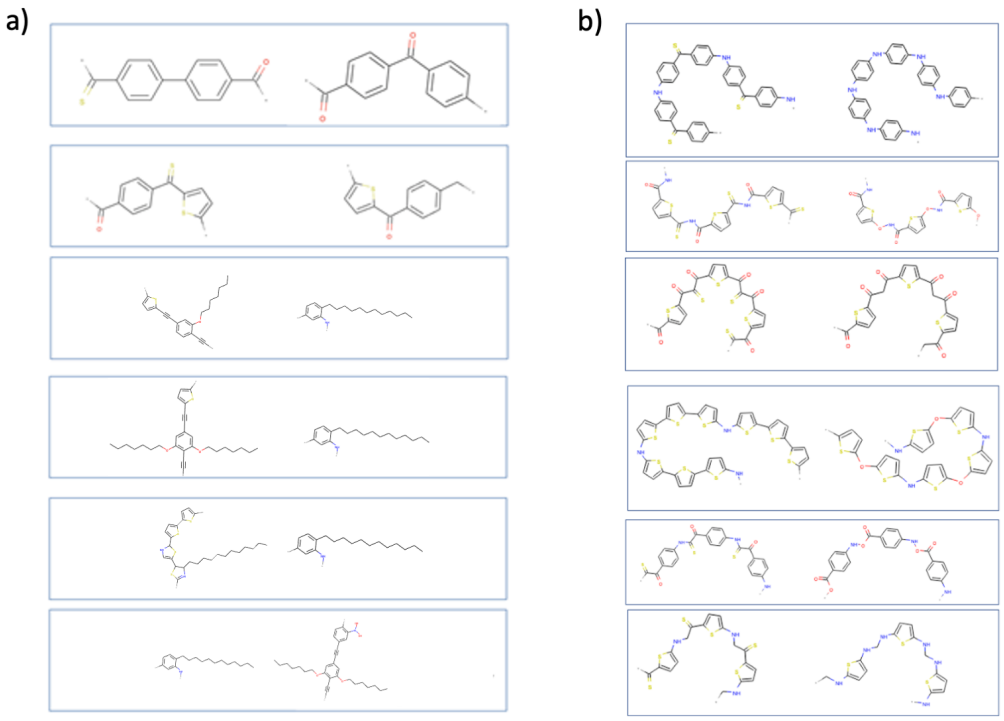


Figure S1: Top six similar polymer pairs according to a) the PG-cosine approach and b) the Morgan-Tanimoto approach

# S3 Computing Polymer Uniqueness

The uniqueness, $U$, of a polymer, $p$, (with respect to all other polymers in a set $\mathcal{P}$) was defined over the Morgan fingerprint, $f$. The formal definition is given below.

$$U(p) = 1 - \max\left[T(f(p), f(q))\right], \forall q \neq p \in \mathcal{P} \tag{1}$$

where $T$ is the Tanimoto similarity. Put simply, if a polymer is not similar (according to the Morgan-Tanimoto approach) to *any* other polymer in $\mathcal{P}$, then it will have a high uniqueness value.

## S4 Hyper-parameter optimization of the classifier

The classifier, a neural network, was implemented using the TensorFlow library in python. The classifier's weights were found using the Adam optimizer to minimize the binary cross-entropy. The best model was chosen during training by monitoring the binary cross-entropy of the validation set. Early stopping, with a patience of 300, was used to determine the number of training epochs. The hyper parameters—initial learning rate and number of dense hidden layers—were optimized using the python package Keras-Tuner. The optimal initial learning rate was found to be $10^{-3}$ while the optimal number of hidden layers was 3. The number of neurons per layer was kept constant at 128. A sigmoid activation was applied following the last dense layer. ReLU activations were used after all preceding layers.

## S5 Feature Descriptions

The following features are among the most important to in determining whether or not a polymer will meet our tri-property objective. Each feature is computed per repeat unit of a given polymers.

- `Ratio: main/side`: The number of atoms in the main chain divided by the number of atoms in the side chain
- `C4_C4_O2`: The frequency of the three-atom fragment containing two 4-fold coordinated carbon atoms and a two-fold coordinated oxygen atom
- `% ring atoms`: The number of atoms in the rings divided by the total number of atoms
- `block_1`: The number of times that the substructure appears. See Figure 4 of the manuscript for an image of the substructure's molecular graph.
- `single bond frequency`: The frequency of carbon-carbon single bonds divided by the total number of atoms

- `4-vertex carbon: main`: The number of 4-vertex carbons in the main chain divided by the total number of atoms in the main chain
- `Is polyamide?`: Is the polymer a polyamide?
- `Is polyacrylate?`: Is the polymer a polyacrylate?
- `Is polyacrylate?`: Is the polymer a polyacrylate?
- `len. largest side chain`: The length of the largest, by number of atoms, side chain
- `3-vertex carbon: side`: The number of 3-vertex non-ring carbon atoms in largest side chain
- `3-vertex carbon: main`: The number of 3-vertex carbons in the main chain divided by the total number of atoms in the main chain

# References


(1) Kim, C.; Batra, R.; Chen, L.; Tran, H.; Ramprasad, R. Polymer design using genetic algorithm and machine learning. *Computational Materials Science* **2021**, *186*, 110067.

(2) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50*, 742–754.