Computational Materials Science 125 (2016) 92–99

Contents lists available at ScienceDirect

Computational Materials Science

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

Multi-objective optimization techniques to design the Pareto front of organic dielectric polymers

Arun Mannodi-Kanakkithodi [a,*], Ghanshyam Pilania [b], Rampi Ramprasad [a], Turab Lookman [c], James E. Gubernatis [c]

[a] Department of Materials Science and Engineering, Institute of Materials Science, University of Connecticut, 97 North Eagleville Road, Storrs, CT 06269, USA
[b] Materials Science and Technology Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
[c] Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

ARTICLE INFO

Article history:
Received 20 July 2016
Accepted 12 August 2016

Keywords:
Materials informatics
Density functional theory
Multi-objective optimization

ABSTRACT

We present two Monte Carlo algorithms to find the Pareto front of the chemical space of a class of dielectric polymers that is most interesting with respect to optimizing both the bandgap and dielectric constant. Starting with a dataset generated from density functional theory calculations, we used machine learning to construct surrogate models for the bandgaps and dielectric constants of all physically meaningful 4-block polymers (that is, polymer systems with a 4-block repeat unit). We parameterized these machine learning models in such a way that the surrogates built for the 4-block polymers were readily extendable to polymers beyond a 4-block repeat unit. By using translational invariance, chemical intuition, and domain knowledge, we were able to enumerate all possible 4, 6, and 8 block polymers and benchmark our Monte Carlo sampling of the chemical space against the exact enumeration of the surrogate predictions. We obtained exact agreement for the fronts of 4-block polymers and at least a 90% agreement for those of 6 and 8-block polymers. We present fronts for 10-block polymer that are not possible to obtain by direct enumeration. We note that our Monte Carlo methods also return polymers close to the predicted front and a measure of the closeness. Both quantities are useful information for the design and discovery of new polymers.

1. Introduction

The search for new materials is generally motivated by the need for finding materials having multiple properties more optimal than those of materials presently known. This multi-objective optimization problem is similar to the one of picking the best possible material from a limited number of possible options to use in a specific application. In the latter problem, the search is often conducted in a very simple manner by using an Ashby plot [1,2].

An Ashby plot is a two-dimensional plot that displays any two important material characteristics for a large number of known materials, for example, the wear-rate coefficient and the hardness for a variety of ceramics, polymers and metals [2]. For a typical application, we require a very hard material that is also highly resistant to wear, meaning we need to maximize both the hardness and the wear rate coefficient. Another example is the Ashby plot for dielectric materials: for large bandgap, high dielectric constant materials [3–7], we could plot the two properties as in Fig. 1. What is clear from this plot is the bandgap and the dielectric constant exhibit an inverse relationship in the sense that increasing the value of one typically decreases the value of the other. Such an inverse relationship is common in multi-objective optimization problems. In general, the task in materials design or materials selection is to propose or choose materials with the best trade-off between two anti-correlated characteristics.

On an Ashby plot, such as Fig. 1, there exists a set of materials that in some sense is more optimal than the others in that they define a boundary. What is characteristic about this boundary is that for every material on it, we can improve neither of its objectives without degrading the other. Such boundary materials define what is called the Pareto front [8,6,9]. This front represents the best trade-off between the two objectives. In Fig. 1, the set of points marked in blue is the Pareto front for the displayed population of materials.

In the current work, we are interested in multi-objectively predicting new members of a specific class of organic dielectric polymers whose dielectric properties are characterized by large

[*] Corresponding author.
E-mail address: arun.mannodi_kanakkithodi@uconn.edu (A. Mannodi-Kanakkithodi).

http://dx.doi.org/10.1016/j.commatsci.2016.08.018
0927-0256/© 2016 Elsevier B.V. All rights reserved.

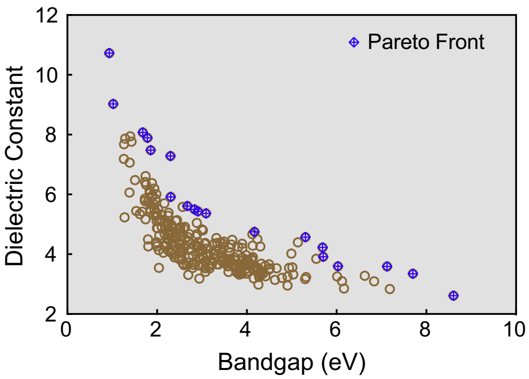


Fig. 1. An Ashby plot to determine materials with the best trade-off between two properties: the bandgap and the dielectric constant. The blue points represent the Pareto front for the entire population. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

bandgaps and high dielectric constants. Fig. 1 contains a number of representative polymers. Evident is the inverse property relationship of the multi-objective problem. With an Ashby plot of proposed or known materials, it is easy to find the Pareto front. The challenge in the search for new materials is meaningfully populating such plots with possible materials that might lie beyond the known front. The difficulty is the number of possible new materials is far too large to identify all candidates by numerical computation or by synthesis and measurement. Here, we propose approaching this complex situation by exploiting our recent work [6], where we generated computational data and used this data to create machine learning (ML) models for the dielectric constant and electronic bandgap of a given chemical space of dielectric polymers.

In the referenced work, we proposed possible organic polymers by using combinations of seven basic chemical building units: CH_2, NH, CO, C_6H_4, C_4H_2S, CS, and O [6,4]. Here, we refer to an n-block polymer as one whose repeat unit is a block of n randomly chosen from this pool of seven. We calculated the bandgaps and the dielectric constants for all possible 4-block systems with Density Functional Theory (DFT) and used machine learning to produce regression models that fitted each computed physical property. We parameterized our fit in such a way that we could use the regression models as surrogates for the same physical properties for 6-block and higher block models. Using these surrogates and their predictions for each polymer in a complete enumeration of possibilities, we could easily obtain the Pareto fronts. However, here we are mainly interested in developing methods to estimate such fronts when complete enumeration is too tedious or not even possible. We present two simple Monte Carlo methods to make such estimates.

Multi-objective optimization is a large active area of engineering research, spanning diverse fields. A variety of techniques exist, examples of which are given in [10–20]. In most cases, the datasets in these applications are much larger than those produced here, and the various methods differ in the manner in which these data are searched for the optimal set. In other cases, the cost of populating the dataset is the challenging factor, and the various methods focus on efficiently adding data to the population. Recently, the use of surrogate models has become prominent for such applications [21,22]. Of the methods we found in the literature, those described by Suman and co-workers [14,15] and by Waldock and Corne [17] most closely resemble ours. We differ from these approaches in part by using surrogates. We feel what we propose is simpler and more appropriate to our use of surrogates.

In Section 2, we summarize our modeling of n-block organic polymers, and in Section 3 we describe two simple Monte Carlo procedures for predicting the Pareto front from the surrogate models learned from our 4-block polymer models. In Section 4, we present our predictions. There, we move from 4 to 6 and then to 8 and 10-block polymers. We benchmark our simulation methods by comparison with their predicted fronts with those generated by complete enumeration for the 4, 6, and 8 block cases. We give a result for a 10 block system that we would be unable to obtain by complete enumeration. For the 6-block case we also give a result where our objectives are optimizing more than two physical properties. Hence, we extend the concept of an Ashby plot to dimensions higher than two. We conclude with a discussion of both how to improve our methods and how to generalize them to other multi-objective materials problems.

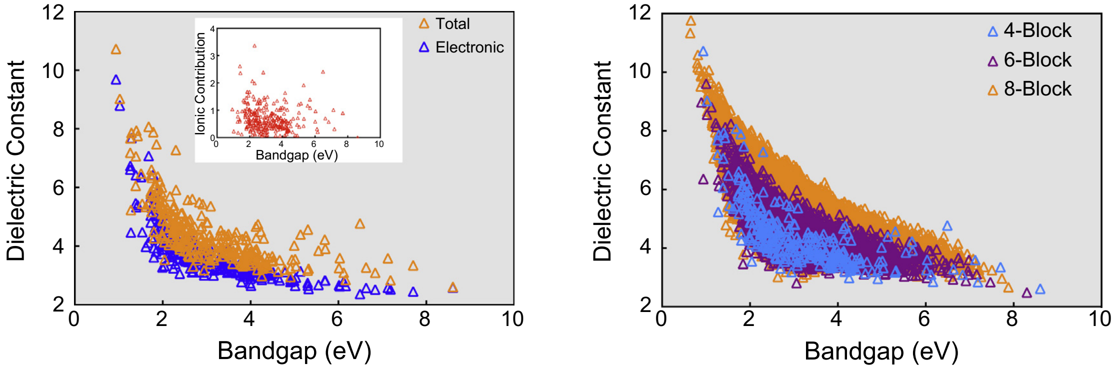
2. Background

We generated a dataset of 4-block polymers via DFT calculations (using VASP [23]) of the dielectric constants (divided into the electronic and ionic parts) and the bandgap values (in eV). The details of these calculations were explained in [6]. If we plot these computed properties against each other (Fig. 2a), we see that while the ionic contribution to the dielectric constant does not correlate well with the bandgap, there is an inverse relationship between the electronic (and consequently, the total) dielectric constant and the bandgap. The Pareto front of the points colored orange[1] in Fig. 2a possesses the polymers of interest when it comes to optimizing both properties.

Whereas 4-block polymers provided convenient atomic structure sizes for DFT calculations, moving to higher block polymers leads to an exponential increase in computational expense. We opted for the following: We developed machine learning (ML) models for the 4-block polymer data, wherein we mapped an intuitive polymer 'fingerprint' to the properties by using a regression algorithm. The fingerprint was a vector of numbers that quantified the different kinds of triplets of blocks (for instance, CH_2-NH-CO constitutes a triplet) present in the given polymer, essentially the unraveled $7 \times 7 \times 7$ matrix, each non-zero element of which marks a specific triplet of blocks. We then used the Kernel Ridge Regression (KRR) method [24] to map this fingerprint to the properties. Training, validation, and testing yielded prediction models for three properties: the electronic dielectric constant, the ionic dielectric constant and the bandgap. The fingerprint chosen was independent of the number of blocks in the repeat unit, meaning the prediction models extend to 6-block, 8-block or even higher block polymers, although we performed the training (fitting) by using only data for 4-block polymers. In [6], we established the accuracy and reliability of block predictions for polymers containing more than 4 blocks in the repeat unit by showing that DFT calculations for a selected assortment of 8-block polymers matched well with the ML predictions.

It was possible to enumerate and fingerprint all possible 6-block and 8-block polymers and to predict the properties for each using the ML models. At first glance, there are 7^n possibilities for an entire population of n-block polymers. However, many are related by translational symmetry; that is, CO-O-C_6H_4-CH_2 is the same polymer as C_6H_4-CH_2-CO-O. We framed our surrogate models to account for this symmetry. Additionally, chemical intuition and domain knowledge tells us that certain adjoining pairs of chemical blocks, such as O-O, CS-CS, CO-CO and NH-NH will form unstable systems, and all such polymers were thus eliminated. These factors led to a reduction in the number of 4-block polymers to 284, 6-block polymers to ~6000, and 8-block polymers to ~130,000.

[1] For interpretation of color in Fig. 2, the reader is referred to the web version of this article.



(a) DFT predicted data for 4-block organic polymers.

(b) ML predicted data for 4, 6 and 8-block polymers.

Fig. 2. Electronic, ionic and total dielectric constants predicted from DFT and ML for 4-block, 6-block and 8-block polymers, plotted against the predicted bandgaps.

Without exploiting translational invariance, chemical intuition and domain knowledge, enumeration all 6-block cases would have been tedious and all 8-block cases not possible. The predicted properties for these much smaller numbers are shown in Fig. 2b. The prediction errors on average are ~0.3 eV for the bandgaps, ~0.3 for the electronic dielectric constants and ~0.2 for the ionic dielectric constants. We established these errors for the 4-block polymers by comparing their surrogate values against their DFT estimated values.

We observed that the requirement of any polymer to occupy the Pareto front is susceptible to these errors: ML predictions varying from the actual DFT results by even with just a 0.5% relative error can make the crucial difference as to whether a given point dominates the remaining points or not. This discrepancy is noticeable even when populating the Pareto front for 4-block polymers. For example, Fig. 3a shows the front obtained by using the DFT computed values and the front obtained with the ML predicted values. DFT calculations yielded 19 polymers on the front whereas the front from ML predictions contained 18 polymers; closer examination reveals that there were only 12 systems present common to the two fronts. This situation is in some ways a weakness of using surrogates to optimize materials for multiple properties; however, the actual errors are still small enough (and acceptable in a statistical treatment) for us to claim that our algorithms find polymers very close to the Pareto front if not actually on them. This is very valuable information for materials design and discovery.

From Fig. 2b, while using just the data generated for 4-block polymers, we also see that since now there is an exponentially larger number of polymers, the Pareto front is much more populated than before. We easily obtained hundreds of Pareto optimal polymer solutions for 6-block and 8-block polymers. However, the enumeration still involves listing and predicting properties of several thousands of polymers, unnecessarily substantial numbers considering our interest is restricted to a mere fraction of them. Thus, the question arises, Is there a way of predicting just the Pareto front for a given n-block polymer family?

3. Methods

Here, we describe two new Monte Carlo based methods for multi-objective optimization in the context of their application to the problem of designing polymers with the optimal dielectric constant and bandgap. We refer to these two methods as ‘Monte Carlo Multi-Objective Optimization Algorithms’ or as MCMO algorithms.

In multi-objective optimization, we have a set of functions

$$\{f_1(x),\ f_2(x),\dots,f_m(x)\},$$

each member of which specifies a material property as a function of the same multi-featured variable $x=(x^1,\ x^2,\dots,x^n)$. In the application at hand, these objective functions are the machine learning fits of the electronic and the ionic contributions to the dielectric constant and the bandgaps of any n-block polymer as described earlier. The variable x is a vector that specifies the n-block polymer with each x^i equal to one of seven possible motifs: CH_2, NH, CO, C_6H_4, C_4H_2S, CS, and O. With respect to determining attractive dielectric polymers, one could define our objective as ‘minimizing the negative of the bandgap’ or ‘minimizing the negative of the dielectric constant’ instead of maximizing such quantities. Given this, we assume a convenient connection to published work by saying that the task is to find an x that minimizes each objective.

In general, a unique solution satisfying all objectives simultaneously does not exist and attention is instead paid to the Pareto optimal solutions. These solutions are based on the following definition of dominance, where a feasible solution x is said to Pareto dominate another solution x' if

$$f_i(x)\leqslant f_i(x'),\ \text{for all } i\in\{1,2,\dots,m\}$$

and

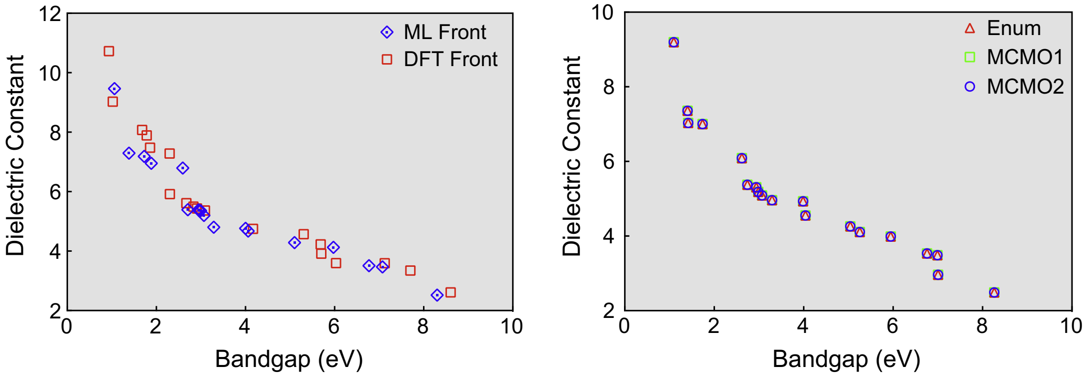
$$f_j(x)<f_j(x'),\ \text{for at least one } j\in\{1,2,\dots,m\}$$

that is, x is as good as x' in all objectives and is strictly better than it in at least one. An x not dominated by any other is called Pareto optimal, and the set of all Pareto optimal solutions constitutes the Pareto front. Our methods are designed to return the Pareto front, as well as additional useful information, as we discuss later. The core of the methods is the notion of an archive of possible members of the front. New polymers x are proposed by a Monte Carlo move and added to the current archive A if

$$f(x)=\max_{x^j\in A,\ j=1,2,\dots,|A|}\left[\min_{i\in\{1,2,\dots,m\}}(f_i(x)-f_i(x^j))\right]<0$$

This is a necessary but not sufficient condition for x being on the Pareto front.

The simulation starts by seeding the archive with a number of polymers whose n blocks are selected randomly. The archive is now scanned one member at a time. For each archive member, every constituent block is scanned, and its current motif is replaced by one randomly chosen. If the proposed polymer is not in the archive and if its $f(x)<0$, then this new polymer is added to the archive. If not added, the next block of the current polymer is randomly changed. If added, the current polymer becomes the one added, and its next block is randomly changed. These processes



(a) DFT front vs ML front.

(b) Fronts from MCMO1, MCMO2 and enumeration.

Fig. 3. Two-objective Pareto front for 4-block polymers obtained from DFT, and from using MCMO1, MCMO2 and enumeration.

are repeated for each block and for each polymer until every member of the original archive is visited once. We call these steps a sweep: A sweep is the attempt to change each unit in the blocks of all members of the current archive. Then a sweep of the new archive is performed, repeating the steps just stated. A few tens of sweeps are performed in this fashion. We refer to this part of the method as its *warm-up phase*.

While admission to the archive requires $f(x) < 0$, once in the archive, the addition of a new member can change whether a previous member still satisfies the condition $f(x) < 0$ (Fig. 4). After the warm-up phase, the archive membership is reset, eliminating those members that do not satisfy $f(x) < 0$. The sweep of the archive is now repeated on the order of a hundred times. The archive size now grows slowly, if at all. We refer to this part of the algorithms as its *equilibrium phase*. When the equilibrium phase is completed, the final front is computed, and the simulation is terminated. We used the cull algorithm to compute the front [25].

The intent of this method (MCMO1) is to maintain an archive of polymers that are on or are near the front, and to identify from them new members that are more optimal than ones chosen randomly. Temporarily including those systems whose $f(x)$ becomes non-negative keeps the archive membership diverse and promotes access to the entire hypersurface of the front.

We also developed a second algorithm (MCMO2) based on the method of simulated annealing. Here, we need to specify temperatures T_{start} and T_{end} such that $T_{\text{start}} > T_{\text{end}}$. We perform the warm-up phase of the simulation at some temperature $T = T_{\text{start}}$. We sweep the archive and change the motifs as before, but this time, if the new polymer x' is not in the archive, we compute $\Delta f = f(x') - f(x)$ and accept or reject the changed polymer according to the Metropolis algorithm. In other words, we accept the new polymer with a probability $\min(1, \exp(-\Delta f/T))$. If accepted, the new polymer is added to the archive; this step can also add polymers with $f(x) > 0$ to the archive. In the equilibrium phase, we perform a number of sweeps of the archive at a given temperature that is chosen at successively lower values, starting with T_{start} and ending with T_{end}. The temperature is lowered according to a power law decay where $T_{\text{decay}} = (T_{\text{start}}/T_{\text{end}})^{N/M}$, where $N < M$ is the number of times we sweep the archive and M is the total number of times the sweeps are performed. This number is roughly the same as the number of sweeps used in the equilibrium phase of MCMO1.

As we discuss in the next section, the fronts predicted by the two methods differ by at most one or two members, if they differ at all. The simulated annealing method creates larger archives and increases the archive sizes (and current front sizes) more rapidly. In general, with this method, the archive sizes can become so large that adding an extra archive reduction step at the end of each temperature step is useful for reducing the computation time.

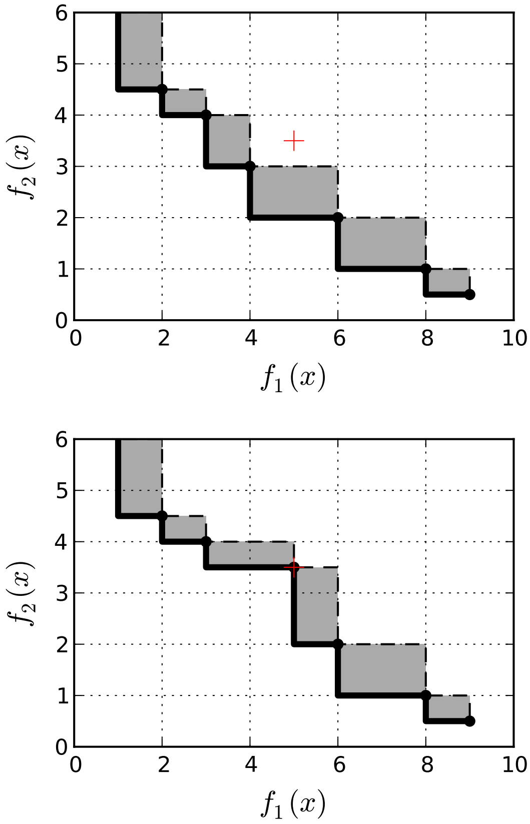


Fig. 4. On the top, the points marked by thick dots constitute a Pareto set for the simultaneous *maximization* of $f_1(x)$ and $f_2(x)$. The solid line connecting the points is the Pareto front. If a point is added to the shaded rectangles, it augments the number of points in the set. If the point marked by + is added, it changes the Pareto set to the one shown in the bottom, causing it to be added to the front and a point to be removed from the previous front.

We remark that a sequence of n motifs specifying a polymer does not directly serve as its "id" because the requirement of translational invariance associates several different sequences with the same polymer. What id's a polymer are its objective values:

transitionally equivalent sequences return the same objective values. Our surrogate models ensure this equivalence within floating point error. In searching the archive to determine whether a proposed polymer would be a new member, we used the objective values as id's. To avoid establishing equivalence via the comparison of floating point numbers, we id'ed each polymer by multiplying its objective values by one million (a somewhat arbitrary number) and truncating the results to integers.

Besides returning the front, the two methods return other useful information. At the end of the simulation, the archive has few, if any, members that have $f(x)$ non-negative. Accordingly, it returns a set possible polymers whose behavior is comparable to those on the front. Additionally, it returns $f(x)$ for each polymer. How negative this quantity is can serve as a rough ranking of the relevance of the members of the archive.

4. Results

In the results presented below, we seeded the archive by randomly generating 100 polymers, adding to the archive only those whose $f(x) < 0$. It is convenient to define other inputs to the algorithms in terms of the number of sweeps. For the warm-up phase we used 25 sweeps, and for the equilibrium phase we used 40 sweeps. In this phase we found it convenient to split this number into 4 batches, where at the end of each batch we reduce the size of the archive. The maximum archive size was 750. If after a sweep this size was exceeded, the archive was reduced to the Pareto front. For 4 and 6 block sizes, the Pareto front rarely grew beyond the warmup phase. For 8 and 10 blocks, it grew by a couple. For MCMO2, we chose $T_{\text{start}} = 1$ and $T_{\text{end}} = 0.0316$. Only for MCMO2 did we observe the archive size exceeding the stated maximum. In general, for MCMO1 the archive size was close to the current front size, while for MCMO2, it generally exceeded the front size. With the exception of the 10-block case, the computations times for the results we present ranged from a few seconds to a few tens of minutes. For 10 blocks the computation time was nearly an hour.

As seen in Fig. 5, the two algorithms did not always predict the same front, and the front from any one algorithm could differ if restarted with a different random number seed. The differences were at most 1 or 2 members. We compensated for these differences by repeating the simulations with each Algorithm 10 times, starting with different random number seeds, combining the fronts from each simulation, and then finding the front of the composite.

Using the two strategies just detailed, we determined the 2 and 3 objective Pareto fronts for 6-block polymers built out of the same seven basic blocks as the 4 block case. The 2 objective front considers the total dielectric constant and the bandgap as the objectives to be optimized, whereas the 3 objective front breaks the dielectric constant into its components and considers the electronic dielectric constant, ionic dielectric constant, and bandgap as the objectives. The 2 objective Pareto fronts obtained for 4-block polymers and 6-block polymers using MCMO1, MCMO2 and simple enumeration (wherein we choose the Pareto optimal polymers out of the blue points and the purple points in Fig. 2b respectively) are plotted in Figs. 3b and 5a respectively. The 3 objective Pareto fronts obtained from MCMO1, MCMO2 and enumeration for 6-block polymers are shown in Fig. 5b.

Whereas MCMO1, MCMO2 and enumeration all yield the exact same fronts for the 4-block polymers, as shown in Fig. 3b, in Fig. 5a, we see that the enumeration strategy reveals 5 points that the optimization algorithms did not determine. These 5 points have very close neighbors on the front, which could lead the MCMO algorithms to miss them while capturing the neighboring points correctly. Regardless, these numbers tell us that the MCMO algorithms have computed the 2 objective Pareto fronts with a ~90% accuracy with respect to complete enumeration, which is encouraging.

The 3 objective front, as shown in Fig. 5b, has a much higher population than the 2-objective front, and is valuable if both components of the dielectric constant need optimization as opposed to just the overall value. Given that the electronic dielectric constant, ionic dielectric constant and the bandgap are all optimized here, we display the 3 objective front by plotting both the electronic and total dielectric constants against the bandgap. Whereas MCMO1 and MCMO2 yield 166 points in the front, enumeration (choosing from the purple points in Fig. 2b) leads to 192 points; this can be seen from the plot in Fig. 5b. Once again, while the points obtained through enumeration are slightly higher in number, the MCMO algorithms produce the front with ~90% accuracy.

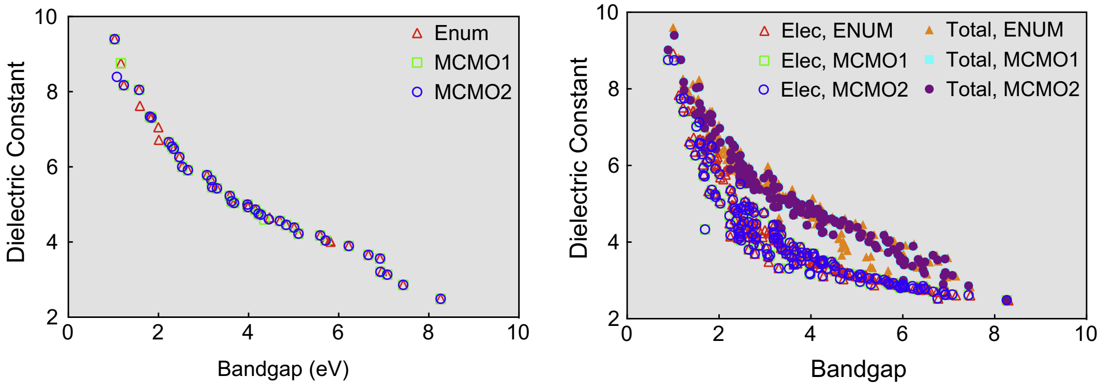
Similarly, we obtained the 2-objective Pareto fronts for 8-block polymers using MCMO1 and MCMO2 as well as using enumeration, and these results are presented in Fig. 6a. We note that there are nearly 130,000 total 8-block polymers, which makes the Pareto frontier population twice as big as for 6-block polymers, as well as a more computationally demanding estimation. The same observations from the 6-block polymers hold here: The two algorithms predict nearly the same Pareto fronts (65 and 66 points respectively) and capture around 90% of the front predicted by enumeration.

Fig. 6b shows the 2-objective Pareto front obtained with the two Monte Carlo algorithms for 10-block polymers, of which there are $\sim 2 \times 10^7$ possibilities. This number is so large that enumeration would take an unreasonably long time, whereas the Pareto front was obtained in a few hours of computing time on a laptop computer. In these simulations we used 50 warmup sweeps, based on our observations that 25 seemed too few for the 8-block case to be at the front size within minor fluctuations.

From 6-block to 8-block to 10-block polymers, the total number of possibilities increases exponentially, and enumeration becomes extremely computationally extensive. However, the fraction of the Monte Carlo steps needed (or consequently, the fraction of time saved) for determining the Pareto front with an approximate >90% confidence is successively reduced, indicating that the MCMO algorithms are indeed very valuable when we go to large populations in material possibilities.

We have thus utilized machine learning prediction models trained using purely 4-block polymer data, in combination with two flavors of multi-objective optimization, to populate the Pareto front of 6-block, 8-block and 10-block polymers. We can similarly obtain Pareto fronts for polymer populations with any number of blocks in the repeat unit, as well as for combinations of many different n-block polymer populations. We can now critically examine the fronts for 4, 6, 8 and 10-block polymers for understanding the kind of polymers (in terms of their constituent blocks) that make Pareto optimal solutions. Given the 28 different types of building block pairs that exist in this chemical space of polymers, an analysis of the Pareto front with respect to the polymer constituent blocks can reveal what kinds of blocks and neighbors are required for optimal polymer solutions.

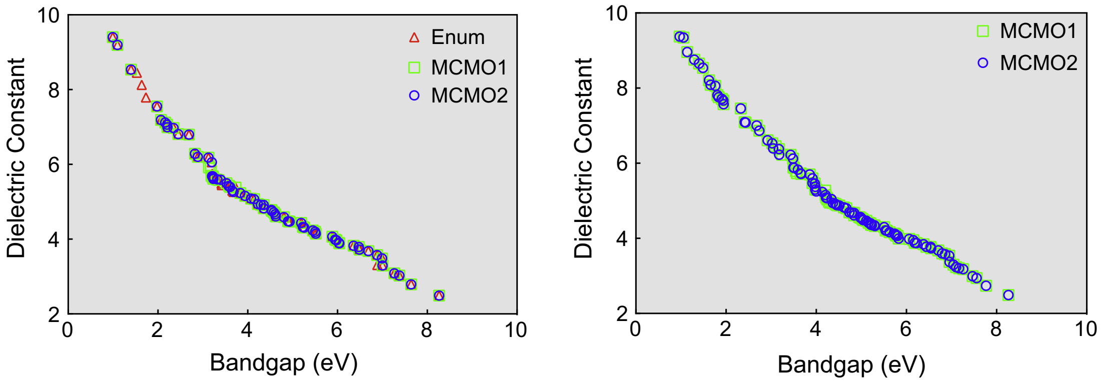
Fig. 7a shows the Pareto fronts for 4, 6, 8 and 10-block polymers; considering an archive with all these polymers from different respective fronts, a combined Pareto front can be obtained by eliminating the specific points that no longer satisfy the front requirements. The combined front has been superimposed on the other fronts in Fig. 7a – this front contains 101 points, which is a reduction of over a 100 points out of the four individual fronts. Further, 80% of these points are 10-block polymers and the remaining are 8, 6 and 4-block polymers, indicating that lower block polymers gradually lose Pareto front prominence as we move towards higher block polymers.



(a) 6-block 2 objective Pareto front.

(b) 6-block 3 objective Pareto front.

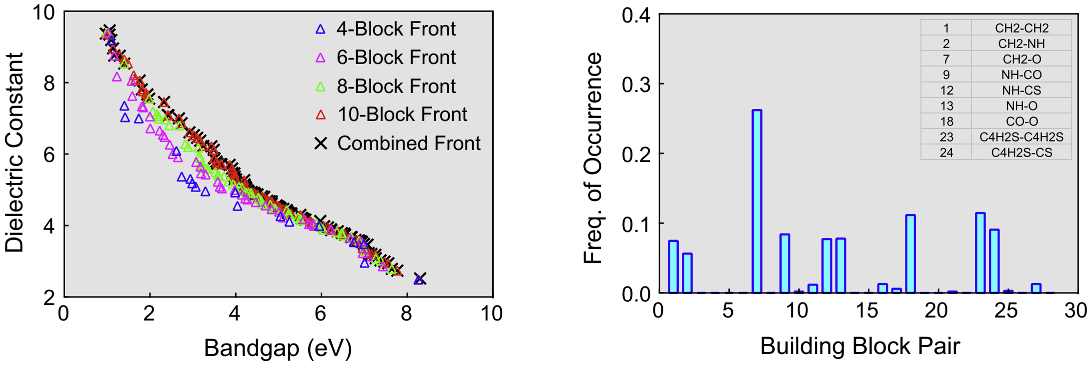
Fig. 5. Two objective and three objective Pareto fronts for 6-block polymers obtained using MCMO1, MCMO2 as well as via enumeration.



(a) 8-block Pareto front.

(b) 10-block Pareto front.

Fig. 6. Two-objective Pareto front for 8-block polymers and 10-block polymers using MCMO1 and MCMO2.



(a) Pareto fronts for 4, 6, 8 and 10-block polymers, and the combined front.

(b) Frequency of occurrence of different building block pairs in the combined Pareto front.

Fig. 7. Examination of the combined two-objective Pareto front of 4-block, 6-block, 8-block and 10-block polymers.

In Fig. 7b, we plotted the fraction of occurrence of different building block pairs in the combined Pareto front of the polymers. The specific pairs of blocks that dominate the front have been shown, such as CH_2-CH_2, CH_2-O, NH-CO, NH-CS, NH-O and C_4H_2S-C_4H_2S. This follows from the correlations drawn in [6] between different block pairs and the properties: the dielectric constants and the bandgaps. While CH_2-CH_2 and CH_2-O pairs contribute to high bandgap values, C_4H_2S-C_4H_2S and C_4H_2S-CS blocks lead to high values of electronic dielectric constant whereas NH-CO, NH-O and NH-CS blocks lead to high ionic dielectric constant values. The Pareto front can said to be occupied by three kinds of polymers: high dielectric constant and low bandgap polymers, high dielectric constant and high bandgap polymers, and low dielectric constant and high bandgap polymers. Thus, it is reasonable that these blocks exist in great numbers in the polymers on the front. The key to further populating the front or regions near it with new *n*-block polymers is in increasing the population of these pairs of blocks in the polymers.

The natural measure of efficiency of our proposed methods is the number of objective function evaluations needed relative to the total number needed to evaluate all possibilities with the surrogates. For MCMO1, we can estimate efficiency in the following manner: In a sweep we have *n-blocks* × *archive-size* evaluation steps. Let us say that the total number of sweeps (warmup and equilibrium phases) is 50. In the warmup phase, the size of the archive is steadily growing but it is nearly constant in the equilibrium phase. Let us assume for each sweep, the archive size is a constant and on average equal to half the size of the Pareto front. For 4-block case, we would need about 2300 evaluations; for the 6-block case, 64,500; and for 8 blocks, 144,000. Thus, direct enumeration is more efficient for 4 blocks and is about the same as the Monte Carlo for 6 and 8 blocks, after symmetry and other factors are used to significantly reduce the number of possibilities. We note that the front found did not include any polymers excluded by these physical arguments. For 10 blocks, the Monte Carlo would require about 500,000 evaluations, clearly besting direct enumeration. For MCMO2, if we were to make similar estimates using the maximum archive size instead of using the front size as the average archive size, we would similarly find that for block sizes 10 and higher the MCMO2 algorithm would be more efficient than direct enumeration. We also note that translational symmetry, chemical intuition or domain knowledge was not used to restrict the size of the space the Monte Carlo sampled.

Efficiency is a different issue than the cost (time) of calculation. We wrote our program in a high level language (Python), and hence a significant reduction in computation time is likely if we instead used the computationally more efficient Fortran 2008 or C++. Likely the biggest reductions in cost would have occurred by using a binary search of the archive sets instead of the linear search used to determine whether the proposed polymer was already a member of the archive and by using a different proposal method for a new archive member. Here, we made the proposal by replacing each unit in the block with one of the seven motifs chosen randomly. Possibly more efficient would have been proposing to replace the motif by the one above or below it in an ordered list of possible motifs. We were mainly concerned with proof of principle and not efficiency.

5. Concluding remarks

We presented two Monte Carlo multi-objective optimization algorithms to find the Pareto front of a selected chemical space of dielectric polymers, the subset of polymers that are of most interest with respect to optimizing both the dielectric constant and the bandgap. While one algorithm involves random modifications of polymer blocks and updating of the Pareto frontier archive based on a dominance definition, the other algorithm uses simulated annealing and updates the archive on the basis of a predefined probability of dominance. The properties of any new polymer are obtained via machine learning models previously trained and developed to predict dielectric constants and bandgaps of these polymers. Combining these learning models with the MCMO algorithms enables us to populate the polymer Pareto front for the larger block polymers in much fewer steps than enumerating all polymer possibilities and predicting their properties.

We benchmarked the effectiveness of the methods by comparing the Pareto fronts they predicted with those the fitted models predicted for the complete enumeration of all possible 4, 6, and 8-block polymers built from our seven basic chemical motifs. We could make this benchmark because we were able to *a priori* reduce the number of possibilities and the machine learning models execute very quickly for a given input. For the 6-block case there were only 5 points from the fitted set not sampled by the Monte Carlo methods, whereas there were 9 such points for the 8-block case.

As previously noted, the prediction errors associated with the machine learning models lead to different Pareto fronts than those found using the actual DFT computed data (Fig. 3). We attempted other regression methods, such as support vector machines, Gaussian processes, and Bayesian kernel ridge regression, to increase the accuracy of the surrogate models without noticeable success. The regression models proved very valuable for providing a quick and efficient determination of optimal polymer solutions even for long polymer chains. Whereas machine learning makes it possible to predict properties of large systems in an on-demand fashion [6,26–29], the MCMO techniques negate the need for traversing through significant portions of polymer chemical subspaces in order to obtain the desired polymers that lie on the Pareto front. We comment that the DFT calculations have unknown errors of their own. Consequently, what is the true Pareto front is likely something that will always be unknown. In some sense, one does not need optimization methods more accurate than the surrogates: One does not need a precise Pareto front but rather a Pareto neighborhood bordering the front. Having methods such as the ones presented, which estimate the front plus materials nearby, is particularly appropriate for materials design and discovery.

In closing, we remark the basic ideas of the presented Monte Carlo methods are useful for multi-objective materials design and discovery for materials other than dielectric polymers. Alloy and solid solutions would be naturals for the use of the approaches presented. As here, the first step would be to generate the surrogates. In contrast to the present case, where the design space variables have discrete values, these problems have continuous variables. Proposed changes in the material would correspond to small changes in the concentrations of the constituents.

Acknowledgements

AMK would like to acknowledge the computational support provided by the Extreme Science and Engineering Discovery Environment (XSEDE) and by Los Alamos National Laboratory (LANL). The work of JEG, TL, and GP was supported by the Department of Energy's Laboratory Directed Research and Development program at LANL.

References

[1] M. Ashby, Materials selection in mechanical design, in: Materials Selection in Mechanical Design, fourth ed., Butterworth-Heinemann, Elsevier, 2010.

[2] G.E. Dieter, Overview of the materials selection process, in: ASM Handbook vol. 20, Materials Selection and Design, ASM International, pp. 243–254.

[3] K. Yim, Y. Yong, J. Lee, K. Lee, H.-H. Nahm, J. Yoo, C. Lee, C. Seong Hwang, S. Han, Novel high-k dielectrics for next-generation electronic devices screened by automated ab initio calculations, NPG Asia Mater. 7 (2015) e190, http://dx.doi.org/10.1038/am.2015.57.

[4] V. Sharma, C. Wang, R.G. Lorenzini, R. Ma, Q. Zhu, D.W. Sinkovits, G. Pilania, A. R. Oganov, S. Kumar, G.A. Sotzing, S.A. Boggs, R. Ramprasad, Rational design of all organic polymer dielectrics, Nat. Commun. 5 (2014), http://dx.doi.org/10.1038/ncomms5845, 4845.

[5] C. Wang, G. Pilania, S. Boggs, S. Kumar, C. Breneman, R. Ramprasad, Computational strategies for polymer dielectrics design, Polymer 55 (4) (2014) 979–988, http://dx.doi.org/10.1016/j.polymer.2013.12.069.

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

[7] A. Mannodi-Kanakkithodi, G.M. Treich, T.D. Huan, R. Ma, M. Tefferi, Y. Cao, G.A. Sotzing, R. Ramprasad, Rational co-design of polymer dielectrics for energy storage, Adv. Mater. 28 (30) (2016) 6277–6291, http://dx.doi.org/10.1002/adma.201600377.

[8] D. Fudenberg, J. Tirole, Nash equilibrium: multiple Nash equilibria, focal points, and Pareto optimality, in: Game Theory, MIT Press, Cambridge, MA, USA, 1991, p. 1823.

[9] T.D. Huan, A. Mannodi-Kanakkithodi, C. Kim, V. Sharma, G. Pilania, R. Ramprasad, A polymer dataset for accelerated property prediction and design, Scient. Data 3 (2016), http://dx.doi.org/10.1038/sdata.2016.12, 160012.

[10] K. Deb, Multi-objective Optimisation Using Evolutionary Algorithms: An Introduction, Springer London, London, 2011, http://dx.doi.org/10.1007/978-0-85729-652-8_1.

[11] K. Miettinen, Nonlinear Multiobjective Optimization, vol. 12, Springer US, New York, NY, USA, 1998, http://dx.doi.org/10.1007/978-1-4615-5563-6.

[12] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A fast and elitist multiobjective genetic algorithm: Nsga-ii, IEEE Trans. Evol. Comput. 6 (2) (2002) 182–197, http://dx.doi.org/10.1109/4235.996017.

[13] J.E. Fieldsend, R.M. Everson, S. Singh, Using unconstrained elite archives for multiobjective optimization, IEEE Trans. Evol. Comput. 7 (3) (2003) 305–323, http://dx.doi.org/10.1109/TEVC.2003.810733.

[14] B. Suman, Study of simulated annealing based algorithms for multiobjective optimization of a constrained problem, Comput. Chem. Eng. 28 (9) (2004) 1849–1871, http://dx.doi.org/10.1016/j.compchemeng.2004.02.037.

[15] B. Suman, P. Kumar, A survey of simulated annealing as a tool for single and multiobjective optimization, J. Operat. Res. Soc. 57 (10) (2006) 1143–1160, http://dx.doi.org/10.1057/palgrave.jors.2602068.

[16] S. Bandyopadhyay, S. Saha, U. Maulik, K. Deb, A simulated annealing-based multiobjective optimization algorithm: Amosa, IEEE Trans. Evol. Comput. 12 (3) (2008) 269–283, http://dx.doi.org/10.1109/TEVC.2007.900837.

[17] A. Waldock, D. Corne, Multi-Objective Probability Collectives, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, http://dx.doi.org/10.1007/978-3-642-12239-2_48 (pp. 461–470).

[18] H.K. Singh, T. Ray, W. Smith, C-psa: constrained Pareto simulated annealing for constrained multi-objective optimization, Inform. Sci. 180 (13) (2010) 2499–2513, http://dx.doi.org/10.1016/j.ins.2010.03.021.

[19] B. Suman, N. Hoda, S. Jha, Orthogonal simulated annealing for multiobjective optimization, Comput. Chem. Eng. 34 (10) (2010) 1618–1631, http://dx.doi.org/10.1016/j.compchemeng.2009.11.015.

[20] X.-B. Hu, M. Wang, Q. Ye, Z. Han, M.S. Leeson, Multi-objective new product development by complete Pareto front and ripple-spreading algorithm, Neurocomputing 142 (2014) 4–15, http://dx.doi.org/10.1016/j.neucom.2014.02.058 ({SI} Computational Intelligence Techniques for New Product Development).

[21] A.I. Forrester, A.J. Keane, Recent advances in surrogate-based optimization, Prog. Aerosp. Sci. 45 (13) (2009) 50–79, http://dx.doi.org/10.1016/j.paerosci.2008.11.001.

[22] F.A.C. Viana, T.W. Simpson, V. Balabanov, V. Toropov, Special section on multidisciplinary design optimization: metamodeling in multidisciplinary design optimization: how far have we really come?, AIAA J 52 (2014) 670–690, http://dx.doi.org/10.2514/1.J052375.

[23] G. Kresse, J. Hafner, Ab initio molecular dynamics for liquid metals, Phys. Rev. B 47 (1993) 558–561, http://dx.doi.org/10.1103/PhysRevB.47.558.

[24] K. Vu, J.C. Snyder, L. Li, M. Rupp, B.F. Chen, T. Khelif, K.-R. Mller, K. Burke, Understanding kernel ridge regression: common behaviors from simple functions to density functionals, Int. J. Quant. Chem. 115 (16) (2015) 1115–1128, http://dx.doi.org/10.1002/qua.24939.

[25] M. Geilen, T. Basten, A calculator for Pareto points, in: 2007 Design, Automation Test in Europe Conference Exhibition, 2007, pp. 1–6, http://dx.doi.org/10.1109/DATE.2007.364605.

[26] T.D. Huan, A. Mannodi-Kanakkithodi, R. Ramprasad, Accelerated materials property predictions and design using motif-based fingerprints, Phys. Rev. B 92 (2015) 014106, http://dx.doi.org/10.1103/PhysRevB.92.014106.

[27] T. Mueller, A.G. Kusne, R. Ramprasad, Machine learning in materials science: recent progress and emerging applications, in: Reviews in Computational Chemistry, John Wiley & Sons, Inc., 2016.

[28] V. Botu, R. Ramprasad, Adaptive machine learning framework to accelerate ab initio molecular dynamics, Int. J. Quant. Chem. 115 (16) (2015) 1074–1083, http://dx.doi.org/10.1002/qua.24836.

[29] G. Pilania, A. Mannodi-Kanakkithodi, B.P. Uberuaga, R. Ramprasad, J.E. Gubernatis, T. Lookman, Machine learning bandgaps of double perovskites, Scient. Rep. 6 (2016), http://dx.doi.org/10.1038/srep19375, 19375.