www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Received: 23 June 2017

Accepted: 22 November 2017

Published online: 14 December 2017

Identifying models of dielectric breakdown strength from high-throughput data via genetic programming

Fenglin Yuan & Tim Mueller

The identification of models capable of rapidly predicting material properties enables rapid screening of large numbers of materials and facilitates the design of new materials. One of the leading challenges for computational researchers is determining the best ways to analyze large material data sets to identify models that can rapidly predict a given property. In this paper, we demonstrate the use of genetic programming to generate simple models of dielectric breakdown based on 82 representative dielectric materials. We identified the band gap E_g and phonon cut-off frequency ω_{max} as the two most relevant features, and new classes of models featuring functions of E_g and ω_{max} were uncovered. The genetic programming approach was found to outperform other approaches for generating models, and we discuss some of the advantages of this approach.

With the ever-increasing power of supercomputers, materials scientists are able to perform high-throughput density functional theory (DFT) calculations[1,2] and build up online databases[3–6] of important materials properties including structure parameters, thermodynamic and transport properties, and electronic structures and properties. Such vast amounts of materials information enables the use of machine learning methods to identify simple predictive models of more complex material properties[7–18]. Such simple models can be used to rapidly screen materials, enabling experimentalists focus on only the most promising candidates and expediting the development of new materials with a reduced cost.

There are two challenges in this approach to materials discovery and design: it is necessary to generate well-organized and high-quality data, and it is necessary to use suitable machine learning algorithms to identify the most relevant models. To address the first challenge, there are several active projects to create broadly accessible databases of high-quality material data[3–6]. The best approach to deal with the second challenge is not clear, as there are a wide variety of machine learning algorithms that could be used and different algorithms are appropriate for different problems[11]. In this paper, we demonstrate the power of a supervised learning algorithm known as "genetic programming", in which an evolutionary algorithm is used to perform symbolic regression, for identifying simple models for dielectric breakdown strength of materials. Mueller *et al.*[19] have previously demonstrated that genetic programming can be used to identify important structural descriptors for hole trap depths in hydrogenated nanocrystalline and amorphous silicon, and here we evaluate the genetic programming approach for the identification of predictive models of dielectric breakdown strength.

Methods

Genetic programming. Genetic programming (GP) is a supervised machine-learning algorithm in which the hypothesis space (the space of functions to be considered) consists of combinations of simple functions and operators (e.g., addition, subtraction, exponentiation, square root, etc.). The goal of the genetic programming algorithm, as in any supervised learning algorithm, is to find a function within this hypothesis space that is able to best predict the output value of interest, which in our case is the calculated dielectric breakdown strength. To search this hypothesis space a genetic or evolutionary algorithm is used to evolve a population of candidate functions (or models) according to natural-selection rules, in which good models are retained, bad models are tossed out, and new models are created by crossover (combining existing models) and mutation (modifying existing models). The fittest models should achieve a balance between complexity, speed and accuracy. This approach has

Department of Materials Science and Engineering, Johns Hopkins University, Baltimore, MD, 21218, USA. Correspondence and requests for materials should be addressed to T.M. (email: tmueller@jhu.edu)

Name	Symbol	Name	Symbol
Band Gap	E_g	Dieletric constant	d_t
Phonon frequency (max)	ω_{max}	Dielectric constant (electron)	d_e
Phonon frequency (mean)	ω_{mean}	Nearest Neighbor Distance	a
Density	ρ	Bulk Modulus	bm

Table 1. Eight feature properties related to dielectric breakdown strength.

been widely applied in multidisciplinary fields, including but not limited to financial market analysis[20], biological science[21,22], software development[23], and identifying interatomic potential models from calculated energies[24–26]. In materials science and engineering, researchers have used genetic programming to develop predictive models for properties of concrete and cement[27–29], asphalt[30], shape memory alloys[31], and heterogeneous catalysts[32]. They have also been used to determine the effects of processing parameters on metal alloys[33–36], predict the impact roughness of cold formed materials[37], optimize productivity for the steel industry[38], and develop models for a variety of problems in structural engineering[39]. Recently genetic programming has been identified as a useful tool for extracting important descriptors of material properties from computational data[19].

Test and training data.

Here we evaluate the effectiveness of genetic programming in predicting dielectric breakdown strength. Dielectric breakdown strength is defined as the maximum external electric field strength that the materials can withstand before turning into a conductor. Materials with high dielectric breakdown strength are used as insulators for applications including high voltage power transmission and capacitors[40–42]. Fundamentally, dielectric breakdown strength is a complex phenomenon involving physical interaction between materials and an electric field. Here, we focus on intrinsic dielectric breakdown strength, which is defined for a defect-free crystal and theoretically is only influenced by materials chemistry. The calculation of intrinsic dielectric breakdown strength is based on von Hippel[43] and Fröhlich[44–46] theory implemented in a DFT framework[47]. Due to the time-consuming nature of such calculations, only 82 representative crystals were calculated. The details of these calculations and the underlying theories can be found in the work by Sun *et al.*[47] and Kim *et al.*[48].

Despite its importance in both academia and industry, dielectric breakdown strength lacks a good predictive model that can be used to rapidly screen new candidate materials. Recently Kim *et al.*[48] provided a case study of searching for relevant models via three supervised machine-learning algorithms: Kernel Ridge Regression (KRR)[49–51], Random Forest Regression (RFR)[49] and Least Absolute Shrinkage and Selection Operator (LASSO)[49,52]. Of these, they determined that the LASSO method was effective for the identification of analytical models, and based on this method they developed several phenomenological models for crystalline dielectric materials. They highlight the following model as being particularly good in terms of simplicity and accuracy:

$$F_b = 24.442e^{0.315\sqrt{E_g\omega_{\mathrm{max}}}}, \tag{1}$$

or, equivalently

$$\ln(F_b) = 3.196 + 0.315\sqrt{E_g\omega_{\mathrm{max}}}, \tag{2}$$

where F_b is the dielectric breakdown strength, E_g is the electronic band gap, and ω_{max} is the maximum phonon frequency.

The genetic programming method used in this paper is in some ways similar to the LASSO method. The LASSO method used by Kim *et al.*[48] assesses linear combinations of elementary terms, where each term is a function of some subset of material properties. The disadvantage to this approach is that a list of possible terms must be provided to the algorithm. (Kim *et al.*[48] combinatorially generated a total of 187,944 terms, each containing functions of up to three properties.). In contrast, genetic programming is capable of dynamically generating such terms by combining properties in non-linear ways, so only the list of known material properties must be provided to the algorithm.

To enable comparison with LASSO and other machine learning algorithms, we have applied the genetic programming approach to the same input dataset generated by Kim *et al.*[48]. This data set is composed of eight feature properties, listed in Table 1, for 82 representative crystalline dielectric materials. The detailed information for 82 crystalline insulators can be found in the supplementary information of Kim *et al.*'s paper[48].

Simulation Details.

We performed genetic programming calculations using the Eureqa software package[53]. In genetic programming, an explicit definition of elementary operators is required to define the hypothesis space, and in this study we chose a collection of four algebraic operators (i.e., plus, minus, division, multiplication) and three function operators (square root, exponential and logarithm functions). For each of these operators we used the default complexity value in Eureqa (see Supplementary Table S1). We used three different but representative metrics to measure the fitness of candidate models: the mean absolute error (MAE), the root mean square error (RMSE) and the Pearson correlation coefficient (PCC) (as defined in the SI). The PCC assesses the degree to which two sets of data are linearly related, regardless of how close they are to each other in value. Thus when the PCC is used as the objective function, the breakdown strengths predicted by the models output by Eureqa need to undergo a linear transformation to enable direct comparisons with the DFT-calculated breakdown strengths (see Supplementary Fig. S1). On the other hand the use of MAE or RMSE as the objective function produces output that is directly comparable to the training data. Although this obviates the need for a linear transformation of the

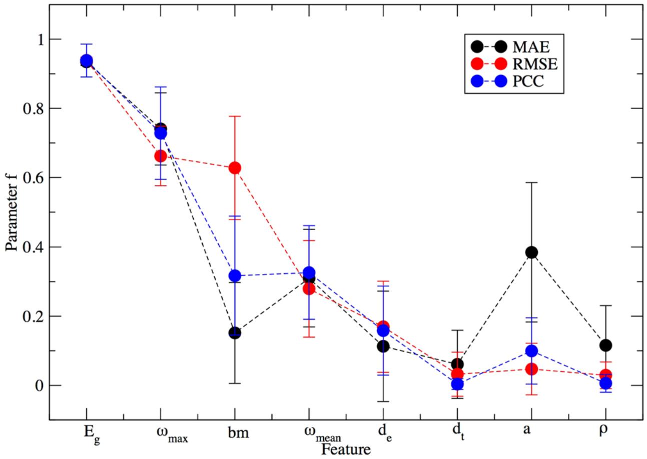

Figure 1. The frequency with which each of the eight features appears on the Pareto frontier (parameter f). Error bars are from eight parallel Eureqa runs.

model output, it imposes an additional requirement on the model output that could make the discovery of simple, accurate models more difficult.

To estimate the predictive performance of the evaluated models, we set the parameters in Eureqa so that in each run the 82 materials in the data set were randomly partitioned into two groups of the same size (i.e., each containing 41 materials): the training and the validation dataset. Model optimization is done using the training set, and the validation set is used to construct a Pareto frontier of models, defined as the set of models for which no other models were both simpler and more accurate. We take the output of a single Eureqa run to be the set of models on the Pareto frontier after total number of evaluated models reached 10^{12}.

Results and Discussion

Evaluation of individual properties and products of properties. To estimate the degree to which each of the eight feature properties is related to the dielectric breakdown strength, we first simply count the number of times each appears in a model found by Eureqa. Because genetic programming is a stochastic method, with randomness in both the evolutionary algorithm and the way in which the validation set is selected, we performed our analysis over eight independent runs for each objective function. We define a parameter f_i by:

$$f_i = N_i/N, \tag{3}$$

where N is the total number of models in the Pareto frontier and N_i is the total number of models in the Pareto frontier that are functions of the the i^{th} material property. Higher f_i values for a particular property suggest that the property is more useful as a descriptor of dielectric breakdown strength.

The calculated values of f_i for the eight material properties are shown in Fig. 1. The properties E_g (the band gap) and ω_{max} (the maximum phonon frequency) have the highest values of f_i. For all three objective functions, more than 60% of the models on the Pareto frontier contain these values. The importance of E_g and ω_{max} in predicting dielectric breakdown strength is in agreement with the results obtained by the machine learning models evaluated by Kim *et al.*[48], as well as a simple correlation analysis (Supplementary Fig. S3). These results indicate that genetic programming is an effective tool for rapidly identifying the most relevant properties, consistent with prior results[19]. Our results also revealed that bm (the bulk modulus) and ω_{mean} (the average phonon frequency) have the next-highest values of f_i, which is not surprising given the degree to which these properties are correlated with ω_{max} (see Supplementary Fig. S3). These three properties (ω_{max}, ω_{mean}, and bm) are associated with the stiffness of the material[54], which is consistent with the proposed physical picture described by Kim *et al.*[48].

After scrutinizing the raw results from Eureqa, it was evident that the multiplication of two features is the most frequent way for features to be combined. Based on this observation, we computed the number of appearances for all possible products of two features (Fig. 2). For all three objective functions the $E_g*\omega_{max}$ term appeared most frequently, providing further evidence of the importance of these two properties.

Performance of the genetic programming models. We first consider a direct comparison between the results of genetic programming and the results obtained by Kim *et al.*[48]. To make this comparison, we have used the logarithm of the dielectric breakdown strength as the output value, to be consistent with their approach[48]. To accelerate the search, we restricted the properties considered to the two that consistently appeared most frequently in models on the Pareto frontier: E_g and ω_{max}, based on Figs 1 and 2. All results in this paper are under this premise unless stated otherwise. Because the two-feature Eureqa runs are faster than the eight-feature runs, for each objective function we gathered statistics for 16 two-feature Eureqa runs. In each of these runs it took approximately 14 hours to evaluate 10^{12} models on a single core of 3 GHz Intel Core i5-2320 CPU.

Averaged over all 16 Eureqa runs as a function of complexity, the performance of the generated models for all three objective functions (MAE, PCC, RMSE) on the total set of input data (i.e., training plus validation sets)

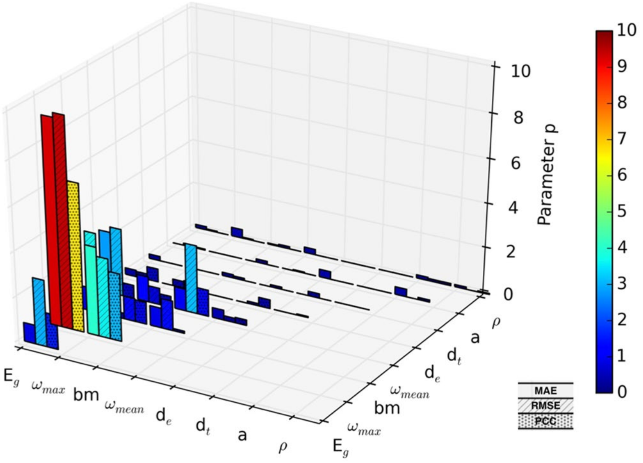

Figure 2. Distribution of the frequency of occurrence (parameter p) of products of two features.

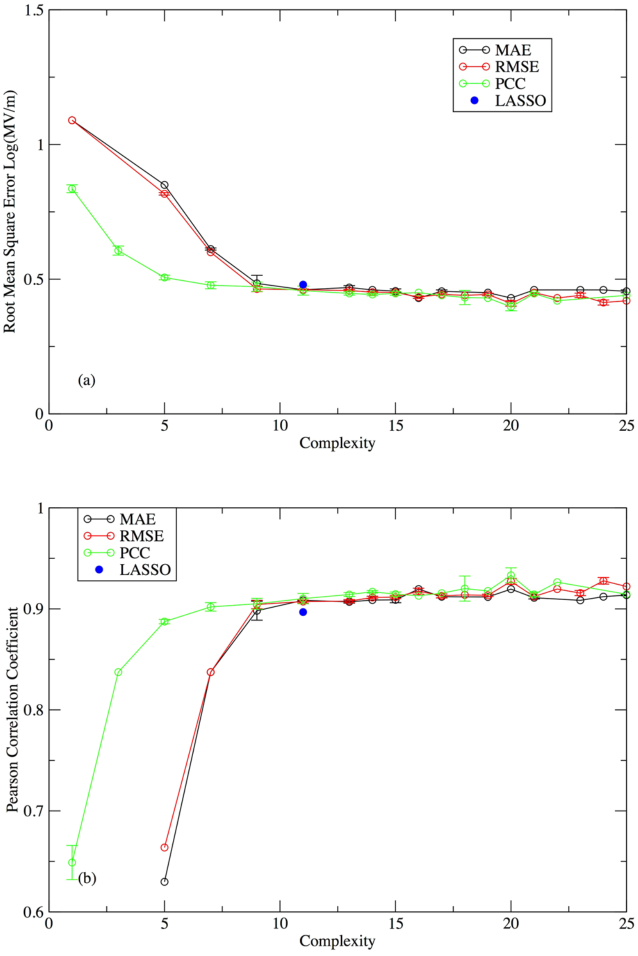

Figure 3. Average (**a**) RMSE and (**b**) PCC performance of models on total input data (training plus validation datasets) by MAE, RMSE, PCC optimizations compared with the LASSO solution. The error bar is the calculated standard deviation from averaging over 16 parallel Eureqa runs.

increases with increasing complexity and levels off when the complexity reaches around 10 (Fig. 3). The LASSO solution (equation (2)) has a complexity of 11, determined via the same complexity measure used for the genetic programming runs. At this level of complexity, the average model found by genetic programming has slightly better performance than the LASSO solution on the training and validation data.

To assess the predictive ability of the generated models, i.e. how well the models on the Pareto frontier perform when exposed to data Eureqa has never seen, we evaluated the models' performance (measured by RMSE and

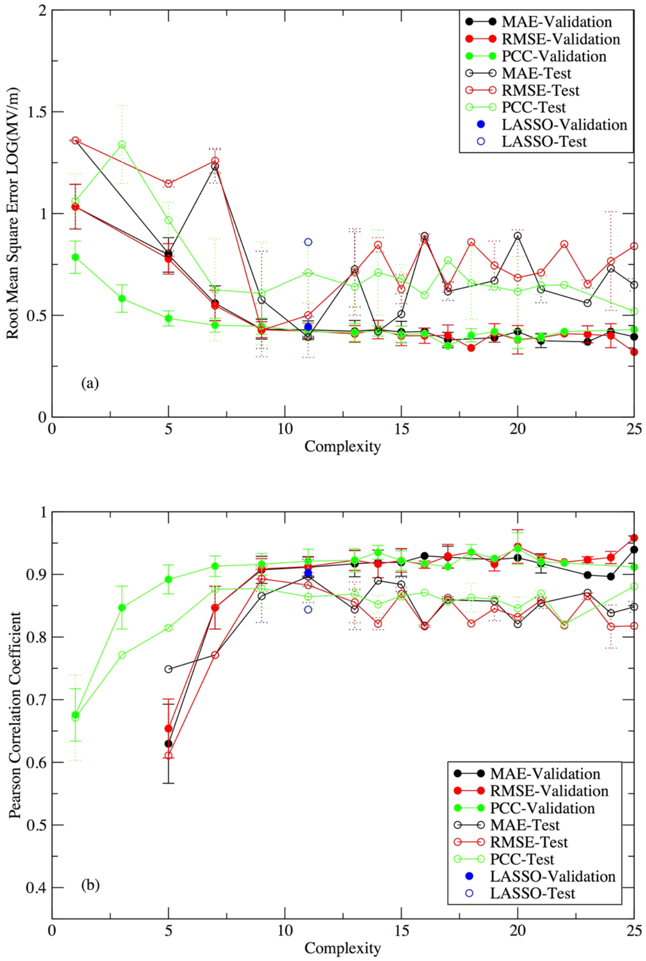

Figure 4. Average (**a**) RMSE and (**b**) PCC performance of models on validation and test sets by MAE, RMSE, PCC optimizations compared with the LASSO solution. The error bar is the calculated standard deviation from averaging over 16 parallel Eureqa runs.

PCC) on an out-of-range test set (Fig. 4). The test set consisted of four cubic crystals[48], Li_2S, Na_2S, $SrCl_2$ and ZrO_2, as well as six perovskite crystals[55], $BaSnO_3$, $CaGeO_3$, $CaSiO_3$, $BSiO_2F$, $BaBO_2F$ and $SrBO_2F$. Detailed information about these ten materials is provided in the works by Kim *et al.*[48,55] and summarized in Supplementary Table S2. The test data was not used by either the genetic programming algorithm or the LASSO algorithm when generating models.

For the models discovered by genetic programming, the average quality of the predictions on the test data improves until about a complexity of about 10, at which point the errors start to increase (Fig. 4). In contrast, the performance of models on the validation data increases with increasing complexity and levels off when the complexity reaches about 10 (Fig. 4). These results indicate that beyond this complexity, the genetic programming algorithm is overfitting the data. Although the LASSO solution has performance similar to the genetic programming models on the validation data, its performance on test data is significantly worse, manifested by higher RMSE and lower PCC values. Its performance on the test data is comparable to the genetic programming models at levels of complexity that overfit the training data, suggesting that the LASSO solution may have also overfit the data. We note that the genetic programming algorithm found the LASSO solution in two of the 48 Eureqa runs, and in both of these runs PCC was the objective function. The PCC runs appear to be better than the RMSE or MAE runs at identifying models with good predictive ability at low levels of complexity, which is understandable given the additional requirements imposed on the generated models when MAE or RMSE is used as an objective function.

Our results suggest that genetic programming is effective at finding models with good predictive ability, provided that the appropriate level of complexity is determined. Models that are too simple are not able to adequately account for the factors that influence dielectric breakdown strength, and models that are too complex overfit the data and have relatively poor predictive ability. The challenge is in determining the appropriate level of complexity to minimize prediction errors. In addition, at a given level of complexity there may be many different models by different genetic programming runs, and it is also necessary to select from these models. One approach to identifying models that are expected to have good predictive accuracy is to simply withhold a set of test data, as we have done here. Approaches in which a set of test data is withheld have the added benefit of allowing for the amount of uncertainty in the predictions for each model to be estimated by evaluating the prediction errors on the withheld data[11]. Similarly, cross-validation could be used to try to identify the optimal level of complexity. Some amount of cross-validation is already included in the genetic programming algorithm, as the data is randomly partitioned into training and validation sets for each Eureqa run. Here we explore two alternative strategies for identifying the best models.

The first strategy we explored is to simply count the number of times a model appears in the different stochastic runs, under the hypothesis that models that appear on the Pareto frontiers more frequently are less likely to have fit the training data well by chance. In each of the 48 Eureqa runs, a different, randomly-selected partition of training and validation data was used. We counted the number of times each model appeared on the 48 Pareto frontiers, considering only the functional relationships between the two feature properties and ignoring differences in the constants (e.g. coefficients), as we found that for the same model the constants identified by Eureqa could differ slightly from run to run. We chose a single set of parameters to plot by using a gradient descent algorithm to identify the locally optimal parameters. (Details are provided in the supporting information.)

A plot of the models with complexity less than 18 on the Pareto frontiers is shown in Fig. 5, and detailed values for each entry in this plot are listed in Supplementary Table S3. We visualize each model's performance on the test data in Fig. 5. One model, $\ln(F_b) = 4.33 + 0.0174E_g\omega_{\max}$, appears on all 48 Pareto frontiers, but has relatively poor predictive performance on the test set. However, the second-most common model on the Pareto frontiers, $\ln(F_b) = \ln(5.45E_g\omega_{\max} + 2.88)$, is one of the best-performing models. It has a lower root-mean-square error on the 82 training and validation materials than the model discovered by LASSO, and it has roughly half the root-mean-square error on the test set.

The relatively weak performance of the simpler model, $\ln(F_b) = 4.33 + 0.0174E_g\omega_{\max}$ (with complexity 7), suggests that it shows up frequently because there are relatively few models to select from at that level of complexity. On the other hand, the better model, $\ln(F_b) = \ln(5.45E_g\omega_{\max} + 2.88)$, shows up nearly as frequently and at a complexity level at which the hypothesis space of possible models is significantly larger. This suggests that one way to search for the models with the best predictive power would be to find the models that show up unusually frequently given their complexity. However this method does not resolve the issue of how to select a single model that is likely to have good predictive ability. In addition, if we repeat this exact same exercise using F_b, rather than $\ln(F_b)$ as the output variable (see Supplementary Fig. S4 and Table S4), we find that the model that performs best on the test set shows up on Pareto frontiers less frequently than some more complex models. It would have been difficult to identify this model as being particularly promising using this first strategy.

The second strategy we used is to create a "universal" Pareto frontier by combining the best models from 48 Pareto frontiers of all Eureqa runs, as evaluated against the validation data (Fig. 6a). The data used to generate this plot is provided in Table S5. When the models in the universal Pareto frontier are benchmarked against test data, we find that at some complexity values (e.g., 9 and 11), the models on the universal Pareto frontier simultaneously have the lowest (or near-lowest) RMSE for both training and test data. However at most complexity values, the best model on the training data is not the best model for the test data. The RMSE on the test data for models on the universal Pareto frontier is similar to the average RMSE over all sixteen Pareto frontiers generated using RMSE as the fitness metric, suggesting that simply appearing on the universal Pareto frontier is not an indicator of low prediction error.

There does appear to be an advantage to using the universal Pareto frontier. There is a large change in slope at a complexity value of 9, which is roughly the optimal complexity value. Beyond this point, the models get only slightly better even as they get significantly more complex. The entries on the universal Pareto frontier at a complexity of 9 and 11 are also two of the best-performing models on the test data (Fig. 6b). A parity plot of the performance of the model with complexity 9 (i.e., $\ln(F_b) = 1.72 + \ln(E_g) + \ln(\omega_{\max})$ or $F_b = 5.58E_g\omega_{\max}$) against the LASSO solution for test, training, and validation data is provided in Fig. 7. A similar result was found when F_b, rather than $\ln(F_b)$, was used as the objective function (see Supplementary Fig. S5 and Table S6). There is a large change in slope on the universal Pareto frontier at a complexity of 10, and the model at this point on the frontier performs very well on the test data (see Supplementary Fig. S5).

It may be fortuitous that the models at the point where the slope changes on the Pareto frontier also happen to perform very well on the test data, as models at other complexities do not perform as well (Fig. 6b). In addition, it may not always be clear what constitutes a "large change in slope." However this change in slope may be an indication of an optimal (or near-optimal) level of complexity. The relatively rapid decrease in the error with increasing complexity up to this point may be an indication that the error is decreasing primarily because of improving model skill. Similarly, the relatively slow decrease in error at high levels of complexity may be an indication that the error is decreasingly primarily by chance; i.e. the increasing size of the hypothesis space makes it easier to find a model that happens to do well on the training and validation data but generalizes poorly.

Predicting dielectric breakdown strength. Using the analysis in the previous section, we can now identify models that are likely to be useful as predictive models of dielectric breakdown strength. Here we report results on predicting the dielectric breakdown strength itself, rather than the natural logarithm of the breakdown strength. This choice effectively places greater importance on making accurate predictions for materials that have a high dielectric breakdown strength, and these are the materials that are often of the most technological interest. We have included results for the natural logarithm of the dielectric breakdown strength in the supporting information (Supplementary Fig. S8 and Supplementary Table S7).

To identify the models that are likely to make accurate predictions, we have used the universal Pareto frontier approach, this time including all data (i.e., training + validation + test data). For comparison, we have also included the LASSO solution, although we note that the LASSO solution was selected without considering the test set and was fit to the natural logarithm of the dielectric breakdown strength, so a direct comparison is not as straightforward as the comparisons in previous section. All models on this Pareto frontier, as well as their performance, are summarized in Table 2. The universal Pareto frontier (Fig. 8) gives a large slope change around

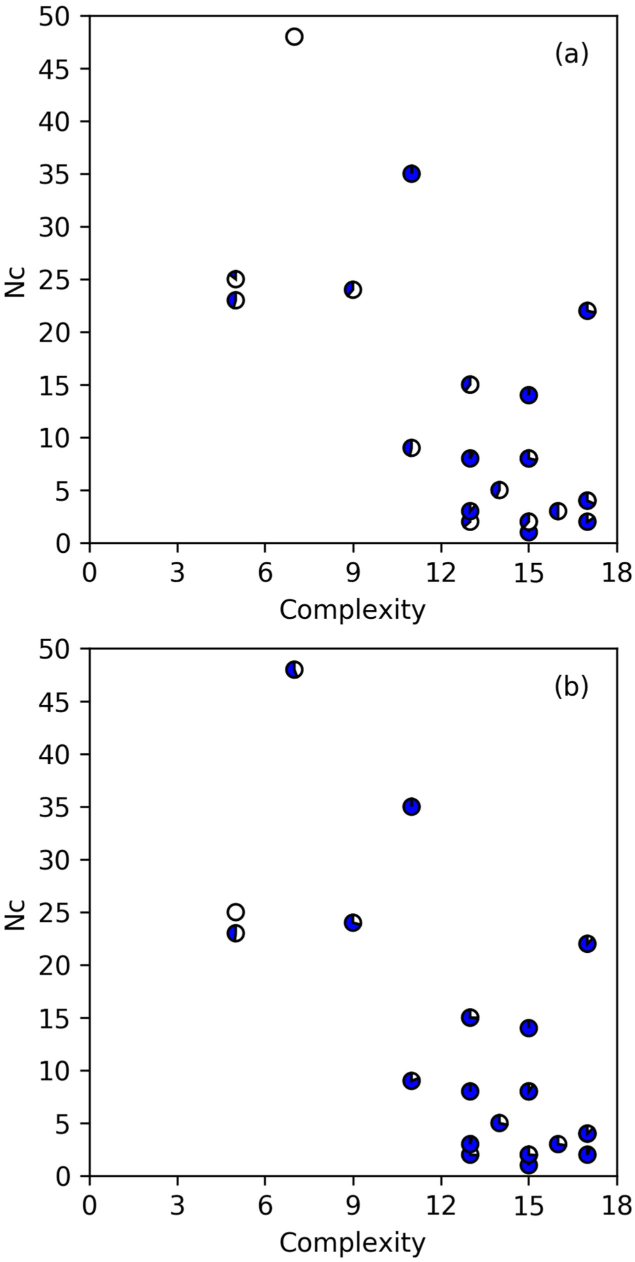

Figure 5. Complexity versus number of appearances (N_c) plots after parameter re-optimization based on training and validation data. Higher coverage indicates a better model as evaluated against the test data. Zero coverage represents (**a**) the highest RMSE, 1.27 ln(MV/m) and (**b**) the lowest PCC, 0.61. Full coverage represents (**a**) the lowest RMSE, 0.39 ln(MV/m) and (**b**) the highest PCC, 0.90.

complexity 8. Based on the analysis in the previous section, this suggests that models at about this level of complexity may have the greatest predictive power.

We have selected from the universal Pareto frontier three models with complexity around 8, labeled S1, S2, and S3 in Table 2, for further analysis. We have also included the LASSO solution for comparison, and label it S4. Parity plots of the values predicted by the models vs. the values predicted by DFT for these four models are provided in Supplementary Figs S9–S11. We emphasize that because the models S1, S2, and S3 were optimized for F_b and the LASSO solution was optimized for $\ln(F_b)$, these plots should not be used to compare the performance of genetic programming vs. LASSO (that comparison was discussed in the previous section). To better understand how each of these four model predicts how F_b will change as a function of E_g and $\omega_{\max}$, we have created plots showing the predicted vs. DFT-calculated values as a function of E_g and $\omega_{\max}$ (Fig. 9). These plots make it clear that although there is a dependence on E_g and $\omega_{\max}$, these variables are not sufficient for a complete description, as there are several pairs of materials in which both materials have similar values for E_g and $\omega_{\max}$ but very different breakdown strengths.

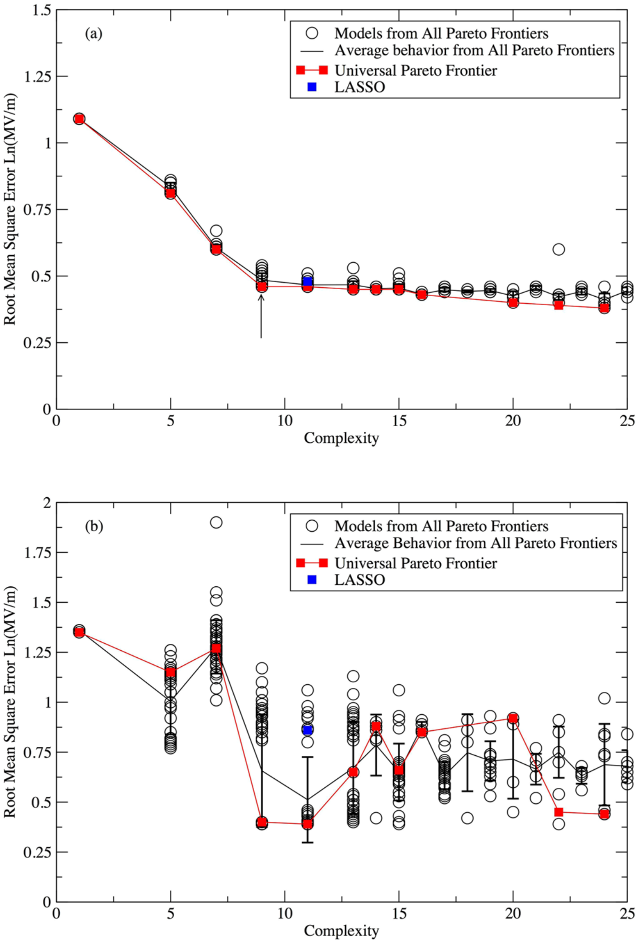

Figure 6. Comparison of RMSE performance of models on (**a**) training + validation and (**b**) test data between LASSO, the universal Pareto frontier and models from all Pareto frontiers. The arrow in (**a**) indicates the point at which there is a relatively large change in the slope along the Pareto frontier.

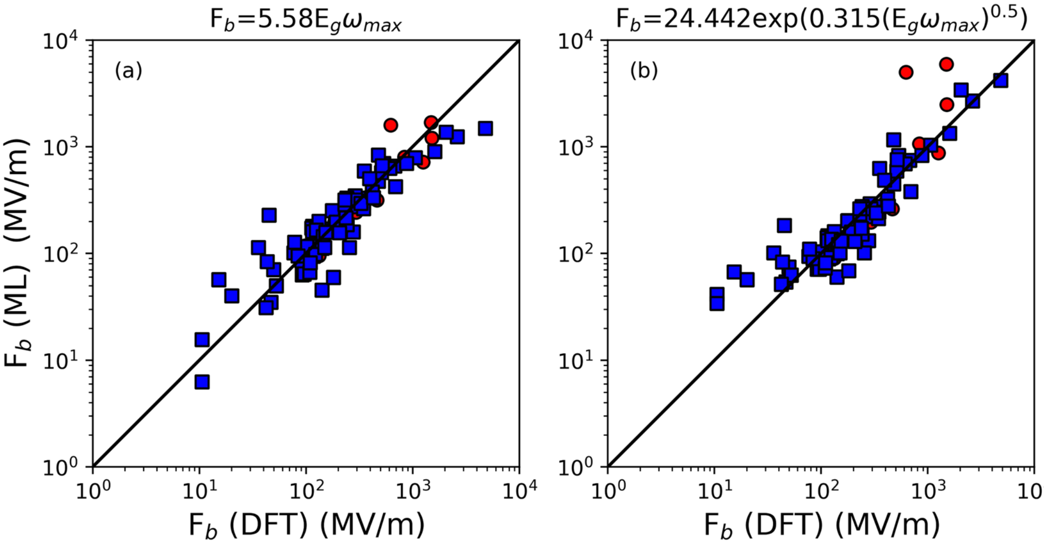

Figure 7. Dielectric breakdown strength F_b predicted by machine learning and density functional theory (DFT) for (**a**) model with complexity 9 at Universal Pareto Frontier, (**b**) the LASSO solution. Blue squares represent the prediction for training and validation data and red circles represent the prediction for test data. The black solid line indicates a perfect match between machine learning and DFT.

Complexity	Model	Benchmark	ε_{al}	ε_{tr}	ε_{va}
1	395	RMSE	639	627 ± 148	614 ± 157
		PCC	N/A	N/A	N/A
3	$34.7\,\omega_{\max}$	RMSE	524	492 ± 183	487 ± 188
		PCC	0.58	0.63 ± 0.17	0.63 ± 0.17
5 (S1)	$E_g^2\omega_{\max}$	RMSE	325	321 ± 48	320 ± 48
		PCC	0.87	0.86 ± 0.05	0.85 ± 0.07
7	$E_g^2(\omega_{\max} - 1.17)$	RMSE	321	319 ± 51	315 ± 53
		PCC	0.87	0.86 ± 0.05	0.86 ± 0.06
8 (S2)	$\frac{262\omega_{\max}}{14.6 - E_g}$	RMSE	248	248 ± 51	238 ± 51
		PCC	0.92	0.91 ± 0.05	0.91 ± 0.05
10 (S3)	$\frac{348\omega_{\max}}{15 - E_g} - 101$	RMSE	235	235 ± 45	226 ± 43
		PCC	0.93	0.91 ± 0.05	0.92 ± 0.05
13	$0.00399e^{E_g} + 5.84E_g\omega_{\max}$	RMSE	233	231 ± 50	223 ± 49
		PCC	0.93	0.92 ± 0.05	0.92 ± 0.05
14	$\frac{184}{2.11\omega_{\max} - 41.3} + 5.92E_g\omega_{\max}$	RMSE	229	227 ± 50	221 ± 48
		PCC	0.93	0.92 ± 0.05	0.92 ± 0.05
15	$4.81 * 10^{-9}e^{E_g} + 6.11E_g\omega_{\max}$	RMSE	227	227 ± 49	217 ± 47
		PCC	0.94	0.92 ± 0.05	0.92 ± 0.05
17	$0.00649e^{E_g} + 20.5\omega_{\max}\ln(E_g)$	RMSE	225	226 ± 44	215 ± 44
		PCC	0.94	0.92 ± 0.04	0.93 ± 0.05
19	$0.0046e^{E_g} + 16.6\omega_{\max}\sqrt{E_g} - 80$	RMSE	224	226 ± 43	214 ± 42
		PCC	0.94	0.92 ± 0.04	0.93 ± 0.05
20	$0.00416e^{E_g} + 5.67E_g\omega_{\max}$ $+21.8/(5.54 - E_g)$	RMSE	217	219 ± 37	210 ± 35
		PCC	0.94	0.93 ± 0.04	0.93 ± 0.05
21	$5.92E_g\omega_{\max} + 138/(40.6 - \omega_{\max})$ $+184/(2.11\omega_{\max} - 41.3)$	RMSE	189	185 ± 38	165 ± 42
		PCC	0.96	0.95 ± 0.04	0.96 ± 0.03
22	$0.0003E_ge^{E_g} + 5.23E_g\omega_{\max}$ $+214/(40.7 - \omega_{\max})$	RMSE	174	184 ± 32	156 ± 39
		PCC	0.96	0.95 ± 0.04	0.96 ± 0.03
25	$15\omega_{\max} + 0.0047E_g^3\omega_{\max}^2$ $-1.12 * 10^{-13}e^{\omega_{\max} - E_g}$	RMSE	169	165 ± 38	166 ± 36
		PCC	0.96	0.96 ± 0.03	0.95 ± 0.03
15 (S4) (LASSO)	$24.442e^{0.315\sqrt{E_g\omega_{\max}}}$	RMSE	692	674 ± 242	614 ± 259
		PCC	0.74	0.78 ± 0.08	0.79 ± 0.10

Table 2. The performance in predicting dielectric breakdown strength on all data ε_{al}, training data ε_{tr}, and validation data ε_{va} for models on the universal Pareto frontier constructed using training, validation, and test data.

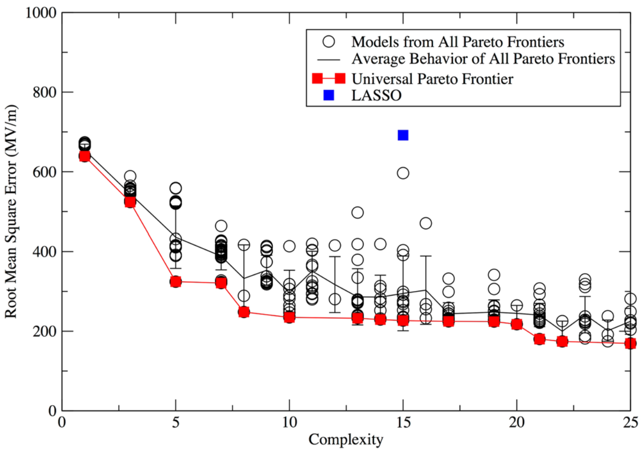

Figure 8. RMSE performance of models on all data (i.e., training + validation + test) for LASSO, the universal Pareto frontier, and models from all Pareto frontiers trained on all data when using dielectric breakdown strength as the output value.

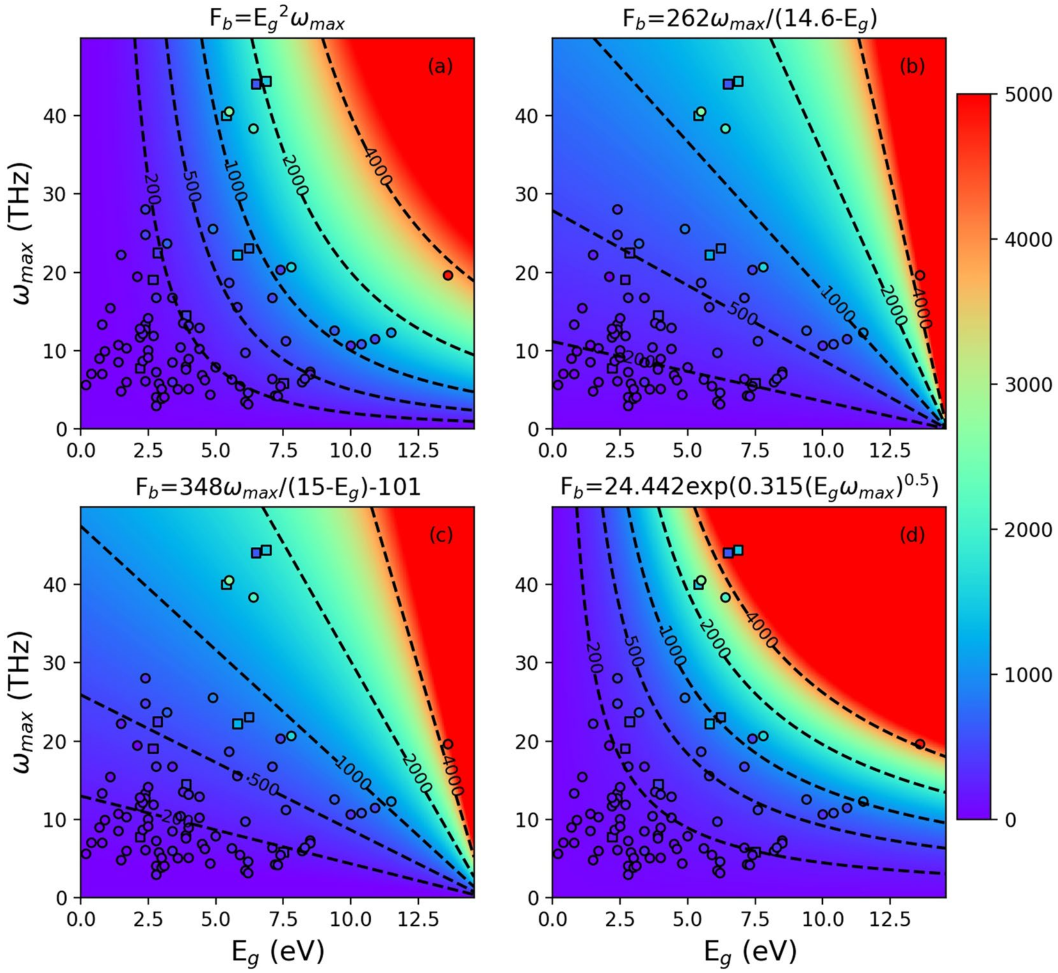

Figure 9. Contour plot of dielectric breakdown strength F_b predicted by machine learning along with scatter plot of the values calculated by density functional theory (DFT) for (**a**) solution S1, (**b**) solution S2, (**c**) solution S3 and (**d**) solution S4 (the LASSO solution). The circles are training and validation data, and the squares are test data. The spheres and squares share the same color-coding scheme as the contour plot. The black dashed lines indicate contour levels labeling machine-learning-predicted F_b values at 200, 500, 1000, 2000, 4000 MV/m.

Of particular interest are models S2 and S3, as these are arguably the best models on the entire set of data in terms of combined simplicity and accuracy. Models S2 and S3 include terms with $14.6 - E_g$ and $15 - E_g$ in their denominators, respectively. These terms will have singularities for materials with band gaps of 14.6 or 15 eV and will become negative for materials with larger band gaps. Fortunately, apart from condensed noble gasses, there is no known material with a band gap larger than that of LiF[56,57], which has a calculated band gap of 13.6 eV. Thus the upper bound on the band gaps in these models corresponds well to the known upper limit in nature for band gaps of materials of the type considered here. Models with similar form showed up on the universal Pareto frontier generated using just training and validation data (Supplementary Table S6), and their persistence when test data was included suggests they have true predictive power. If we generate a universal Pareto frontier using only models discovered using the training and validation data, but with error (the y-axis) evaluated against all data, the models with forms similar to S2 and S3 have the best performance (Supplementary Fig. S12 and Supplementary Table S8).

Models S2 and S3, and similar models, are also of interest because they exist outside of the hypothesis space that was searched by the LASSO algorithm[48]. This highlights a problem with methods that attempt to enumerate all possible solutions in the hypothesis space: the combinatorial space of even relatively simple functions is very large and difficult to comprehensively enumerate. The advantage of an approach such as genetic programming is that it can effectively search this hypothesis space without the need to explore the entire space; it naturally focuses on the regions of the space with the most promising (i.e. "fit") solutions. As discussed in this paper, care must be taken to avoid overfitting the training data, but that will be a problem with any algorithm that searches a similarly-sized hypothesis space, including LASSO.

Conclusion

In this paper, we demonstrated that genetic programming is an effective way to search a large hypothesis space of simple functions of known material properties. For the specific property of the dielectric breakdown strength of materials, we identified a new family of models based on $\omega_{\max}(\sim 15\text{eV} - E_g)^{-1}$ that performed well on the training and validation data and then again on the test data. Our results indicate that there is a substantial risk to overfitting the training and validation data, both with genetic programming and with the LASSO approach. We explored different techniques to mitigate this risk and facilitate the use of genetic programming to discover models with good predictive power. The more effective of these appears to be finding the model(s) at or near the point at which the Pareto frontier starts to level off. It can be helpful to consider the number of times a model shows up in repeated genetic programming runs, but this approach appears to be less reliable in identifying models with good predictive power. We believe further exploration of these and related approaches will make genetic programming a more practically useful tool for researchers. There are a number of additional potential areas for improvement, including how to best define "complexity" and how to best partition the known data. Despite the room for further improvement, the relative success of the genetic programming approach in identifying simple models of dielectric breakdown strength provides additional evidence that it is a valuable tool for descriptor identification in materials science and engineering.

References

1. Hohenberg, P. & Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **136**, B864–&, https://doi.org/10.1103/PhysRev.136.B864 (1964).
2. Kohn, W. & Sham, L. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **140**, 1133 (1965).
3. Curtarolo, S. *et al*. AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations. *Comput. Mater. Sci.* **58**, 227–235, https://doi.org/10.1016/j.commatsci.2012.02.002 (2012).
4. Landis, D. D. *et al*. The Computational Materials Repository. *Comput. Sci. Eng.* **14**, 51–57, https://doi.org/10.1109/MCSE.2012.16 (2012).
5. Jain, A. *et al*. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002, https://doi.org/10.1063/1.4812323 (2013).
6. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM* **65**, 1501–1509, https://doi.org/10.1007/s11837-013-0755-4 (2013).
7. 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, https://doi.org/10.1103/PhysRevLett.114.105503 (2015).
8. Huan, T. D., Mannodi-Kanakkithodi, A. & Ramprasad, R. Accelerated materials property predictions and design using motif-based fingerprints. *Phys. Rev. B* **92**, 014106, https://doi.org/10.1103/PhysRevB.92.014106 (2015).
9. Wicker, J. G. P. & Cooper, R. I. Will it crystallise? Predicting crystallinity of molecular materials. *Cryst. Eng. Comm.* **17**, 1927–1934, https://doi.org/10.1039/C4CE01912A (2015).
10. Xu, H., Liu, R., Choudhary, A. & Chen, W. A Machine Learning-Based Design Representation Method for Designing HeterogeneousMicrostructures. *J. Mech. Des* **137**, 051403-051403–051410, https://doi.org/10.1115/1.4029768 (2015).
11. Mueller, T., Kusne, A. G. & Ramprasad, R. In *Reviews in Computational Chemistry* (eds Abby L. Parrill & Kenny B. Lipkowitz) 186–273 (John Wiley & Sons, Inc, 2016).
12. Jong, M. D. *et al*. A Statistical Learning Framework forMaterials Science: Application to Elastic Moduli of k-nary Inorganic Polycrystalline Compounds. *Sci. Rep.* **6**, srep34256, https://doi.org/10.1038/srep34256 (2016).
13. Pilania, G., Balachandran, P. V., Kim, C. & Lookman, T. Finding New Perovskite Halides via Machine Learning. *Comput. Mater. Sci.* 19, https://doi.org/10.3389/fmats.2016.00019 (2016).
14. Ghiringhelli, L. M. *et al*. Learning physical descriptors for materials science by compressed sensing. *New J. Phys.* **19**, 023017, https://doi.org/10.1088/1367-2630/aa57bf (2017).
15. Goldsmith, B. R., Boley, M., Vreeken, J., Scheffler, M. & Ghiringhelli, L. M. Uncovering structure-property relationships of materials by subgroup discovery. *New J. Phys.* **19**, 013031, https://doi.org/10.1088/1367-2630/aa57c2 (2017).
16. Seko, A., Hayashi, H., Nakayama, K., Takahashi, A. & Tanaka, I. Representation of compounds for machine-learning prediction of physical properties. *Phys. Rev. B* **95**, 144110, https://doi.org/10.1103/PhysRevB.95.144110 (2017).
17. Takahashi, K. & Tanaka, Y. Unveiling descriptors for predicting the bulk modulus of amorphous carbon. *Phys. Rev. B* **95**, 054110, https://doi.org/10.1103/PhysRevB.95.054110 (2017).
18. Xue, D. *et al*. An informatics approach to transformation temperatures of NiTi-based shape memory alloys. *Acta Mater.* **125**, 532–541, https://doi.org/10.1016/j.actamat.2016.12.009 (2017).
19. Mueller, T., Johlin, E. & Grossman, J. C. Origins of hole traps in hydrogenated nanocrystalline and amorphous silicon revealed through machine learning. *Phys. Rev. B* **89**, 115202, https://doi.org/10.1103/PhysRevB.89.115202 (2014).
20. Neely, C., Weller, P. & Dittmar, R. Is Technical Analysis in the Foreign Exchange Market Profitable? A Genetic Programming Approach. *J. Financ. Quant. Anal.* **32**, 405–426, https://doi.org/10.2307/2331231 (1997).
21. Muttil, N. & Chau, K.-W. Neural network and genetic programming for modelling coastal algal blooms. *Int. J. Environ. Pollut.* **28**, 223–238, https://doi.org/10.1504/IJEP.2006.011208 (2006).
22. Schmidt, M. D. *et al*. Automated refinement and inference of analytical models for metabolic networks. *Phys. Biol.* **8**, 055011, https://doi.org/10.1088/1478-3975/8/5/055011 (2011).
23. Burgess, C. J. & Lefley, M. Can genetic programming improve software effort estimation? A comparative evaluation. *Inf. Soft. Technol.* **43**, 863–873, https://doi.org/10.1016/S0950-5849(01)00192-6 (2001).
24. Slepoy, A., Peters, M. D. & Thompson, A. P. Searching for globally optimal functional forms for interatomic potentials using genetic programming with parallel tempering. *J. Comput. Chem.* **28**, 2465–2471, https://doi.org/10.1002/jcc.20710 (2007).
25. Brown, W. M., Thompson, A. P. & Schultz, P. A. Efficient hybrid evolutionary optimization of interatomic potential models. *J. Chem. Phys.* **132**, 024108, https://doi.org/10.1063/1.3294562 (2010).
26. Kenoufi, A. & Kholmurodov, K. Symbolic Regression of Interatomic Potentials via Genetic Programming. *Biol. Chem. Res.* **2**, 1–10 (2015).
27. Baykasoğlu, A., Dereli, T. & Tanış, S. Prediction of cement strength using soft computing techniques. *Cem. Concr. Res.* **34**, 2083–2090, https://doi.org/10.1016/j.cemconres.2004.03.028 (2004).
28. Ozbay, E., Gesoglu, M. & Güneyisi, E. Empirical modeling of fresh and hardened properties of self-compacting concretes by genetic programming. *Constr. Build. Mater.* **22**, 1831–1840, https://doi.org/10.1016/j.conbuildmat.2007.04.021 (2008).
29. Gandomi, A. H., Alavi, A. H. & Sahab, M. G. New formulation for compressive strength of CFRP confined concrete cylinders using linear genetic programming. *Mater. Struct.* **43**, 963–983, https://doi.org/10.1617/s11527-009-9559-y (2009).
30. Alavi, A. H., Ameri, M., Gandomi, A. H. & Mirzahosseini, M. R. Formulation of flow number of asphalt mixes using a hybrid computational method. *Constr. Build. Mater.* **25**, 1338–1355, https://doi.org/10.1016/j.conbuildmat.2010.09.010 (2011).
31. Eskil, M. & Kanca, E. A new formulation for martensite start temperature of Fe–Mn–Si shape memory alloys using genetic programming. *Comput. Mater. Sci.* **43**, 774–784, https://doi.org/10.1016/j.commatsci.2008.01.042 (2008).
32. Baumes, L. A. *et al*. Using Genetic Programming for an Advanced Performance Assessment of Industrially Relevant Heterogeneous Catalysts. *Mater. Manuf. Process.* **24**, 282–292, https://doi.org/10.1080/10426910802679196 (2009).

33. Brezocnik, M. & Kovacic, M. Integrated Genetic Programming and Genetic Algorithm Approach to Predict Surface Roughness. *Mater. Manuf. Process.* **18**, 475–491, https://doi.org/10.1081/AMP-120022023 (2003).
34. Brezocnik, M., Kovacic, M. & Ficko, M. Prediction of surface roughness with genetic programming. *J. Mater. Process. Technol.* **157–158**, 28–36, https://doi.org/10.1016/j.jmatprotec.2004.09.004 (2004).
35. Kovacic, M., Uratnik, P., Brezocnik, M. & Turk, R. Prediction of the Bending Capability of Rolled Metal Sheet by Genetic Programming. *Mater. Manuf. Process.* **22**, 634–640, https://doi.org/10.1080/10426910701323326 (2007).
36. Dimitriu, R. C., Bhadeshia, H. K. D. H., Fillon, C. & Poloni, C. Strength of Ferritic Steels: Neural Networks and Genetic Programming. *Mater. Manuf. Process.* **24**, 10–15, https://doi.org/10.1080/10426910802539796 (2008).
37. Gusel, L. & Brezocnik, M. Modeling of impact toughness of cold formed material by genetic programming. *Comput. Mater. Sci.* **37**, 476–482, https://doi.org/10.1016/j.commatsci.2005.11.007 (2006).
38. Kovačič, M. & Šarler, B. Application of the Genetic Programming for Increasing the Soft Annealing Productivity in Steel Industry. *Mater. Manuf. Process.* **24**, 369–374, https://doi.org/10.1080/10426910802679634 (2009).
39. Gandomi, A. H. & Alavi, A. H. A new multi-gene genetic programming approach to nonlinear system modeling. Part I: materials and structural engineering problems. *Neural Comput. Appl.* **21**, 171–187, https://doi.org/10.1007/s00521-011-0734-z (2012).
40. Chen, G. *et al*. Review of high voltage direct current cables. *CSEE J. Power Energy Syst.* **1**, 9–21, https://doi.org/10.17775/CSEEJPES.2015.00015 (2015).
41. Sarjeant, W. J., Zirnheld, J. & MacDougall, F. W. Capacitors. *IEEE Trans. Plasma Sci.* **26**, 1368–1392, https://doi.org/10.1109/27.736020 (1998).
42. Fillery, S. P. *et al.* Nanolaminates: Increasing Dielectric Breakdown Strength ofComposites. *ACS Appl. Mater. Interfaces* **4**, 1388–1396, https://doi.org/10.1021/am201650g (2012).
43. Von Hippel, A. Electric Breakdown of Solid and Liquid Insulators. *J. Appl. Phys.* **8**, 815–832, https://doi.org/10.1063/1.1710258 (1937).
44. Frohlich, H. Theory of Electrical Breakdown in IonicCrystals. *Proc. R. Soc. Lond. A: Math. Phys. Eng. Sci.* **160**, 230–241, https://doi.org/10.1098/rspa.1937.0106 (1937).
45. Fröhlich, H. Theory of dielectric breakdown. *Nature* **151**, 339–340, https://doi.org/10.1038/151339a0 (1943).
46. Fröhlich, H. On the theory of dielectric breakdown in solids. *Proc. R. Soc. Lond. A* **188**, 521–532, https://doi.org/10.1098/rspa.1947.0023 (1947).
47. Sun, Y., Boggs, S. A. & Ramprasad, R. The intrinsic electrical breakdown strength of insulators from first principles. *Appl. Phys. Lett.* **101**, 132906, https://doi.org/10.1063/1.4755841 (2012).
48. Kim, C., Pilania, G. & Ramprasad, R. From Organized High-Throughput Data to Phenomenological Theory using Machine Learning: The Example of Dielectric Breakdown. *Chem. Mater.*. https://doi.org/10.1021/acs.chemmater.5b04109 (2016).
49. Hastie, T., Tibshirani, R. & Friedman, J. *The elements of statistical learning: data mining, inference, and prediction, Second Edition*. 2nd ed. 2009. Corr. 7th printing 2013 edition edn, (Springer, 2011).
50. Muller, K. R., Mika, S., Ratsch, G., Tsuda, K. & Scholkopf, B. An introduction to kernel-based learning algorithms. *IEEE Trans. Neural Netw.* **12**, 181–201, https://doi.org/10.1109/72.914517 (2001).
51. Hofmann, T., Schölkopf, B. & Smola, A. J. Kernel Methods in Machine Learning. *Ann. Stat.* **36**, 1171–1220 (2008).
52. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B* **58**, 267–288 (1994).
53. Schmidt, M. & Lipson, H. Distilling free-form natural laws from experimental data. *Science* **324**, 81–85, https://doi.org/10.1126/science.1165893 (2009).
54. Siethoff, H. & Ahlborn, K. The dependence of the debye temperature on the elastic constants. *Phys. Status Solidi B* **190**, 179–191, https://doi.org/10.1002/pssb.2221900126 (1995).
55. Kim, C., Pilania, G. & Ramprasad, R. Machine learning assisted predictions of intrinsic dielectric breakdown strength of ABX_3 perovskites. *J. Phys. Chem. C* **120**, 14575–14580, https://doi.org/10.1021/acs.jpcc.6b05068 (2016).
56. Roessler, D. M. & Walker, W. C. Electronic spectrum of crystalline lithium fluoride. *J. Phys. Chem. Solids* **28**, 1507–1515, https://doi.org/10.1016/0022-3697(67)90280-6 (1967).
57. Weber, M. J. *Handbook of optical materials*. 1 edition edn, (CRC Press, 2002).

Acknowledgements

We thank Ramamurthy Ramprasad for helpful discussions and providing training, validation, and test data. This work was supported by the Office of Naval Research under grant N000141512665. Computing resources were provided by the Maryland Advanced Research Computing Center (MARCC).

Author Contributions

T.M. designed the research and F.Y. carried out the modeling and calculations. Both T.M. and F.Y. performed data analysis and wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-017-17535-3.

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.