Polymer Dielectrics Design
Using First Principles Computations and Machine Learning

Chenchen Wang

Department of Chemical, Materials and Biomolecular Engineering,
Institute of Materials Science,
University of Connecticut

Principal Advisor: Prof. R. Ramprasad
Associate Advisor: Prof. A. M. Dongare
Associate Advisor: Prof. G. Rossetti, Jr.

Abstract

The current standard for capacitive energy storage applications is biaxially-oriented polypropylene (BOPP) which has a large band gap and a high breakdown strength, but a small dielectric constant. To meet the growing needs for high energy density capacitors, we use high throughput density functional theory (DFT) calculations in combination with machine learning (ML) methods to design new polymer dielectrics with high dielectric constant and large band gap. Different polymer classes are considered, from C-based organic polymers to novel Si-, Ge-, and Sn-based polymers. Newly developed high throughput DFT methods were used first to accurately determine the dielectric constant and band gap of different polymer systems for a set of limited compositions and configurations. ML methods were then used to predict the properties of systems spanning a much larger part of the configurational and compositional space. Based on this strategy, we are able to identify a set of new polymers with simultaneously high dielectric constant and large band gap.
1 Introduction

Dielectric materials with high energy density are of great importance for a broad range of applications such as food preservation, nuclear test simulations, electric propulsion of ships, and hybrid electric vehicles [1, 2, 3, 4]. The maximum possible stored energy density of a capacitor is given by $U = 0.5\varepsilon_0\varepsilon E_b^2$, where $\varepsilon_0$, $\varepsilon$, and $E_b$ are, respectively, the permittivity of free space, the dielectric constant, and the dielectric breakdown field. A major subset of high energy density capacitors utilizes polymeric materials as the dielectrics.

Compared with ceramic dielectrics, polymer-film capacitors can provide several attractive properties, including ease of processability, high breakdown field, low loss, low cost, and graceful failure [1]. Commercial capacitor polymers include biaxially oriented polypropylene (BOPP) [5], polycarbonate (PC) [6], polyester (PET) [6], polypolychloroprene-sulfide (PPS) [6], polyimide [7], etc. Figure 1 summarizes the dielectric and energy storage properties of such dielectric polymers. Among these polymers, BOPP is the current state-of-the-art polymer dielectric film due to its high breakdown strength and low loss. However, the low dielectric constant ($\sim 2.2$) of BOPP limits its energy density [6, 8, 9].

Continuous miniaturization in electrical power systems demands further increases in energy density of dielectric materials since these capacitors contribute significant volume and weight to these systems. Attempts to surpass the properties of BOPP has been mainly focused on poly(vinylidene fluoride) (PVDF) and its copolymers. However, the relatively high remanent polarization significantly limits the energy density in PVDF [10]. Later, it was demonstrated that ferroelectric PVDF polymers can be converted to relaxor ferroelectric polymers with very small remanent polarization by copolymerization. The modified PVDF polymers such as P(VDF-CTFE) and P(VDF-HFP) (CTFE: chlorotrifluoroethylene; HFP: hexafluoropropylene) can have very high energy densities at high electric fields [11, 12]. Unfortunately, these polar polymers with strongly coupled dipoles exhibit pronounced polarization hysteresis at high electric fields which leads to high loss. Another area of active investigation is in developing polymer nanocomposite to increase the dielectric constant and hence the energy density of capacitors [13, 14, 15]. The challenge for the nanocomposite approach is achieving a low dielectric loss while raising the dielectric constant with inorganic nanofillers.

Given the limited success of the above mentioned approaches, high energy density polymer dielectrics research has begun to shift towards the development of new classes of dielectric polymers. One example is the aromatic polythiourea [16, 17], which is shown to have high breakdown strength, low dielectric loss, and high energy density. The success of these new dielectric polymers immediately bring up the question - can we design more new polymers with the same, or even better dielectric properties? While this discovery process has largely been guided by experimental efforts in the past, the in silico design of new materials using electronic-structure calculations is increasingly being considered as an attractive alternative option.

2 Proposed Research Plan

2.1 General Objectives

Our overarching goal of this work is to design new polymer dielectrics with high energy density using electronic-structure calculations, namely, density functional theory (DFT). As mentioned before, the stored energy density
of a capacitor is given by $U = 0.5e_0\epsilon E_b^2$, where $e_0$, $\epsilon$ and $E_b$ are, respectively, the permittivity of free space, the dielectric constant and the dielectric breakdown field. In order to increase the energy density, the dielectric constant and/or the breakdown field need to be increased. While first principles based computation of $E_b$ is quite involved and computationally demanding, we note that $E_b$ is correlated with the band gap [18], since it requires more energy to excite an electron from an occupied energy level to an unoccupied level if the band gap is larger. Therefore, the larger the band gap, more difficult it is to produce mobile electrons and thus larger is the $E_b$. Thus we use the band gap as a descriptor for the breakdown strength, and then our goal becomes to be designing new polymers with both high dielectric constant and large band gap.

2.2 Methodologies

DFT [19, 20] based quantum mechanical electronic structure methods have reached a state that allows atomic level interactions in diverse chemical environments to be studied accurately. Specifically, DFT-based methods provide a reliable tool to determine static (low frequency) and optical (electronic) dielectric constants for a given configuration of atoms. In this work, all DFT calculations are/will be performed using the plane wave code VASP [21]. In order to finish all the simulations in a tractable amount of time, calculations are performed in the Linux parallel computing environments located at IMS, and the National Science Foundation (NSF) Teragrid Resource Allocation.

Owing to the staggering compositional and configurational degrees of freedom possible in materials, it is fair to assume that the chemical space of even a restricted subclass of materials is far from being exhausted, and an enormous number of new materials with useful properties are yet to be discovered. Given this formidable chemical landscape, a fundamental bottleneck to an efficient materials discovery process is the lack of suitable methods to rapidly and accurately predict the properties of a vast array of new yet-to-be-synthesized materials. The standard approaches adopted thus far involve either expensive and lengthy synthesis-testing experimental cycles, or laborious and time-intensive computations, performed on a case-by-case manner. In this work, a radically different paradigm, namely, machine learning, is used for materials property predictions. Machine learning is a topic central to cognitive/decision/game theory [22], pattern recognition [23, 24, 25], and event forecasting [26, 27]. Such learning methods can be used to establish a mapping between a suitable representation of a material (i.e., its fingerprint or its profile) and any or all of its properties using known historic, or intentionally generated data. Once the profile ⇔ property mapping has been established, the properties of a vast number of new materials within the same subclass may then be directly predicted (and correlations between properties may be unearthed) at negligible computational cost, thereby completely by-passing the conventional laborious approaches towards material property determination and understanding alluded to above.

2.3 Specific Objectives

To achieve our overarching goal, work aimed at the following specific objectives is planned and is currently underway.

1. **High throughput method for polymer dielectric constant prediction.** Although accurate dielectric properties of materials are readily accessible today through DFT, such a conventional approach requires an knowledge of the appropriate crystal structure for the materials to be investigated. If one is interested in a large part of the chemical space, determination of a global minimum on the potential energy surface for each of the new systems to be investigated may not be a practically feasible solution. Towards this end, we plan to develop a new method to rapidly estimate the dielectric constant of polymers. This approach requires only a knowledge of the individual chain structure, but not the manner in which the chains are packed together. Aided with this single-chain approach, we will then develop a high throughput method which allows us to explore the polymer chemical space in a relatively short period of time.
2. Modification of existing polymer dielectrics. While the discovery of completely new classes of polymers for capacitor dielectric applications is exciting in its own right and may prove to be transformative, the evolutionary strategy of enhancing the properties of existing standard materials may be a reasonable risk-mitigation strategy. In this part of the work, we plan to study functionalized polyethylene (PE) and identify how the functional groups, e.g., -OH, -NH\textsubscript{2}, -SH, etc, will affect the dielectric properties of PE.

3. New organic polymers. In theory, there is a huge number of different organic polymers, but only a limited amount has been synthesized. Out of the synthesized polymers, only a few are considered as dielectric materials. We plan to explore the organic polymer chemical space using the high throughput method and search for new polymers with both high dielectric constant and large band gap. The promising polymers identified from our approach will then be subjected to detailed studies.

4. Novel hybrid polymers. The \(\sigma\) and \(\pi\) conjugation along the polymer backbone controls the extent of electron delocalization and hence the electronic contribution to the dielectric constant. By introducing Si, Ge, and Sn into the C-based polymer backbones, the \(\sigma\) conjugation along the backbone will get enhanced, which will make this class of polymers interesting candidates for polymer dielectrics. Exploration of the chemical space of Group IV element-based hybrid polymers will also be conducted, from which promising polymers will be identified.

5. Machine learning approach for larger scale polymer chemical space exploration. The materials discovery process can be significantly expedited and simplified if we can learn effectively from past knowledge. This part of the work involve the development of a similarity-based machine learning approach that can help predict the band gap and dielectric constant of a new polymer in a minuscule fraction of the time necessitated by a typical DFT computation. The machine learning approach will be validated and used to explore a larger scale polymer chemical space.

3 Completed Work

3.1 High-throughput method for polymer dielectric constant prediction [28]

The investigation of polymers is confounded by the necessity of a knowledge of the appropriate crystal structure to be used in DFT computations. Moreover, even if such information is available (or can be "guessed"), each computation may be time consuming. In order to reduce the time involved in each such computation, and to obtain rapid estimates of the dielectric constant and band gap values in the absence of reliable crystallographic information, we have developed a method that is based on purely single chain computations.

As shown in Figure 2(a), we consider an isolated infinite chain of a polymer placed in a supercell volume \(V_{\text{tot}}\), and use density functional perturbation theory (DFPT) to compute its dielectric constant. We note that the dielectric constant calculated from DFPT for such a supercell includes the contributions from the polymer as well as from the vacuum region of the supercell. Treating the supercell as a vacuum-polymer composite, effective medium theory may then be used to estimate the dielectric constant of just the polymer [28, 29].

According to the Maxwell-Garnett equation, the principal components of dielectric constant of a vacuum-filler composite \((\epsilon_{ii})\) containing a volume fraction \(\delta\) of polarizable fillers (polymer chain in this case) with dielectric constant \((\epsilon_{ii}^{\text{polymer}})\) can be written as [30]:

\[
\frac{\epsilon_{ii} - 1}{1 + (\epsilon_{ii} - 1)P_i} = \delta \frac{\epsilon_{ii}^{\text{polymer}} - 1}{1 + (\epsilon_{ii}^{\text{polymer}} - 1)P_i}
\]

Here \(i\) represents the cartesian axes x, y or z, and \(P_i\) is a geometry-dependent depolarizing factor [31, 32, 33]. In our case, assuming that the polymer chain is along the z direction, \(P_z = 0\), and \(P_x = P_y = 0.5\). This leads to the following formula for the axial and off-axis components of the dielectric constant [34, 35]:
Figure 2: (a) Single-chain model represented as a polymer-vacuum composite. (b) Comparison of dielectric constant results from single-chain approach and experiments for common polymers. (c) Comparison of the single-chain based results vs the full crystal results for several XY₂ polymers in different types of crystal structure. Polymers SiF₂, SiCl₂, and GeF₂ can be stabilized in two different types of crystal structures which are represented as Type A and Type B.

\[ \epsilon_{zz} - 1 = \delta(\epsilon_{polymer}^{zz} - 1) \]

\[ \frac{\epsilon_{xx} - 1}{\epsilon_{xx} + 1} = \delta \frac{\epsilon_{polymer}^{xx}}{\epsilon_{polymer}^{xx} + 1} \]

In these equations, \( \delta = V_{polymer}/V_{tot} \) is the volume fraction of the polymer in the supercell as shown in Figure 2(a). In order to use this method to estimate the dielectric constant of just the polymer, i.e., \( \epsilon_{ii}^{polymer} \), the volume occupied by the polymer chain in the supercell is needed. Here, a procedure based on charge density cutoffs is used to estimate the volume. If the electronic charge density is larger than a cutoff value at a particular location, then this location is deemed occupied by the polymer. In order to determine the charge density cutoff value, we have considered several polymers for which experimental volumes (or densities) are available to determine the charge density cutoff value that would result in the experimental density. The polymers used for this analysis include polyethylene, polypropylene, polyacetylene, polythiophene, polypyrrole, polydimethylsiloxane, etc. In general, we find that the charge density cutoff needs to be in the 0.003 electron/Å³ to 0.007 electron/Å³ range in order to reproduce the experimental densities. We note that this range of charge density cutoffs translates to a range of volumes, \( V_{polymer} \), and hence to error bars in the calculated dielectric constants using Eq. 2 and 3.

To validate the single-chain approach, we choose a few common polymers with known dielectric constant and compare the results from this method with those from the experiments [36]. As shown in Figure 2(b), the estimated dielectric constants from our single-chain approach are in relatively good agreement with the experimental values if considering the associated error bars in our method. But the dielectric constants of common polymers are only in the range of around 2 to 4. In order to validate that the single-chain approach can also make a good prediction for the systems with high dielectric constant, we considered homopolymers with a XY₂ repeat unit, where X = C, Si or Ge and Y = H, F, Cl [28]. The dielectric constants of these polymers span a large range from around 2 to 45. We compare the dielectric constant results of our single-chain approach with those from the three-dimensional crystal based calculations as shown in Figure 2(c). Overall, the agreement between the two methods is good, given that interchain interactions are completely neglected in the single-chain computations. The complete neglect of interchain interactions in the single-chain approach may account...
for the consistent underestimation of the dielectric constant with respect to the crystal results (as interchain interaction tend to ”soften” the intrachain phonon modes, and, consequently, increase the dielectric constant values). Although there are some discrepancies, the correct trends are mostly captured by the single-chain approach. Hence, we believe that, this approach can offer a practical strategy for rapidly screening candidate polymer chain structures within high throughput chemical space exploration efforts.

3.2 Modification of existing polymer dielectrics [37, 38]

The single-chain approach described above enables us to study the dielectric properties of different polymer systems, which is essential given that our goal is to design new polymer dielectrics with both high dielectric constant and large band gap. While the discovery of completely new polymers is exciting, the strategy of enhancing the dielectric properties of existing standard materials is also a reasonable risk-mitigation approach.

Polypropylene (PP) is the current standard polymer dielectrics, and any improvement to it will be beneficial. One possible strategy to increase the dielectric constant is adding polarizable functional groups to PP, such as -OH, -NH₂, -SH, etc. Recent experiments have shown that PP with just 4-6 mol% hydroxyl (-OH) side groups displays an almost two-fold increase in the dielectric constant, with minimal impact on the dielectric loss and morphology [39]. The experimental results also imply that the PP-OH contains roughly 0.5 water molecule per OH moiety on the chain.

Following this interesting experimental finding, we investigated the role played by OH functional groups and the trapped moisture. Given that DFT computations are time-intensive, we considered two short-chain polyethylene (PE) oligomers (o-PE), each with 11 carbon atoms. The two chains are arranged in a head-to-tail configuration, and a H chain end atom is substituted by an OH group. Additionally, to explore the effect of water molecules, o-PE-OH with one and two water molecules (o-PE-OH-H₂O and o-PE-OH-2H₂O, respectively) were also studied. Figure 3 shows snapshots of the optimized geometries. We note that owing to the proper account of van der Waals (vdW) interactions, secondary bonding phenomena such as H-bonding are adequately captured in the DFT computations. For instance, Figure 3(d) shows the relaxed geometry of the o-PE-OH-2H₂O system, where the formation of an H-bonded ring containing
two OH groups and 2H₂O molecules is clearly evident. The dielectric constant of these systems is calculated using the single-chain method. The average value of the total dielectric constant (i.e., the trace of the dielectric constant tensor, including both ionic and electronic contributions) for the o-PE, o-PE-OH, o-PE-OH-H₂O, and o-PE-OH-2H₂O systems were, respectively, 2.4, 3.3, 3.7, and 4.3 (Figure 4). For pure PE, the dielectric constant, which is almost entirely due to electronic contributions, is close to the accepted experimental value. The increase in the dielectric constant due to the OH groups (without and with water molecules) is entirely due to the ionic contributions.

Although our calculations involve higher -OH concentration than was considered in the experiments, the main findings pertaining to hydrogen bonding, water trapping and dielectric constant enhancement are expected to persist under the scrutiny of more realistic computations. Indeed, recent detailed force-field based molecular dynamics (MD) simulations in which the morphological details, the concentration of OH, and trapped water are taken into account confirm our results. Figure 4 summarize the DFT, MD, and experimental dielectric constant results. The MD simulations were conducted by our collaborators (Prof. Kumar in Columbia University). In the MD simulations, PE-OH with 4.2 mol% -OH groups was considered and varying amounts of water were added into PE-OH system. From Figure 4, we can see that the addition of water results in a significant increase in dielectric constant. For 0.5 water per OH group, the MD result closely match the experimental value of 3.4, and larger amounts of water further increase the dielectric constant. Also, if considering that the effective concentration of OH in the DFT calculations is rather large (36 mol%), the agreement between the DFT calculations and experiments is rather respectable. Most importantly, the results of the DFT computations are entirely consistent with the MD simulations in that successive increases in dielectric constant can be expected due to OH groups and water molecules. Based on the DFT, MD and experimental results, it can be concluded that the incorporation of a small amount of highly polar functional groups, such as hydroxyls, significantly improves the dielectric properties of a saturated hydrocarbon polymer. These findings are significant, as they point towards a rational pathway for the tunable control of dielectric properties of polymers via creative utilization of molecular functionalization (and moisture).

3.3 New organic polymers [40]

Section 3.2 has proved that by proper molecular functionalization, the dielectric properties of current standard polymer dielectrics, PP, can be enhanced. But how high the dielectric constant of modified PP can be is still limited by the small dielectric constant of PP. Thus, the natural next step should be discovering new polymer dielectrics.

In this part of the work, we seek to design new classes of organic polymers which can be applied as polymer dielectrics by exploring organic polymer chemical space. Figure 5(a) captures the computational model adopted in our strategy. In essence, we consider an all-trans single polymer chain containing four independent blocks with periodic boundary conditions along the chain axis. We then allow each block in the polymer backbone to be one of the following motifs: -CH₂-, -NH-, -C(=O)-, -C₆H₄- (benzene), -C₄H₂S- (thiophene), -C(=S)-, and -O-. These are commonly seen motifs in the backbone of organic polymers, and different combinations of these motifs form traditional polymers, including polyesters, polyamides, polyethers, polyureas, etc. This scheme results in 267 symmetry unique systems.

With the aid of the single-chain approach, we conducted high throughput DFT computations to obtain the band gap and dielectric constant values of all the 267 polymer systems. Figure 5(b) and (c) shows the relationship between the electronic (Figure 5(b)) and total (Figure 5(c)) dielectric constant and the band gap for the 267 polymer systems. A near perfect inverse pareto-optimal front relationship between the band gap and the electronic dielectric constant can be seen from Figure 5(b), which imposes a theoretical limit on the electronic part of the dielectric constant that one may be able to achieve (a limit that may be understood by regarding the electronic part of the dielectric response as a sum over electronic transitions from occupied to unoccupied
Figure 5: (a) Model showing our computation strategy, in which we substitute different functionalities into four blocks, and DFT computed (b) electronic and (c) total dielectric constant as a function of the band gap. The axes are in logarithmic scale.

Figure 6: DFT calculated band gaps and dielectric constants of -(CH$_2$)$_n^-$, -(SiF$_2$)$_n^-$, -(SiCl$_2$)$_n^-$, -(GeF$_2$)$_n^-$, -(GeCl$_2$)$_n^-$, -(SnF$_2$)$_n^-$, and -(SnCl$_2$)$_n^-$.}

states). Figure 5(c) shows the variation of the total dielectric constant with the band gap. The total dielectric constant derives its contributions from the electronic and ionic contributions. From Figure 5(c), a set of the most promising polymers may be selected based on the criteria that the band gap > 3 eV, and the total dielectric constant > 4. This promising region is shown in Figure 5(c), and corresponds to systems composed of at least one aromatic group and at least one of the following three groups: -NH-, -C(=O)-, -O-. Polymers containing these groups can provide a number of useful constraints for determining the best motifs to incorporate into a polymer backbone in order to achieve good dielectric properties. Guided by this finding, our experimental collaborators (Prof. Sotzing in UCONN) have successfully synthesized several new aromatic polyureas and polyurethanes. The dielectric constants of these polymers are either equal to or higher than BOPP, the current standard polymer dielectric material.

So far we have developed a pathway to design new polymer dielectrics by carefully planned high throughput computations. For organic polymers, the results allow us to identify promising motifs and provide guidance to experiments. But by careful scrutiny of the band gap and dielectric constant values of the organic polymers we have studied (Figure 5(b) and (c)), we find that the total dielectric constants of these polymers are mostly dominated by the electronic contribution which is inversely related to the band gap. It means that the dielectric constant values that we can achieve within this set of polymer are limited. Thus in order to realize higher dielectric constant values, we need to explore polymers which result in combinations of improved ionic and electronic dielectric constant.

3.4 Novel hybrid polymers [28, 41, 42]

As mentioned above, we need to explore other types of polymers besides the C-based organic polymers. The $\sigma$ and $\pi$ conjugation along the polymer backbone controls the extent of electron delocalization and hence the electronic contribution to the dielectric constant. By introducing other Group IV elements, Si, Ge, and Sn into the
C-based polymer backbones, the $\sigma$ conjugation along the backbone will get enhanced, which probably will make this class of polymers interesting candidates for polymer dielectrics. Also, substituting C with Si, Ge and Sn can help ensure the chemical compatibility by preserving the local chemical environment and bonding. In order to achieve high ionic contribution to the dielectric constant, large dipole moments that should easily response to the applied electric field need to be created in the polymers, which can be realized by introducing small atoms with high electronegativity such as F and Cl to the side chain. By considering all of these factors, we decided to consider a few Si-, Ge- and Sn-based polymers. These polymers are $-(\text{SiF}_2)_n$, $-(\text{SiCl}_2)_n$, $-(\text{GeF}_2)_n$, $-(\text{GeCl}_2)_n$, $-(\text{SnF}_2)_n$, and $-(\text{SnCl}_2)_n$. Single-chain approach was then applied to estimate their dielectric constants with the assumption that they have the same chain structure as polyethylene $-(\text{CH}_2)_n$. The results are shown in Figure 6, from which we can see that in general, the Si-, Ge- and Sn-based polymers exhibit higher dielectric constant and lower band gap compared with polyethylene, and the increase in dielectric constant mostly results from the ionic contribution, especially for Ge- and Sn-based polymers.

The Si-, Ge-, and Sn-base homopolymers show large dielectric constants, but have smaller band gaps. On the other hand, pure polyethylene has a large band gap, but suffers from a relatively smaller dielectric constant. A natural next step is to "mix" polyethylene with the Si-, Ge-, and Sn-base homopolymers to identify optimal compositions and configurations. The investigation of the myriad of the heteropolymer possibilities was conducted using the same high throughput method we developed for organic polymers. Infinitely long isolated all-trans polyethylene chains containing four independent CH$_2$ units (c.f. Figure 5(a)) in a supercell were simulated using periodic boundary conditions. While block $Y_1$ was fixed to be CH$_2$ unit in the backbone to break the extent of $\sigma$ conjugation along the backbone, three places in the backbone (schematically depicted as $Y_2$, $Y_3$, $Y_4$ in Figure 6(a)) were filled by drawing from the following "pool" of seven possibilities: CH$_2$, SiF$_2$, SiCl$_2$, GeF$_2$, GeCl$_2$, SnF$_2$ and SnCl$_2$. This scheme resulted in 175 symmetry unique systems which were subjected to DFT band gap and dielectric constant computations within our high throughput approach.

Figure 7(a) and (b) show the DFT calculated band gap, electronic and total dielectric constant of 175 polymers. Figure 7(a) portrays the dependence of electronic dielectric constant on the band gap. While the pristine polyethylene chain has the highest calculated band gap among all the systems explored, additions of the larger Group IV elements to the chain lead to a progressive decrease in the band gap. The relationship between electronic dielectric constant and band gap is the similar as that of organic polymers. The total dielectric constant of the Group IV element-based polymers spans over a large range between 2.5 to 47, with the smallest and largest values corresponding to $-(\text{CH}_2)_n$ and $-\text{CH}_2-(\text{SnF}_2)_3$, respectively. In general, it is found that for a given composition and configuration, as the backbone atoms varies from C to Si to Ge to Sn (while all the other units in the chain are held fixed), both the electronic and total dielectric constant increase and the band gap decreases. This
result is in line with the fact that among all the seven units considered, the SnF$_2$ has the largest dipole moment; at the same time Sn-Sn bond rotation has the lowest barrier among all the X-X' (with X or X' = C, Si, Ge, and Sn) bond rotation barriers. Compared with organic polymers, the Group IV element-based hybrid polymers can achieve higher dielectric constant without sacrificing too much of the band gap, which makes them valuable to pursue.

### 3.5 Machine learning approach for larger scale polymer chemical space exploration [43]

Previously, the high throughput DFT strategy was used for the computation of relevant dielectric properties, but it is practically limited to polymer systems with a small periodic chain length. For longer chains (required to screen a larger part of the compositional and configurational space), the computational cost associated with DFT computations rises quite rapidly. Furthermore, as the system size increases, the number of candidates to be explored grows exponentially leading to a combinatorial explosion.

Take the Group IV element-based hybrid polymers for example, doubling the supercell size along the chain direction so as to include 8 distinct building units in a periodic repeating unit (while retaining every fourth unit to be CH$_2$ to be consistent with the 4-unit structures considered above) leads to a total of 29,365 different symmetry unique systems. Clearly, exploration of such a vast chemical space using the present first principles based approach may be impractical. Hence, we investigate this large class of systems using a machine learning approach.

![Figure 8: Schematic of the machine (or statistical) learning methodology employed.](image)

The essential ingredients of our machine learning approach are captured schematically in Figure 8. The first step in the approach prescribed in the panels of Figure 8 is to reduce each material system under inquiry to a string of numbers—we refer to this string as the fingerprint vector. For the polymer chains composed of seven possible building blocks, the following coarse-level chemo-structural fingerprint vector was considered first: $\langle f_1, \ldots, f_6, g_1, \ldots, g_7, h_1, \ldots, h_7 \rangle$, where $f_i$, $g_i$, and $h_i$ are, respectively, the number of building blocks of type $i$, number of $i-i$ pairs, and number of $i-i-i$ triplets, normalized to total number of units (note that $f_7$ is missing in above vector as it is not an independent quantity owing to the relation: $f_7 = 1 - \sum_{i=1}^{6} f_i$).

Next, a suitable measure of chemical distance is defined to allow for a quantification of the degree of (dis)similarity between any two fingerprint vectors. Consider two systems $a$ and $b$ with fingerprint vectors $F^a$ and $F^b$. The similarity of the two vectors may be measured in many ways, e.g., using the Euclidean norm of the difference between the two vectors, $|F^a - F^b|$, or the dot product of the two vectors $F^a \cdot F^b$. In the present work, we use the former, which we refer to as $|F_{ab}|$ (Figure 9). Clearly, if $|F_{ab}| = 0$, materials $a$ and $b$ are equivalent (insofar as we can conclude based on the fingerprint vectors), and their property values $P^a$ and $P^b$ are the same. When $|F_{ab}| \neq 0$, materials $a$ and $b$ are not equivalent, and
Figure 9: Parity plots comparing property values computed using DFT against predictions made using learning algorithms trained using chemo-structural fingerprint vectors. Pearson’s correlation index is indicated in each of the panels to quantify the agreement between the two schemes.

\[ P^a - P^b \] is not necessarily zero, and depends on \[ | \vec{F}_{ab} | \]. This observation may be formally quantified when we have a prior materials-property dataset, in which case we can determine the parametric dependence of the property values on \[ | \vec{F}_{ab} | \]. In the present work, we use the statistical learning algorithm referred to as kernel ridge regression (KRR). Using this scheme, the property of a new system \( b \) is given by a sum of weighted Gaussians,

\[
P^b = \sum_{a=1}^{N} \alpha_a \exp\left( -\frac{1}{2\sigma^2} | F_{ab} |^2 \right).
\]

where \( a \) runs over the systems in the previously known dataset. The coefficients \( \alpha_a \)'s and the parameter \( \sigma \) are obtained by ‘training’ the above form on the systems \( a \) in the previously known dataset, while adhering to the best practices of statistical and cross-validation methods [44, 45].

The above scheme has been applied to the Group IV element-based hybrid polymers. As mentioned above, the initial dataset was generated using DFT for systems with repeat units containing 4 distinct building blocks. Of the total 175 such systems, 130 were classified to be in the ‘training’ set (used in the training of the KRR model, Equation (4)), and the remainder in the ‘test’ set. To validate the machine learning approach, besides the band gap and dielectric constant, five more properties (atomization energy, formation energy, \( c \) lattice parameter, spring constant, and electron affinity) were considered. Figure 9 shows the agreement between the predictions of the learning model and the DFT results for the training and the test sets. Furthermore, we considered several chains composed of 8-block repeat units (in addition to the 175 4-block systems), performed DFT computations on these, and compared the DFT predictions of the 8-block systems with those predicted using our learning scheme. As can be seen, the level of agreement between the DFT and the learning schemes is uniformly good for all eight properties across the 4-block training and test set, as well as the somewhat out-of-sample 8-block test set (regardless of the variance in the property values). Overall, the high fidelity nature of the learning predictions is particularly impressive, given that these calculations take a minuscule fraction of the time necessitated by a typical DFT computation.

While the good agreement between the machine learning and the DFT results is exciting, the real power of this approach lies in the possibility of exploring a much larger chemical space than is practically possible using DFT calculations or laborious experiments. As mentioned before, merely expanding the 4-block Group IV element-based hybrid polymers to 8-block will result in 29,365 symmetry unique systems. The machine
learning approach we just validated can make the study of this staggeringly large number of systems possible. Figure 10 shows the predictions of band gap, electronic and total dielectric constant for the 29,365 systems. The inverse correlation of the band gap with the electronic dielectric constant is confirmed once again from Figure 10(a). It is rather surprising that a rough inverse correlation is seen between the total dielectric constant and the band gap (as shown in Figure 10(b)), although clear deviations from this inverse behavior can be seen. This extensive search then allows us to identify polymers with high dielectric constant and large band gap, which leads to those systems that are at the top part of Figure 10(b) (corresponding to the ‘deviations’ from the inverse correlation, and indicated by a circle in Figure 10(b)). These are systems that contain 2 or more contiguous SnF$_2$ units, but with an overall CH$_2$ mole fraction of at least 25%. Such organo-tin systems may be particularly appropriate for polymer dielectrics. Our experimental collaborators (Prof. Sotzing in UCONN) are now actively pursuing the synthesis of organo-tin systems.

4 Proposed Work

4.1 Materials design

So far, I have been able to provide new understanding of the dielectric properties of functionalized PE, and develop new computational methodologies to search the polymer chemical space for polymers which have both high dielectric constant and large band gap. The extension of this part of the work is planned as followed:

- We have developed a model to study the effect of -OH functionalization on the dielectric properties of PE. The natural next step is exploring other functional groups, such as -NH$_2$, -SH, etc, using the same approach. This will allow us to identify how the dielectric properties of PE will change when modified with different functional groups, and which functional group will result in the highest dielectric constant increase of PE.

- By screening a large number of polymers, a set of promising polymers with simultaneously high dielectric constant and large band gap were identified. We plan to perform detailed studies of these polymers. Our experimental collaborators (Prof. Sotzing in UCONN) will provide information to us about which ones of these identified promising polymers are synthesizable or what modifications are needed to make them...
synthesizable. With these feedbacks, we can focus on the promising polymers that can be synthesized and conduct more thorough DFT calculations on these polymers.

- The chemical space of two different classes of polymers, namely, organic polymers and hybrid polymers, has been explored. Seven motifs, CH$_2$, NH, C(=O), C$_6$H$_4$ (benzene), C$_4$H$_2$S (thiophene), C(=S), and O, were used for organic polymers exploration, and to search the Group IV element-based hybrid polymers, motifs CH$_2$, SiF$_2$, SiCl$_2$, GeF$_2$, GeCl$_2$, SnF$_2$ and SnCl$_2$ were considered. In the future, we plan to identify more motifs which allow us to extend the chemical space exploration.

### 4.2 New robust machine learning approach

One question remaining in our machine learning approach is related to the proper choice of the components of the fingerprint vector. If we can identify a subset of the fingerprint vector components that are most important in determining a property, this knowledge can be used to perform targeted searches. It is also conceivable that some of the fingerprint vector components may be hurtful rather than helpful in predictions. For this reason, we propose to develop a scheme to reduce our fingerprint vector.

In this work, we plan to use genetic algorithm (GA) for the fingerprint vector reduction. GA is a search and optimization algorithm premised on the evolutionary ideas of natural selection and genetic. Unlike the optimization problems, here we are facing ‘yes or no’ decisions regarding whether each component of the fingerprint vector should be included. The flowchart of our planned GA approach is shown in Figure 11. The basic idea is that, the initial fingerprint vector is represented by an array of $n$ ones and zeros, indicating whether a particular component is, or is not, included. Based on this array, we can convert our initial fingerprint vector to a reduced one which we will use to run the machine learning prediction. If the prediction accuracy is the same or gets increased, it means that the eliminated components are not essential for a particular property; if the accuracy decreases, we probably lose some important components in the fingerprint vector. To search for the best reduced fingerprint vector, we need to go through the GA loop. As shown in Figure 11, first we need to
create a random initial generation, and each one of the arrays of ones and zeros is used to evaluate the prediction accuracy. Then the new generation is created by selecting and reproducing, namely, taking the initial generation as the parents and reproducing children by mutation and crossover. If the highest accuracy in the new generation is equal or higher than the set accuracy, the GA stops; if not, the next generation will be created and the loop goes on.

For the chemo-structural fingerprint vector used in our machine learning approach, actually one may generalize the above vector to include all possible $i - j$ pairs, $i - j - k$ triplets, $i - j - k - l$ quadruplets, etc. For the Group IV element-based hybrid polymers, we intuitively chose the $i - i$ pairs, $i - i - i$ triplets as the components of the fingerprint vector. But for the organic polymers, it is not clear which ones are important for a certain property and should be included in the fingerprint vector. Thus, we plan to apply the GA scheme to the organic polymers. In this way, the fingerprint vector for organic polymers can be reduced to contain the components that are essential for determining a property.

5 Impact of Work

- The scheme we developed allows us to design new polymer dielectrics via high throughput computations and machine learning, and provide guidance to experiments. This is a new, but rapidly growing field in material science [46, 47, 48].

- The new polymer dielectrics designed using our approach have both high dielectric constant and large band gap. Not only do they find applications in high energy density capacitors, these new polymers can also be applicable to organic electronics such as organic field effect transistors, where the high dielectric constant materials are needed.

- We have demonstrated how theoretical computer science (machine learning approach) can be effectively employed in materials science where not only machine learning makes the study of a huge number of cases possible, it also allows for a search for correlations between properties in a systematic manner (as shown in Figure 12, which shows the correlations between eight properties shown in Figure 9). Correlation diagram such as the ones in Figure 12 offer a pathway to ‘design’ systems that meet a given set of property requirements. Although we mainly focus on polymer dielectrics design in our work, the scheme we developed is broadly applicable to other systems as well and can contribute to the significant acceleration of the discovery of new materials.
Figure 12: (a) Pearson’s correlation index for each pair of the eight properties, (b) Correlations between the band gap and six properties.

References


