CHEMISTRY OF MATERIALS

Article

pubs.acs.org/cm

From Organized High-Throughput Data to Phenomenological Theory using Machine Learning: The Example of Dielectric Breakdown

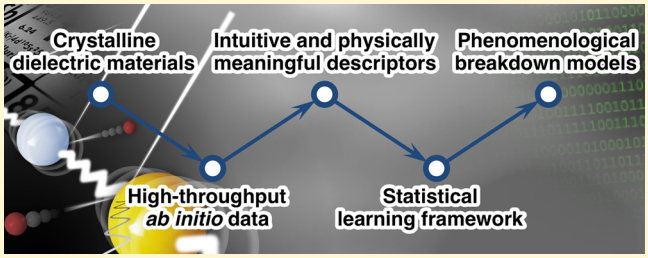
Chiho Kim,[†] Ghanshyam Pilania,[‡] and Ramamurthy Ramprasad*[,†]

[†]Department of Materials Science & Engineering, and Institute of Materials Science, University of Connecticut, 97 North Eagleville Road, Storrs, Connecticut 06269-3136, United States

[‡]Materials Science and Technology Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, United States

Supporting Information

ABSTRACT: Understanding the behavior (and failure) of dielectric insulators experiencing extreme electric fields is critical to the operation of present and emerging electrical and electronic devices. Despite its importance, the development of a predictive theory of dielectric breakdown has remained a challenge, owing to the complex multiscale nature of this process. Here, we focus on the intrinsic dielectric breakdown field of insulators—the theoretical limit of breakdown determined purely by the chemistry of the material, i.e., the elements the material is composed of, the atomic-level structure, and the bonding. Starting from a benchmark data set (generated from laborious first-principles computations) of the intrinsic dielectric breakdown field of a variety of model insulators, simple predictive phenomenological models of dielectric breakdown are distilled using advanced statistical or machine learning schemes, revealing key correlations and analytical relationships between the breakdown field and easily accessible material properties. The models are shown to be general, and can hence guide the screening and systematic identification of high electric field tolerant materials.



INTRODUCTION

Scientific inquiry begins with empirical or observational data, which leads to an initial hypothesis. When consequences of the hypothesis fail to agree with available data, the hypothesis is revised. Successive iterations between the data and hypothesis spaces lead to progressive refinement of the working hypothesis, ultimately culminating in fundamental and precise theories, such as quantum mechanics, gravitation, and electrodynamics.[1−3] Alternatively, data can also lead to phenomenological theories via correlations revealed by statistical analysis. Examples of such developments within the materials sciences include the Hume−Rothery rules of solid solubility[4] and the Hall−Petch relation for materials strengthening.[5,6] Although not as precise as fundamental theories, phenomenological models can be immediately used to screen and design materials with practical value.[7,8] Such models may also be viewed as approximations to precise theories, in that several not-so-relevant, unimportant, and unnecessarily complex features of a fundamental theory are "integrated" out. The data that power the discovery of phenomenological theories may be obtained either from empirical sources, or from fundamental theories (via modern high-throughput computational methods). These notions are schematically captured in Figure 1.

The present contribution provides a systematic and inductive approach by which a phenomenological theory of dielectric breakdown is developed. The behavior of a material

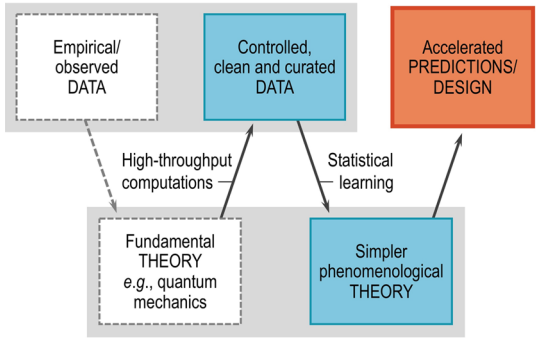


Figure 1. Role of data in building predictive theories and phenomenological models. For the purpose of identifying correlations and building phenomenological models, data may be prepared in a controlled manner, for instance via high-throughput methods, followed by the analysis of the accumulated data using best-practices statistical methods.

experiencing enormous electric fields has long defied the creation of a predictive theory. This is largely because the dielectric degradation and breakdown process in real materials is complex—it is the result of the interplay between the

Received: October 23, 2015
Revised: February 2, 2016
Published: February 2, 2016

DOI: 10.1021/acs.chemmater.5b04109
Chem. Mater. 2016, 28, 1304−1311

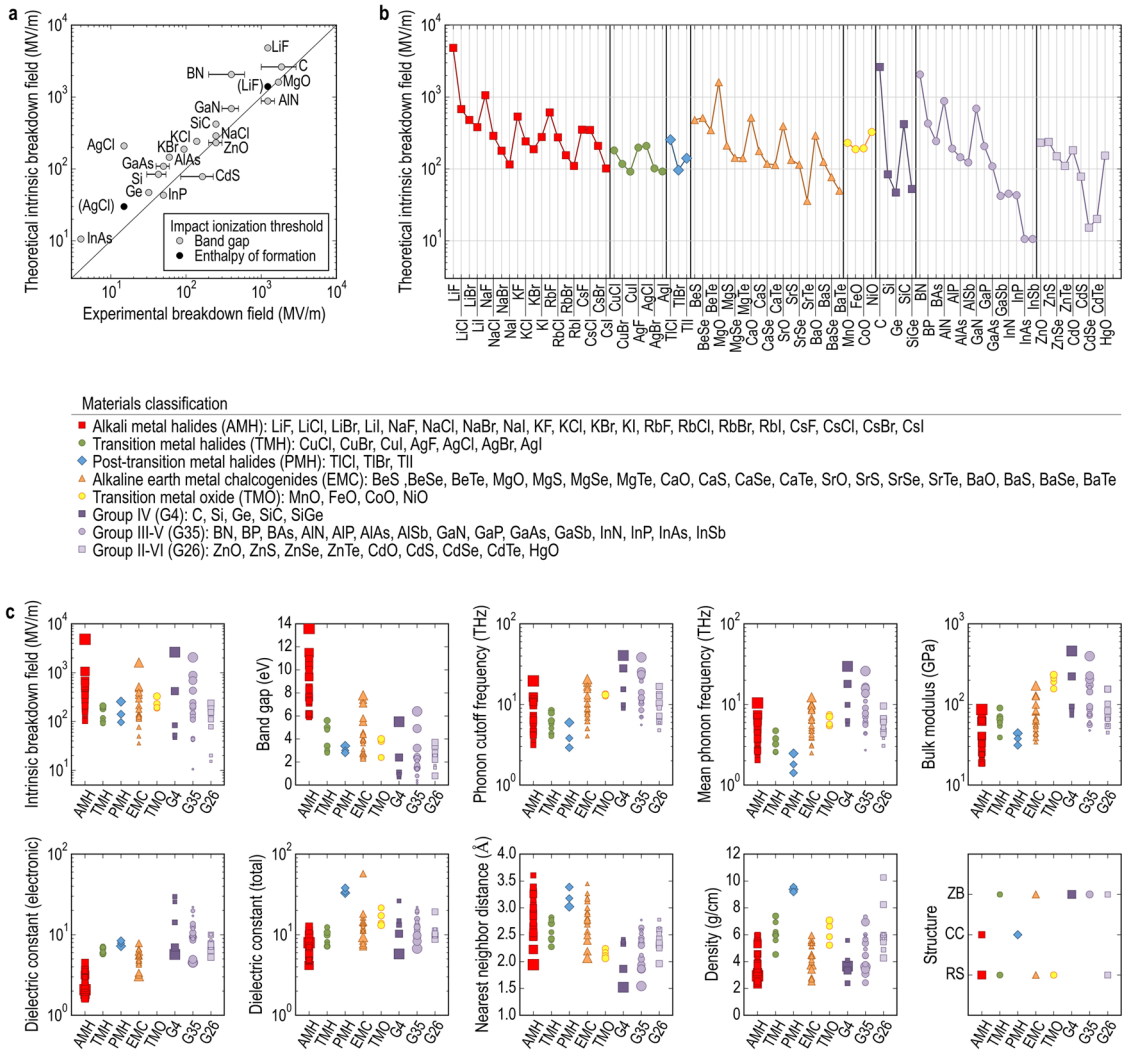


Figure 2. Intrinsic dielectric breakdown field of elemental and binary insulators. (a) Validation of the computational data generation method by comparison of the theoretical (DFT-computed) intrinsic dielectric breakdown field with available experimental results. Error bars span the minimum and maximum known experimental values. In the cases of LiF and AgCl, the enthalpy of formation is much lower than the band gap indicating that bond breakage will occur before impact ionization. When the enthalpy of formation is used as the impact ionization threshold for these materials (instead of the band gap), the computed intrinsic dielectric breakdown field is in good agreement with experiments. (b) DFT-computed intrinsic dielectric breakdown field for 82 reference insulators (including 79 binary compounds and 3 elemental materials). (c) DFT-computed intrinsic dielectric breakdown field, the 8 primary features (expected to bear a cause-effect relationship with the breakdown field), and the structure of the 82 insulators plotted separately for various materials subclasses studied. All properties are obtained from first-principles calculations except the band gap and structure for which experimental results are used. The size of the symbols in each plot is proportional to the magnitude of the intrinsic dielectric breakdown field, to visually reveal correlations. It can be seen that larger dielectric breakdown field values are correlated with larger band gap and phonon frequencies, and smaller dielectric constant and nearest neighbor distances. In the panel showing crystal structures, symbol sizes are determined by the average magnitude of the intrinsic dielectric breakdown field for materials in the same structure and materials subclass. No particular correlation of the breakdown field with structure appear to exist.

magnitude of the electric field, the time span of imposition of the field, the temperature, and the state of the material (i.e., its defect content and morphology). Dielectric degradation (which eventually culminates in breakdown) is difficult to track empirically, and is essentially the progressive creation and accumulation of atomic and nanoscale defects. The primary focus of this contribution though is the intrinsic dielectric breakdown field, determined solely by the material's chemistry. This quantity has a special significance as it is the highest possible electric field that a "perfect" (i.e., defect-free) material can tolerate; it is hence the theoretical limit of dielectric breakdown.

The starting point for the present development is the fundamental theory of intrinsic dielectric breakdown as formulated by von Hippel[9] and Fröhlich,[10−12] and recently implemented within a first-principles computational framework.[13] The intrinsic dielectric breakdown field, computed from first-principles, for a benchmark set of 82 insulators provides the data set for the discovery of a phenomenological dielectric breakdown model (see Figure 2). This starting point—based on computed data—is justified because experimental values of the breakdown field are available only for a small number of cases (and even for those cases, samples and measurement methods are prone to enormous statistical variation). It is also worth noting that the first-principles

DOI: 10.1021/acs.chemmater.5b04109
Chem. Mater. 2016, 28, 1304−1311

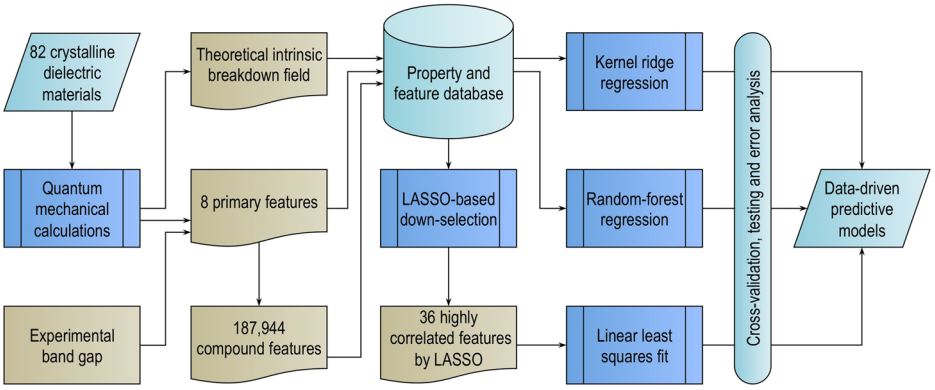


Figure 3. Schematic workflow used in the data-driven discovery of a phenomenological model of intrinsic dielectric breakdown. Of the three learning models adopted, kernel ridge regression, and random forest regression attempt to predict the intrinsic dielectric breakdown field of a material given a set of 8 material property features (but without resorting to actual functional forms), whereas the linear least-squares pathway discovers the functional relationship between the intrinsic dielectric breakdown field and a set of compound (nonlinear) features identified by the least absolute shrinkage and selection operator (LASSO). See the Supporting Information S2 for detailed workflow of model discovery.

computational approach to predicting the intrinsic dielectric breakdown field of an insulator is extremely computation time intensive, and does not automatically reveal key correlations, hence warranting a targeted search for simpler models.

A principled approach is then adopted to correlate the computed breakdown field to a variety of far more easily accessible material properties. The rationale for such a search for correlations is provided by fairly (visually) obvious patterns (see Figure 2b) and relationships between the breakdown field and properties such as the band gap, dielectric constant, phonon spectra, etc. (see Figure 2c). Data-driven models employing advanced statistical learning routines, inspired by Big Data concepts,[14−19] were utilized to discover simple predictive models of dielectric breakdown. Such techniques have risen in popularity and have recently been applied successfully to address several important materials and chemical science problems.[20−28] The present work differs from these past efforts in the manner in which materials are represented; "higher-level" attributes of the system (as mentioned above and discussed in detail below) are used here, in contrast atomic or molecular level attributes used in much of these recent past studies. The models of dielectric breakdown discovered here using data-driven approaches are then tested, and validated, on new materials not in the original data set. This development is as revealing as it is powerful, and has the potential to guide the development of new electric field tolerant materials with high breakdown strength.

METHODS

Computation of the Intrinsic Dielectric Breakdown Field. At low applied electric fields and nonzero temperatures, the conduction electron energy distribution reaches a steady state, since the energy gain from the external electric field is balanced by energy loss from collisions with phonons. However, at sufficiently large electric fields, the electrons in the conduction band incessantly gain kinetic energy from the external electric field until a threshold—equal to the band gap energy of the insulator—is reached when a high-energy electron, via impact ionization, leads to carrier multiplication signaling breakdown. According to the Fröhlich-von Hippel dielectric breakdown criterion,[9−12] the breakdown field is the lowest external field at which the average electron energy gain from the field is greater than the average energy loss to phonons for all electron energies less than those which can give rise to carrier multiplication. Thus, the electron−phonon interactions provide the only relevant scattering mechanism in this theory. The Fröhlich-von Hippel criterion for intrinsic dielectric breakdown has been recently implemented within a first-principles density functional theory (DFT) framework.[13] This criterion can be written as

$$A(E, F) > B(E) \text{ for all } E \text{ in } \{CBM, E_i\} \quad (1)$$

where $A(E,F)$ is the rate of the energy gain of an electron of energy E at an electric field F, and $B(E)$ is the rate of energy loss. The threshold energy for impact ionization, E_i, is assumed to be $CBM + E_g$, where CBM is the conduction band minimum and E_g is the band gap (see Figure S1). The rate of energy gain of the electron can be evaluated as

$$A(E, F) = \frac{e^2\tau(E)F^2}{3m} \quad (2)$$

where e and m are the electronic charge and mass, respectively. $\tau(E)$ is the electron relaxation time due to phonon scattering. Determination of both $\tau(E)$ and $B(E)$ requires a knowledge of the electron−phonon coupling function as explained in previous studies.[13,29] Both $\tau(E)$ and $B(E)$ were evaluated at a temperature of 300 K. All relevant quantities and the intrinsic dielectric breakdown field were computed using DFT within the local density approximation (LDA) and norm conserving pseudopotentials[30,31] as implemented in the Quantum ESPRESSO code.[32] Electron−phonon coupling function was computed in the linear response regime using density functional perturbation theory (DFPT). A Monkhorst−Pack k-point mesh of 32 × 32 × 32 (to sample the electronic states) and q-point mesh of 4 × 4 × 4 (to sample the phonon states) was used for all materials to obtain converged results.[33]

Computation of Properties Expected to Be Correlated to Breakdown. In an effort to identify correlations between the breakdown field and other more easily accessible properties, we considered eight material properties that are expected to bear a cause-effect relationship with the breakdown field. This set of "primary features" included the band gap, the average and maximum (cutoff) phonon frequency, the dielectric constant (electronic and total), and attributes that may control the behavior of phonons and their scattering propensity (such as the nearest neighbor distance, mass density, and the bulk modulus).

The average phonon frequency, phonon cutoff frequency, nearest neighbor distance, mass density, and bulk modulus were computed using Quantum ESPRESSO, and the electronic part of the dielectric constant and the total dielectric constant were determined using the Vienna Ab initio Simulation Package,[34] using projector augmented wave (PAW)[35,36] frozen core potentials. All DFT computations for all materials except the 4 transition metal oxides MnO, FeO, CoO, and NiO were performed at the LDA level of theory. For these oxides, spin-polarized LDA+U[37] calculations were performed, with the effective U parameters being 2.1, 4.3, 7.0, and 7.1 for Mn, Fe, Co, and Ni, respectively. Owing to the significant uncertainties in the DFT

DOI: 10.1021/acs.chemmater.5b04109
Chem. Mater. 2016, 28, 1304−1311

band gap predictions (regardless of the exchange-correlation interaction treatment), experimental band gap data was used uniformly for all cases.

Statistical Learning Methods. Our goal is to discover general mathematical relationships between the property of interest P_i (i.e., the intrinsic dielectric breakdown field) of material i, and an Ω-dimensional representation (or descriptor or feature vector) $\mathbf{d}_i$ of the material i. A suitable initial choice of $\mathbf{d}_i$, and the rules by which this initial choice is contracted and manipulated to lead to predictive models $P_i(\mathbf{d}_i)$ is at the heart of modern machine learning methods. Here, we pursue three fundamentally distinct strategies for model discovery. The overall workflow adopted is outlined in Figure 3.

The first of these schemes is kernel ridge regression (KRR), capable of handling complex nonlinear relationships.[16,38,39] The KRR method works on the principle of similarity. By comparing $\mathbf{d}_i$ of material i with those of a set of reference cases for which the property values are known (say, via a distance measure such as the Euclidean norm), an interpolative prediction of P_i can be made.

The second learning scheme adopted is random forest regression (RFR), which involves creation of an ensemble of decision trees. Predictions from the forest are then made by averaging over predictions from the individual trees. In both learning approaches, the components of $\mathbf{d}_i$ were drawn from the set of the eight primary property features, namely, band gap, average phonon frequency, phonon cutoff frequency, electronic part of the dielectric constant, total dielectric constant, nearest neighbor distance, mass density, and bulk modulus.

The third adopted learning strategy—namely, least absolute shrinkage and selection operator (LASSO)-based model selection—was fundamentally different from the first two, and in the end, proved to be the most revealing, powerful, and accurate. Although KRR and RFR provided predictions of the dielectric breakdown field (i.e., P_i), and identified the most relevant subset of the 8 primary features that determine the dielectric breakdown field, they are not constructed to provide the actual functional relationship between P_i and the relevant features. LASSO, on the other hand, offers a pathway for this explicit functional relationship determination. In order to exploit this capability, conjunctive—or compound—features were built explicitly in a controlled manner, starting with the 8 primary features in the following way: 12 prototypical functions, namely, x, x^{-1}, $x^{1/2}$, $x^{-1/2}$, x^2, x^{-2}, x^3, x^{-3}, $\ln x$, $(\ln x)^{-1}$, e^x, and e^{-x}, with x being one of the 8 primary features were considered. This immediately leads to 96 features. Cross multiplying these features of single functions taken either two or three at a time leads to additional 4480 and 183 368 features, respectively. The sum total of these provided us with 187 944 compound features, each of which is a function involving up to three primary features.

From this large set of compound features, our goal was to look for the Ω-dimensional (preferably, for a small Ω value) descriptor that gives the best linear fit with the intrinsic dielectric breakdown field. Here, by Ω-dimensional descriptor we mean a descriptor composed of Ω number of compound features. The LASSO algorithm helps solve this nondeterministic polynomial-time (NP)-hard problem (whose computational solution is infeasible), by recasting it into a convex minimization problem.[16,40] This strategy of creating a large number of initial compound features and down-selecting to the most relevant ones using LASSO has recently been successfully employed for the first time to identify new descriptors to classify binary AB-octet crystal structures.[20] A somewhat related recent development involved identification of lower-dimensional representations of alloy cluster expansions.[41] Further details on the KRR, RFR, and LASSO methodology are provided in the Supporting Information S3.

RESULTS AND DISCUSSION

Intrinsic Dielectric Breakdown Field and Feature Properties. Figure 2a compares our computed theoretical intrinsic dielectric breakdown field with available experimental values for a number of elemental and binary dielectrics. As can be seen, the breakdown field spans 3 orders of magnitude for this set of materials considered, with favorable agreement between calculations and experiments over this entire range. The theoretical intrinsic dielectric breakdown field for a much broader data set of 82 elemental and binary dielectric materials that occur in a variety of crystal structures (but all with two atoms per primitive cell) including the zinc blende (ZB), rock salt (RS) and cesium chloride (CC) structures is shown in Figure 2b, subdivided in terms of materials subclasses. Clear self-evident periodic trends in the values of the breakdown field can be seen, with more ionic binary compounds (i.e., those with a highly electronegative anion) displaying larger breakdown field values. See, for instance, the monotonic decrease in the breakdown field along the LiF, LiCl, LiBr, LiI series, followed by a jump at NaF, and then another pattern of monotonic decrease, etc. This is the first clue that elementary correlations between the breakdown field and other attributes of the systems exist.

All eight properties along with the computed intrinsic dielectric breakdown field for the 82 compounds of our data set are tabulated in Supporting Information S4. The ranges of possible values the breakdown field and each of the 8 properties can take (clustered in terms of materials subclasses) are captured in Figure 2c. Each plot in this figure corresponds to a specific property, and the size of the symbol used for a material is chosen to be proportional to its breakdown field value, to reveal correlations. It can be seen immediately that larger dielectric breakdown field values are correlated with larger band gap and phonon frequencies, and smaller dielectric constant and nearest neighbor distances. The advantage of this feature set of eight properties over properties such as the ionicity discussed above is that the former applies to any material (i.e., not just binaries), and so any learning or phenomenological model of dielectric breakdown that one may arrive at on the basis of these features is potentially generalizable to a new material (and thus testable).

Data-Driven Model Discovery. Perhaps the most important (and practical) reason for seeking to develop a phenomenological model of intrinsic dielectric breakdown is that the DFT approach to predicting this property, although general and accurate, is exceedingly computation-time intensive even for the simple elemental or binary dielectrics (composed of just two atoms per primitive cell) considered here. The DFT route will be impractical for widespread studies of systems involving larger unit cells. Furthermore, a brute-force application of the DFT methodology will not automatically allow us to understand the underlying factors that control breakdown. Hence, we aim to discover quantitative and verifiable relationships between the breakdown field and other easily accessible (or computable) materials properties (such as the eight properties discussed above) using the data accumulated for the 82 benchmark compounds. Indeed, such attempts have been undertaken in the past based on empirical breakdown data,[42,43] but have met with marginal success because of the significant uncertainties and statistical variation associated with the data, and the limited repertoire of data-mining methods used.

The performance of the three adopted learning models, namely, KRR, RFR and LASSO, was evaluated by their ability to "learn" the intrinsic dielectric breakdown field using an Ω-dimensional descriptor, which is composed of either the primary features (for KRR and RFR models) or the compound features (for LASSO-based model). Since the breakdown field spans almost 3 orders of magnitude (see Figure 2b), the logarithm of this quantity was used in all learning models as the

DOI: 10.1021/acs.chemmater.5b04109
Chem. Mater. 2016, 28, 1304−1311

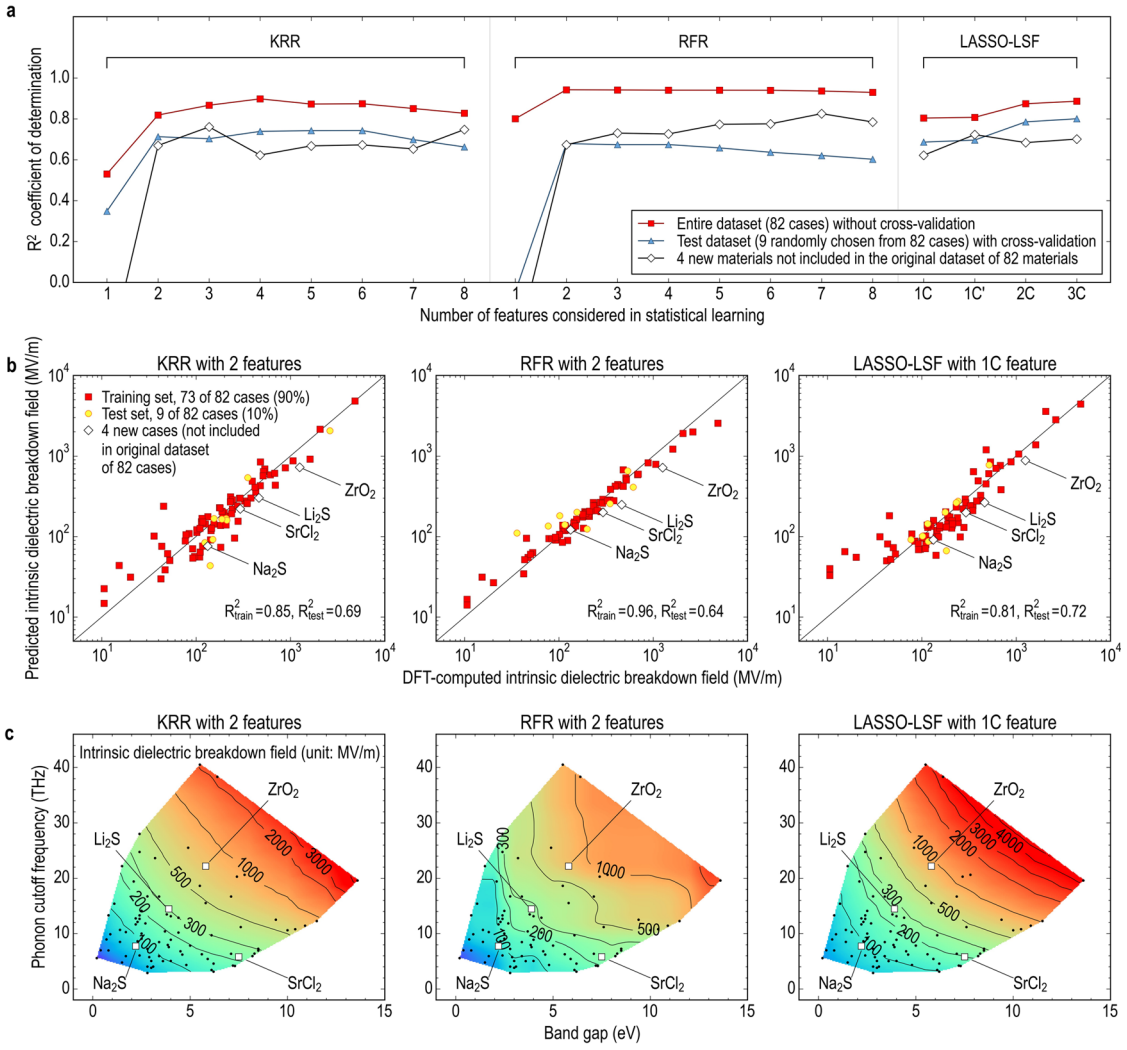


Figure 4. Performance and utility of phenomenological prediction models obtained by data-driven statistical learning. (a) Coefficient of determination (R^2) of the models with and without cross-validation, for Ω-dimensional descriptors (or feature vectors). Ω ranges from 1 to 8 in the case of the kernel ridge regression (KRR) and random forest regression (RFR) models, while it ranges from 1 to 3 in the case of the least absolute shrinkage and selection operator based least-squares fit (LASSO-LSF) model. In order to clarify that the LASSO-LSF model utilizes compound features, the number of features for this case is represented as nC with n = 1−3. The LASSO-LSF procedure leads to two different 1-dimensional feature vector with equal predictive power, and these are indicated as 1C and 1C′. By definition, $R^2 = 1 - \Sigma (y_i - y_{i,\,\mathrm{pred}})^2 / \Sigma (y_i - y_{\mathrm{avg}})^2$ with y_i, $y_{i,\,\mathrm{pred}}$ and y_{avg} being the DFT computed intrinsic dielectric breakdown field, model-predicted breakdown field and average of DFT values in the training or test set. Estimation of R^2 for 4 new compounds (Li_2S, Na_2S, $SrCl_2$, and ZrO_2) not already included in original data set are shown as well. (b) Parity plots comparing DFT-computed intrinsic dielectric breakdown field against predicted intrinsic breakdown field obtained by KRR, RFR and LASSO-LSF, for the training and test sets (drawn from the set of 82 original compounds), as well as for the 4 new compounds not included during the model development stage. (c) Design maps for the prediction of intrinsic dielectric breakdown field using the band gap and phonon cutoff frequency. The corresponding values of these two properties of the 82 benchmark materials are indicated using dots, and further highlighted by the shading. The 4 new compound materials, namely, Li_2S, Na_2S, $SrCl_2$ and ZrO_2, are also included. The DFT-computed intrinsic dielectric breakdown field are 462.5, 132.7, 293.4, and 1253.0 MV/m, respectively, for these cases.

target property. In the two models that employ primary features, Ω was varied from 1 to 8. For each choice of Ω, all possible Ω-tuples of the primary features were considered. In the third strategy, LASSO was used to down-select from the ∼190 000 compound features to a set of the top 36 compound features. Again, for each choice of Ω (which in this case was allowed to range from 1 to 3), all possible Ω-tuples of the contracted 36 compound features were considered. These were then used in a linear least-squares model to fit to the intrinsic dielectric breakdown field.

To quantify learning performance, we performed a systematic analysis in terms of various metrics. One such metric, namely, the coefficient of determination (R^2) is shown (and defined) in Figure 4a for each of the three learning models for various choices of Ω, both with and without cross-validation. Cross-validation—an essential statistical tool to combat overfitting and to ensure model generality[16]—was accomplished by breaking up the data set of 82 compounds into random training (90%) and test (10%) set splits, training the learning models on the training set, applying the model on the test set, and determining the corresponding R^2. For each choice of Ω, results corresponding to the best performing Ω-tuple is shown. Other performance measures were also considered (presented in the Supporting Information S5 and S6) but provided similar results.

DOI: 10.1021/acs.chemmater.5b04109
Chem. Mater. 2016, 28, 1304−1311

It can be concluded based on Figure 4a that the performance of the models generated by KRR and RFR with the best 1-dimensional descriptor is not impressive. In contrast to the 1-dimensional descriptor, the performance of the best 2-dimensional descriptor for both KRR and RFR is significantly better, and almost as good as descriptors with higher dimensionality. The best performing 2-dimensional descriptors within both the KRR and RFR involved the band gap and the phonon cutoff frequency. In the case of LASSO-based prediction model, two different 1-dimensional compound descriptors, namely $\sqrt{E_g\omega_{max}}$ and $\omega_{max}\sqrt{E_g}/\ln(\omega_{mean})$, were equally good, where E_g, ω_{max}, and ω_{mean} are, respectively, the band gap, phonon cutoff frequency, and mean phonon frequency. Both of these 1-dimensional descriptors were competitive in performance compared to descriptors with higher dimensionality. We favor the simpler of the two, namely, $\sqrt{E_g\omega_{max}}$. Interestingly, all three learning models single out the same two properties, namely, E_g and ω_{max}, as most relevant—in the end, not surprising, as the former is related to electronic excitations and the latter to the scattering of the excited electrons by phonons. Moreover, the LASSO approach, via a least-squares fit (LSF) of the logarithm of the breakdown field, F_b, versus $\sqrt{E_g\omega_{max}}$, provides an explicit functional form (after cross-validation),

$$F_b = 24.442\exp(0.315\sqrt{E_g\omega_{max}}) \quad (3)$$

with F_b, E_g, and ω_{max} specified in units of MV/m, eV, and THz, respectively. The LASSO-LSF procedure is reminiscent of symbolic regression, a technique in which such functional forms are discovered by searching the mathematical space of symbolic operations involving the variables (or features, in the current parlance) via evolutionary algorithms.[44] See the Supporting Information S7 for other functional forms with 1-dimensional compound descriptors.

Figure 4b shows the performance of our three prediction models via parity plots by comparing the DFT-calculated and model-predicted intrinsic dielectric breakdown field. The predictions using the KRR and RFR models were performed by taking just the $\{E_g, \omega_{max}\}$ as the features, whereas eq 3 was used to make the LASSO-LSF model predictions. The parameters that enter all three prediction models were obtained after cross-validation. Contour plots that directly reveal the dependence of the breakdown field on $\{E_g, \omega_{max}\}$ are shown in Figure 4c. Although the general aspects of all three prediction models, as captured in Figure 4c, are qualitatively similar, the RFR model appears to lead to significant corrugations of the breakdown field contours (the LASSO-LSF model, on the other hand, leads to smoother contours owing to its simple and clearly defined functional form, namely, eq 3). These contour plots may be viewed as "design maps" that can aid in the rapid screening and identification of new materials with high dielectric breakdown strength, provided that their band gap and phonon cutoff frequencies are known and fall within the ranges of the data set materials. Given the similar errors underlying the predictions of the three models, we recommend that all three models be consulted, and the results used as estimates of the intrinsic breakdown field of a new material. Still, the LASSO-LSF approach has an edge in that an analytical form (eq 3) to predict the intrinsic breakdown field is at hand. This outlook is consistent with the notion that 'all models are wrong, but some are useful'.[45]

How Generalizable Is the Model? The strength of the above claim, namely, that the learned phenomenological models will have broad applicability, may be tested by considering new compounds not already part of the original data set of 82 insulators, whose band gaps and phonon cutoff frequencies lie in the range used for statistical learning (see the shaded regions of Figure 4c). Toward that end, the intrinsic dielectric breakdown field of Li_2S, Na_2S, $SrCl_2$, and ZrO_2 that consist of 3 atoms in $Fm\bar{3}m$ structure was computed using DFT (see the Supporting Information S8 for these results, along with the corresponding 8 primary property features). On the basis of the primary property features of these insulators, their breakdown field was predicted as well using the three learning models. Figure 4 captures the performance of these predictions. The R^2 corresponding to these new compounds is comparable to that of the original test set within all three models for all choices of the descriptor dimensionality (see Figure 4a). Comparison of predicted breakdown field (using the 2 features, E_g and ω_{max}) with the DFT results also shows fair agreement as shown in Figure 4b, meaning that these 2 features play an important role in determining the dielectric breakdown field of the new compounds as well. For completeness, and to highlight the utility afforded by the contour maps of Figure 4c, the results corresponding to the 4 new compounds are also placed in those plots.

CONCLUSION

The present development has specific and broad implications. It has been demonstrated that the intrinsic dielectric breakdown field—the theoretical limit of catastrophic breakdown of a material subjected to high electric fields—may be computed from first-principles for a variety of insulators. More importantly, statistical or machine learning based protocols have been developed to systematically and rationally distill out materials property features that are most correlated with the intrinsic dielectric breakdown field. These efforts have led to predictive and analytical phenomenological models of dielectric breakdown that are revealing, accurate, and enormously faster than the fundamental first-principles approach. The focal point of this work has been the intrinsic dielectric breakdown field determined purely by the chemistry of the material (i.e., the elements the material is composed of, the atomic-level structure, and the bonding); extrinsic factors (such as defects and their dynamical evolution) expected to play important roles in determining real engineering breakdown catalyzed by gradual dielectric degradation have not been considered. Nevertheless, the models developed here may be used in a first line of screening aimed at identifying high electric field tolerant materials. It is worth noting that the present choices of materials, dimensions and fields encountered in electrical and electronic systems in this age of ultraminiaturization are limited by the dielectric breakdown of the insulations. More broadly, we have demonstrated that the power of high throughput computing (for data generation) and the available repertoire of Big Data analytics tools may be harnessed and aimed at the discovery of simple models of complex materials problems.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.chemmater.5b04109.

DOI: 10.1021/acs.chemmater.5b04109
Chem. Mater. 2016, 28, 1304−1311

Breakdown mechanism; phenomenological model discovery work flow; statistical learning methods; computed intrinsic dielectric breakdown field and primary features; error analysis for prediction models; Ω-dimensional features with prediction errors; prediction models found by LASSO-LSF with 1-dimensional compound features; data set used for validation of prediction models, and references (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: rampi@uconn.edu.

Author Contributions

C.K. and G.P. contributed equally to the work and manuscript, and R.R. designed and supervised the study. Specifically, the high-throughput DFT computations were performed by C.K., and the learning models were developed by G.P. All authors discussed the results, wrote, and shaped the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This paper is based on work supported by the Office of Naval Research through grants N00014-10-1-0944 and N00014-15-1-2665, the former being a Multidisciplinary University Research Initiative (MURI) grant. 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 grant (20140679PRD3) as a Director's postdoctoral fellowship. Luca Ghiringhelli, Jan Vybiral and Matthias Scheffler are acknowledged for detailed guidance on using the LASSO-LSF method, and a critical reading of the manuscript. Ying Sun and Clive Bealing are acknowledged for a prior post-Quantum ESPRESSO code development effort to compute the intrinsic dielectric breakdown field. Discussions with Kenny Lipkowitz and Paul Armistead, and a critical reading of the manuscript by Erik Nykwest, Venkatesh Botu, and Huan Tran are also acknowledged.

REFERENCES

(1) Zeilinger, A. The Quantum Centennial. *Nature* **2000**, *408*, 639−641.

(2) Cohen, I. B. Newton's Discovery of Gravity. *Sci. Am.* **1981**, *244*, 166−179.

(3) Mahon, B. How Maxwell's Equations Came to Light. *Nat. Photonics* **2015**, *9*, 2−4.

(4) Hume-Rothery, W.; Smallman, R. E.; Haworth, C. W. *The Structure of Metals and Alloys*; The Institute of Metals: London, 1988.

(5) Hall, E. O. *Yield Point Phenomena in Metals and Alloys*; Plenum Press: New York, 1970.

(6) Petch, N. J. The Cleavage Strength of Polycrystals. *J. Iron Steel Inst.* **1953**, *174*, 25−32.

(7) Crowson, R. A. Science and Phenomenology. *Nature* **1969**, *223*, 1318−1319.

(8) Tolédano, J.-C.; Tolédano, P. *The Landau Theory of Phase Transitions*; World Scientific: Singapore, 1987.

(9) Von Hippel, A. Electric Breakdown of Solid and Liquid Insulators. *J. Appl. Phys.* **1937**, *8*, 815−832.

(10) Fröhlich, H. Theory of Electrical Breakdown in Ionic Crystals. *Proc. R. Soc. London, Ser. A* **1937**, *160*, 230−241.

(11) Fröhlich, H. Theory of Dielectric Breakdown. *Nature* **1943**, *151*, 339−340.

(12) Fröhlich, H. On the Theory of Dielectric Breakdown in Solids. *Proc. R. Soc. London, Ser. A* **1947**, *188*, 521−532.

(13) Sun, Y.; Boggs, S. A.; Ramprasad, R. The Intrinsic Electrical Breakdown Strength of Insulators from First Principles. *Appl. Phys. Lett.* **2012**, *101*, 132906.

(14) Ghahramani, Z. Probabilistic Machine Learning and Artificial Intelligence. *Nature* **2015**, *521*, 452−459.

(15) LeSar, R. Materials informatics: An Emerging Technology for Materials Development. *Stat. Anal. Data Min.* **2009**, *1*, 372−374.

(16) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: New York, 2009.

(17) Witten, I. H.; Frank, E.; Hall, M. A. *Data Mining: Practical Machine Learning Tools and Techniques*; Elsevier: Amsterdam, 2011.

(18) Fischer, C. C.; Tibbetts, K. J.; Morgan, D.; Ceder, G. Predicting Crystal Structure by Merging Data Mining with Quantum Mechanics. *Nat. Mater.* **2006**, *5*, 641−646.

(19) Srinivasan, S.; Rajan, K. Property Phase Diagrams" for Compound Semiconductors Through Data Mining. *Materials* **2013**, *6*, 279−290.

(20) 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.* **2015**, *114*, 105503.

(21) Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J. E.; Doak, J. W.; Thompson, A.; Zhang, K.; Choudhary, A.; Wolverton, C. Combinatorial Screening for New Materials in Unconstrained Composition Space with Machine Learning. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *89*, 094104.

(22) Schütt, K. T.; Glawe, H.; Brockherde, F.; Sanna, A.; Müller, K. R.; Gross, E. K. U. How to Represent Crystal Structures for Machine Learning: Towards Fast Prediction of Electronic Properties. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *89*, 205118.

(23) Fernandez, M.; Boyd, P. G.; Daff, T. D.; Aghaji, M. Z.; Woo, T. K. Rapid and Accurate Machine Learning Recognition of High Performing Metal Organic Frameworks for CO_2 Capture. *J. Phys. Chem. Lett.* **2014**, *5*, 3056−3060.

(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.* **2015**, *115*, 1094−1101.

(25) Faber, F.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R. Machine Learning Energies for 2 M Elpasolite (ABC_2D_6) Crystals. http://arxiv.org/abs/1508.05315.

(26) Pilania, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R. Accelerating Materials Property Predictions using Machine Learning. *Sci. Rep.* **2013**, *3*, 2810.

(27) Pilania, G.; Mannodi-Kanakkithodi, A.; Uberuaga, B. P.; Ramprasad, R.; Gubernatis, J. E.; Lookman, T. Machine Learning Bandgaps of Double Perovskites. *Sci. Rep.* **2016**, *6*, 19375.

(28) Mueller, T.; Kusne, A. G.; Ramprasad, R. *Machine Learning in Materials Science: Recent Progress and Emerging Applications*; Parrill, A. L.; Lipkowitz, K. B., Ed.; Reviews in Computational Chemistry; Wiley: New York, 2016.

(29) Sjakste, J.; Vast, N.; Tyuterev, V. *Ab initio* Method for Calculating Electron-Phonon Scattering Times in Semiconductors: Application to GaAs and GaP. *Phys. Rev. Lett.* **2007**, *99*, 236405.

(30) Perdew, J. P.; Wang, Y. Accurate and Simple Analytic Representation of the Electron-gas Correlation Energy. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1992**, *45*, 13244−13249.

(31) Ceperley, D. M.; Alder, B. J. Ground State of the Electron Gas by a Stochastic Method. *Phys. Rev. Lett.* **1980**, *45*, 566−569.

(32) Giannozzi, P.; Baroni, S.; Bonini, N.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Chiarotti, G. L.; Cococcioni, M.; Dabo, I.; Dal Corso, A.; de Gironcoli, S.; Fabris, S.; Fratesi, G.; Gebauer, R.; Gerstmann, U.; Gougoussis, C.; Kokalj, A.; Lazzeri, M.; Martin-Samos, L.; Marzari, N.; Mauri, F.; Mazzarello, R.; Paolini, S.; Pasquarello, A.; Paulatto, L.; Sbraccia, C.; Scandolo, S.; Sclauzero, G.; Seitsonen, A. P.; Smogunov, A.; Umari, P.; Wentzcovitch, R. M. QUANTUM ESPRESSO: a Modular and Open-source Software Project for

DOI: 10.1021/acs.chemmater.5b04109
Chem. Mater. 2016, 28, 1304−1311

Quantum Simulations of Materials. *J. Phys.: Condens. Matter* **2009**, *21*, 395502.
(33) Monkhorst, H. J.; Pack, J. D. Special Points for Brillouin-zone Integrations. *Phys. Rev. B* **1976**, *13*, 5188−5192.
(34) Kresse, G.; Furthmüller, J. Efficiency of Ab-initio Total Energy Calculations for Metals and Semiconductors using a Plane-wave Basis Set. *Comput. Mater. Sci.* **1996**, *6*, 15−50.
(35) Blöchl, P. E. Projector Augmented-wave Method. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1994**, *50*, 17953−17979.
(36) Kresse, G.; Joubert, D. From Ultrasoft Pseudopotentials to the Projector Augmented-wave Method. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1999**, *59*, 1758−1775.
(37) Dudarev, S. L.; Botton, G. A.; Savrasov, S. Y.; Humphreys, C. J.; Sutton, A. P. Electron-energy-loss Spectra and the Structural Stability of Nickel Oxide: An LSDA+U Study. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1998**, *57*, 1505−1509.
(38) Müller, K. R.; Mika, S.; Ratsch, G.; Tsuda, K.; Scholkopf, B. An Introduction to Kernel-based Learning Algorithms. *IEEE Trans. Neural Networks* **2001**, *12*, 181−201.
(39) Hofmann, T.; Scholkopf, B.; Smola, A. J. Kernel Methods in Machine Learning. *Ann. Statist.* **2008**, *36*, 1171−1220.
(40) Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc., Ser. B* **1996**, *58*, 267−288.
(41) Nelson, L. J.; Hart, G. L. W.; Zhou, F.; Ozoliņš, V. Compressive Sensing as a Paradigm for Building Physics Models. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *87*, 035125.
(42) McPherson, J.; Kim, J.; Shanware, A.; Mogul, H.; Rodriguez, J. Proposed Universal Relationship between Dielectric Breakdown and Dielectric Constant. *IEDM Technol. Dig.* **2002**, 633−636.
(43) McPherson, J.; Kim, J.; Shanware, A.; Mogul, H.; Rodriguez, J. Trends in the Ultimate Breakdown Strength of High Dielectric-Constant Materials. *IEEE Trans. Electron Devices* **2003**, *50*, 1771−1778.
(44) Schmidt, M.; Lipson, H. Distilling Free-Form Natural Laws from Experimental Data. *Science* **2009**, *324*, 81−85.
(45) Box, G. E. P.; Draper, N. R. *Empirical Model-Building and Response Surfaces*; John Wiley & Sons; New York, 1987.

DOI: 10.1021/acs.chemmater.5b04109
Chem. Mater. 2016, 28, 1304−1311