Pareto Optimization of Oligomer Polarizability and Dipole Moment using a Genetic Algorithm

Danielle C. Hiener[†] and Geoffrey R. Hutchsion[*,†,‡]

†*Department of Chemistry, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, United States*

‡*Department of Chemical and Petroleum Engineering, University of Pittsburgh, 3700 O'Hara Street, Pittsburgh, PA 15261*

E-mail: geoffh@pitt.edu

Abstract

High performance electronic components are highly sought after in order to produce increasingly smaller and cheaper electronic devices. Drawing inspiration from inorganic dielectric materials, in which both polarizability and polarization contribute, organic materials can also maximize both. For a large set of small molecules drawn from PubChem, a Pareto-like front appears between polarizability and dipole moment indicating the presence of an apparent trade-off between these two properties. We tested this balance in π-conjugated materials by searching for novel conjugated hexamers with simultaneously large polarizabilities and dipole moments with potential use for dielectric materials. Using a genetic algorithm (GA) screening technique in conjunction with an approximate density functional tight binding method (GFN2-xTB) for property calculations, we were able to efficiently search chemical space for optimal hexamers. Given the scope of chemical space, using the GA technique saves considerable time and resources by speeding up molecular searches compared to a systematic search. We also explored the underlying structure-function relationships, including sequence and monomer properties, that characterize large polarizability and dipole moment regimes.

Introduction

The impact of electronics on twenty-first century life is hard to overstate, as they drive everything from personal communication devices to battery-powered automobiles. In the current age of rapid technological advancement, electronic components are continuously improved to facilitate the production of smaller, cheaper, faster, and more efficient products. Component improvement is facilitated in large part by the discovery and use of novel high dielectric materials.[1,2] While inorganic dielectrics often possess both high polarization and high polarizability, organic materials often focus on either high polarizability or high polarization. For example, the well-known organic polymer dielectric, polyvinylidene difluoride (PVDF)

and its copolymers, exhibits a relatively large dielectric constant (ϵ $10-12$) due to polarization due to aligned C-F bonds across the chain and in polar domains.[3–5]By maximizing both properties to yield high dielectric constants in organic materials would be advantageous to the development of high-performance capacitors,[6] transistors,[7] and organic photovoltaics (OPVs),[8] among other applications.[1,2]

Recent computational work has successfully used various strategies, including high throughput screening[9,10] and machine learning,[11,12] to examine and predict polymer dielectric properties. Computational discovery of novel high dielectric materials has employed a variety of methods,[13] often either systematically altering known molecular structures,[14–16] or using some combination of inverse design strategies, including high-throughput screening, evolutionary algorithms, and/or machine learning.[17–19] Similar computational discovery in the related field of polarizable materials has been conducted through the lens of discovery of materials with high refractive indices through similar inverse design approaches, including machine learning[20]and high throughput virtual screening.[21] Other computational studies searching specifically for novel OPV conjugated polymer structures tend to focus structural motifs optimizing properties such as energy levels and band gaps.[22–24]

In this study, we searched for novel high dielectric constant materials, focusing on π conjugated materials because of their high polarizability and proven utility for this application. While much of the previous work in the field of dielectric materials discovery has focused on directly optimizing dielectric constant or the related polarizability, we approached this search from a dual property optimization perspective. Because the macroscopic dielectric constant of a *polar* material is related to both the polarizability and the dipole moment by the Debye equation, we sought molecular candidates which maximized both of these properties simultaneously for integration into intrinsically polar organic materials.

In an isotropic or non-polar medium, the relative dielectric is related to the molecular polarizability per unit volume through the Clausius-Mossotti relation. With a polar solid, the polarization from the dipole moment is included. Thus, as a starting point, we computed

the isotropic polarizability and dipole moment terms from the Debye equation using the semi-empirical tight-binding method GFN2-xTB[25] for the PubChemQC data set.[26]

As shown in Figure 1, there is an absence of species possessing both both high polarizability and high dipole moment terms. Perhaps for a molecule with high polarizability per unit volume, a high permanent dipole moment is unlikely, and for a molecule with high permanent charge separation (e.g., a zwitterion), polarizability would cause an induced dipole moment in opposition. These plausible arguments suggest a potential Pareto trade-off in these known small molecules.

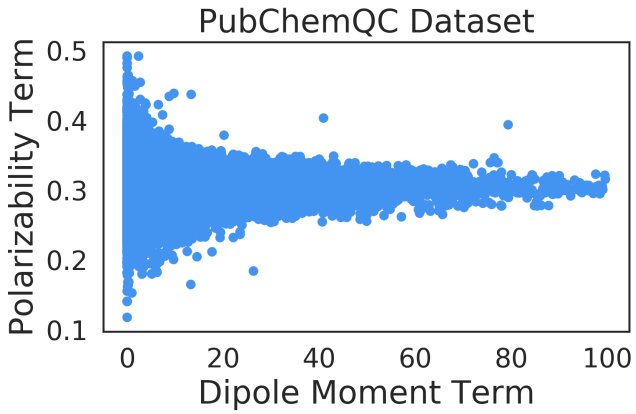


Figure 1: An apparent trade-off is evident between the Debye equation polarizability and dipole moment terms calculated for the PubChemQC data set.

Using this example of the PubChemQC data set as a motivator, we examined whether the perceived Pareto front between the isotropic polarizability and dipole moment terms among small molecules hold for optimized conjugated oligomers. Perhaps a targeted search through a large chemical "design space" can find candidates which simultaneously maximize both properties. To make our search efficient and effective, we employed the principles of inverse design through a genetic algorithm.

Computational Methods

Genetic Algorithm

In the field of computational chemistry, genetic algorithms (GAs) implement inverse design principles by applying concepts from evolutionary biology to find molecular structures with increasingly more desirable features by evolving generations of structures over time.[27,28] In an evolutionary scheme, a population of possible solutions to an optimization problem are generated and then run through selection, crossover, and mutation operators to produce a new population of children.[29] Each successive population is known as a generation and can contain increasingly better solutions; generations are generated until a level of convergence is reached among the top solutions. GAs have been proven as efficient methods for finding molecules with a wide variety of tailored properties in the vastness of chemical space by exploring a small subset.[23,30–34] This is in large part due to the many paths a GA may take through chemical space to a target molecule, and therefore the high probability of quickly finding a species on one of these paths by random chance.[35] GA searches through chemical space have also been found to be faster in some cases than generative machine learning models.[36]

The GA we implemented is capable of searching for oligomers with optimal molecular properties of given specifications: length, number of monomer types, and specific end groups. In this work, we evaluated hexamers containing one or two monomer types, and chose to use one of three pairs of end groups for each run.

The workflow of the GA (Figure 2) begins by initializing the first of each 32-member generation with a randomly-constructed population of oligomer candidates. The selection operator determines each oligomer candidate's property values and uses a fitness function to rank the candidates. The top half proceeds to the next operator, while the bottom half is discarded. The crossover operator takes the top half of the current generation and uses them as "parents" which are combined to produce "children" that replace the previously

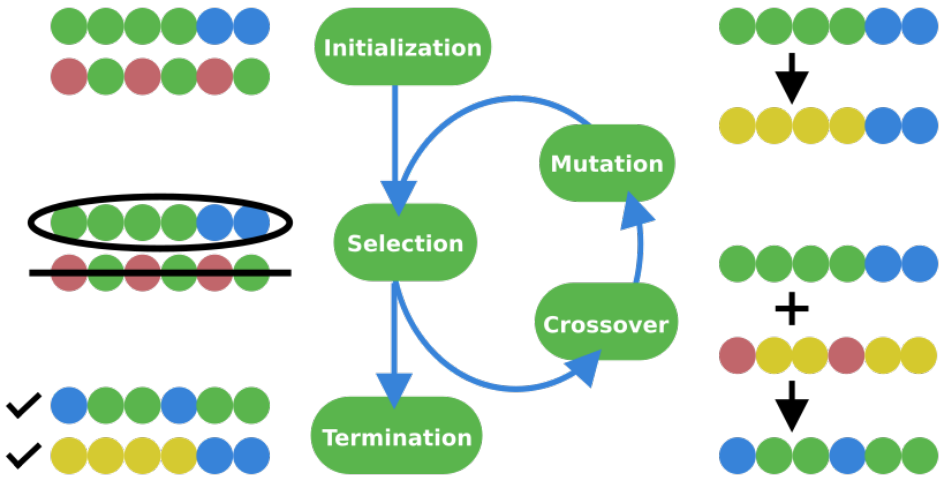


Figure 2: The genetic algorithm (GA) uses operators borrowed from Darwinian evolution to find optimal hexamers, as demonstrated by this simplified schematic.

discarded half of the population. A child is produced by randomly choosing two parents, then randomly selecting monomer species from one or both parents and randomly selecting the sequence of one parent. The mutation operator performs point mutations on a subset of the population determined by a 40% mutation rate. Candidates randomly selected for mutation have either one of their monomer species or their sequence randomly chosen and changed to another randomly selected monomer species or sequence, respectively. After the mutation operation, the cyclic transformation of one generation into another is complete. In each run, the GA performs 400 cycles, after which the data is analyzed to determine whether general convergence has been reached.

Oligomer Composition and Representation

Within the GA, each oligomer candidate is represented as a list containing a tuple of digits indicating its sequence and two integers representing the indices of each of its monomer species from a master list of monomers. For example, [(0,1,0,1,0,1), 10, 20] would be a co-oligomer with an alternating sequence of the monomer at index 10 and the monomer at index 20 in the monomer list.

The monomer data set used by the GA is composed of 1,235 SMILES strings representing various small monomers selected from literature reports and their obvious synthetic modifications.[37] The monomers, chosen because of their likely utility in organic photovoltaics, represent a broad array of aromatic and conjugated species and primarily contain combinations of the elements C, H, N, O, S, and F and are encoded with explicit polymerization sites. All SMILES are included in the supporting information.

The oligomer sequences used by the GA consist of all 64 possible sequences of hexamers made up of one or two monomer species. This includes some redundant possibilities such as 111111 or 000000.

In each run, every candidate is assigned the same endgroups, either amino and nitro, methoxy and cyano, or dimethyl amino and trifluoromethyl. These endgroup pairs were chosen to qualitatively maximize molecular dipole moment along the polymer chain, with each pair containing an electron donating group paired with an electron withdrawing group. Using three different endgroup pairs allowed us to determine if different endgroups produced a substantial effect on the GA search and resulting properties, as discussed below..

Electronic Property Calculations

The three-dimensional structure of each oligomer candidate is geometrically optimized prior to running property calculations, ensuring the most probable physical conformations are represented. After each generation, candidates are converted into SMILES strings, then into a 3D geometry using Open Babel, followed by optimization using MMFF94[38] or UFF,[39] then further optimized using the semi-empirical tight-binding method GFN2-xTB.[25]

Property calculations are also performed with GFN2-xTB, chosen for its time efficiency. As discussed below, comparing the GFN2 method to calculated polarizabilities using the density functional ωB97X in the cc-pVTZ basis set, the GFN2-xTB polarizabilities tend to be an underestimate, likely due to the method's minimum basis set. Our work suggests that this underestimation is fairly systematic and generally increases with increasing π-system

size (see Figures S1 and S2). The relative ordering of molecules based on their polarizability appears to remain similar to the density functional predictions. For this reason, we believe for this study GFN2-xTB provides sufficient electronic property values to search for oligomer candidates with relatively high polarizability and dipole moment terms compared to one another. A more detailed investigation of GFN2-xTB polarizabilities and dipole moments using the large π-conjugated systems from this work is in progress.

Because a single point calculation providing both polarizability and dipole moment values can be performed immediately following geometry optimization in GFN2-xTB, only one job external to the GA script is necessary for each candidate. To increase efficiency, optimized geometries and associated property values are retained after the first time they are calculated. Since each successive generation retains half of the members of the previous generation, after the first generation the number of calculations necessary is no greater than half the population size.

Candidate Evaluation

A key component of the GA selection step is the fitness function, which provides each member of a generation a quantitative score by which it can be ranked. Our premise is to search for oligomers with optimized dielectric constants by simultaneously optimizing their isotropic polarizability and dipole moment terms as given in the Debye equation. This equation modifies the Clausius-Mossotti relation to make it appropriate for polar molecules by including an additional term for the effects of dipole moment. Thus it relates both isotropic polarizability (α) and dipole moment (μ) to the bulk phase dielectric constant (ε_r), where N is the number of molecules, ε_0 is vacuum permittivity, V is the volume, k_b is Boltzman's constant and T is temperature:[40–42]

$$\frac{\varepsilon_r - 1}{\varepsilon_r + 2} = \frac{N}{V}\frac{\alpha}{3\varepsilon_0} + \frac{N}{V}\frac{\mu^2}{9\varepsilon_0 k_b T} \tag{1}$$

As shown by the Debye equation, to maximize the dielectric constant in a polar solid, both polarizability per unit volume and dipole moment must likewise be maximized. In principal, the full polarizability tensor should be used, though the isotropic polarizability will be an underestimate. Moreover the gas-phase molecular dipole moment will be an overestimate of the dipole moment in a polar, polarizable solid, but should correlate well with the latter. Using the sum of the Debye equation polarizability and dipole moment terms then as the scoring mechanism for the GA fitness function, both calculated properties are represented by a single quantitative value, facilitating optimization.

Results and Discussion

The main barrier to generating novel materials for electronic application is finding the best molecular structures in the vastness of chemical space, estimated to include 10^{20} to 10^{60} possible structures.[43] Coupled with the need to run quantum calculations for electronic properties which take minutes to hours per structure, this makes a systematic brute-force search impossible, with a lower time bound on the scale of millions of hours. Like many previous dielectric materials searches, we employed an inverse design strategy to make our search through chemical space efficient and effective. Inverse design is an alternative approach in which target features are determined first, then molecular structures which optimize these features are found. It has proven useful through many implementations which use optimization, sampling, and search procedures to efficiently traverse chemical space to find ideal molecular targets.[44,45] Implementing inverse design using a GA allowed us to perform global optimization without the guarantee of stopping in local extrema traps,[27] it did not require the start-up time of database compilation or machine learning training, it was straightforward to implement, and it provided a proven speed up of 6,000-8,000 times over an exhaustive search method.[37]

Initial Runs

For our initial searches, we performed three 400 generation GA runs. To examine the convergence behavior of the GA, we observed the polarizability and dipole moment terms of the highest scoring candidate from each generation over time for each run (Figure 3). The dipole moment term of the highest scoring candidate gradually increased and appeared to approach asymptotic behavior over 400 generations, while the polarizability term of the same candidate did not have the same increasing trend across the three runs. This is likely due to the difference in order of magnitude of the two terms. The dipole moment term is two to three orders of magnitude larger than the polarizability term and, therefore, dominates the scoring function (the simple sum of these two terms) in the initial runs to favor the dipole moment. The difference in order of magnitude is likely due to evaluating the dipole moment of an individual oligomer molecule, whereas in the bulk phase, the net dipole moment of the material would likely be less.

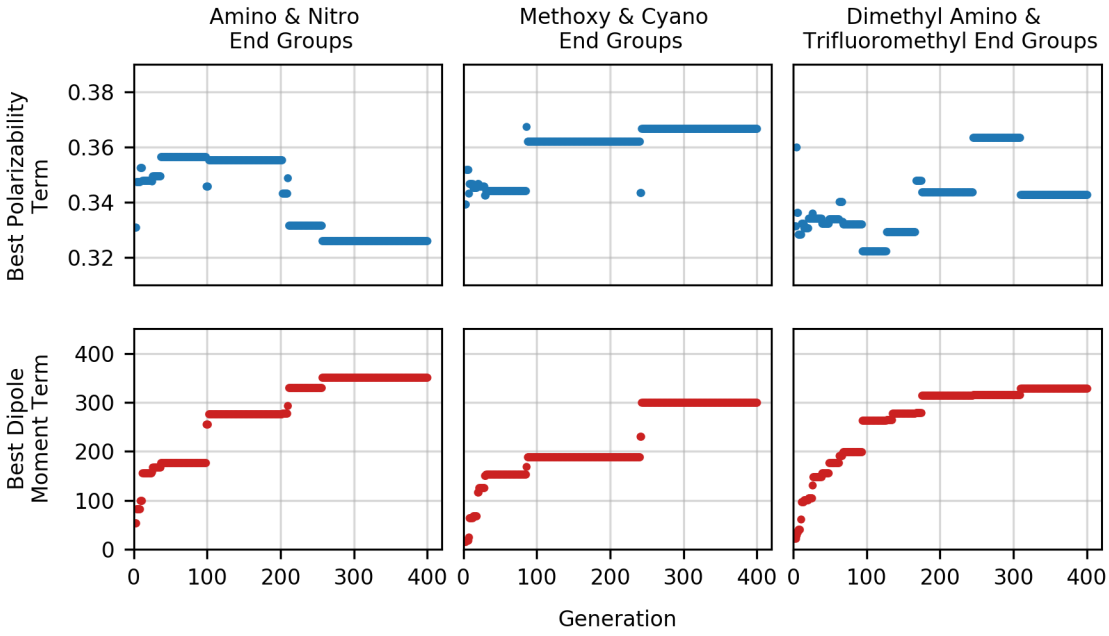


Figure 3: Initial 400 Generation Runs: The top row of plots shows the Debye equation polarizability term value for the best scoring candidate for each generation. The bottom row of plots shows the Debye equation dipole moment term for the same best scoring candidate for each generation.

While the three runs with different end groups yielded slightly different results, all produce a similar shape when the Debye equation polarizability term of each member of each generation was plotted against the dipole moment term of each member of each generation (Figure 4). In the top row of Figure 4, the points representing candidates are rendered translucent to display the scale of the quantity of data. The bottom row of the figure shows the data color-coded by generation, where earlier generations are darker and later generations are lighter. While even in the most recent generations, the candidates found vary greatly in property values, over time it is possible to see the GA pushing further into the search space, especially in maximizing the dipole moment which dominated the scoring of these initial runs.

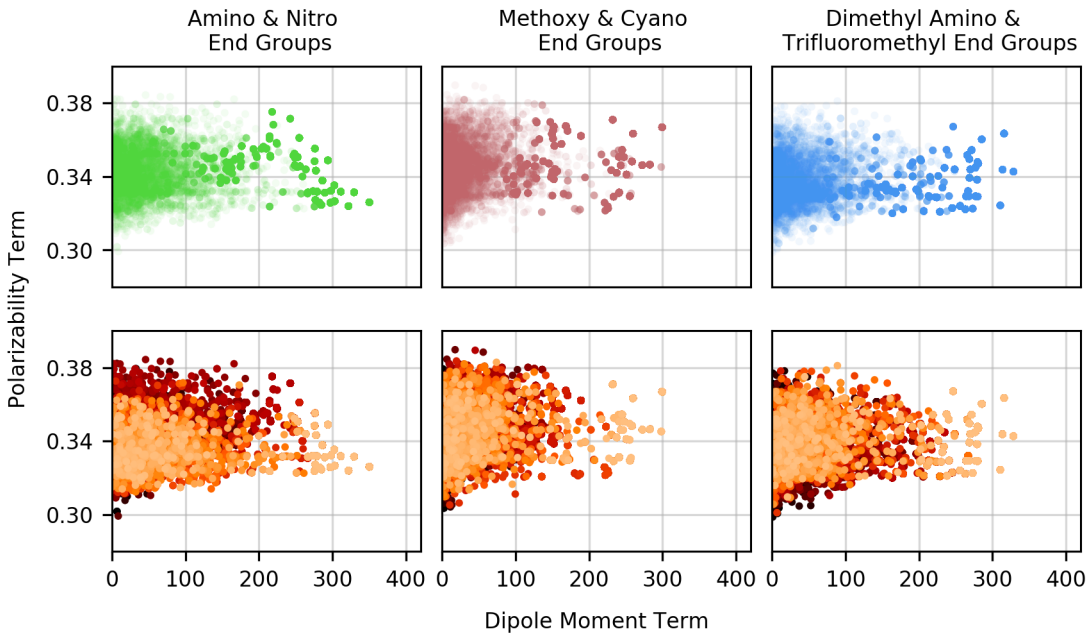


Figure 4: Initial 400 Generation Runs: In the top row, each point represents a candidate's polarizability and dipole moment terms from the Debye equation. In bottom row the same data is color coded to show the age of various candidates where darker is older and lighter is newer.

Among all three of the runs, no Pareto front trade-off was immediately obvious in the plots of polarizability term vs. dipole moment term. To analyze these plots, we created a binning scheme in which each plot was divided into six equally sized bins. The bins were

created relative to the data itself, using the median polarizability term value as the horizontal division between the three upper and three lower bins and placing two vertical divisions in equal thirds between 0 and the maximum dipole moment term value. The bins were labeled left to right C, B, A above the median line and left to right F, E, D below the median line. This labelling scheme emphasizes our interest in the top right-hand bin, A, where candidates have the largest simultaneous polarizability and dipole moment terms. Setting the bins relative to each plot's data allowed us to monitor the region where both properties are maximized, regardless of the size of the search space the GA explored at any given point in its run.

In early generations, the vast majority of polymer candidates found did not fall into bin A, where both terms of the Debye equation are maximized. This seemed consistent with the presence of a potential Pareto front in the data. By the 400th generation, however, all three runs showed multiple candidates in the top bin A, indicating that the GA was able to break through the perceived front and find candidates optimized to its criteria. The progression of both the size and content of the bins in the run with dimethyl amino and trifluoromethyl endgroups is shown as an example in Figure S3.

Focusing on the best regions (bin A) of the initial run plots at the end of 400 generations, we examined the trends in popular sequences and monomers. Of the 64 possible sequences available to the GA for hexamers composed of one or two monomer species, only four or five sequences were present in any of the runs' bin A regions (Figure S6). The top three sequences which accounted for the majority of candidates in each of the bin A regions were all "segregated", with one monomer species on one side of the polymer and the other monomer species on the other side. The limited number of sequences and the similarity of the most numerous sequences in bin A across end groups suggests that sequence plays a strong role in determining the value of a candidate's properties. Because the dipole moment term was favored in the scoring mechanism of these runs, the popularity of segregated monomer species supports the naïeve design principle that separation of monomer species aids in separation

of partial charges to form a strong permanent dipole.

Due to the large number of individual monomer species found in candidates in the bin A region, the top ten most numerous monomer species were analyzed and found to share similarities (Figure S4). In all three runs, one monomer species accounted for 40% of all monomer instances. Of the 30 total "top" monomer species across the three initial runs, 12 were shared top monomers between one or more of the runs, and of those, three top monomers were found in all three runs. These existence of common top performers in multiple independent runs supports the GA's ability to quickly and consistently locate candidate components which optimize the desired properties.

Among the top monomer species, two distinct groups were found: those that preferentially occur on the side of candidates with an electron donating end group (amino, methoxy, or dimethyl amino), and those that preferentially occur on the side with an electron withdrawing end group (nitro, cyano, or trifluoromethyl). The monomer species themselves fall into the same category as the end group toward which they gravitate (Figure 5). Paired with the prevalence of monomer species segregation among sequences, this reinforces the design principle that separating donor and acceptor monomer types reinforces the molecule's permanent dipole moment.

Eleven candidates were found in one or more of the bin A regions of the initial runs, differing only by the end groups specific to each run (Table S3). These candidates reinforce the desirable qualities of segregated monomer species and grouping electron donating and withdrawing groups (Figure S5). Finding common candidates structures varying only by end groups indicates that specific end groups do not have as great an effect on the electronic properties of polymer candidates as do their constituent monomer species and sequences. This also indicates that the GA is effective at finding top candidates given specific properties to optimize, since it is capable of finding the same top candidates in independent runs. Given the tiny proportions of candidates found in the bin A region of each run as compared with the total number of candidates found, as well as the even smaller proportion of candi-

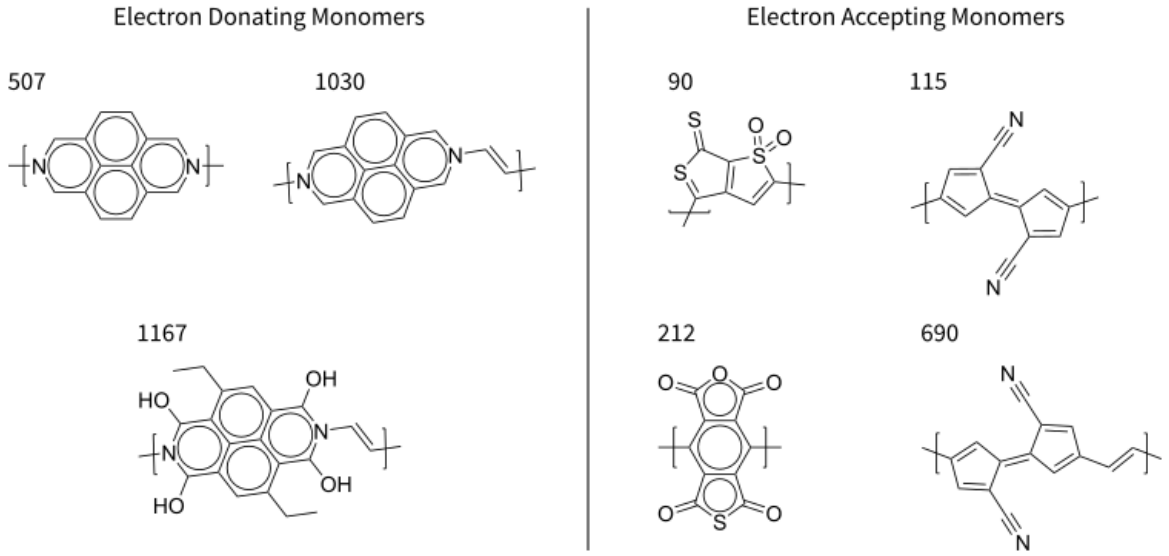


Figure 5: Common monomer species found among top candidates from initial GA runs.

dates found by each run compared to the total possible structures in the search space, the GA's effectiveness at implementing the inverse design approach to quickly find proven best solutions is further supported (Table S1).

Reweighting the Fitness Function

After observing the difference in magnitude between Debye equation polarizability and dipole moment terms, we performed two sets of runs. In both sets, the dipole moment term was re-weighted in the fitness function to explore the effect on the generated candidates. Each run was assigned one of the same end group pairs used in the initial experiments, and similar analysis procedures were performed on these data sets to compare their results to those of the initial runs. Note that the Debye equation terms were only weighted in the scoring mechanism of the fitness function to guide the GA's search, and the terms presented in the following plots and analysis are unweighted.

Equalized Scoring Mechanism Runs

In the first set of weighted runs, the dipole moment term was downweighted by multiplying by a coefficient of 0.389/111, to roughly equalize the effects of both terms on the score assigned to candidates by the fitness function. This coefficient was reached by dividing the largest polarizability term by the largest dipole moment term that had been observed up to that point in the data collection. While larger terms were later observed in subsequent generations, the orders of magnitude for the largest terms remained the same, so this coefficient remained valid.

Observing the polarizability and dipole moment terms of the highest scoring candidate from each generation over time for each run, the top candidate's dipole moment term again generally increased with increasing generation, albeit more quickly (Figure 6). The top candidate's polarizability term also lacked a specific trend among runs for about the first hundred generations. Unlike the initial runs, however, it established a trend of maintaining or increasing during the last 300 generations across all runs which supported the reasoning that re-weighting the dipole term in the fitness function allows both properties to more equally contribute to top candidate scores. Both the top polymer candidate polarizability and dipole moment terms also reached plateaus that remained constant for at least the last 200 generations on all three runs, suggesting that the GA converged faster than in the initial runs.

When comparing the polarizability and dipole moment terms of all oligomer candidates found over the generations, the trends of each of the three runs were similar (Figure 7). The bottom row of the figure with data color-coded by generation shows that even in the most recent generations the candidates found vary greatly in property values. Again, over time it is possible to see the GA pushing further into the search space.

No Pareto front trade-off was immediately obvious between the polarizability term and the dipole moment term. The same binning scheme as in the initial runs was used for analysis. Very few polymer candidates fell into bin A in the initial generations, while later

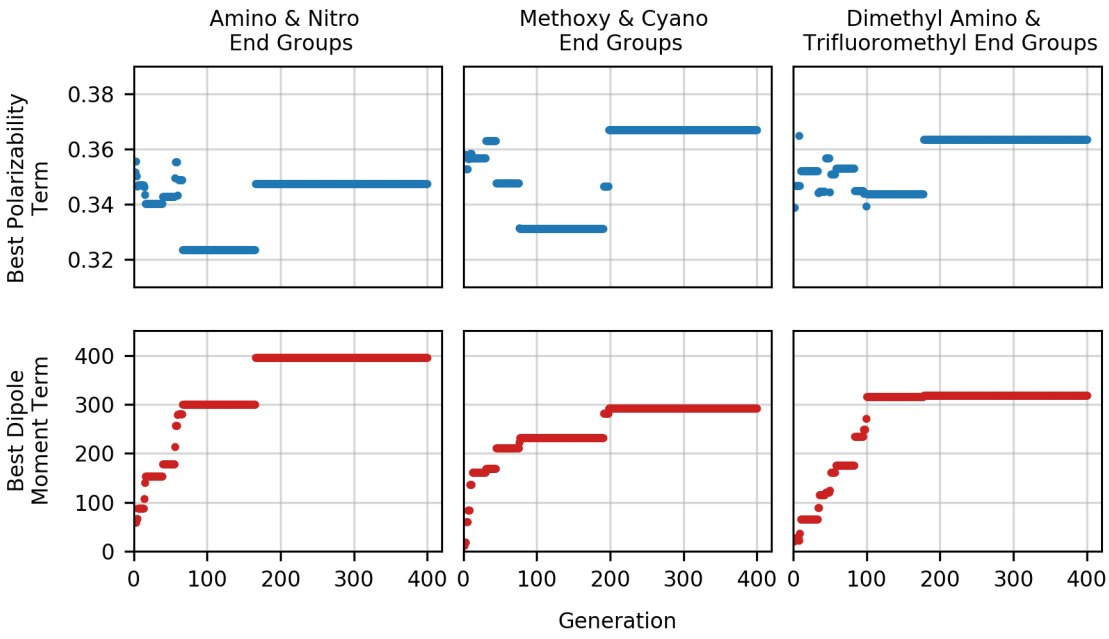


Figure 6: Equalized Scoring Mechanism 400 Generation Runs: The top row of plots shows the Debye equation polarizability term value for the best scoring candidate for each generation. The bottom row of plots shows the Debye equation dipole moment term for the same best scoring candidate for each generation.

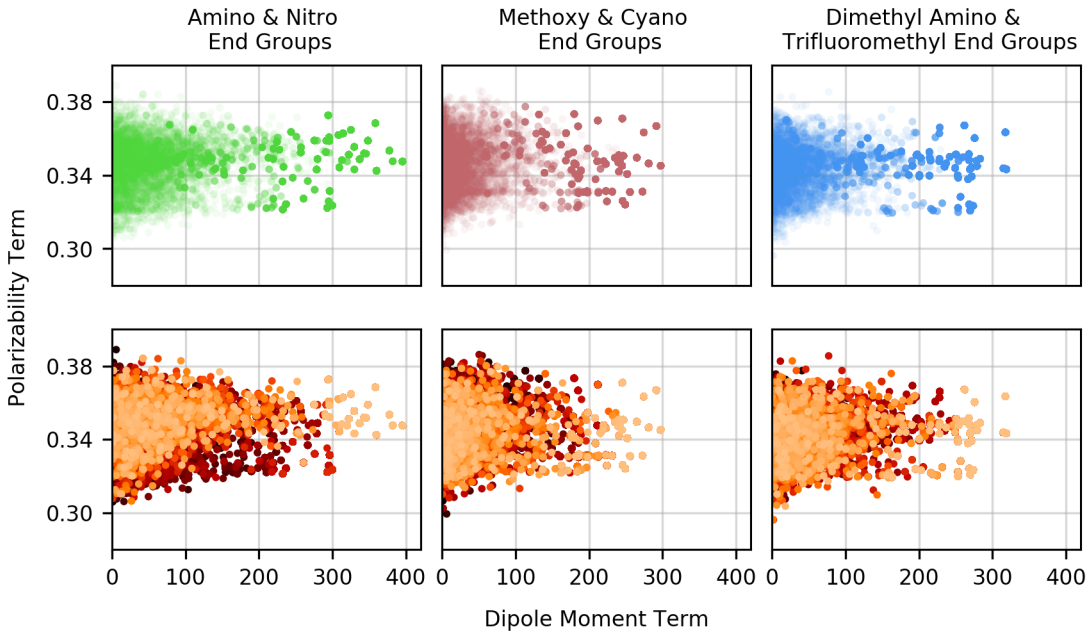


Figure 7: Equalized Scoring Mechanism 400 Generation Runs: In the top row, each point represents a candidate's polarizability and dipole moment terms from the Debye equation. In the bottom row the same data is color coded to show the age of candidates where darker is older and lighter is newer.

generations showed multiple candidates in bin A. This indicates that the GA was able to maximize both the polarizability and dipole moment terms. The progression of both the size and content of the bins in the run with dimethyl amino and trifluoromethyl endgroups for the equalized scoring mechanism runs is shown as an example in Figure S7.

We examined the trends in popular sequences in monomers in the bin A regions of the equalized scoring mechanism plots after 400 generations. Only four of the possible 64 sequences were represented in bin A in each of the runs (Figure S10). The top three sequences accounting for the majority of the bin A populations were segregated by monomer species. This suggests that even when both property terms are weighted equally in the scoring mechanism, the separation of donor and acceptor monomer species, seemingly beneficial to increasing the polymer candidate's permanent dipole moment, is preferred.

The top ten most numerous monomer species found in the top region of the equalized scoring mechanism plots were again found to share similarities (Figure S8). All three equalized scoring mechanism plots had a clear top performing monomer species, which was the same species across all three runs. Within the 30 total "top" monomer species across the three runs, seven are shared top monomers between one or more of the runs, and of those, three top monomers are found in all three runs.

Top monomer species can be divided into those which are electron donating and those which are electron accepting, preferentially occurring on the side of polymer candidates with the same respective type of end group (Figure 8). Additionally, the equalized scoring mechanism bin A sequences favor monomer species segregation suggesting that for both maximized polarizability and dipole moment, separating partial charges is an advantage.

Eleven candidates were found in one or more of the bin A regions of the initial runs, with three candidates occurring in all three runs, all differing only by the end groups specific to each run (Table S4). These common candidates reinforce the desirable qualities of segregated monomer species and grouping electron donating and withdrawing groups (Figure S9). The number of unique polymer candidates found in bin A compared to the total found in each of

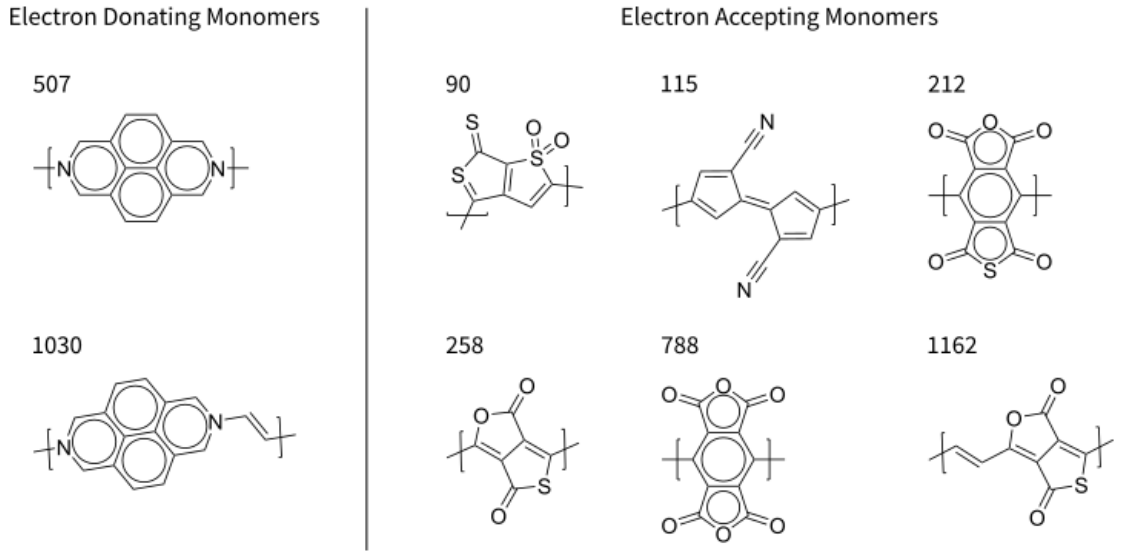


Figure 8: Common monomer species found among top candidates from equalized scoring mechanism runs.

the runs again reinforces the GA's ability to find the top candidates consistently (Table S2).

Five very similar candidates were found in two different runs in both the initial set of runs and the equalized scoring mechanism runs (Table S5). The structure of these oligomers and their presence as top candidates in both sets of runs suggests that the physical and chemical properties supporting a strong permanent dipole moment in a polymer also naturally support large polarizabilities.

Polarizability Favored Runs

In a second set of re-weighted runs, the dipole moment term was decreased by three orders of magnitude below that of the polarizability term. In this case, the dipole moment term was multiplied by a coefficient of $(0.389/111)^2$, reflecting the largest polarizability and dipole moment terms observed by that point in the data collection as discussed above. For these "polarizability favored" runs, the polarizability and dipole moment terms of the highest scoring candidates for each generation were observed over time (Figure 9). The top candidate's dipole moment term was substantially lower in magnitude than either the initial or

equalized scoring mechanism runs. The top candidate's polarizability term increased quickly with increasing generation and reached maximum values greater than in either of the two previous sets of runs. Both of these observations support the weighted scoring mechanism guiding the GA to target the search space where polarizability alone was maximized. Both terms plateaued and remained constant for at least the last two hundred generations of all three runs, suggesting relatively strong convergence behavior.

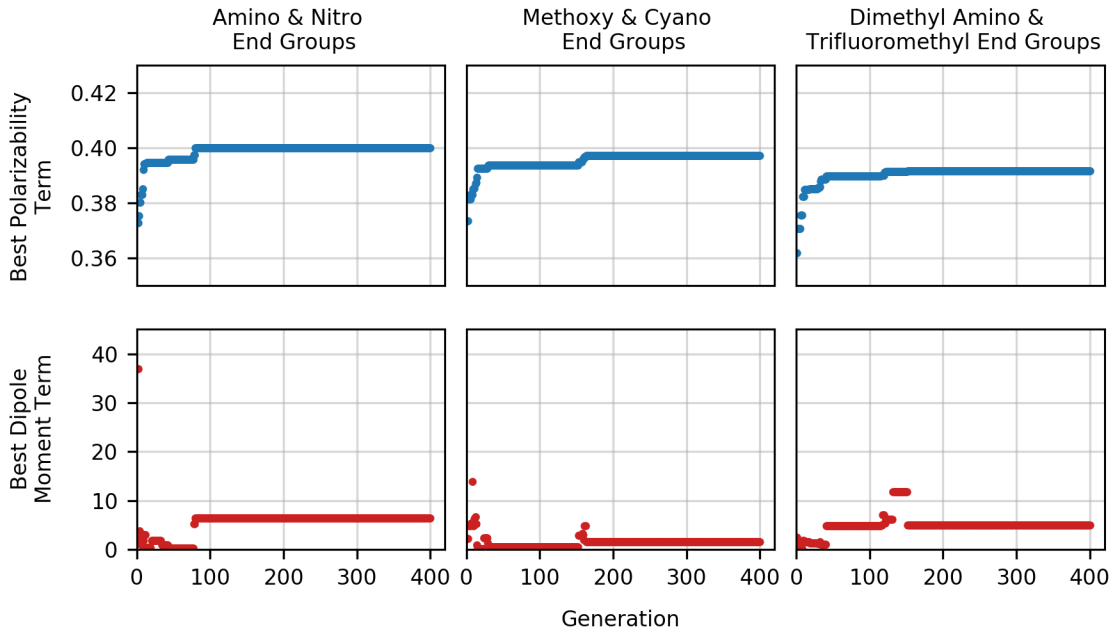


Figure 9: Polarizability Favored 400 Generation Runs: The top row of plots shows the Debye equation polarizability term value for the best scoring candidate for each generation. The bottom row of plots shows the Debye equation dipole moment term for the same best scoring candidate for each generation.

While all three polarizability runs displayed similar results when observing the Debye equation polarizability and dipole moment terms of each member in each generation, they varied significantly from either of the two previous sets of runs (Figure 10). Unlike either of the previous sets of runs, the plots in Figure 10 show only a much small, more concentrated area of data density, suggesting that the GA was able to locate the desirable area of the search space almost immediately. The bottom row with data color-coded by generation shows that even in the most recent generations the candidates found vary widely in polarizability term

values, but comparitively little in dipole moment term values. These plots show that the GA is capable of probing different areas of the search space when tasked with optimizing different properties. These results also suggest that the GA was able to locate candidates with maximized polarizability terms more easily than those with either maximized dipole moment terms or with both property terms maximized, since far less of the search space was probed thoroughly before locating the ideal search area.

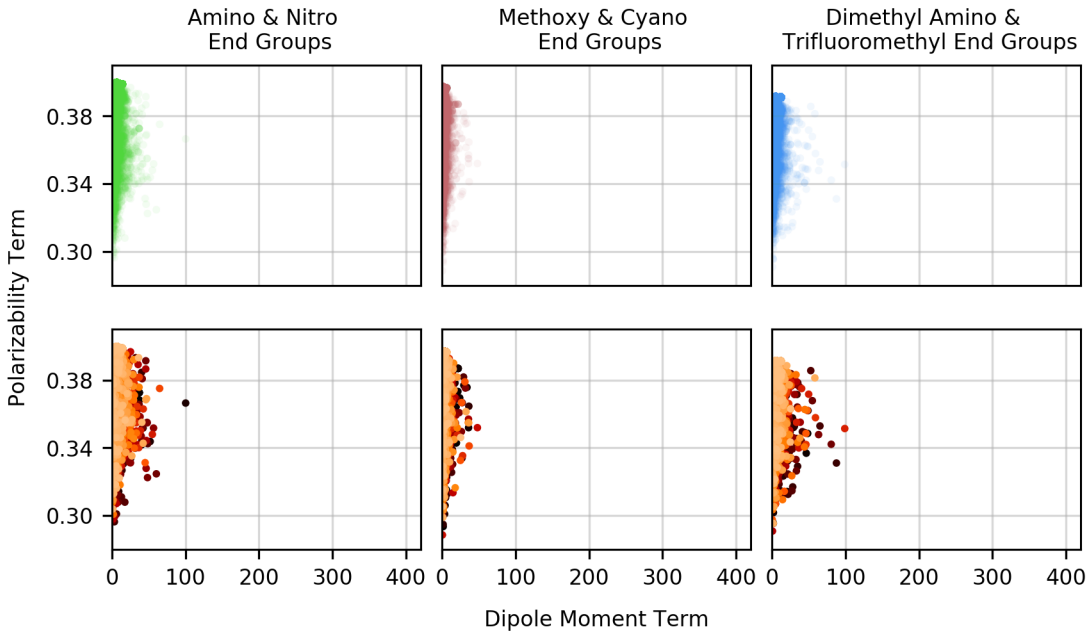


Figure 10: Polarizability Favored 400 Generation Runs: In the top row, each point represents a candidate's polarizability and dipole moment terms from the Debye equation. In the bottom row the same data is color coded to show the age of candidates where darker is older and lighter is newer.

Since the polarizability favored runs had vastly different plot characteristics from either of the previous sets of runs, we used different binning scheme for analysis. For these runs, we divided the plots of Debye equation polarizability term vs. dipole moment term into two bins using a horizontal divider set at the median polarizability term value. Since we were interested only in high polarizability candidates, we focused the analysis on upper bin.

We examined the trends in popular sequences in monomer species in the best regions of the polarizability favored plots after 400 generations. Among the top ten most numerous se-

quences in the upper bin of all three plots, no one sequence clearly dominated (Figure S12). Instead, primarily homopolymer sequences or sequences with only one differing monomer species were most prevalent. This suggests that homopolymer, or near-homopolymer, sequences are the best for maximizing polarizability. (By simple statistics, far more candidates exist with one differing monomer than pure homogeneous sequences.)

The top ten most numerous monomer species found in the upper bin of the polarizability favored plots were found to share similarities (Figure S11). In two of the three runs, there were two clear top performing monomer species. This is likely due to the prevalence of candidates in these runs using the same two monomer species with all of the sequence variations of five of one monomer species, one of the other. Examining the 30 total "top" monomer species across the three runs, 13 are shared top monomers between one or more of the runs, and of those, four top monomers are found in all three runs.

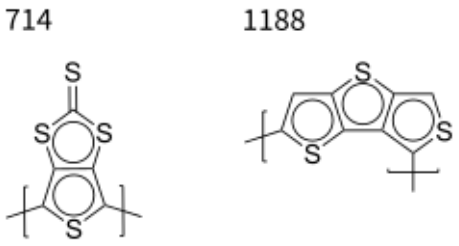


Figure 11: Common monomer species found in top homopolymer and near-homopolymer species from polarizability favored runs.

There was overlap among top candidates found in multiple runs, differing only by their end groups. In all, 215 polymer candidates appeared in two or more runs, and 18 appeared in all three runs. Those common candidates substantially shared monomers, with only 25 unique monomer species found among them.

To further investigate the homopolymer-like trend among top candidates, we constructed all 64 possible sequenced versions of a candidate with amino and nitro end groups and the top two most numerous monomer species (referenced by index as 714 and 98) from that polarizability favored run and calculated their polarizability term values from GFN2-xTB isotropic

polarizability calculations. Sorted from highest to lowest polarizability term value, the homopolymer of monomer species 714 is ranked first, while the homopolymer of monomer species 98 is ranked last. The six polymer candidates with sequences resembling the homopolymer of monomer species 714 with one substitution of monomer species 98 rank 2-7 (Table S6). This provides evidence that polarizability is best maximized by homogeneous sequences, despite the greater synthetic diversity of co-polymers. It also suggests that the reason homopolymer sequences do not always occur with the highest freqeuency in the upper bin is because single substitution near-homopolymer sequences have polarizability term values close to the true homopolymer and are six times more likely to occur by random chance. These sequences therefore appear to occur with near equal frequency to both each other and the true homopolymer.

Conclusions and Future Work

In this study, the GA search for novel conjugated materials was able to break through an apparent Pareto front between polarizability and dipole moment terms of the Debye equation, as seen in the small-molecule PubChemQC data set. Both the initial and equalized scoring mechanism runs performed similarly, in terms of the number and quality of the hexamers found in the region of the search space with simultaneously large polarizability and dipole moment terms. Since the initial runs unintentionally favored the dipole moment term, it is important to note that those runs still found candidates with large polarizability terms in addition to candidates where their dipole moment were maximized at the expense of their polarizability. This is in contrast to runs in which the dipole moment was intentionally down-weighted, which almost exclusively found candidates with large polarizability terms at the expense of their dipole moment. Similarly, candidates with both properties maximized found in runs in which both dipole moment and polarizability terms were roughly equal, had physical properties (segregated sequences, monomer species of both the electron donating

and electron accepting types) which tended to fit the design principles of polymers with maximized dipole moment, rather than with maximized polarizability. This suggests that the polarizability term is more easily (though not automatically) maximized in conjunction with maximizing the dipole moment term, than vice versa. Because the same polymer candidates were found in multiple runs with different end groups in all of the data sets, this suggests that for conjugated hexamers, end groups do not play a large role in determining electronic property value but rather in determining the optimal arrangement of monomer species within the polymer structure.

This work further proved the utility of a genetic algorithm for inverse design of novel materials. The GA successfully found candidates which maximized the Debye equation polarizability term, dipole moment term, or both, proving its ability to explore and target areas within the chemical search space. Additionally, the GA was able to find the same candidates in different independent runs, illustrating the reliability of this method, despite the stochastic nature.

Given the emerging links between structure and electronic properties this study indicates, future work needs to be done examining the underlying chemical phenomena to better understand the link between polarizability and dipole moment when both are maximized simultaneously in conjugated polymeric materials. The role of charge transfer in particular is a concept that can be investigated, potentially with energy decomposition analysis, symmetry adapted perturbation theory (SAPT),[46] or other computational tools, to better understand the relationship between polarizability and electrostatic moments ultimately to design better materials.

Given the general trend that GFN2-xTB tends to substantially underestimate isotropic polarizabilities for molecules with increasingly larger π-systems, a more thorough exploration of the integrity of this method is warranted. An investigation of GFN2-xTB and DFT polarizabilities is currently underway using the set of highly polarizable compounds generated in this work.

While plots of the properties of the best scoring candidates in all of the data sets show convergence, further work to improve the efficiency of genetic algorithm molecular search methods can focus on improving convergence rates, determination of convergence, and combination with other search and generative algorithms.

Acknowledgement

We acknowledge support from Department of Energy-Basic Energy Sciences Computational and Theoretical Chemistry (Award DE-SC0019335) and the University of Pittsburgh Center for Research Computing through the computational resources provided.

References

(1) Ortiz, R. P.; Facchetti, A.; Marks, T. J. High-kOrganic, Inorganic, and Hybrid Dielectrics for Low-Voltage Organic Field-Effect Transistors. *Chemical Reviews* **2010**, *110*, 205–239.

(2) Wang, B.; Huang, W.; Chi, L.; Al-Hashimi, M.; Marks, T. J.; Facchetti, A. High-kGate Dielectrics for Emerging Flexible and Stretchable Electronics. *Chemical Reviews* **2018**, *118*, 5690–5754.

(3) R. Gregorio, J.; Ueno, E. M. Effect of crystalline phase, orientation and temperature on the dielectric properties of poly (vinylidene fluoride) (PVDF). *Journal of Materials Science* **1999**, *34*, 4489–4500.

(4) Dong, R. First-principles simulations of PVDF copolymers with high dielectric energy density: PVDF-HFP and PVDF-BTFE. *Physical Review B* **2016**, *94*.

(5) Sharma, M.; Madras, G.; Bose, S. Process induced electroactive β-polymorph in PVDF: effect on dielectric and ferroelectric properties. *Physical Chemistry Chemical Physics* **2014**, *16*, 14792.

(6) Chen, Q.; Shen, Y.; Zhang, S.; Zhang, Q. M. Polymer-Based Dielectrics with High Energy Storage Density. *Annu. Rev. Mater. Res.* **2015**, *45*, 433–458.

(7) Wang, Y.; Huang, X.; Li, T.; Li, L.; Guo, X.; Jiang, P. Polymer-Based Gate Dielectrics for Organic Field-Effect Transistors. *Chem. Mater.* **2019**, *31*, 2212–2240.

(8) Brebels, J.; Manca, J. V.; Lutsen, L.; Vanderzande, D.; Maes, W. High dielectric constant conjugated materials for organic photovoltaics. *J. Mater. Chem. A* **2017**, *5*, 24037–24050.

(9) Huan, T. D.; Mannodi-Kanakkithodi, A.; Kim, C.; Sharma, V.; Pilania, G.; Ram-

prasad, R. A polymer dataset for acceleratedproperty prediction and design. *Sci. Data* **2016**, *3*.

(10) Mannodi-Kanakkithodi, A.; Huan1, T. D.; Ramprasad, R. Mining Materials Design Rules from Data: The Example of Polymer Dielectrics. *Chem. Mater.* **2017**, *29*, 9001–9010.

(11) Mannodi-Kanakkithodi, A.; Pilania, G.; Ramprasad, R. Accelerating materials property predictions using machine learning. *Sci. Rep.* **2013**, *3*, 2810.

(12) Mannodi-Kanakkithodi, A.; Pilania, G.; Ramprasad, R. Critical assessment of regression-based machine learning methods for polymer dielectrics. *Comput. Mater. Sci.* **2016**, *125*, 123–135.

(13) Wang, C. C.; Pilania, G.; Boggs, S. A.; Kumar, S.; Breneman, C.; Ramprasad, R. Computational strategies for polymer dielectrics design. *Polymer* **2014**, *55*, 979–988.

(14) Heitzer, H. M.; Marks, T. J.; Ratner, M. A. Computation of Dielectric Response in Molecular Solids for High Capacitance Organic Dielectrics. *Acc. Chem. Res.* **2016**, *49*, 1614–1623.

(15) Dyck, C. V.; Marks, T. J.; Ratner, M. A. Chain Length Dependence of the Dielectric Constant and Polarizability in Conjugated Organic Thin Films. *ACS Nano* **2017**, *11*, 5970–5981.

(16) Armin, A.; Stoltzfus, D. M.; Donaghey, J. E.; Clulow, A. J.; Nagiri, R. C. R.; Burn, P. L.; Gentle, I. R.; Meredith, P. Engineering dielectric constants in organic semiconductors. *J. Mater. Chem. C* **2017**, *5*, 3736–3747.

(17) Choudhary, K.; Garrity, K. F.; Sharma, V.; Biacchi, A. J.; Walker, A. R. H.; Tavazza, F. High-throughput density functional perturbation theory and machine learning predic-

tions of infrared, piezoelectric, and dielectric responses. *npj Comput. Mater.* **2020**, *6*, 64.

(18) Mannodi-Kanakkithodi, A.; Pilania, G.; Huan, T. D.; Lookman, T.; Ramprasad, R. Machine Learning Strategy for Accelerated Design of Polymer Dielectrics. *Sci. Rep.* **2016**, *6*, 20952.

(19) Sharma, V.; Wang, C.; Lorenzini, R. G.; Ma, R.; Zhu, Q.; Sinkovits, D. W.; Pilania, G.; Oganov, A. R.; Kumar, S.; Sotzing, G. A. et al. Rational design of all organic polymer dielectrics. *Nat. Commun.* **2014**, *5*, 4845.

(20) Haghighatlari, M.; Vishwakarma, G.; Afzal, M. A. F.; Hachmann, J. *A Physics-Infused Deep Learning Model for the Prediction of Refractive Indices and Its Use for the Large-Scale Screening of Organic Compound Space*; 2019.

(21) Afzal, M. A. F.; Haghighatlari, M.; Ganesh, S. P.; Cheng, C.; Hachmann, J. Accelerated Discovery of High-Refractive-Index Polyimides via First-Principles Molecular Modeling, Virtual High-Throughput Screening,and Data Mining. *J. Phys. Chem. C* **2019**, *123*, 14610–14618.

(22) Risko, C.; McGehee, M. D.; Bredas, J.-L. A quantum-chemical perspective into low optical-gap polymers for highly-efficient organic solar cells. *Chem. Sci.* **2011**, *2*, 1200–1218.

(23) Kanal, I. Y.; Owens, S. G.; Bechtel, J. S.; Hutchison, G. R. Efficient Computational Screening of Organic Polymer Photovoltaics. *J. Phys. Chem. Lett.* **2013**, *4*, 1613–1623.

(24) Zhuang, W.; Lundin, A.; Andersson, M. R. Computational modelling of donor–acceptor conjugated polymers through engineered backbone manipulations based on a thiophene–quinoxaline alternating copolymer. *J. Mater. Chem. A* **2014**, *2*, 2202–2212.

(25) Bannwarth, C.; Ehlert, S.; Grimme, S. *GFN2-xTB - an Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions*; 2018.

(26) Nakata, M.; Shimazaki, T. PubChemQC Project: A Large-Scale First-Principles Electronic Structure Database for Data-Driven Chemistry. *J. Chem. Inf. Model.* **2017**, *57*, 1300–1308.

(27) Leardi, R. Genetic algorithms in chemometrics and chemistry: a review. *J. Chemometrics* **2001**, *15*, 559–569.

(28) Supady, A.; Blum, V.; Baldauf, C. First-Principles Molecular Structure Search with a Genetic Algorithm. *J. Chem. Inf. Model.* **2015**, *55*, 2338–2348.

(29) Forrest, S. Genetic algorithms: principles of natural selection applied to computation. *Science* **1993**, *261*, 872–878.

(30) McGarrah, D. B.; Judson, R. S. Analysis of the Genetic Algorithm Method of Molecular Conformation Determination. *J. Comput. Chem.* **1993**, *14*, 1385–1395.

(31) Kanters, R. P. F.; Donald, K. J. CLUSTER Searching for Unique Low Energy Minima of Structures Using a Novel Implementation of a Genetic Algorithm. *J. Chem. Theory Comput.* **2014**, *10*, 5729–5737.

(32) Curtis, F.; Li, X.; Rose, T.; Vazquez-Mayagoiti, Á.; Bhattacharya, S.; Ghiringhelli, L. M.; Marom, N. GAtor: A First Principles Genetic Algorithm for Molecular Crystal Structure Prediction. *J. Chem. Theory Comput.* **2018**, *14*, 2246–2264.

(33) Capecchi, A.; Zhang, A.; Reymond, J.-L. Populating Chemical Space with Peptides Using a Genetic Algorithm. *J. Chem. Inf. Model.* **2020**, *60*, 121–132.

(34) Verhellen, J.; der Abeele, J. V. Illuminating elite patches of chemical space. *Chem. Sci.* **2020**, *11*, 11485–11491.

(35) Jensen, J. H. A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space. *Chem. Sci.* **2019**, *10*, 3567–3572.

(36) Henault, E. S.; Rasmussen, M. H.; Jensen, J. H. Chemical space exploration: how genetic algorithms find the needle in the haystack. *PeerJ Physical Chemistry* **2020**, *2*, e11.

(37) Kanal, I. Y.; Hutchison, G. R. *Rapid Computational Optimization of Molecular Properties using Genetic Algorithms: Searching Across Millions of Compounds for Organic Photovoltaic Materials*; 2017.

(38) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490–519.

(39) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; III, W. A. G.; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024–10035.

(40) Debye, P. *Polar Molecules*; Chemical Catalog Co., 1929.

(41) Bottcher, C. J. F. *Theory of Electric Polarization*, 2nd ed.; Elsevier, 1973; Vol. 1.

(42) Dang, Z.-M. *Dielectric Polymer Materials for High-Density Energy Storage*; William Andrew, 2018.

(43) van Deursen, R.; Reymond, J.-L. Chemical Space Travel. *ChemMedChem* **2007**, *2*, 636–640.

(44) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018**, *361*, 360–365.

(45) Freeze, J. G.; Kelly, H. R.; Batista, V. S. Search for Catalysts by Inverse Design: Artificial Intelligence, Mountain Climbers, and Alchemists. *Chem. Rev.* **2019**, *119*, 6595–6612.

(46) Stasyuk, O. A.; Sedlak, R.; Guerra, C. F.; Hobza, P. Comparison of the DFT-SAPT and Canonical EDA Schemes for the Energy Decomposition of Various Types of Non-covalent Interactions. *J. Chem. Theory Comput.* **2018**, *14*, 3440–3450.

Supplementary Information

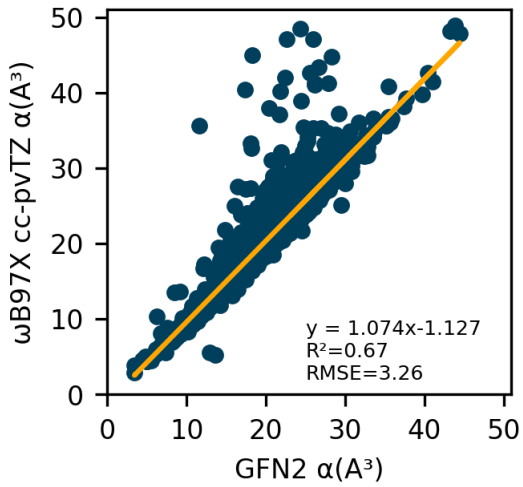


Figure S1: Comparison of isotropic polarizabilities calculated with GFN2 and DFT ωB97X cc-pVTZ for 8415 PubChemQC molecules.

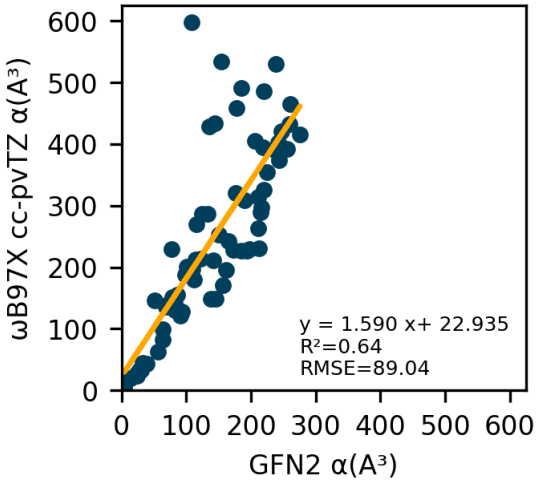


Figure S2: Comparison of isotropic polarizabilities calculated with GFN2 and DFT ωB97X cc-pVTZ for 70 conjugated species.

Table S1: Number of Candidates Found in Initial 400 Generation Runs

End Group Run	Maximum Candidates Generated by a Random Search in 400 Generations*	Unique Polymer Candidates Found	Unique Polymer Candidates Found in Bin A
Amino/Nitro	12800	4676	20
Methoxy/Cyano	12800	4737	18
Dimethyl Amino/ Trifluoromethyl	12800	4471	24

Table S2: Number of Candidates Found in Equalized Scoring Mechanism 400 Generation Runs

End Group Run	Maximum Candidates Generated by a Random Search in 400 Generations*	Unique Polymer Candidates Found	Unique Polymer Candidates Found in Bin A
Amino/Nitro	12800	4669	21
Methoxy/Cyano	12800	4542	19
Dimethyl Amino/ Trifluoromethyl	12800	4434	27

*Based on a maximum of 32 candidates per generation over a 400 generation simple random search.

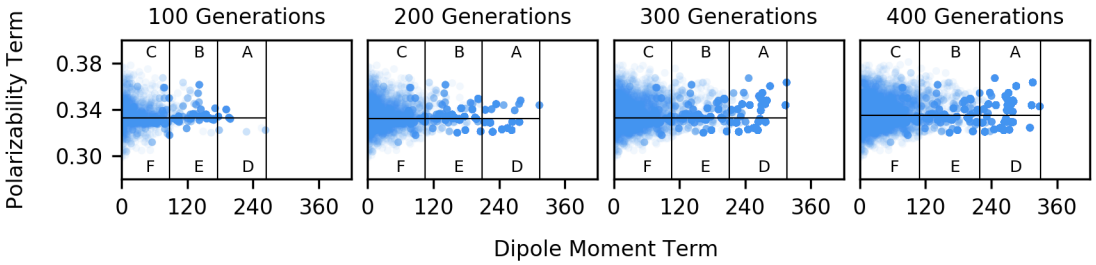


Figure S3: Initial 400 Generation Run with Dimethyl Amino and Trifluoromethyl End Groups: Bin size and population progression is seen with snapshots every 100 generations.

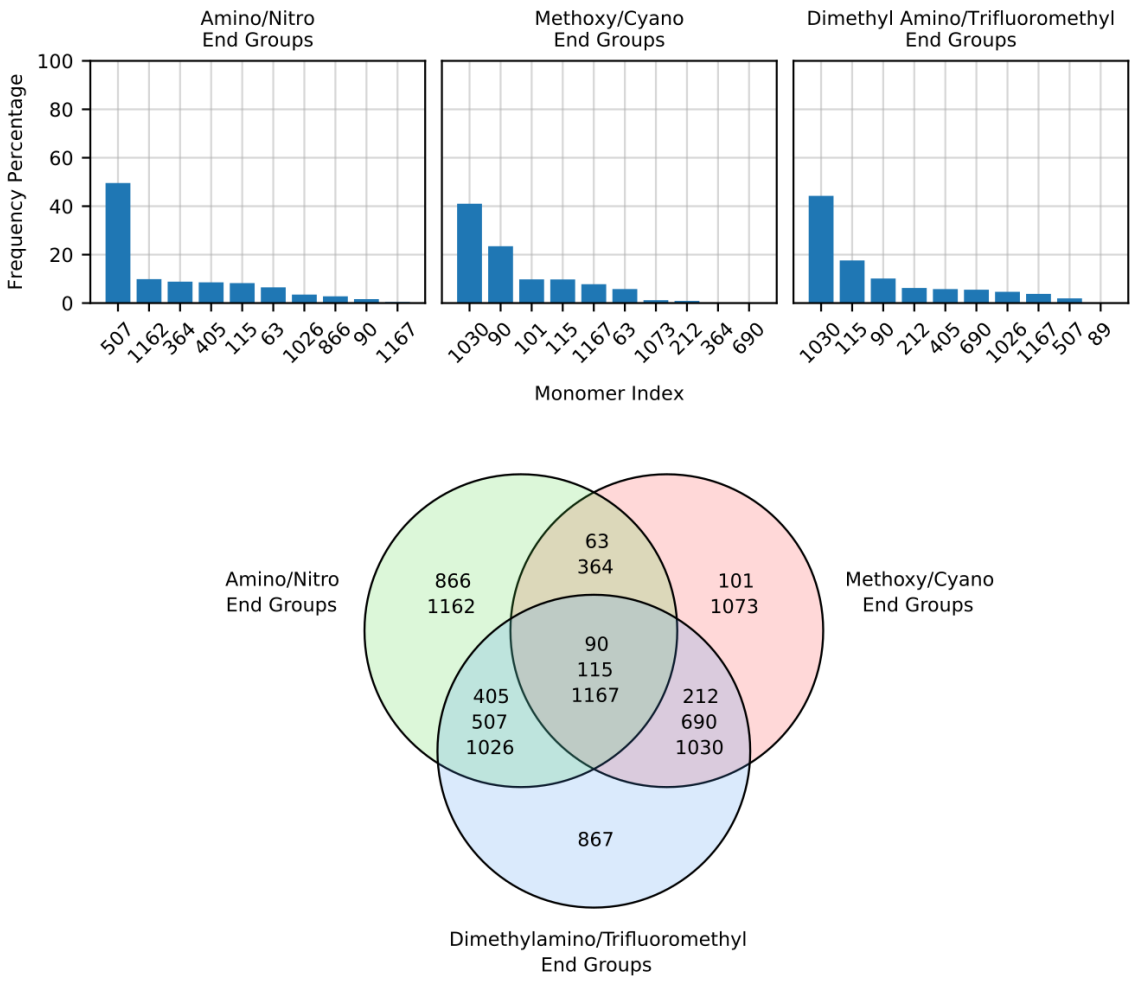


Figure S4: Initial 400 Generation Runs: The top set of plots shows frequency percentages of the top ten most numerous monomer species (identified by index in monomer set) from polymer candidates found in the bin A region of polarizability term vs dipole moment term plots. The bottom Venn diagram displays the overlap in top monomer species between independent runs.

Table S3: Common Polymer Candidates Found in Multiple Initial Runs

Polymer Candidate			Runs Containing Polymer Candidate		
Monomer 0	Monomer 1	Sequence	Amino/ Nitro	Methoxy/ Cyano	Dimethylamino/ Trifluoromethyl
507	115	000111	X		X
1167	90	000111	X	X	X
1030	90	000111		X	X
1030	115	000111		X	X
1030	90	001111		X	X
1030	90	000011		X	X
1030	115	000011		X	X
1167	90	001111		X	X
1030	212	000111		X	X
1030	212	000011		X	X
1030	690	000011		X	X

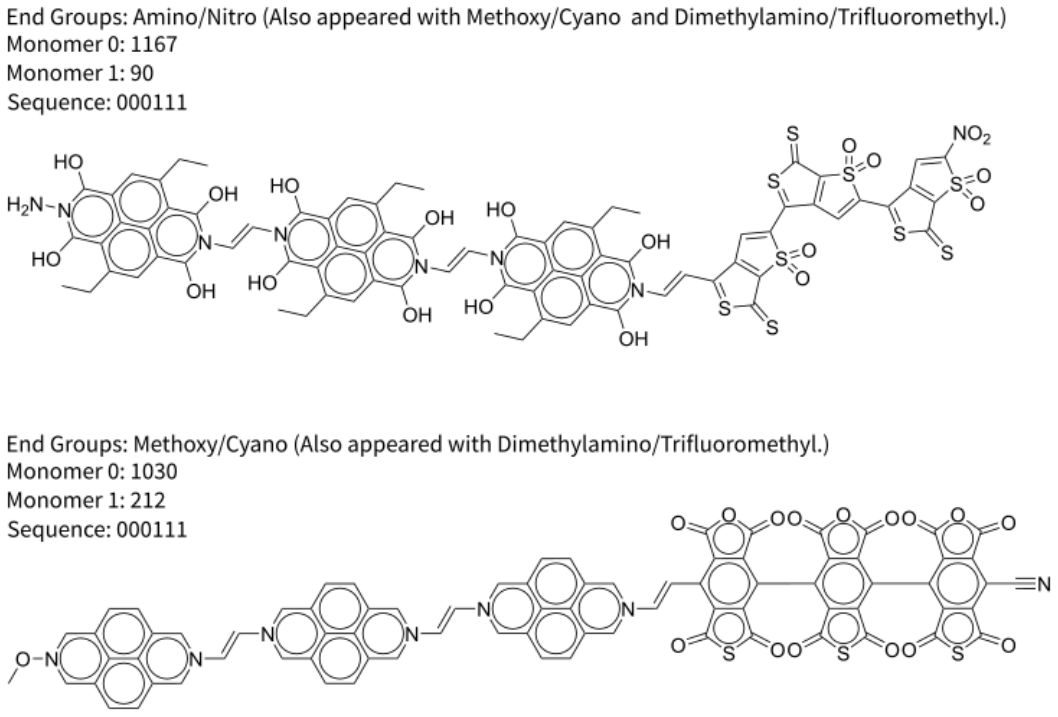


Figure S5: Common polymer candidates found in the bin A regions of multiple initial runs.

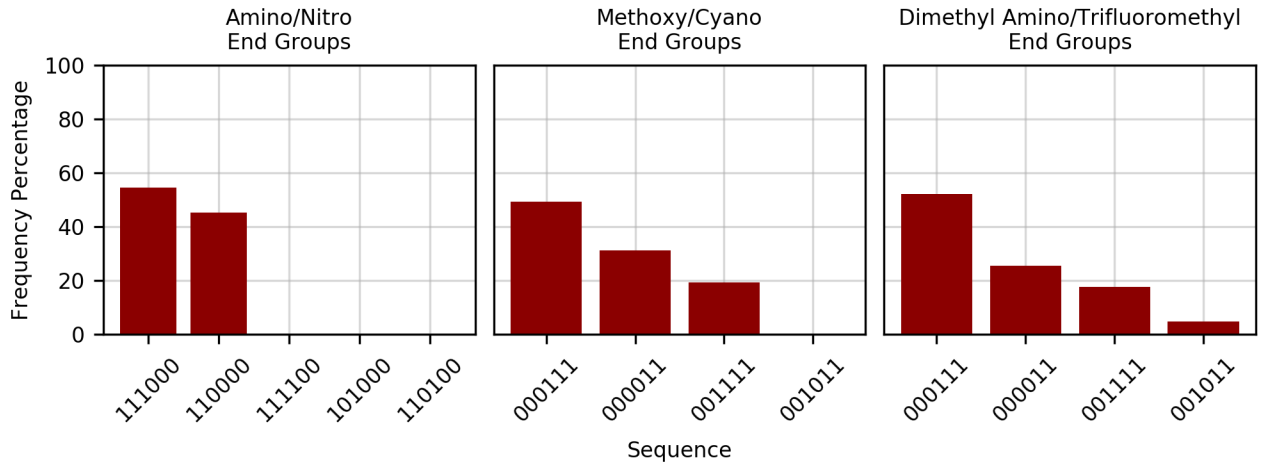


Figure S6: Initial 400 Generation Runs: Frequency percentages of sequences from polymer candidates found in the bin A region of polarizability term vs dipole moment term plots.

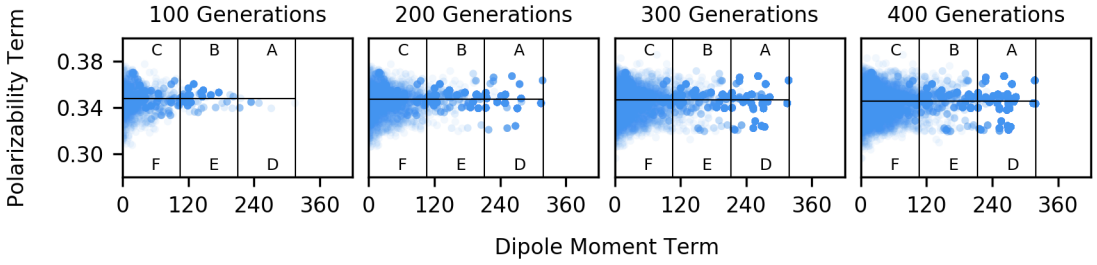


Figure S7: Equalized Scoring Mechanism 400 Generation Run with Dimethyl Amino and Trifluoromethyl End Groups: Bin size and population progression is seen with snapshots every 100 generations.

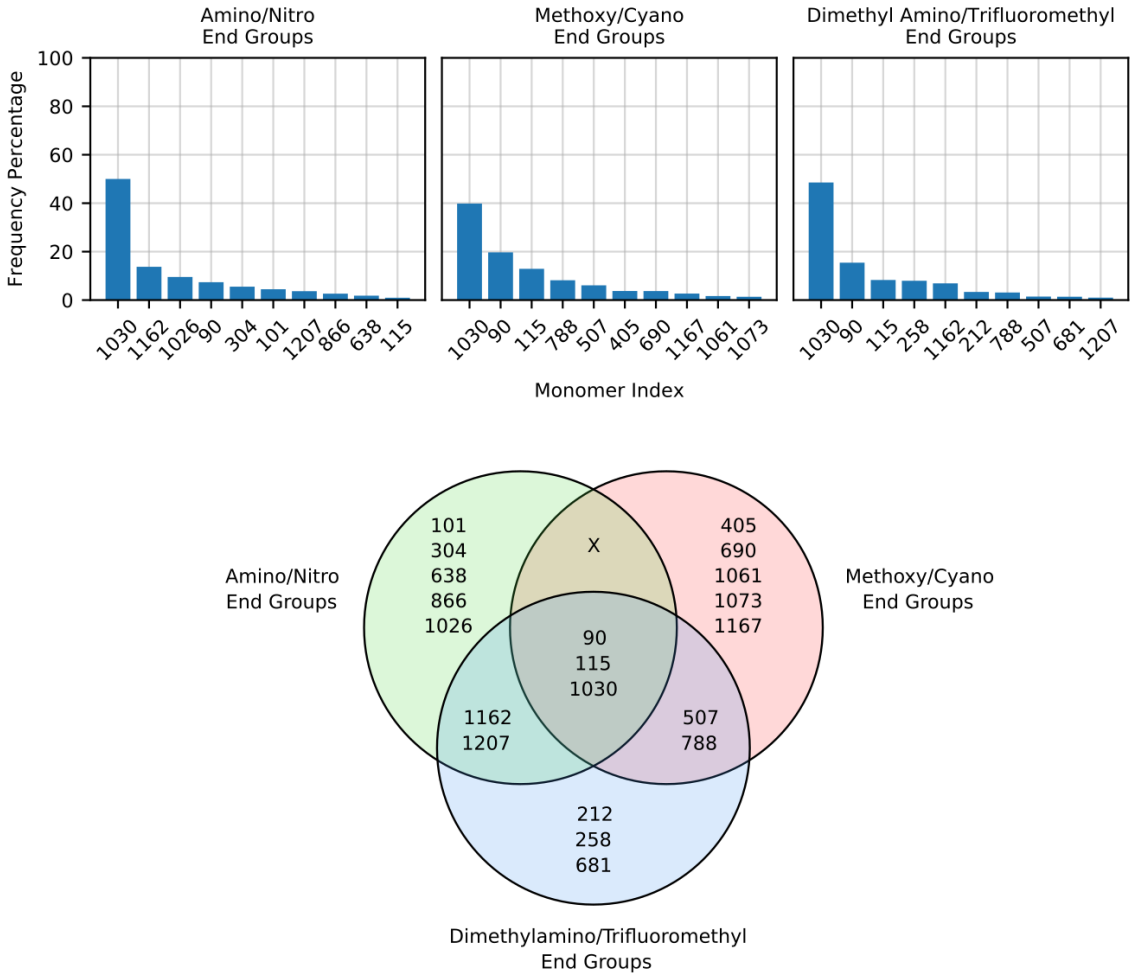
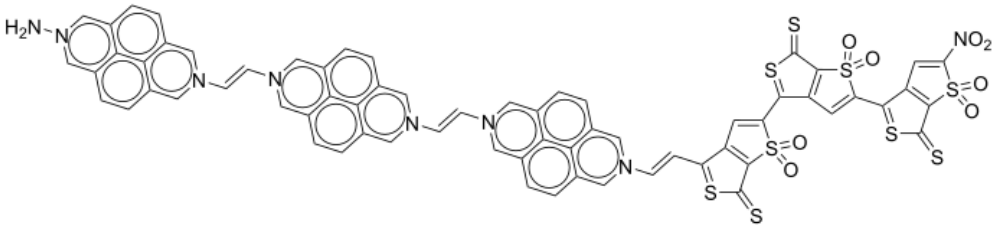


Figure S8: Equalized Scoring Mechanism 400 Generation Runs: The top set of plots shows frequency percentages of the top ten most numerous monomer species (identified by index in monomer set) from polymer candidates found in the bin A region of polarizability term vs dipole moment term plots. The bottom Venn diagram displays the overlap in top monomer species between independent runs.

Table S4: Common Polymer Candidates Found in Multiple Equalized Scoring Mechanism Runs

Polymer Candidate			Runs Containing Polymer Candidate		
Monomer 0	Monomer 1	Sequence	Amino/ Nitro	Methoxy/ Cyano	Dimethylamino/ Trifluoromethyl
1030	1162	000111	X		X
1030	1162	000011	X		X
1030	1162	001111	X		X
1030	90	001111	X	X	X
1030	90	000011	X	X	X
1030	90	000111	X	X	X
1030	212	000111	X		X
507	90	000111		X	X
1030	115	000011		X	X
1030	788	000011		X	X
1030	258	000111		X	X

End Groups: Amino/Nitro (Also appeared with Methoxy/Cyano and Dimethylamino/Trifluoromethyl.)
Monomer 0: 1030
Monomer 1: 90
Sequence: 000111



End Groups: Dimethylamino/Trifluoromethyl (Also appeared with Amino/Nitro.)
Monomer 0: 1030
Monomer 1: 1162
Sequence: 000111

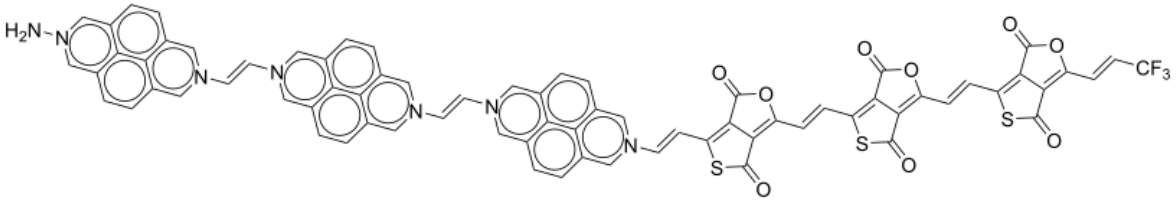


Figure S9: Common polymer candidates found in the bin A regions of multiple equalized scoring mechanism runs.

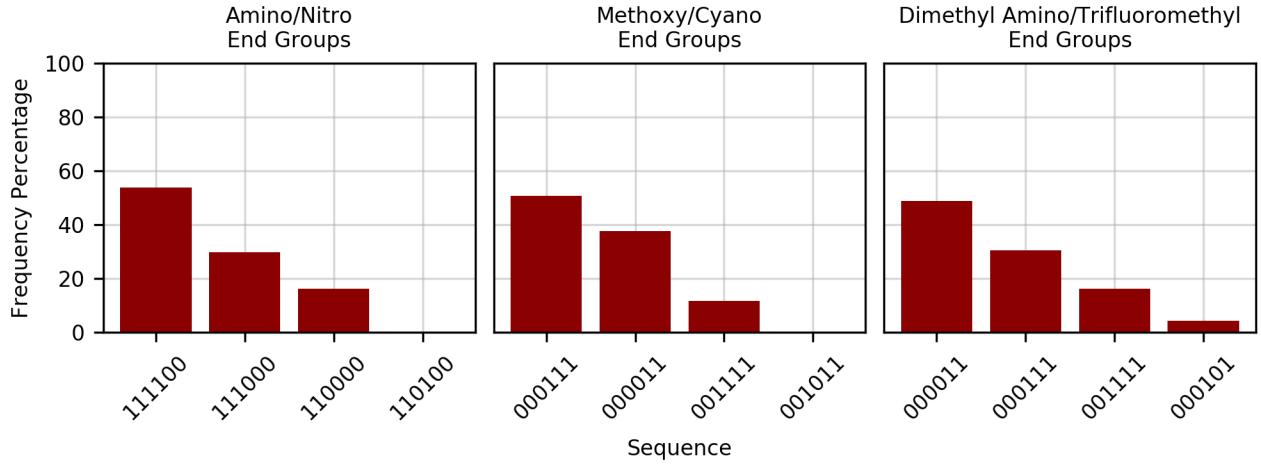


Figure S10: Equalized Scoring Mechanism 400 Generation Runs: Frequency percentages of sequences from polymer candidates found in the bin A region of polarizability term vs dipole moment term plots.

Table S5: Common Polymer Candidates Found in Both Initial and Equalized Scoring Mechanism Runs

Polymer Candidate			Runs Containing Polymer Candidate		
Monomer 0	Monomer 1	Sequence	Amino/ Nitro	Methoxy/ Cyano	Dimethylamino/ Trifluoromethyl
1030	90	000111		X	X
1030	90	001111		X	X
1030	90	000011		X	X
1030	115	000011		X	X
1030	212	000111		X	X

Table S6: Top Seven Hexamers Constructed from Monomer Species 714 and 98, Ranked by Polarizability Term.

Monomer 0	Monomer 1	Sequence	Polarizability Term
714	98	000000	0.39992
714	98	100000	0.39944
714	98	010000	0.39941
714	98	001000	0.39939
714	98	000010	0.39936
714	98	000100	0.39934
714	98	000001	0.39931

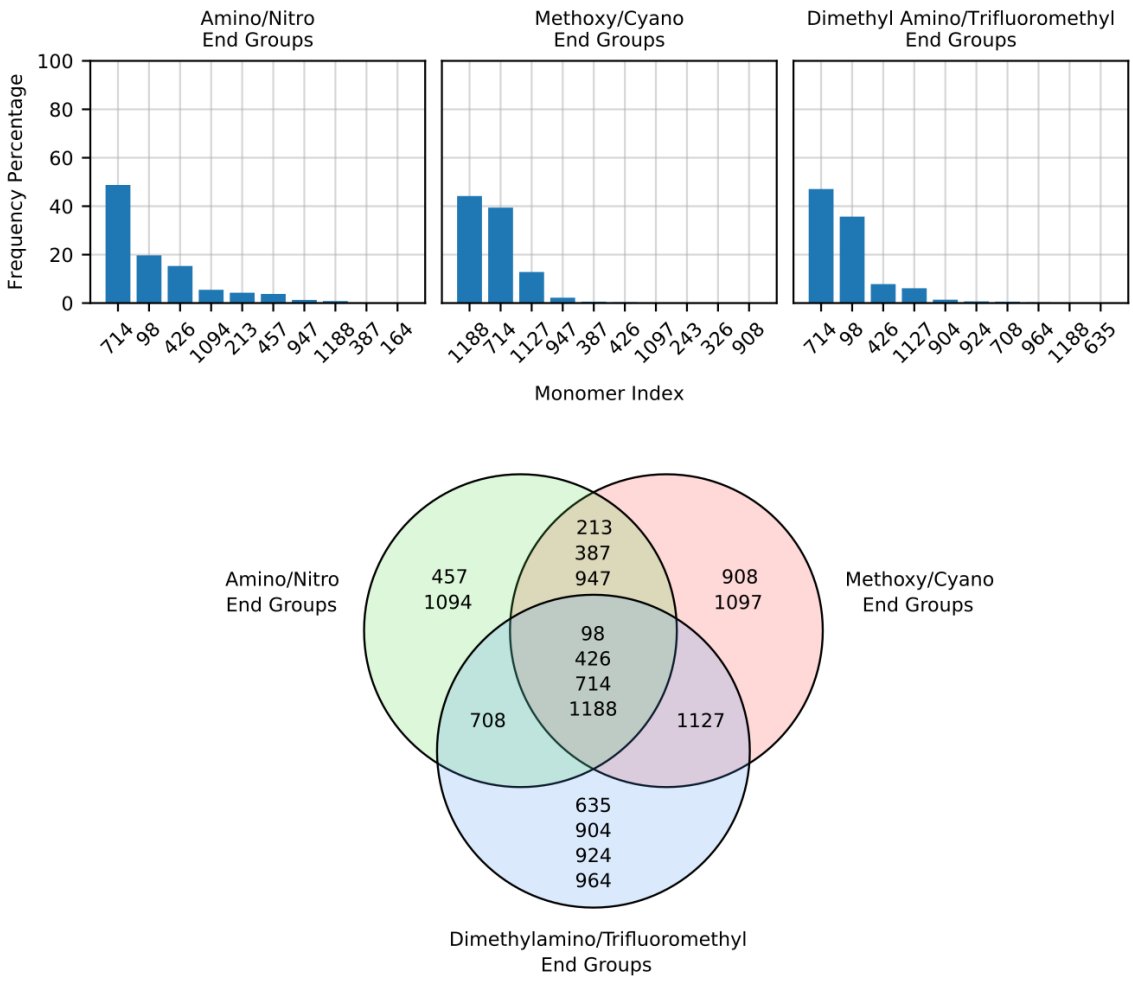


Figure S11: Polarizability Favored 400 Generation Runs: The top set of plots shows frequency percentages of the top ten most numerous monomer species (identified by index in monomer set) from polymer candidates found in the upper bin of polarizability term vs dipole moment term plots. The bottom Venn diagram displays the overlap in top monomer species between independent runs.

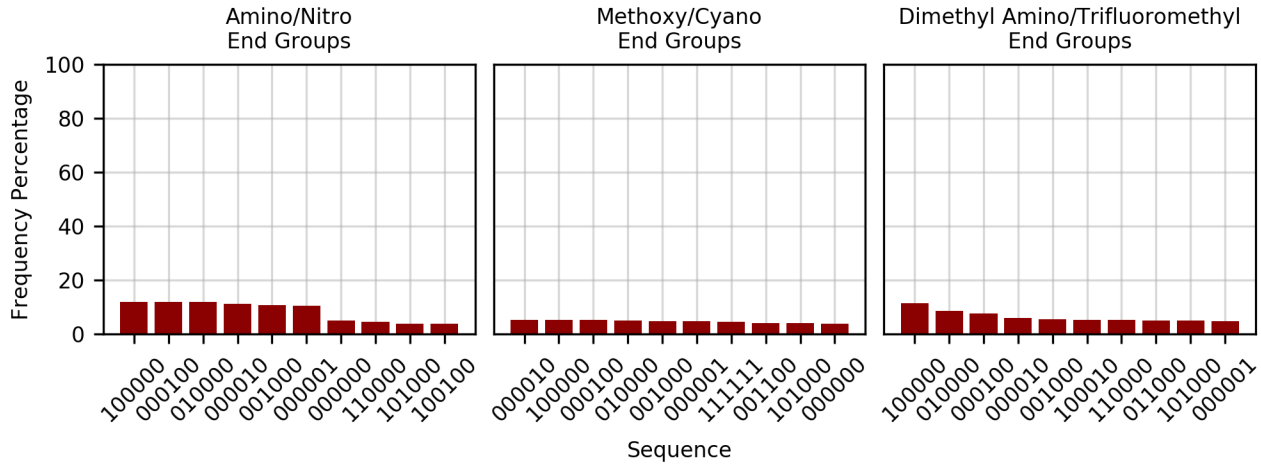


Figure S12: Polarizability Favored 400 Generation Runs: Frequency percentages of sequences from polymer candidates found in the bin A region of polarizability term vs dipole moment term plots.

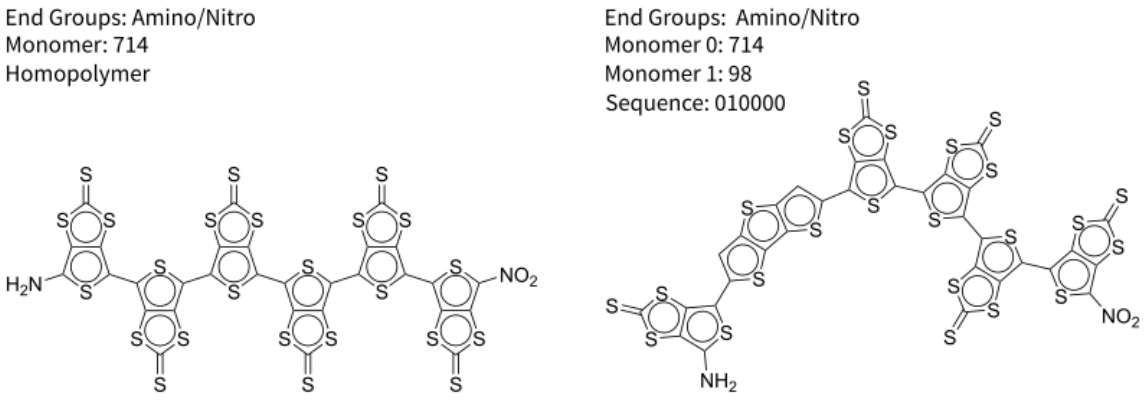


Figure S13: Example high-polarizability homopolymer and near-homopolymer structures.