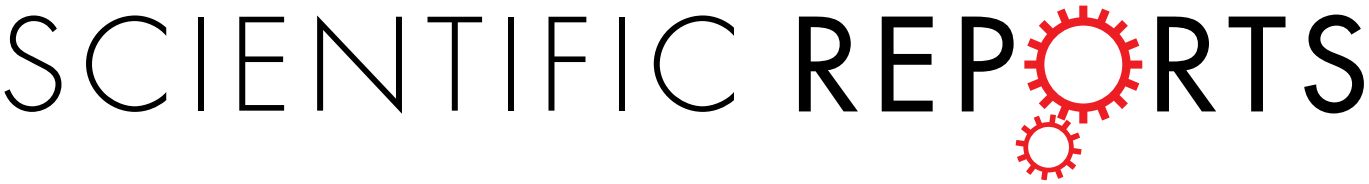

OPEN

Machine Learning Strategy for Accelerated Design of Polymer Dielectrics

Received: 30 September 2015
Accepted: 13 January 2016
Published: 15 February 2016

Arun Mannodi-Kanakkithodi[1], Ghanshyam Pilania[2], Tran Doan Huan[1], Turab Lookman[3] & Rampi Ramprasad[1]

The ability to efficiently design new and advanced dielectric polymers is hampered by the lack of sufficient, reliable data on wide polymer chemical spaces, and the difficulty of generating such data given time and computational/experimental constraints. Here, we address the issue of accelerating polymer dielectrics design by extracting learning models from data generated by accurate state-of-the-art first principles computations for polymers occupying an important part of the chemical subspace. The polymers are 'fingerprinted' as simple, easily attainable numerical representations, which are mapped to the properties of interest using a machine learning algorithm to develop an on-demand property prediction model. Further, a genetic algorithm is utilised to optimise polymer constituent blocks in an evolutionary manner, thus directly leading to the design of polymers with given target properties. While this philosophy of learning to make instant predictions and design is demonstrated here for the example of polymer dielectrics, it is equally applicable to other classes of materials as well.

The materials design process requires the identification of materials that meet a desired application or property need. The traditional routes adopted thus far to meet such design goals involve the determination of the relevant properties of a large number of potential candidates, via high-throughput experiments or computations, and choosing the best cases for further studies and optimisation[1–4]. While powerful and successful, this strategy suffers from two primary drawbacks. First, the consideration of each material in a case-by-case manner is laborious and time-consuming, especially if one were to ignore the availability of past data on the same or similar candidate materials. Second, the prevalent strategy addresses the materials design problem in an 'inverted' manner, i.e., instead of approaching the "desired properties → suitable materials" design problem (previously referred to as *inverse* design[5–8]), the "materials → properties" problem is tackled, and the former design aspect is addressed indirectly through enumeration, i.e., explicit consideration of a large number of candidate materials. Confronting both these hurdles is critical to accelerate, streamline and focus the materials design process.

In the present contribution, strategies are presented to overcome such design challenges for the example of polymer dielectrics—essential components in several applications such as electrical insulation[9], capacitive energy storage[1,10–13], organic photovoltaics[14,15], and flexible, stretchable and wearable electronics[16,17]. While some polymer dielectrics options are available for these applications, given the vastness of the polymer chemical space, it is extremely likely that significant untapped opportunities remain hidden. A more diverse spectrum (than currently available) of new, better and more suitable candidates will constantly be needed to meet growing future needs mandated by performance measures, amenability to synthesis and compatibility with other parts of devices. Rational and accelerated polymer design strategies and solutions would thus be enormously useful.

Our starting point in the present work is the generation of reference property data (using first principles computations) for a benchmark set of polymers spanning a particular chemical subspace. Interpolative statistical learning concepts[18–25] are then used to train an *on-demand* instant property prediction model using this initial dataset, via an intermediate (and critical) 'fingerprinting' step that converts every polymer to a numerical string (c.f. Fig. 1). The prediction scheme produces accurate results for cases not used in the training phase (but falling within the same chemical subspace), as demonstrated by comparisons with more laborious first principles

[1]Department of Materials Science and Engineering, Institute of Materials Science, University of Connecticut, 97 North Eagleville Road, Storrs, Connecticut 06269, USA. [2]Materials Science and Technology Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA. [3]Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA. Correspondence and requests for materials should be addressed to R.R. (email: rampi.ramprasad@uconn.edu)

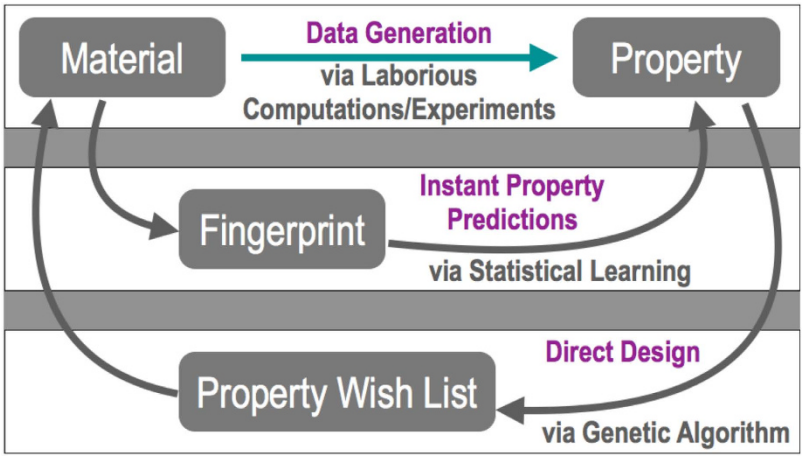

Figure 1. The overall outline of this work. This work is divided into three stages: the data generation stage, the instant property prediction stage and the direct design stage.

computations and experimental measurements. Such a model via an enumeration scheme can be used to predict the properties of a plethora of new candidate materials in an attempt to search for cases meeting a particular set of property needs. Furthermore, one can make rapid go/no-go decisions on whether a new synthesisable polymer is worth pursuing or not.

The enumeration approach to materials design is not the most efficient one, as it involves consideration of an enormous number of cases, most of which will not be viable in the end (thus leading to low success rates). A better approach is to use the on-demand property prediction scheme within a genetic algorithm[26,27], to directly tackle the "desired properties → suitable materials" design problem. Several polymers that meet a property requirement criterion are designed directly here using such a strategy at a minuscule fraction of the time required for enumeration. The predicted property results of the designed polymers are validated by explicit first principles computations.

The suite of tools and strategies that emerge from this effort take us a step closer to rational, accelerated and direct design of materials in general, and polymer dielectrics in particular. These strategies can also be extended to larger polymer chemical and property subspaces. The essential ingredients of this effort are illustrated in Fig. 1, and described in detail in the following.

Data Generation

Polymers in our chemical subspace contain a number of linearly repeating chemical building blocks chosen from the following pool: CH_2, NH, CO, C_6H_4, C_4H_2S, CS and O. These blocks are commonly found in much of the known polymer space, like polyethylene, polyureas, polythioureas, polyesters, etc. Polymers built from this same pool of blocks were considered by us in the past[1], but the data from this previous work apply to just individual (i.e., isolated) polymer chains, thus leading to (extrapolated) property data with significant uncertainties. Here, we go beyond the past work, and determine the 3-dimensional packing and crystal structure of polymers arising from these building blocks. Properties computed for such 3-dimensional structures constitute a robust dataset.

For the creation of the initial dataset via first principles computations, we restrict ourselves to 4-block polymers, that is, polymers built with 4 blocks in the repeating unit (with each of these drawn from the pool of 7 building blocks). As described below, the goal of the learning models developed here is to use this dataset to predict the properties of polymers with arbitrarily long repeat units. A total of 406 symmetry-unique 4-block polymers can be formed using the 7 building blocks, of which only 284 were considered here. This reduced number is because chemical intuition and prior knowledge dictates that some combinations of adjoining chemical blocks make for unstable systems, leading to the elimination of all polymers consisting of O-O, CS-CS, CO-CO and NH-NH pairs. The crystal structures of all 284 4-block polymers were determined using the minima hopping method[28,29], with the necessary total potential energies and atomic forces computed using density functional theory (DFT). The structure prediction and DFT details are provided in the Methods section; all the DFT predicted data for the 4-block polymers is provided in the Supplementary Information.

With the 3-dimensional structure of all 284 polymers determined, their relevant properties were calculated. In the present work, we focus on the bandgap (E_{gap}), computed using hybrid electron exchange-correlation functionals, and the electronic (ϵ_{elec}), ionic (ϵ_{ionic}) and total ($\epsilon_{total} = \epsilon_{elec} + \epsilon_{ionic}$) dielectric constant, computed using density functional perturbation theory, as described in the Methods section. In the case of dielectrics, the bandgap and dielectric constant are the primary properties of interest, generally used in an initial screening stage, regardless of the specific applications[1,30–32].

The workflow underlying the data generation step is depicted in Fig. 2a, and the DFT results are portrayed in Fig. 2b. It can be seen from Fig. 2b that ϵ_{elec} (shown in purple) seems to follow an inverse relationship to E_{gap}, whereas ϵ_{ionic} (shown in yellow) has no particular relationship with E_{gap}[1,32]. Given the larger range of values of ϵ_{elec} $(2-10)$ than those of ϵ_{ionic} $(0-3)$, this effect translates to ϵ_{total}, and we can see an overall inverse relationship between ϵ_{total} (shown in red) and E_{gap} as well. For high dielectric constant polymer insulator applications, we are interested in polymers that simultaneously show high ϵ_{total} and large E_{gap}. Indeed, based on this notion, our past

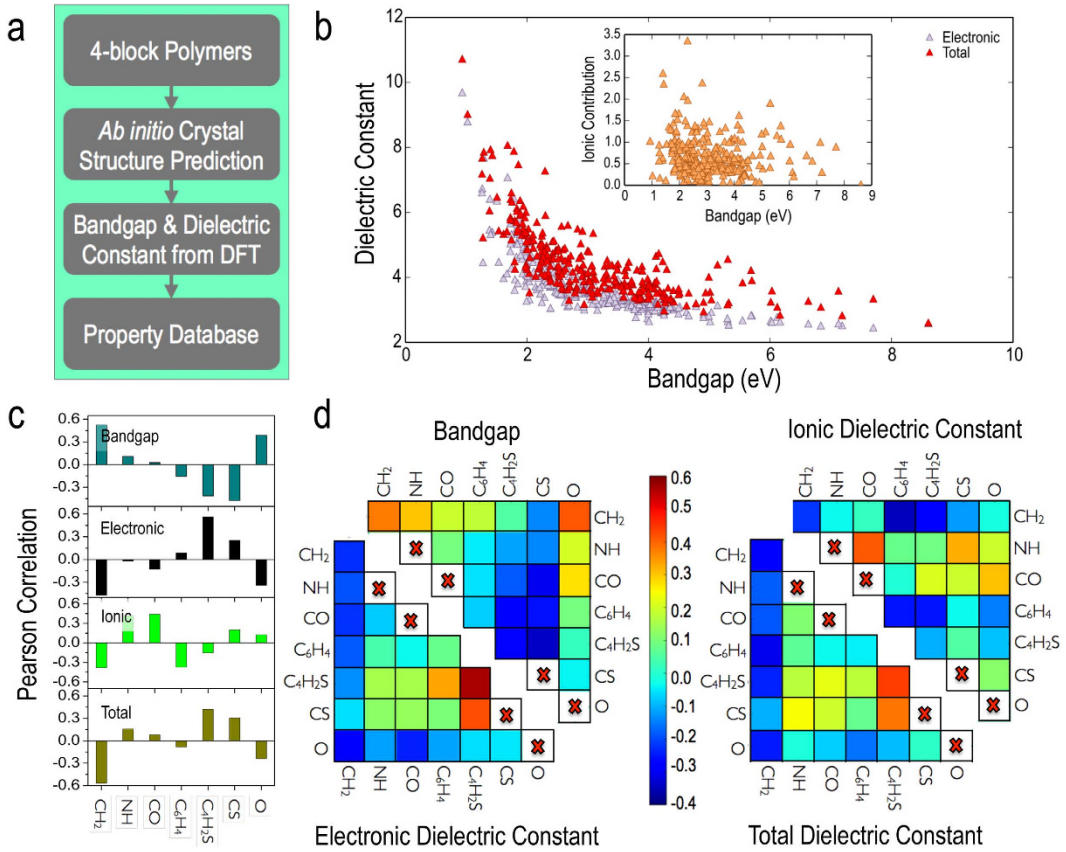

Figure 2. Data generation from DFT and origins of properties. (**a**) The different steps involved in generating a database of the properties of 4-block polymers. (**b**) The DFT computed electronic, ionic and total dielectric constants plotted vs bandgaps for the 4-block polymers. (**c**) Pearson correlation coefficients between fingerprint M_I and the 4 properties. (**d**) Correlations between fingerprint components of M_{II} and the properties shown in the form of heat maps. Red crosses represent the components which lead to unstable polymers and were not considered in the present study (see text for details).

work—although it dealt with isolated polymer chains and estimates of ϵ_{total} and E_{gap} in the absence of crystal structure information—has lead to new polymer dielectric solutions[1,10].

Fingerprinting Polymers

While high-throughput data generation efforts can provide useful 'lead candidates' with desired properties, the natural question that arises is whether one can understand the origins of the attractive behavior, and harness this understanding to search for other suitable options. Within the context of polymeric materials under investigation here, the origins should be traceable to the identities of the chemical building blocks. This comes from the theory that electronic and dielectric properties of organic polymers can be effectively expressed in terms of a sum of contributions from different constituent groups[33]. These contributions are in the form of polarisabilities and dipole-dipole interactions from the groups, with different weights attached to different groups. In the case of our polymers, some building blocks, or some combination of blocks, are expected to have a particular influence on the properties being studied.

Thus, if we can numerically represent—or fingerprint—our polymers based on their building block identities, correlations can potentially be established between the fingerprints (or parts of it) and properties. Indeed, numerically representing molecules and materials is emerging as an active topic of inquiry within materials science, physics and chemistry in recent years[34–37]. Descriptors such as this have historically been used in cheminformatics and related fields like medicinal chemistry and drug discovery[38,39]. Key requirements of such representations are that the fingerprints should be intuitive, easily computable, invariant with respect to translations and rotations of the material, invariant to permutations of like atoms or motifs, and generalisable to all cases within the same chemical subspace. Such an approach was recently implemented by us for fingerprinting organic molecules and crystals in terms of constituent atom types (analogous to the chemical building blocks in case of our polymers here)[6]. Singles, doubles and triples of different atom types (based on atom identity and coordination of bonds) were successfully correlated with a number of calculated properties.

A simple polymer fingerprint could therefore be a count of the number of different types of building blocks (e.g., the number of CH_2 blocks, the number of C_6H_4 blocks, etc.), normalised by the total number of blocks in the repeat unit. This would give rise to a 7 dimensional vector, each component of which corresponds to one of the blocks and is related to the number of times it appears in the given polymer repeat unit. We call this fingerprint

M_I. While it is a simple and elegant way of representing a polymer, M_I does not take the effects of neighbouring blocks into account. Thus, we go a step higher in complexity and propose fingerprint M_{II}, which is a count of the number of different types of pairs of building blocks in the polymer, normalised again by the total number of blocks in the repeat unit. M_{II} is defined as a 7×7 matrix, every component of which corresponds to any one pair of two neighbouring blocks (eg. CH_2-NH pairs, CS-O pairs, etc.). Similarly, a fingerprint M_{III} can be defined which would be a $7 \times 7 \times 7$ matrix each component of which refers to any triplet of blocks (CH_2-NH-CO triplets, C_4H_2S-C_6H_4-CS triplets, etc.).

In this fashion, we could go to higher dimensional fingerprints with more information added at every step; in the limit that we consider *n*-tuple block combinations, we can uniquely represent any polymer out of an *n*-block polymer repeat unit chemical space. We note that this general fingerprinting concept was presented before by us[32], but only a subset of the fingerprint components (namely, the diagonal) was considered earlier. We argue here that there is no reason to choose such a restricted fingerprint, and moreover, show in the Methods section that the full tensorial representation satisfies key sum rules. With the present prescription, the fingerprint for any given n-block polymer is populated by assigning a certain score to every block or pair of blocks or triplet of blocks that is encountered, with the counting done from either end of the polymer repeat unit to take periodicity and inversion into account. The scores are always averaged and normalised by the total number of blocks in the repeat unit. The averaging step ensures that sum rules are satisfied, and normalisation assures that the fingerprints are generalisable to repeat units of arbitrary length. It should be noted that this polymer fingerprint does not take into account spatial degrees of freedom or other structural factors, and would thus not distinguish between two polymers with the same repeat unit but different crystal structural arrangements.

For ease of initial discussion, we consider the fingerprints M_I and M_{II}. Correlations between the different components of fingerprint M_I and 4 properties (ϵ_{elec}, ϵ_{ionic}, ϵ_{total} and E_{gap}) are shown in Fig. 2c. The coefficients plotted on the *y*-axes were obtained using the Pearson correlation analysis, which gives us values between -1 and $+1$ showing the degree of negative or positive correlation between any property and any component of the fingerprint vector. The opposite behaviour of ϵ_{elec} and E_{gap} can be ascertained by observing their respective plots: CH_2 and O blocks make notable positive contributions to E_{gap} and negative contributions to ϵ_{elec}, whereas C_4H_2S and CS contribute positively to ϵ_{elec} and negatively to E_{gap}. The same effects largely translate to ϵ_{total} as well while for ϵ_{ionic}, CO and NH blocks contribute the most.

Results for a similar Pearson correlation analysis between M_{II} and the 4 properties are shown in Fig. 2d in the form of half-matrix heat maps. The shade of the colour in any matrix component (based on the adjoining colour scale) shows how positively or negatively that particular pair of blocks is correlated with the given property. Once again, it can be seen how the heat map for E_{gap} is really opposite to that of ϵ_{elec} or ϵ_{total} in terms of the spectrum of colours (dark blue to dark red). While C_6H_4-C_4H_2S, C_4H_2S-C_4H_2S and C_4H_2S-CS pairs make the most positive contributions to ϵ_{elec} and CH_2-O and CO-O pairs make the most negative contributions, the roles of these pairs are just reversed when considering their contributions to E_{gap}. In case of ϵ_{ionic}, NH-CO, NH-CS and CO-O pairs contribute to its increase while CH_2-C_6H_4 and CH_2-C_4H_2S pairs have the opposite effect. It is now possible for us to come up with educated combinations of different kinds of pairs of building blocks targeted towards increasing the dielectric constant or the bandgap or indeed, both. In light of these insights, it is not surprising that polymers with [-NH-CO-NH-C_6H_4-], [-NH-CS-NH-C_6H_4-], and [-NH-CO-NH-C_6H_4-] repeat units were singled out in past work as promising dielectrics for energy storage applications[1].

On-Demand Property Prediction

While qualitative notions such as discussed above are useful, a quantitative property prediction model that is fast (because it by-passes the DFT route to property predictions) would satisfy several practical needs. Following previous work, we use kernel ridge regression (KRR)[40] to establish a quantitative mapping between the polymer fingerprints on the one hand and the relevant properties (namely, E_{gap}, ϵ_{elec}, and ϵ_{ionic}) on the other. KRR is a statistical or machine learning algorithm capable of handling non-linear relationships[6,32]. By comparing the fingerprint, say M_{III}, of a new polymer with those of a set of reference cases for which property values are known, an interpolative prediction of the property of the new polymer may be obtained. In practice, the machine learning prediction model is developed for a subset of the available dataset, referred to as the training set, and the performance of the model is tested on the remainder of the dataset, referred to as the test set. Model development based on the training set also included internal cross-validation to minimise over-fitting and ensure model generality. In the present work, about 90% of the 284 4-block polymer dataset was taken to be the training set, and the remaining 10% was placed in the test set. The optimal training set size was determined by studying the ML model performances for different training set sizes; this data is presented in the Supplementary Information. Further details of the KRR method and specifics of the model development are provided in the Methods section.

The plots in Fig. 3 show E_{gap}, ϵ_{elec} and ϵ_{ionic} as predicted using the KRR-based machine learning (ML) model (and fingerprint M_{III}) versus the respective DFT values. The insets also show the relative error distribution for each property prediction, indicating that the average error for all three properties is of the order of 10% or less. We thus have a model in our hands that will convert a fingerprint (M_{III}, in the present illustration) to property values with errors that are reasonable (given the efficiency of the prediction process relative to DFT). Prediction performances using M_I and M_{II} for KRR are shown in the Supplementary Information for completeness.

The true power of such a property prediction model is its ability to instantly predict E_{gap}, ϵ_{elec} and ϵ_{ionic} for a polymer with arbitrarily long repeat unit (but with the building blocks drawn from the same pool of 7), without needing to pursue the cumbersome approach of structure prediction and DFT. The workflow involved in predicting the properties of new *n*-block polymers is depicted in Fig. 4a. If one were to pursue the enumeration approach, it is straightforward to list all possible *n*-block polymers for any given *n*, as long as *n* is a small enough number. To illustrate this, we came up with all the possible symmetry-unique 6-block polymers (~6000 in number) and 8-block polymers (~150000 in number), determined their respective fingerprints, and estimated their

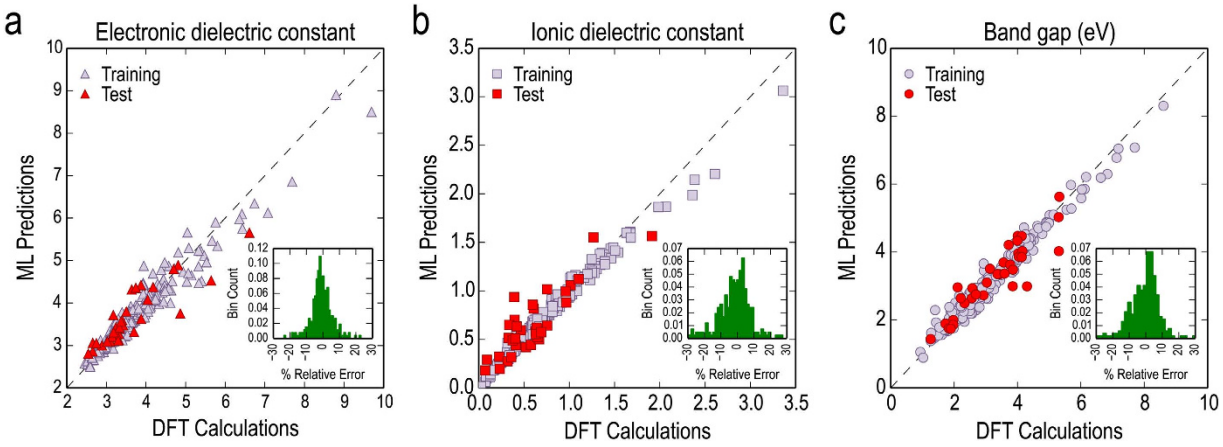

Figure 3. Prediction model performances. Comparison of the KRR property predictions with DFT evaluated properties for the prediction models for ϵ_{elec}, ϵ_{ionic} and E_{gap} respectively.

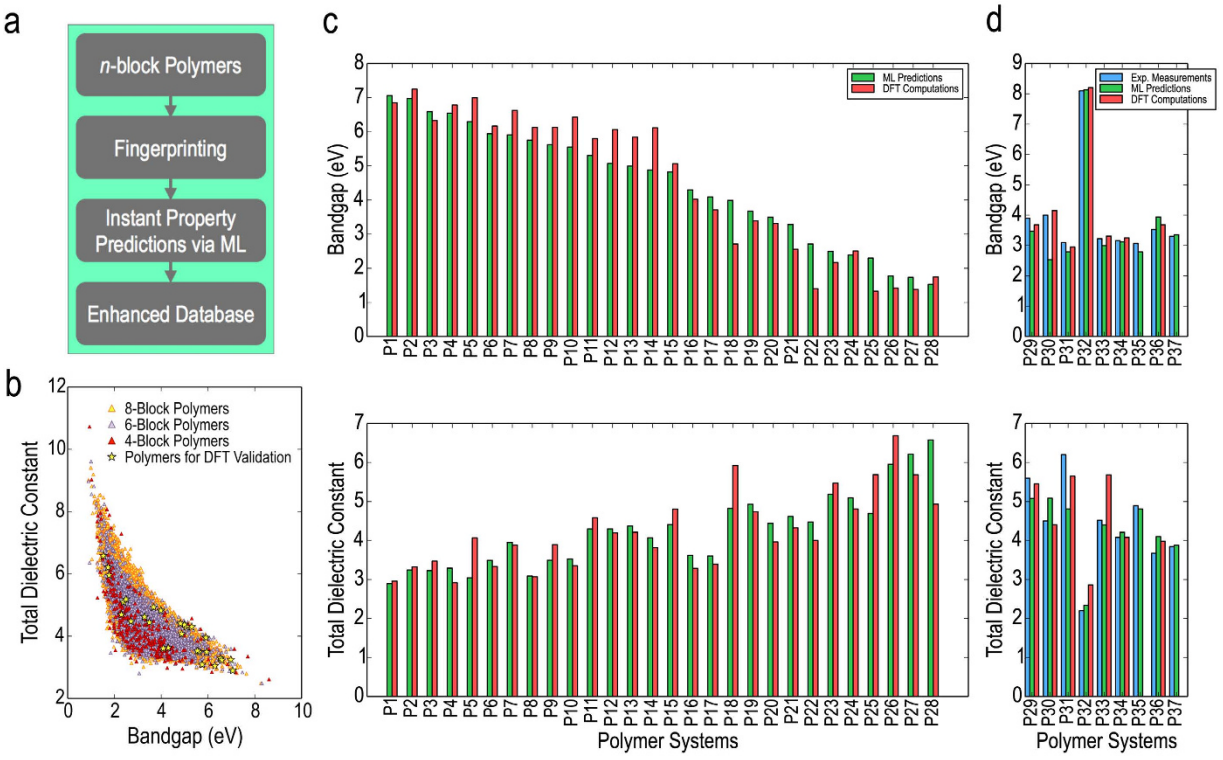

Figure 4. On-demand property prediction of polymers. (**a**) The steps involved in predicting properties of any given *n*-block polymer using the instant prediction models. (**b**) Dielectric constants and bandgaps from the prediction models plotted against each other for all 6-block polymers and 8-block polymers, with the computational data for 4-block polymers also shown for reference. (**c**) Machine learning predicted and DFT computed properties of 28 polymers obtained by applying the direct design scheme to different ranges of dielectric constants and bandgaps. (**d**) The machine learning predicted, DFT computed and experimentally measured properties of some previously synthesised polymers.

properties using our ML model. Figure 4b shows the predicted ϵ_{total} ($=\epsilon_{\text{elec}}+\epsilon_{\text{ionic}}$) plotted against the predicted E_{gap} values for all the 6-block polymers and 8-block polymers, as well as for the considerably smaller number of the 4-block polymers. Figure 4b is a demonstration of how one may use interpolative statistical learning methods to densify the population within a chemical subspace. We thus have a lot more options to choose from.

The predictive performance of our model can be put to test in two ways: by comparing our predictions with actual DFT calculations, and by comparing them with available laboratory measurements. First, we validate our ML model against DFT calculations. A selection of 8-block polymers ranging from low (high) to high (low) values of ϵ_{total} (E_{gap}) was chosen out of Fig. 4b (shown by stars in figure; incidentally, these were also the cases

Label	Polymer Repeat Unit	Label	Polymer Repeat Unit
P1	CH_2-O-CH_2-O-CH_2-CH_2-CH_2-CH_2	P20	O-C_6H_4-CO-C_4H_2S-CO-NH-O-CO
P2	CH_2-O-CH_2-O-CH_2-CH_2-CH_2-O	P21	CH_2-CH_2-O-CS-NH-CS-C_6H_4-NH
P3	CH_2-NH-CH_2-CH_2-CH_2-O-CH_2-O	P22	C_6H_4-C_6H_4-CH_2-CS-C_4H_2S-CS-CH_2-O
P4	CH_2-CH_2-O-CO-O-CH_2-CH_2-O	P23	C_6H_4-NH-C_6H_4-CS-NH-C_4H_2S-CO-NH
P5	CO-O-CH_2-CH_2-CH_2-CH_2-CH_2-O	P24	CO-C_4H_2S-NH-CS-O-C_4H_2S-NH-C_4H_2S
P6	CH_2-CH_2-O-CO-NH-CH_2-CH_2-O	P25	CS-CO-CH_2-CH_2-NH-C_6H_4-CS-C_6H_4
P7	CH_2-NH-CO-NH-CH_2-O-CH_2-O	P26	C_6H_4-NH-C_4H_2S-C_4H_2S-CS-C_4H_2S-C_4H_2S-NH
P8	CH_2-CH_2-CH_2-CH_2-NH-CO-CH_2-CH_2	P27	C_4H_2S-C_4H_2S-C_4H_2S-CS-C_4H_2S-NH-CS-NH
P9	CO-NH-O-CH_2-CH_2-CH_2-CH_2-O	P28	C_4H_2S-CS-C_4H_2S-CS-CO-NH-C_6H_4-C_4H_2S
P10	CH_2-O-CO-NH-CH_2-CH_2-NH-CH_2	P29	NH-CO-NH-C_6H_4
P11	CH_2-NH-CH_2-NH-CO-NH-CO-NH	P30	CO-NH-CO-C_6H_4
P12	CH_2-NH-CO-O-NH-CO-NH-O	P31	NH-CS-NH-C_6H_4
P13	CO-NH-CO-O-CO-NH-CH_2-NH	P32	CH_2-CH_2-CH_2-CH_2
P14	CO-NH-CO-NH-CH_2-CH_2-CH_2-NH	P33	NH-CS-NH-C_6H_4-NH-CS-NH-C_6H_4-O-C_6H_4
P15	CO-NH-CO-CH_2-NH-CO-O-NH	P34	NH-CS-NH-C_6H_4-NH-CS-NH-C_6H_4-CH2-C_6H_4
P16	C_6H_4-O-CO-CH_2-CO-CH_2-CH_2-O	P35	NH-CS-NH-C_6H_4-NH-CS-NH-C_6H_4
P17	CH_2-CH_2-CO-O-CO-CH_2-C_6H_4-C_6H_4	P36	NH-CS-NH-C_6H_4-NH-CS-NH-$[CH_2]_6$
P18	CO-NH-O-NH-CO-NH-C_4H_2S-NH	P37	NH-CS-NH-C_6H_4-CH_2-C_6H_4
P19	CO-NH-CO-NH-CO-NH-C_4H_2S-NH		

Table 1. Polymer repeat units denoted by the labels P1 to P37 in Fig. 4c,d.

identified by our genetic algorithm, discussed in the next section, but the same examples serve the present purpose of ML model validation). The stable crystal structures of these 8-block polymers were determined, following which their properties were calculated using DFT. Figure 4c compares the ML prediction with the corresponding DFT results. *As can be seen, the agreement is impressive indicating that the prediction model trained on 4-block cases is transferable to polymers with repeat units of arbitrary size.*

Next, in Fig. 4d, we compare the on-demand predictions with experimental values for polymers synthesised and tested in the recent past[1,10], as well as the corresponding DFT results, for completeness. These polymers were synthesised following the earlier work[1] on high-throughput computational data generation using the isolated polymer chains model; this means we have available experimental as well as computational quantification of ϵ_{total} and E_{gap} for a number of polymers which are predictable with our prediction models. Clearly, again, the performance of the ML model is impressive. The closeness of our predictions with first principles as well as with actual experiments allows us to state with some confidence that we have the means to instantly, and with reasonable accuracy, predict the properties of any *n*-block polymer belonging to the chemical subspace under consideration. All the polymers plotted in Fig. 4c,d are denoted by some labels, and the polymer corresponding to each label is mentioned in Table 1. The ML predictions are always close to the experimental values, validating our claim of accelerating property prediction for arbitrarily long polymer chains.

On-Demand Direct Design

Although the entire expanse of the chemical space can be covered using enumeration, it is essentially a brute-force search for suitable polymers, and as such not the best possible design strategy. For instance, enumerating for 8-block, 10-block and 12-block polymers will lead to $\sim 1.5 \times 10^5$, 5×10^6 and 5×10^7 systems respectively, which are unreasonably large numbers considering the property domain of interest may restrict us to a small fraction of that. We thus attempted to find an efficient way of obtaining specific *n*-block polymers that simultaneously show a certain desirable dielectric constant and a desirable bandgap, without having to individually consider every possible polymer. Such a model would make the "desired properties → suitable materials" route an instant, on-demand reality[5–8].

We applied a genetic algorithm (GA) approach as the means to optimise the polymers given the target properties. It has been shown that GA is a very efficient approach in searching for materials with desired properties when compared to other approaches like random search and even chemical-rules based search[26]. The idea here is to start with a random initial population of *n*-block polymers (for any given *n*) and let them undergo evolution (in terms of constituent blocks and their neighbours) based on the principles of GA, finally yielding a set of polymers with properties closest to the provided targets. At any step, the properties of the polymers are computed instantly using the on-demand prediction ML model we developed and explained in the previous section. The series of steps followed in this method are shown in Fig. 5a. In an earlier work[6], we implemented the same philosophy but used a simulated annealing approach instead of GA for designing organic molecules with specific target properties.

Given the target ϵ_{total} and E_{gap}, and the number of blocks in the polymer repeat unit (the value of *n*), the algorithm generates a list of 300 *n*-block polymers which serves as the first generation. Based on the predicted property values, a fitness score (explained in detail in Methods) is assigned to every polymer and all the polymers are ranked according to this score. While polymers with satisfactory fitness scores survive (this is called elitism), the rest undergo different kinds of evolution, namely crossover and mutation (again, explained in Methods). New

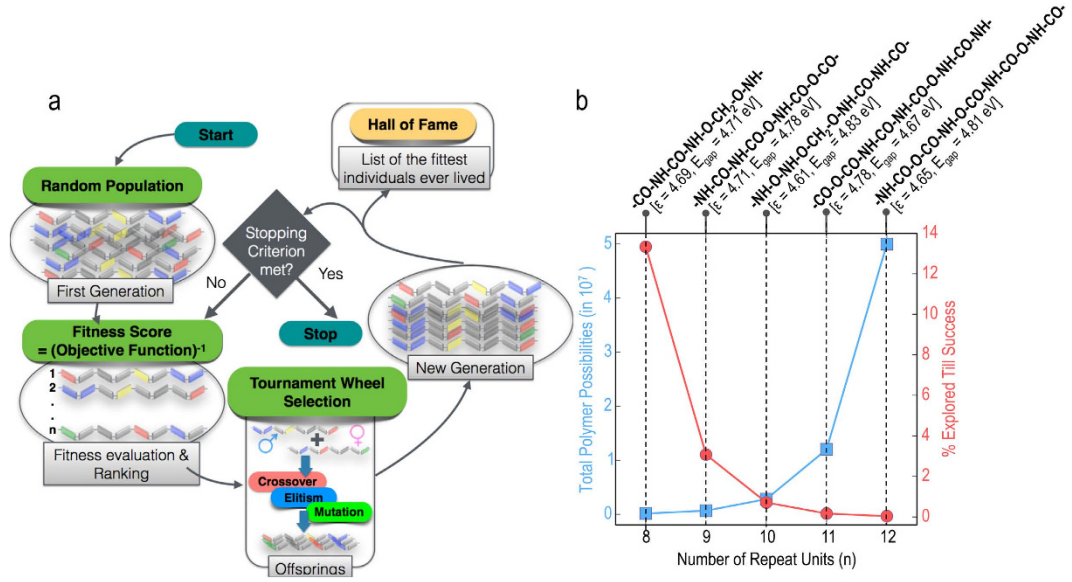

Figure 5. On-demand direct design of polymers. (**a**) The steps involved in the genetic algorithm (GA) approach leading to direct design of polymers. (**b**) The exponential increase in total polymer possibilities for increasing number of repeating blocks, and the simultaneous decrease in the percentage of points to be explored till success. Also shown are one optimal polymer each for each case for a target dielectric constant and bandgap of 5 and 5 eV respectively.

generations of polymers are produced in this manner; a stopping criterion is provided based on the fitness score, and once polymers with suitable fitness scores are obtained, the algorithm stops. From every generation, the polymers with fitness scores above a certain threshold are compiled as the list of best solutions. At the end of the algorithm, this list contains the final set of optimal polymers showing the desired ϵ_{total} and E_{gap}.

For a demonstration and validation of this approach, we restricted our initial search to 8-block polymers only, as this provides us with a substantial population of systems to explore while ensuring the system size does not become so large as to render subsequent first principles validation extremely expensive. We took 6 different (ϵ_{total}, E_{gap}) combinations as the targets, and allowed the algorithm to search for suitable 8-block polymers showing each combination of properties. Figure 4c gives a glimpse of the results: we show a few polymers each obtained for the different targets we provided. The ML model predicted property values for these polymers are always close but not exactly the same as the target values—but these are the polymers showing the highest fitness scores for the given targets. A comparison between the solutions obtained from the genetic algorithm approach and the solutions obtained from the enumerated list of all 8-block polymers, is shown in the Supplementary Information. It is seen that for a given target property set, this scheme does indeed determine a number of optimal solutions, if not all of them.

To understand exactly how valuable the direct design scheme is, we need to quantify the speed of the GA approach when compared to enumeration. Taking the example of 8-block polymers, while there are a total possible ~150000 such systems, GA is able to traverse a small percentage of the points in determining the required polymer(s). Upon going to higher block systems, like 9-block or 10-block polymers, the total possibilities are exponentially higher but the percentage of points the algorithm needs to explore is even smaller. Figure 5b shows that despite the exponential increase in total polymer possibilities, as the number of repeating units n increases, a smaller and smaller percentage of points need to be considered by the algorithm in order to obtain the optimal polymer(s). Also shown in Fig. 5b are n-block polymers obtained for different values of n for a target ϵ_{total} of 5 and a target E_{gap} of 5 eV. Thus, with actual polymer outputs (with arbitrarily long chains) as well as a quantification of the speed-up, we have in our hands an efficient polymer design model that negates the need for enumeration followed by down-selection of desired systems.

Summary

Given a material, we want to have the means to instantly estimate its properties, and thus make a quick decision on its suitability for an application. We have demonstrated how carefully created and curated materials data can be used to train statistical learning models. These models, following testing and validation, require merely the fingerprint of a new material to output its properties. We have further shown how a genetic algorithm can be combined with the learning models to determine specific materials that possess a certain set of desired properties. In this manner, using the example of polymer dielectrics, we have successfully tackled both the "desired properties → suitable materials" and the "materials → properties" problems. The materials design philosophy applied in this work can be used for any class of materials, as long as there is sufficient data available for training, and an intuitive and easily attainable material fingerprint can be proposed.

We have at our disposal a polymer database, generated from first principles and expanded substantially using the on-demand prediction model, which enables us to recommend a number of new polymers previously not considered for synthesis and testing. While this fulfills one of the specific aims we set out to achieve, that is expand on the list of promising dielectric polymers, the prediction and design models can prove to be very valuable tools for polymer synthesists. Not only can they make an instant go/no-go decision on any polymer, but they can actively seek specific polymers that would suit their requirements. This adds a new, very useful dimension to the field of dielectric polymer design.

There are of course limitations to the models, not least in terms of the chemical blocks currently considered. The information in the models, and thus the possible guidance, is restricted to combinations of only the 7 basic building blocks. For an extension to more blocks, sufficient amount of data on polymers containing those blocks needs to be generated, and the models retrained. Further, the fingerprints currently used only take into account the constituent blocks in the polymers, and exclude information on how the polymer chains are stacked against each other, and other factors that may affect the properties. While this is indeed a limitation, we argue that such a fingerprint still functions very well, and points to the fact that materials can typically be boiled down to fairly simple numerical representations. Lastly, it must be noted that all property predictions from the on-demand prediction model come with some uncertainties, which are inevitable in any statistical learning enterprise. Nevertheless, we have established a useful materials design protocol that should help accelerate the design and discovery of polymers encompassing a much larger chemical subspace, or non-polymeric but quasi-1D systems (e.g., superlattice heterostructures).

Methods

First principles computations. A unit cell is set up containing 2 polymer chains stacked next to each other. The Minima Hopping structure prediction algorithm[28,29,41,42] was applied on the starting polymer geometry, leading to the exploration of many low energy crystal structure arrangements which were ranked according to their relative energies. For each polymer, the lowest energy crystal structure thus obtained was taken for DFT property calculations. DFT[43] as implemented in the Vienna ab initio software package (VASP)[44] was applied, and relaxation was performed using the rPW86 functional wherein the DFT-DF2 vdW correction is applied[45] to capture the van der Waals interactions in the polymer correctly[46]. We used projector-augmented wave (PAW)[47] pseudopotentials and imposed a tight energy convergence criterion of 10^{-8} eV and an energy cut-off of 500 eV. The relaxed geometry thus obtained went as input into a subsequent density functional perturbation theory (DFPT)[48] calculation, which provided us with the dielectric constant tensor that includes the electronic component[49] as well as the ionic (lattice) component[50]. The reported dielectric constant values were obtained by determining the trace of the respective dielectric tensor (a 3×3 matrix in this case). Further, the Heyd-Scuseria-Ernzerhof (HSE)[51] functional was used on the relaxed geometries to obtain the HSE bandgap values, which are known to be more reliable[52].

Fingerprint. M_I, M_{II} and M_{III} are characterised by a number of key mathematical constraints which have been listed below-

1. The sum of all the elements in any fingerprint should be equal to the total number of blocks in the polymer (N). Thus: $\sum_{i=1}^{7} M_I^i = N$, $\sum_{i,j=1}^{7} M_{II}^{ij} = N$ and $\sum_{i,j,k=1}^{7} M_{III}^{ijk} = N$.
2. The sum of elements in any row or column of M_{II} should be equal to the total number of blocks of that kind in the polymer. This can be written as: $\sum_{j=1}^{7} M_{II}^{ij} = M_I^i$. Similarly, the sum of elements in any given 7×7 matrix plane in M_{III} should be equal to the total number of blocks of that kind in the polymer, which can be written as: $\sum_{j,k=1}^{7} M_{III}^{ijk} = \sum_{j=1}^{7} M_{II}^{ij} = M_I^i$.
3. The periodic symmetry in the polymer dictates that the fingerprint matrix diagonal acts as a mirror; the corresponding elements on either side of it should be equal. That is, $M_{II}^{ij} = M_{II}^{ji}$ and $M_{III}^{ijk} = M_{III}^{kji}$.
4. The diagonal elements in any fingerprint matrix should be integer values, that is, M_{II}^{ii} and $M_{III}^{iii} \in$ the set of non-negative integers.

Regression. Kernel Ridge Regression (KRR) was applied to develop a similarity-based model, where the Euclidean distances between fingerprints are used to compute a distance Kernel. In this work, a Gaussian Kernel was used. A given property is then expressed as a weighted sum of the Gaussians. The different parameters that go into the training of such a model— the Gaussian width parameter, the regularisation parameter (which helps to prevent overfitting in the data)[40], and the coefficients of the Gaussians— are changed in a systematic manner so as to achieve maximum closeness between the weighted sum of the kernels and the property. From our database of 284 4-block polymers, all the points were randomly divided into two sets— the training set (250 points) and the test set (34 points). The training set was used to train the KRR model and thus come up with the prediction model with the minimum error in property prediction. The best models thus obtained were used to predict the properties on a test set in order to evaluate their true out-of-sample performance. To ensure the best possible training in an unbiased manner, a cross-validation technique was used where the training set itself is divided into two sets and one set is used for preliminary training with validation done on the other.

Genetic Algorithm. Based on the target dielectric constant and bandgap, an objective function was defined as the following-

$$W = [\epsilon - \epsilon^{target}]^2 + \left[E_{gap} - E_{gap}^{target}\right]^2$$

where ϵ^{target} and E_{gap}^{target} are the target dielectric constant and bandgap values respectively, while ϵ and E_{gap} are the dielectric constant and bandgap of the polymer undergoing optimisation. This function would be minimised when the difference between either property of the polymer and the respective target property is the least. Further, a *Fitness Score* was defined as the inverse of the objective functional value, and acted as the measure of suitability of any system. We devised a polymer encoding system that converted any *n*-block polymer into an *n*-component vector, assigning a number between 0 and 6 to each of the 7 motifs respectively. Using completely random values for this vector, an initial population of 300 polymers was generated. Properties were instantly calculated for all these polymers using the on-demand prediction models, and the fittest polymers (showing the highest fitness scores) were selected. Mating is performed between these individuals using a combination of crossovers, elitism and mutation[26], giving rise to the 'offspring' polymers that then go forth to the next generation of polymers. In crossover, some of the vector components of the parent polymers were simply exchanged to generate the children. Elitism means preserving a few of the fittest parent polymers in the next iteration, whereas with mutation, we changed some of the vector components of the parents randomly to obtain the children. Thus, generation after generation of polymers was studied and those with the highest fitness scores at every generation went into the list of best solutions. In the end, this list would contain the best individuals that ever lived (that is, the polymers with properties closest to the target values ϵ^{target} and E_{gap}^{target}), and these would be our solutions.

References

1. Sharma, V. *et al.* Rational design of all organic polymer dielectrics. *Nature Commun.* **5,** 4845 (2014).
2. Jain, A. *et al.* A high-throughput infrastructure for density functional theory calculations. *Comp. Mat. Sc.* **50,** 22952310 (2011).
3. Strasser, P. *et al.* High throughput experimental and theoretical predictive screening of materials A comparative study of search strategies for new fuel cell anode catalysts. *J. Phys. Chem. B* **40,** 1101311021 (2003).
4. Greeley, J., Jaramillo, T. F., Bonde, J., Chorkendorff, I. & Norskov, J. K. Computational high-throughput screening of electrocatalytic materials for hydrogen evolution. *Nat. Mater.* **5,** 909–913 (2006).
5. Yu, L., Kokenyesi, R. S., Keszler, D. A. & Zunger, A. Inverse design of high absorption thin-film photovoltaic materials. *Adv. En. Mat.* **3,** 4348 (2013).
6. Huan, T. D., Mannodi-Kanakkithodi, A. & Ramprasad, R. Accelerated materials property predictions and design using motif-based fingerprints. *Phys. Rev. B* **92,** 014106 (2015).
7. Hautier, G., Fischer, C. C., Jain, A., Mueller, T. & Ceder, G. Finding natures missing ternary oxide compounds using machine learning and density functional theory. *Chem. Mat.* **22,** 37623767 (2010).
8. d'Avezac, M., Luo, J., Chanier, T. & Zunger, A. Genetic-algorithm discovery of a direct-gap and optically allowed superstructure from indirect-gap Si and Ge semiconductors. *Phys. Rev. Lett.* **108** 027401 (2012).
9. Mller, K., Paloumpa, I., Henkel, K. & Schmeisser, D. A polymer high-k dielectric insulator for organic field-effect transistors. *J. App. Phys* **98,** 056104 (2005).
10. Ma, R. *et al.* Rational design and synthesis of polythioureas as capacitor dielectrics. *J. Mater. Chem. A*. **3,** 14845 (2015).
11. Baldwin, A. F. *et al.* Poly(dimethyltin glutarate) as a prospective material for high dielectric applications. *Adv. Mat.* **27,** 346351 (2015).
12. Baldwin, A. F. *et al.* Rational design of organotin polyesters. *Macromolecules* **48,** 24222428 (2015).
13. Pilania, G. *et al.* New group IV chemical motifs for improved dielectric permittivity of polyethylene. *J. Chem. Inf. and Modeling* **53,** 879886 (2013).
14. Facchetti, A. Pi-conjugated polymers for organic electronics and photovoltaic cell applications. *Chem. Mat.* **23,** 733758 (2011).
15. Chan, S.-H. *et al.* Synthesis, characterization, and photovoltaic properties of novel semiconducting polymers with thiophenephenylenethiophene (TPT) as coplanar units. *Macromolecules* **41,** 55195526 (2008).
16. Ling, Q.-D. *et al.* Polymer electronic memories: Materials, devices and mechanisms. *Prog. in Pol. Sc.* **33,** 917978 (2008).
17. Yan, H. *et al.* A high-mobility electron-transporting polymer for printed transistors. *Nature* **457,** 679–686 (2009).
18. Dayan, P. *The MIT Encyclopedia of the Cognitive Sciences* (1999).
19. Mueller, T., Kusne, A. G. & Ramprasad, R. Machine Learning in Materials Science: Recent Progress and Emerging Applications *Rev. Comput. Chem.* (2015).
20. Bishop, C. M. *Pattern Recognition and Machine Learning* Springer (2006).
21. Rajan, K. *Informatics for Materials Science and Engineering* Elsevier (2013).
22. Ghahramani, Z. Probabilistic machine learning and artificial intelligence. *Nature* **521,** 452459 (2015).
23. Schuett, K. T. *et al.* How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Phys. Rev. B* **89,** 205118 (2014).
24. Faber, F., Lindmaa, A., Von Lilienfeld, O. A. & Armiento, R. Crystal structure representations for machine learning models of formation energies. *Int. J. Quantum Chem.* **115,** 10941101 (2015).
25. Faber, F., Lindmaa, A., Von Lilienfeld, O. A. & Armiento, R. Machine learning energies of 2 M elpasolite (ABC2D6) crystals. http://arxiv.org/pdf/1508.05315.pdf (2015) (21/08/2015).
26. Jain, A., Castelli, I. E., Hautier, G., Bailey, D. H. & Jacobsen, K. W. Performance of genetic algorithms in search for water splitting perovskites. *J. Mat. Sc.* **48,** 6519–6534 (2013).
27. Dudiy, S. V. & Zunger, A. Searching for alloy configurations with target physical properties: Impurity design via a genetic algorithm inverse band structure approach. *Phys. Rev. Lett.* **97,** 046401 (2006).
28. Goedecker, S. Minima hopping: an efficient search method for the global minimum of the potential energy surface of complex molecular systems. *J. Chem. Phys.* **120,** 9911–7 (2004).
29. Amsler, M. & Goedecker, S. Crystal structure prediction using the minima hopping method. *J. Chem. Phys.* **133,** 224104 (2010).
30. Wang, C. C., Pilania, G. & Ramprasad, R. Dielectric properties of carbon-, silicon-, and germanium-based polymers: A first-principles study. *Phys. Rev. B* **87,** 035103 (2013).
31. Mannodi-Kanakkithodi, A., Wang, C. C. & Ramprasad, R. Compounds based on Group 14 elements: building blocks for advanced insulator dielectrics design. *J. Mat. Sc.* **50,** 801–807 (2015).
32. Pilania, G., Wang, C. C., Jiang, X., Rajasekaran, S. & Ramprasad, R. Accelerating materials property predictions using machine learning. *Sci. Rep.* **3,** 2810 (2013).
33. Miller, R. L. *Crystallographic Data and Melting Points for Various Polymers* John Wiley and Sons Inc. (2003).
34. Rupp, M., Tkatchenko, A., Muller, K. & von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **108,** 058301 (2012).
35. Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C. & Scheffler, M. Big data of materials science: critical role of the descriptor. *Phys. Rev. Lett.* **114,** 105503 (2015).
36. Meredig, B. *et al.* Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B* **89,** 094104 (2014).

37. Botu, V. & Ramprasad, R. Adaptive machine learning framework to accelerate ab initio molecular dynamics. *Int. J. Quantum Chem.* **115,** 10741083 (2015).
38. Todeschini, R. & Consonni, V. *Handbook of Molecular Descriptors*, 2nd edition, Wiley (2009).
39. Mannhold, R., Kubinyi, H. & Folkers, G. *Methods and Principles in Medicinal Chemistry*, 41, Wiley (2003).
40. Vu, K. *et al.* Understanding kernel ridge regression: Common behaviors from simple functions to density functionals. *Int. J. Quantum Chem.* **115,** 1115–1128 (2015).
41. Amsler, M., Botti, S., Marques, M. A. L. & Goedecker, S. Conducting Boron sheets formed by the reconstruction of the alpha-Boron (111) surface. *Phys. Rev. Lett.* **111,** 136101 (2013).
42. Huan, T. D., Sharma, V., Rossetti, G. A. & Ramprasad, R. Pathways towards ferroelectricity in hafnia. *Phys. Rev. B* **90,** 064111 (2014).
43. Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev. B* **136,** B864 (1964).
44. Kresse, G. & Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* **47,** 558 (1993).
45. Klime, J., Bowler, D. R. & Michaelides, A. Chemical accuracy for the van der Waals density functional. *J. Phys. Cond. Matt.* **22,** 022201 (2010).
46. Liu, C.-S., Pilania, G., Wang, C. C. & Ramprasad, R. How critical are the van der Waals interactions in polymer crystals? *J. Phys. Chem. A* **116,** 93479352 (2012).
47. Blochl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50,** 17953 (1994).
48. Baroni, S., de Gironcoli, S. & Corso, A. D. Phonons and related crystal properties from density-functional perturbation theory. *Rev. Mod. Phys.* **73,** 515 (2001).
49. Bernardini, F., Fiorentini, V. & Venderbilt, D. Polarization-based calculation of the dielectric tensor of polar crystals. *Phys. Rev. Lett.* **79,** 3958 (1997).
50. Zhao, X. & Vanderbilt, D. Phonons and lattice dielectric properties of zirconia. *Phys. Rev. B* **65,** 075105 (2002).
51. Heyd, J., Scuseria, G. E. & Ernzerhof, M. Hybrid functionals based on a screened Coulomb potential. *J. Chem. Phys.* **118,** 8207 (2003).
52. Heyd, J., Peralta, J. E., Scuseria, G. E. & Martin, R. L. Energy band gaps and lattice parameters evaluated with the Heyd-Scuseria-Ernzerhof screened hybrid functional. *J. Chem. Phys.* **123,** 174101 (2005).

Acknowledgements

This paper is based upon work supported by a Multidisciplinary University Research Initiative (MURI) grant (N00014-10-1-0944) from the Office of Naval Research. Computational support was provided by the Extreme Science and Engineering Discovery Environment (XSEDE) and the National Energy Research Scientific Computing Center (NERSC). G.P. acknowledges the support of the U.S. Department of Energy through the LANL/LDRD Program through a Director's postdoctoral fellowship. AMK would further like to thank LANL for providing computational resources.

Author Contributions

R.R. designed and supervised the study. The high-throughput DFT computations were performed by A.M.K. and T.D.H. and the learning models were developed by A.M.K. with help from G.P. and T.L. G.P. implemented the direct design scheme with genetic algorithm. All authors discussed the results, wrote and shaped the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Mannodi-Kanakkithodi, A. *et al.* Machine Learning Strategy for Accelerated Design of Polymer Dielectrics. *Sci. Rep.* **6**, 20952; doi: 10.1038/srep20952 (2016).