

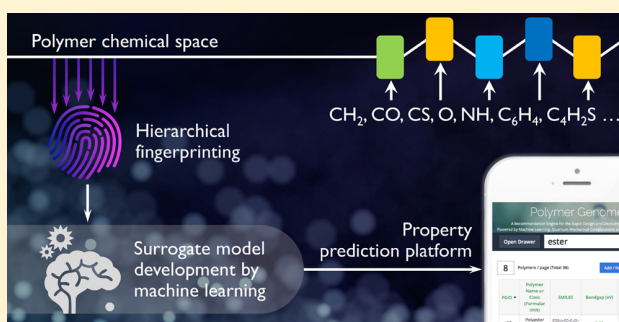
Polymer Genome: A Data-Powered Polymer Informatics Platform for Property Predictions

Chiho Kim,^{†,§} Anand Chandrasekaran,^{†,§} Tran Doan Huan,^{‡,§} Deya Das,[†] and Rampi Ramprasad^{*,†,§}[†]School of Materials Science and Engineering, Georgia Institute of Technology, 771 Ferst Drive NW, Atlanta, Georgia 30332, United States[‡]Department of Materials Science and Engineering and Institute of Materials Science, University of Connecticut, 97 North Eagleville Road, Storrs, Connecticut 06269-3136, United States

S Supporting Information

ABSTRACT: The recent successes of the Materials Genome Initiative have opened up new opportunities for data-centric informatics approaches in several subfields of materials research, including in polymer science and engineering. Polymers, being inexpensive and possessing a broad range of tunable properties, are widespread in many technological applications. The vast chemical and morphological complexity of polymers though gives rise to challenges in the rational discovery of new materials for specific applications. The nascent field of polymer informatics seeks to provide tools and pathways for accelerated property prediction (and materials design) via surrogate machine learning models built on reliable past data.

We have carefully accumulated a data set of organic polymers whose properties were obtained either computationally (bandgap, dielectric constant, refractive index, and atomization energy) or experimentally (glass transition temperature, solubility parameter, and density). A fingerprinting scheme that captures atomistic to morphological structural features was developed to numerically represent the polymers. Machine learning models were then trained by mapping the fingerprints (or features) to properties. Once developed, these models can rapidly predict properties of new polymers (within the same chemical class as the parent data set) and can also provide uncertainties underlying the predictions. Since different properties depend on different length-scale features, the prediction models were built on an optimized set of features for each individual property. Furthermore, these models are incorporated in a user-friendly online platform named Polymer Genome (www.polymergenome.org). Systematic and progressive expansion of both chemical and property spaces are planned to extend the applicability of Polymer Genome to a wide range of technological domains.



1. INTRODUCTION

The past few years have seen a surge in the application of data-driven techniques in a plethora of research and development fields, ranging from image-recognition¹ to drug-discovery.^{2,3} Sophisticated machine learning techniques, initially within the purview of computer science researchers mainly, are now becoming ubiquitous in many other branches of science and engineering and have the potential to spur technological innovations.

In materials science, the increasing availability of large amounts of data (both computational and experimental) has led to the prominent field of materials informatics over the past few years.^{4–21} The strategic visions and plans of the Materials Genome Initiative (USA)²² and the recent developments of essential infrastructures for materials informatics such as the NOMAD Laboratory: a European Centre for Excellence (EU),²³ and the Materials Research by Information Integration Initiative (Japan) are expected to lead to a paradigm shift in the discovery of novel functional materials.²⁴

Polymers form an important (and challenging) materials class. They are pervasive with applications ranging from daily

products, e.g., plastic packaging and containers, to state-of-the-art technological components, e.g., high-energy density capacitors, electrolytes for Li-ion batteries, polymer light-emitting diodes, and photovoltaic materials. Their chemical and morphological spaces are immensely vast and complex,²⁵ leading to fundamental obstacles in polymer discovery. Some recent successes in rationally designing polymer dielectrics via experiment–computation synergies^{4,5,13,17,26–33} indicate that there may be opportunities for machine learning and informatics approaches in this challenging research and development area.

The biggest hurdle of the machine learning approach to polymer discovery is of both scientific and nonscientific in nature. The properties of a polymer are strongly dependent on distinctive factors such as branching, molecular weight distribution, copolymerization, additives, and processing conditions. These factors, along with issues such as nonstandard naming

Received: March 27, 2018

Revised: June 27, 2018

Published: July 13, 2018

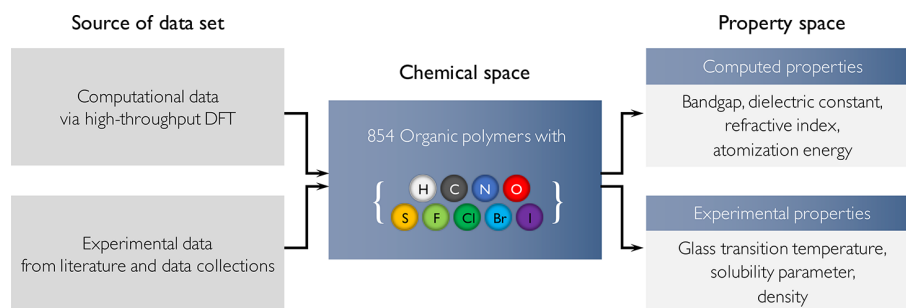


Figure 1. Overview of our polymer data set used for development of property prediction models.

conventions, have made it exceedingly difficult to create a universal polymer database upon which one may base a polymer informatics framework. A detailed analysis of the challenges faced in this front is presented in a recent review paper by Audus and de Pablo.³⁴

We have created an informatics platform capable of predicting a variety of important polymer properties on-demand. This platform utilizes surrogate (or machine learning) models, which link key features of polymers to properties, trained on high-throughput DFT calculations and experimental data from literature and existing databases. This data set of 854 polymers and the properties considered thus far are summarized in Figure 1. Certain properties, like the atomization energy, depend mainly on the atomic constituents and short-range bonding, whereas other properties, such as the glass transition temperature (T_g), are strongly influenced by morphological characteristics like the chain-stiffness and branching. We have constructed a hierarchical and automated fingerprinting scheme to identify the most important set of features to accurately describe a particular polymer property. The features span multiple length scales and range from 3-atom long fragments to descriptors such as the ratio of side-chain and main-chain atoms. The description of polymers in terms of these fundamental chemical and morphological “building blocks” is what inspired the coinage of the term “Polymer Genome”. Machine learning algorithms, specifically those based on Gaussian process regression (GPR),³⁵ were used to generate predictive models to correlate the polymer’s “genome” to its associated properties. The property prediction models have been implemented in an online platform (www.polymergenome.org), to guide polymer choices for further investigation via synthesis.

This article is organized as follows. In section 2, we describe the curation of high-throughput computational data for polymers built using a set of predefined chemical “blocks” and experimental data, which were obtained from existing databases.^{36,37} In section 3, we describe the hierarchy of descriptors used to fingerprint the polymers. In section 4, various aspects of prediction model development, e.g., the fingerprint dimensionality reduction schemes and machine learning algorithms, are described in detail. The predictive accuracy of the model, using the aforementioned hierarchical fingerprinting scheme, is demonstrated for the particular case of T_g . The performance of the final models for all the properties are summarized in section 5. In section 6, we provide an overview of our online polymer property prediction platform. Details of how this platform may be used (including how polymers may be queried using a customized SMILES string language) are provided separately in the Appendix.

2. DATA SET

Two strategic tracks were followed for the creation of our data set (see Figure 1): (1) via high-throughput computation using density functional theory (DFT) as presented earlier,^{26,38,39} and (2) by utilizing experimentally measured properties from literature and data collections.^{36,37} The overall data set includes 854 polymers made up of a subset of the following species: H, C, N, O, S, F, Cl, Br, and I. Seven different properties were included in the present study. The bandgap, dielectric constant, refractive index and atomization energy were determined using DFT computations, and T_g , solubility parameter and density were obtained from measurements.

All the computational data was generated through a series of studies related to advanced polymer dielectrics.^{26,38,39} The computational data set includes polymers containing the following building blocks, CH_2 , CO , CS , NH , C_6H_4 , $\text{C}_4\text{H}_2\text{S}$, CF_2 , CHF , and O .^{13,16,38–40} Repeat units contained 4–8 building blocks, and 3D structure prediction algorithms were used to determine their structure.^{26,38,39} The building blocks considered in the data set are found in common polymeric materials including polyethylene (PE), polyesters and polyureas, and could theoretically produce an enormous variety of different polymers. The bandgap was computed using the hybrid Heyd–Scuseria–Ernzerhof (HSE06) electronic exchange–correlation functional.⁴¹ Dielectric constant and refractive index (the square root of the electronic part of the dielectric constant) were computed using density functional perturbation theory (DFPT).⁴² The atomization energy was computed for all the polymers following previous work.^{28–31,39,40,43–48}

The T_g , solubility parameter and density data was obtained from existing databases of experimental measurements.^{36,37} T_g , which is an indication of the transition point between the glassy and supercooled liquid phases in an amorphous polymer, is important in many polymer applications because the structural characteristics (and, consequently, other properties) of the polymer changes dramatically at this point. The solubility parameter of a polymer is typically used to determine a suitable solvent to use during polymer synthesis. In this particular study we consider the Hildebrand solubility parameter, which is also useful to make quantitative estimation of polymer–solvent interaction.^{49–51}

We have determined the chemical formula and the associated topological structure from the name of polymers listed in the literature. The data set contains a total of 854 organic polymers composed of nine frequently found atomic species; i.e., C, H, O, N, S, F, Cl, Br, and I with properties listed in the right side panel of Figure 1. Figure 2 shows a summary of the property space for the polymer data set, including the range of property values, distribution, standard

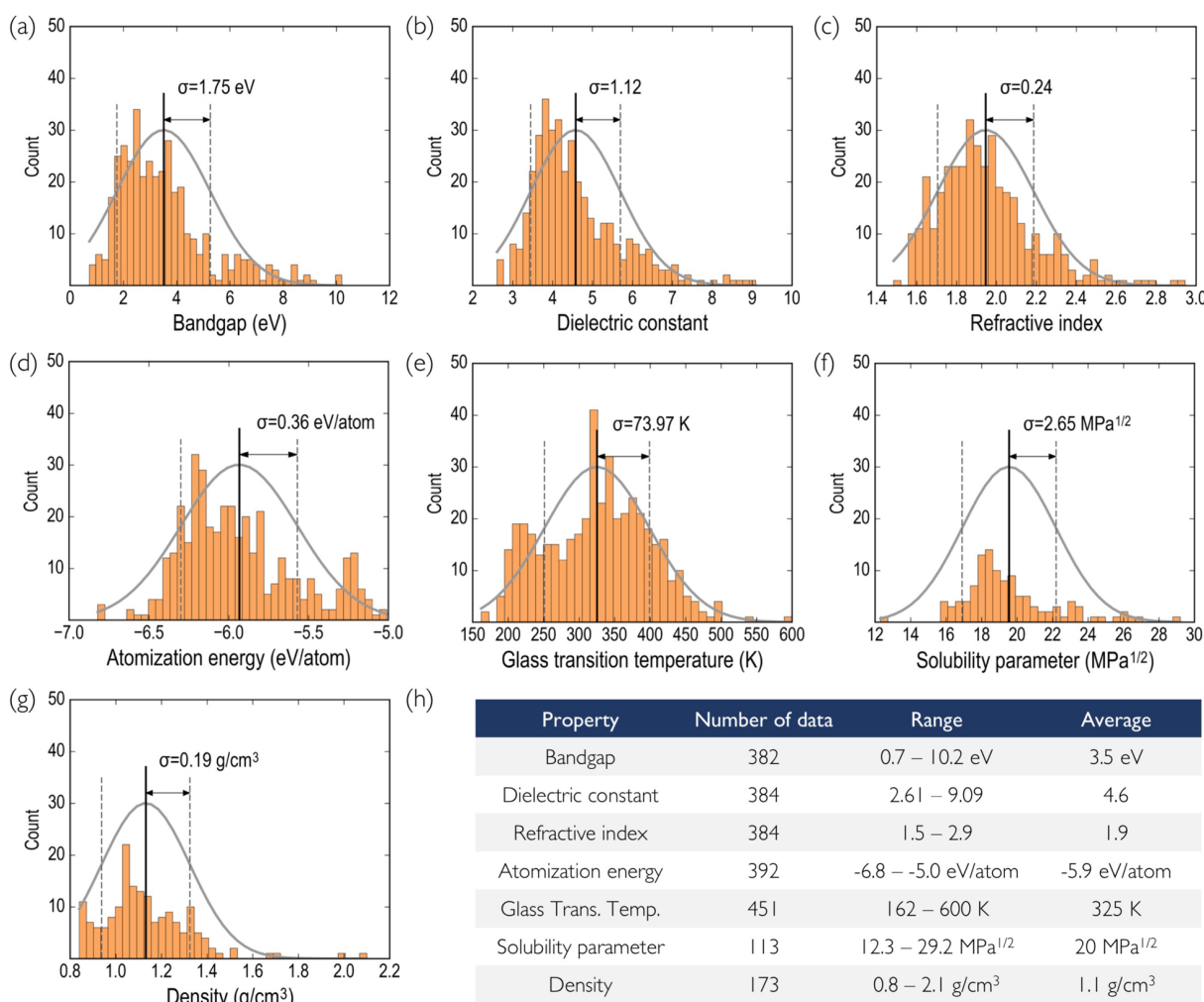


Figure 2. Property space of Polymer Genome data set. The seven properties considered in this study were the bandgap, dielectric constant, refractive index, atomization energy, T_g , solubility parameter, and density.

deviation and the number of polymers associated with each property.

3. HIERARCHICAL FINGERPRINTING

Fingerprinting is a crucial step of our data-driven machine learning approach. In this step, the geometric and chemical information on the polymers is converted to a numerical representation. To comprehensively capture the key features that may control the diversity of properties of interest, we consider three hierarchical levels of descriptors spanning different length scales. At the atomic-scale, the occurrence of a fixed set of atomic fragments (or motifs) are tracked.⁵² An example of such a fragment is O1–C3–C4, made up of three contiguous atoms, namely, a one-fold coordinated oxygen, a 3-fold coordinated carbon, and a 4-fold coordinated carbon, in this order. Such a series of predefined “triplets” has been shown to be a good fingerprint for a diverse range of organic materials.^{17,52} A vector of such triplets form the fingerprint components at the lowest hierarchy. For the polymer classes under study, there are 108 such components.

Next in the hierarchy of fingerprint components are larger length-scale descriptors of the quantitative structure–property relationship (QSPR) type, often used in chemical and biological sciences, and implemented in the RDKit Python library.^{53–55}

Examples of such descriptors are van der Waals surface area,⁵⁶ the topological polar surface area (TPSA),^{57,58} the fraction of atoms that are part of rings (i.e., the number of atoms associated with rings divided by the total number of atoms in the formula unit), and the fraction of rotatable bonds. TPSA is the sum of surfaces of polar atoms in the molecule and we observed this descriptor to be strongly correlated to the solubility. Descriptors such as the fraction of ring atoms and fraction of rotatable bonds strongly influenced properties such as T_g and density. Such descriptors, 99 in total, form the next set of components of our overall fingerprint.

The highest length-scale fingerprint components we considered may be classified as “morphological descriptors”. These include features such as the shortest topological distance between rings, fraction of atoms that are part of side-chains, and the length of the largest side-chain. Properties such as T_g strongly depend on such features which influence the way the chains are packed in the polymer. For instance, if two rings are very close, the stiffness of the polymer backbone is much higher than if the rings were separated by a larger topological distance. Both the number and the length of the side-chains strongly influence the amount of free volume in the polymeric material and therefore directly influence T_g . The larger the

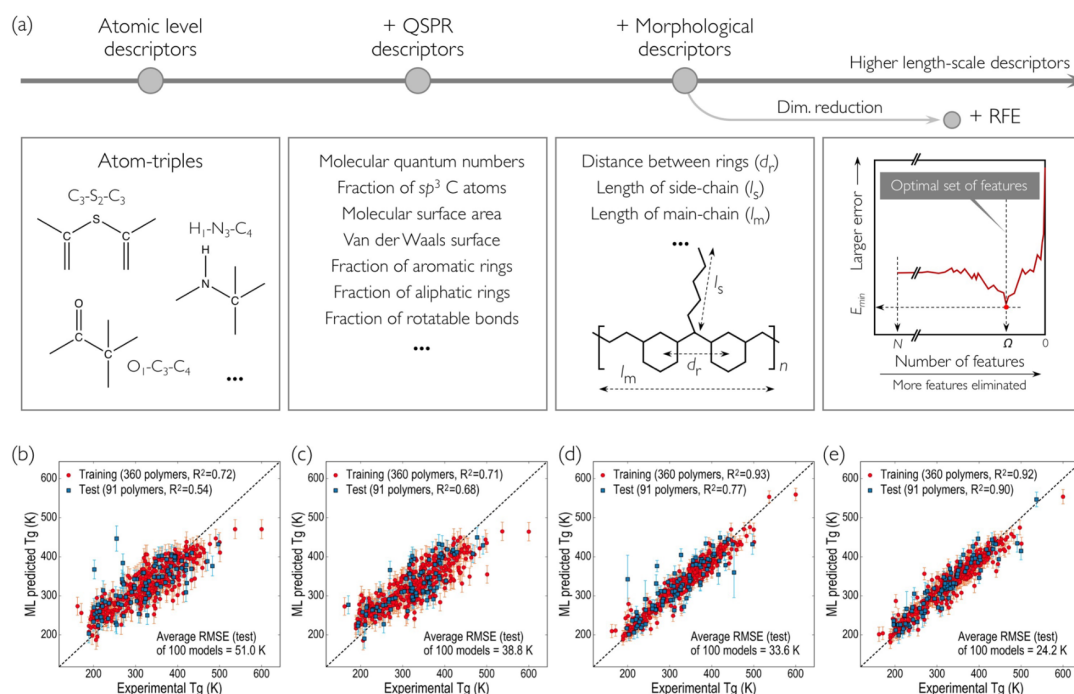


Figure 3. Hierarchy of descriptors used to fingerprint the polymers, and an example demonstration for the systematic improvement of model performance depending on the type of fingerprint considered. (a) Classification of descriptors according to the physical scale and chemical characteristics are shown with representative examples. Dimension of the fingerprint in each level can be reduced by a recursive feature elimination (RFE) process. In the “+RFE” panel, N , Ω , and E_{\min} are total number of features in fingerprint, optimal number of features determined by RFE, and minimum error of prediction model, respectively. Plots at the bottom panel show the performance of machine learning prediction models for glass transition temperature (T_g) with (b) only atomic level descriptors, (c) atomic level and QSPR descriptors, and (d) entire fingerprint components including morphological descriptors. (e) How the optimal subset selected by RFE improves the prediction model for T_g .

free-volume, the lower the T_g . We include 22 such morphological descriptors in our overall fingerprint.

Figure 3a shows the hierarchy of polymer fingerprint, including atomic level, QSPR and morphological descriptors. The overall fingerprint of a polymer is constructed by concatenating the three classes of fingerprint components. In total, this leads to a fingerprint with 229 components. Subsequently, we show that the dimensionality of the fingerprints needs to be reduced to improve prediction performance. Also, during performance assessment, we use different combinations of fingerprint components. For clarity of that discussion, we introduce some nomenclature. The atom triples fingerprint, QSPR descriptors, and morphological descriptors are, respectively, denoted by “A”, “Q”, and “M”. Therefore, “AQ” implies a combination of just the atom-triples and QSPR descriptors.

In order to visualize the chemical diversity of polymers considered here, we have performed principal component analysis (PCA) of the complete fingerprint vector. PCA identifies orthogonal linear combinations of the original fingerprint components that provide the highest variance; the first few principal components account for much of the variability in the data.⁷ Figure 4 displays the data set with the horizontal and vertical axes chosen as the first two principal components, PC_1 and PC_2 . Molecular models of some common polymers are shown explicitly, and symbol color, symbol size and symbol type are used to represent the fraction of sp^3 -bonded C atoms, fraction of rings, and TPSA of polymers, respectively. As an example from the figure, PE is composed of only sp^3 -bonded C without any rings in the chain, while poly(1,4-phenylene sulfide) contains no sp^3 -bonded C atoms, and more than 90% of its atoms are part of rings. As a

result, these two polymers are situated far from each other in 2D principal component space.

4. SURROGATE (MACHINE LEARNING) MODEL DEVELOPMENT

4.1. Recursive Feature Elimination. As alluded to earlier, our general fingerprint is rather high in dimensionality, and not all of the components may be relevant for describing a particular property. In fact, irrelevant features often lead to a poor prediction capability. On the practical side, large fingerprint dimensionality also implies longer training times. There is thus a need to determine the optimal subset of the complete fingerprint necessary for the prediction of a particular property (i.e., different properties may require different subsets of the fingerprint vector). Rather than manually deciding which fingerprint components to use, one may utilize a wide variety of dimensionality reduction techniques to automatically select a set of features that best represent a particular property. In the current work, we utilize the recursive feature elimination (RFE) algorithm to sequentially eliminate the least important features for a given property.⁶⁰ The RFE is an iterative procedure for reducing the number of features by recursively repeating the estimation of feature ranking (importance) and elimination of the least important feature. The rightmost panel of Figure 3a demonstrates how the optimal set of features were determined as the best fingerprint components by RFE. A simple linear model was used to rapidly remove unwanted features and the final set of features is passed forward to the nonlinear machine learning algorithm described next in section 4.2. The final set of features selected by RFE can also be used to obtain an

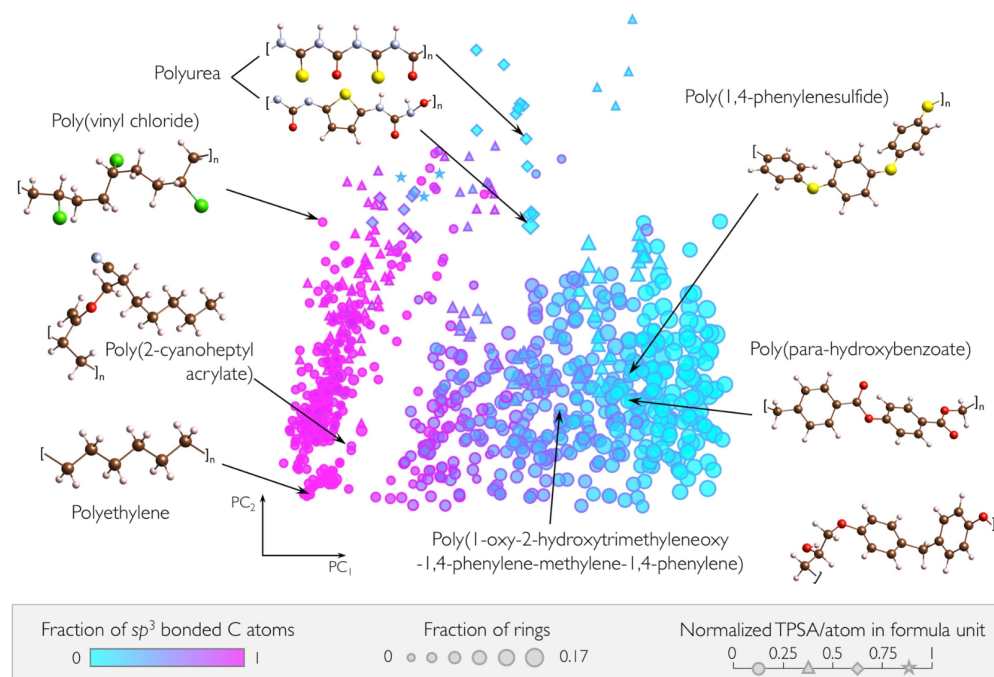


Figure 4. Graphical summary of chemical space of polymers considered. 854 chemically unique organic polymers generated by structure prediction method (minima-hopping⁵⁹) and experimental sources^{36,37} distributed in 2D principal component space. Two leading components, PC₁ and PC₂, are produced by principal component analysis, and assigned to axes of the plot. Fraction of sp³-bonded C atoms, fraction of rings, and normalized TPSA per atoms in a formula unit are used for color code, size, and symbol of each polymer. A few representative structures with various number of aromatic and/or aliphatic rings and their position on the map are shown.

intuitive understanding of how certain key fingerprint components influence particular materials properties.

4.2. Gaussian Process Regression. In our past work,^{6,13,26} we have successfully utilized kernel ridge regression (KRR)⁶¹ to learn the nonlinear relationship between a polymer's fingerprint and its properties. However, in this work we utilize GPR because of two key benefits. First, GPR learns a generative, probabilistic model of the target property and thus provides meaningful uncertainties/confidence intervals for the prediction. Second, the optimization of the model hyperparameters is relatively faster in GPR because one may perform gradient-ascent on the marginal likelihood function as opposed to the cross-validated grid-search which is required for KRR. We use a radial basis function (RBF) kernel defined as

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \exp \left[\frac{-(\mathbf{x}_i - \mathbf{x}_j)^2}{2l^2} \right] + \sigma_n^2 \delta(\mathbf{x}_i, \mathbf{x}_j) \quad (1)$$

where σ , l , and σ_n are hyperparameters to be determined during the training process (in the machine learning parlance, these hyperparameters are referred to as signal variance, length scale parameter and noise level parameter, respectively). \mathbf{x}_i and \mathbf{x}_j are the fingerprint vectors for two polymers i and j . \mathbf{x}_i is an m dimensional vector with components $x_i^1, x_i^2, x_i^3, \dots, x_i^m$, determined and optimized by the RFE step described above. Performance of the model was evaluated based on the root-mean-square error (RMSE) and the coefficient of determination (R^2). During the surrogate model development step using GPR, including substeps for selection of the best combination of fingerprint types, and optimization of dimensionality of the fingerprint by recursive feature elimination (RFE), 80% of the data was used for training and the remaining 20%

was set aside as a test set. Learning curves for machine learning models for each property are shown in Figure S1.

4.3. Optimization of Fingerprint Vector. Certain properties, like the DFT-computed atomization energy, depend mainly on the connectivity of the atomic species and only weakly on the morphology. As a result, we were able to achieve a test-RMSE of 0.01 eV using just “AQ” components. However, properties such as T_g are dependent not only on the atomic species and bonding but they are also strongly influenced by the morphology of the polymer. In Figure 3b, we see how using just the atom-triplet fingerprint components results in a poor model for the T_g with a significantly high RMSE of 51 K. However, adding the QSPR descriptors results in a visible improvement of the model performance, resulting in an RMSE of 39 K (Figure 3c). Two examples of QSPR descriptors which are highly correlated with T_g are the fraction of rotatable bonds (Pearson correlation -0.66) and the fraction of ring atoms (Pearson correlation $+0.63$). As the fraction of rotatable bonds is increased, the polymer chains become more flexible, thus resulting in a decrease in T_g . The fraction of rings is positively correlated with T_g due to the fact that rings increase the stiffness of the polymer chain thus reducing interchain mobility. The addition of the morphological fingerprints like the number of side-chains and the shortest topological distance between rings further improves the predictive capability of the model (Figure 3d). However, after including all hierarchical levels of the fingerprint, the dimensionality of the fingerprint vector becomes unnecessarily large (229). Subjecting these combined set of fingerprints to RFE brings down the dimensionality to 69. Through this systematic process of fingerprint development the final test-RMSE for T_g is brought down to 24.2 K (Figure 3e).

Similarly, other experimental properties like the solubility parameter and density showed a strong dependence on the “Q”

Table 1. Summary of Fingerprint Used for Development of Machine Learning Prediction Model and the Performance of Prediction for Each Property^a

property	best fingerprint	dimension of fingerprint	RMSE
bandgap	AQM + RFE	88	0.30 eV
dielectric constant	AQ + RFE	35	0.48
refractive index	AQM + RFE	19	0.08
atomization energy	AQ	207	0.01 eV/atom
glass transition temperature	AQM + RFE	69	18 K
solubility param	AQM + RFE	24	0.56 MPa ^{1/2}
density	AQ + RFE	9	0.05 g/cm ³

^aThe best fingerprint is selected based on average RMSE of test-set for 100 models (A, atomic level descriptors; Q, QSPR descriptors; M, morphological descriptors; + RFE, subject to the RFE process).

or “M” type fingerprints. For some cases, the feature elimination process reduces the number of fingerprints to no more than a dozen or so. For instance, in the case of the refractive index, 19 fingerprint components are sufficient to obtain a good model.

5. MODEL PERFORMANCE VALIDATION

The final machine learning models for each of the properties under consideration here were constructed using the entire polymer data set for each property. To avoid overfitting the data, and to ensure that the models are generalizable, we employed 5-fold cross-validation, wherein the data set is divided into five different subsets and one subset was used for testing while remaining sets were employed for training. Table 1 summarizes the best fingerprint, dimension of fingerprint vector, and performance based on RMSE for the entire data set.

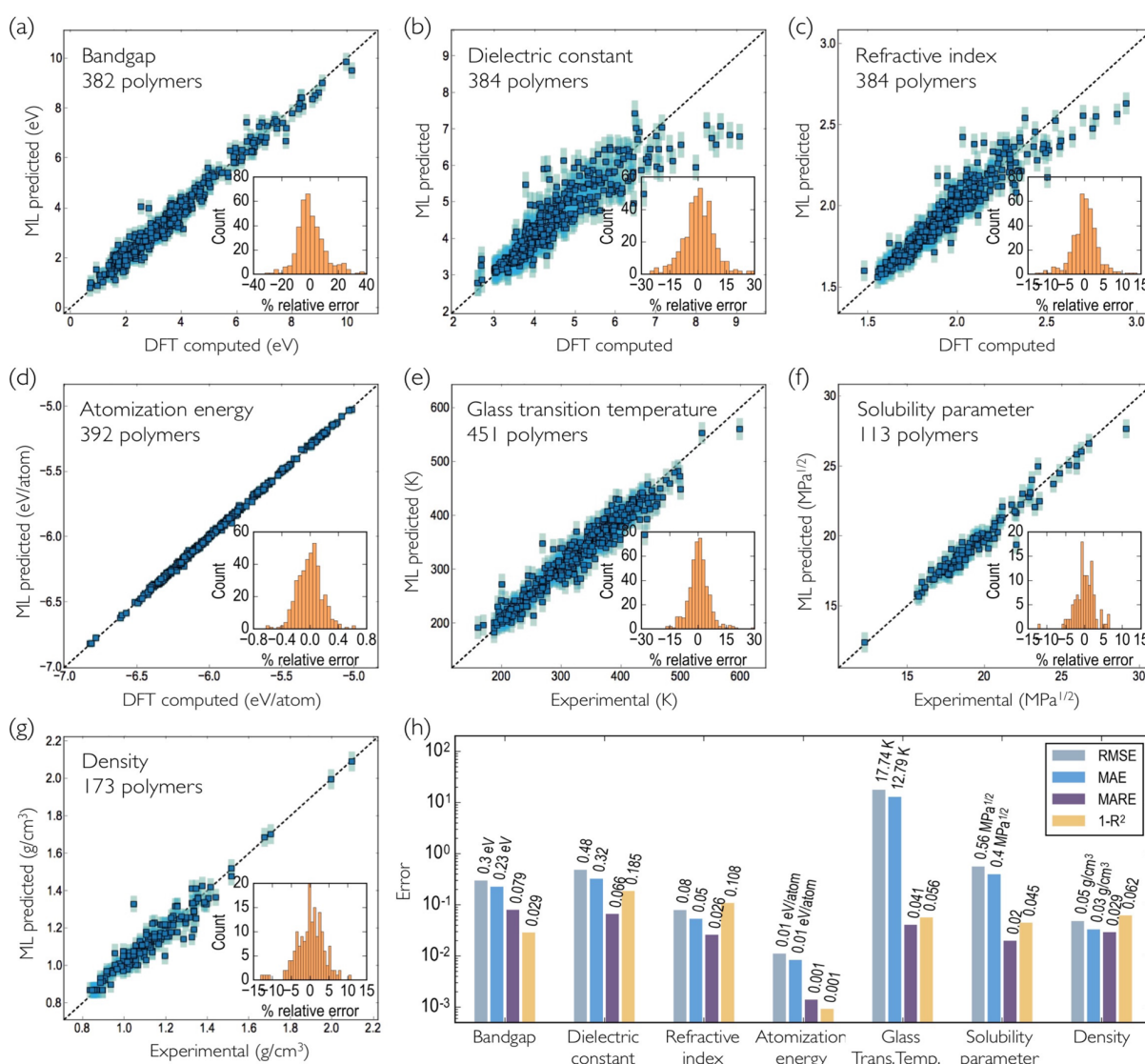
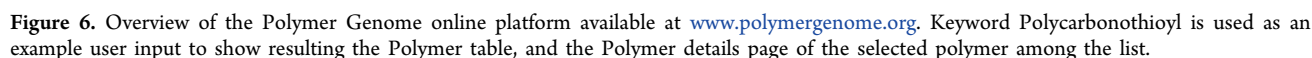


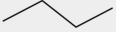
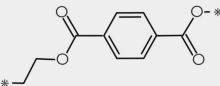
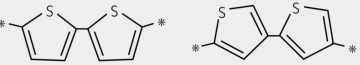
Figure 5. Performance of the cross-validated machine learning models developed by GPR with combination of RBF and white noise kernels. Comparison of DFT computed (a) bandgap, (b) dielectric constant, (c) refractive index, (d) atomization energy, experimental (e) T_g , (f) Hildebrand solubility parameter, and (g) density for the predicted values are shown with inset of distribution of % relative error, $(y - Y)/Y \times 100$ where Y is DFT computed or experimental value, and y is the machine learning predicted value. Other error metrics including RMSE, mean absolute error (MAE) and mean absolute relative error (MARE) and $1 - R^2$ are summarized in part h.



Once the user input is delivered to Polymer Genome by the user, property predictions (with uncertainty) are made, and the results are shown in an organized table. The names of polymers (if there are more than one meeting the search criteria) with SMILES and repeat unit are provided with customizable collection of properties. Upon selection of any polymer from this list, comprehensive information is reported. This one-page report provides the name and class of the polymer, 3D visualization of the structure with atomic coordinates (if such is available), and properties determined using our machine learning models. A typical user output of Polymer Genome is captured in Figure 6.

Going forward, the process of inverse design, or the autonomous suggestion of materials candidates with user-requested properties, would be an invaluable addition to any materials informatics platform and preliminary progress on this front has been reported in an earlier work.¹³ Systematic pathways to achieve such expansions are presently being examined to

Table 2. Example Polymers Specified by Multiple Input Types

Input type	Example		
Polymer name	Polyethylene, PE, C ₂ H ₄	Polyethylene terephthalate, PETE, C ₁₀ H ₈ O ₄	Polythiophene, C ₄ H ₂ S
Repeat unit	CH ₂ , CH ₂ -CH ₂ , ...	CH ₂ -CH ₂ -O-CO-C ₆ H ₄ -CO-O	C ₄ H ₂ S, C ₄ H ₂ S-C ₄ H ₂ S, ...
SMILES	CC	$\text{CCOC(=O)C(C=C_1)=CC=C_1C(=O)O}$	C(S_1)=CC=C_1 $[*]\text{C(S_1)=CC([*])=C_1}$
Sketch			

extend the applicability of the polymer informatics paradigm to a wide range of technological domains.

8. CONCLUSIONS

The Materials Genome Initiative and similar other initiatives around the world have provided the impetus for data-centric informatics approaches in several subfields of materials research. Such informatics approaches seek to provide tools and pathways for accelerated property prediction (and materials design) via surrogate models built on reliable past data. Here, we have presented a polymer informatics platform capable of predicting a variety of important polymer properties on-demand. This platform utilizes surrogate (or machine learning) models that link key features of polymers (i.e., their “fingerprint”) to properties. The models are trained on high-throughput DFT calculations (of the bandgap, dielectric constant, refractive index and atomization energy) and experimental data from polymer data handbooks (on the glass transition temperature, solubility parameter and density). Certain properties, like the atomization energy, depend mainly on the atomic constituents and short-range bonding, whereas other properties, such as the glass transition temperature, are strongly influenced by morphological characteristics like the chain-stiffness and branching. Our polymer fingerprinting scheme is thus necessarily hierarchical and captures features at multiple length scales ranging from atomic connectivity to the size and density of side chains. The property prediction models are incorporated in a user-friendly online platform named Polymer Genome (www.polymergenome.org), which utilizes a custom Python-based machine learning and polymer querying framework.

■ APPENDIX

User Input Interface

The Polymer Genome online platform accepts multiple types of user inputs. These user inputs are converted to fingerprints which are in turn used by the surrogate models to obtain property predictions. The first type of user input is the name of the polymer. We have attempted to preserve compatibility with many different naming conventions such as the formula unit, common name, structure based name, and other commonly accepted abbreviations for each polymer. For instance, polyethylene, can be queried by submitting any one of the following names: PE, C₂H₄, and polyethylene. It is also convenient to represent polymers through a string of concatenated building blocks which make up its repeat-unit. For example, PE and polyethylene terephthalate (PETE) can be written as CH₂-CH₂ and C₆H₄-CO-O-CH₂-CH₂-O-CO, respectively. Here we use the symbol “-” to imply the connection of neighboring blocks. The current version of the

Polymer Genome application can handle the following building blocks: CH₂, CH, O, CS, CO, NH, C₆H₄, C₄H₂S, C₅H₃N, C₄H₃N, CF₂, CF, CHF, CC₁₂, CC₁, CHCl, CBr₂, CBr, CHBr, Cl₂, Cl, and CHI (Figure S2).

There are several other schemes, proposed in earlier studies, to represent the structure of polymers, such as Wiswesser line-formula notation (WLFN),^{63,64} SMARTS,⁶⁵ MDL Molfiles,⁶⁶ etc. Among these representations, Polymer Genome is capable of handling the SMILES format, a powerful representation to describe the topological structure of molecular systems.⁶⁷ Since SMILES was originally developed for isolated molecules, we have introduced the following custom variations to extend its applicability to polymers: 1) If not specified, the first and last atoms in the main-chain of the SMILES are the linking atoms of the polymer chain. 2) Atoms other than the first and last can also be assigned as the linking atoms by adding the special symbol, [*], next to the atom symbol. For example, C(S₁)=CC=C₁ is obviously the SMILES of polythiophene with interconnection between the first and the last C atoms, while [*]C(S₁)=CC([*])=C₁ has different connectivity between neighboring rings. This difference of connectivity for these two variants of polythiophene can be seen in Table 2. Here, symbol “=” is used to indicate a double bond between adjacent atoms. Following conventional SMILES notation, triple bonds are specified by “#”. Branches/side-chains are specified using brackets (). Numbers beside the atom symbols indicate that the atoms associated with the same number are connected to each other. S₁ and C₁ from the above examples are connected ring-atoms, thus showing how the thiophene ring can be illustrated through the linear connection of atom symbols in SMILES.

The most convenient way for users to search/query any polymer is to draw the repeat-unit using the sketcher tool (structure editor) provided by Polymer Genome. This 2D topological sketch can be converted on-the-fly to the SMILES format with help of the JSME Molecule Editor toolkit⁶⁸ as implemented in the search-interface of Polymer Genome. Four types of acceptable user input are summarized with examples of PE, PETE, and two different structures of polythiophene in Table 2. The sketching utility and SMILES are especially useful for crafting polymers with complex connectivity or for those cases in which the predefined repeat-units are not sufficient to represent the desired polymers. Having the flexibility to choose from multiple user input formats, makes the Polymer Genome online platform a versatile digital assistant for polymer applications.

■ ASSOCIATED CONTENT

Supporting Information

This material is available free of charge via the Internet at <http://pubs.acs.org/>. The Supporting Information is available

free of charge on the ACS Publications website at DOI: 10.1021/acs.jpcc.8b02913.

Figure S1, learning curves constructed from the RMSE of the machine learning models, and Figure S2, building blocks implemented in Polymer Genome for constructing polymer repeat units (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: rampi.ramprasad@mse.gatech.edu (R.R.).

ORCID

Tran Doan Huan: 0000-0002-8093-9426

Rampi Ramprasad: 0000-0003-4630-1565

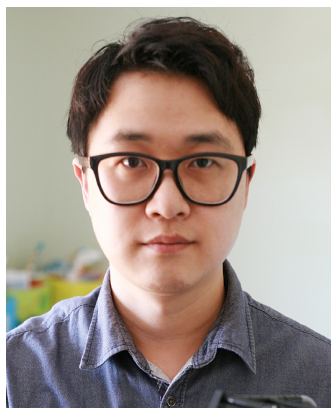
Author Contributions

[§]C.K. and A.C. contributed equally to this work. All authors participated in the writing of the manuscript.

Notes

The authors declare no competing financial interest.

Biographies



Chiho Kim is a research engineer in the School of Materials Science and Engineering, Georgia Institute of Technology. He obtained a Ph.D. from Hanyang University, South Korea, in 2009, and worked for Samsung Electronics until 2014. From 2014 to 2018, he was a postdoctoral fellow and research associate at the University of Connecticut. His current research interests are in applying machine learning techniques to a variety of material classes to accelerate the design and discovery of advanced materials.



Anand Chandrasekaran is a postdoctoral fellow in the School of Materials Science and Engineering, Georgia Institute of Technology. He received his Ph.D. from EPFL, Switzerland, in 2016 for his thesis on the ab initio and experimental study of defects and domain walls in ferroelectric oxides. From 2016 to 2018, he was a postdoc at the

University of Connecticut. His current interests include applying state-of-the-art machine learning techniques to a wide range of materials science challenges.



Tran Doan Huan received his Ph.D. in Physics from Florida State University in 2010. He is now a postdoc fellow at the University of Connecticut after completing a postdoc at the University of Basel (Switzerland) where he started his current interest in advanced computational approaches (electronic structure and machine learning) in materials science



Deya Das is a postdoctoral fellow in the School of Materials Science and Engineering, Georgia Institute of Technology. She received her Ph.D. degree from the Materials Research Center, Indian Institute of Science, India, for her thesis on Li storage studies in graphitic materials through tuning of electronic structures. Her current work is focused towards understanding of Li diffusion in polymer electrolyte for Li ion batteries.



Professor Rampi Ramprasad is presently the Michael E. Tennenbaum Family Chair and the Georgia Research Alliance Eminent Scholar in

Energy Sustainability in the School of Materials Science and Engineering, Georgia Institute of Technology. Prior to joining Georgia Tech, he held positions at the University of Connecticut and Motorola's R&D Laboratories at Tempe, AZ. Prof. Ramprasad received his B.Tech. in Metallurgical Engineering at the Indian Institute of Technology, Madras, India, an M.S. degree in Materials Science & Engineering at Washington State University, and a Ph.D. degree also in Materials Science & Engineering at the University of Illinois, Urbana—Champaign. Prof. Ramprasad's area of expertise is in the development and application of atomistic and data-driven materials computational tools, and more broadly in the utilization of such methods for the design and discovery of new materials, especially dielectrics and catalysts. Prof. Ramprasad is a Fellow of the American Physical Society, an elected member of the Connecticut Academy of Science and Engineering, and the recipient of an Alexander von Humboldt Fellowship and a Max Planck Society Fellowship for Distinguished Scientists.

■ ACKNOWLEDGMENTS

The authors acknowledge support of this work by the Office of Naval Research (Award Numbers N00014-16-1-2580 and N00014-10-1-0944) and the Toyota Research Institute through the Accelerated Materials Design and Discovery program. Computational support was provided by the Extreme Science and Engineering Discovery Environment (XSEDE).

■ REFERENCES

- (1) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Proceedings of the 25th International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Lake Tahoe, NV, 2012; Volume 1; pp 1097–1105.
- (2) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions: Moving Beyond Fingerprints. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 595–608.
- (3) Kadurin, A.; Nikolenko, S.; Khrabrov, K.; Aliper, A.; Zhavoronkov, A. druGAN: An Advanced Generative Adversarial Autoencoder Model for De Novo Generation of New Molecules with Desired Molecular Properties in Silico. *Mol. Pharmaceutics* **2017**, *14*, 3098–3104.
- (4) Mannodi-Kanakkithodi, A.; Chandrasekaran, A.; Kim, C.; Huan, T. D.; Pilania, G.; Botu, V.; Ramprasad, R. Scoping the Polymer Genome: A Roadmap for Rational Polymer Dielectrics Design and Beyond. *Mater. Today* **2017**, DOI: 10.1016/j.mattod.2017.11.021.
- (5) Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine Learning in Materials Informatics: Recent Applications and Prospects. *npj Comput. Mater.* **2017**, *3*, 54.
- (6) Mannodi-Kanakkithodi, A.; Pilania, G.; Ramprasad, R. Critical Assessment of Regression-Based Machine Learning Methods for Polymer Dielectrics. *Comput. Mater. Sci.* **2016**, *125*, 123–135.
- (7) Mueller, T.; Kusne, A. G.; Ramprasad, R. Machine Learning in Materials Science. *Reviews in Computational Chemistry* **2016**, *29*, 186–273.
- (8) Hautier, G.; Fischer, C. C.; Jain, A.; Mueller, T.; Ceder, G. Finding Nature's Missing Ternary Oxide Compounds Using Machine Learning and Density Functional Theory. *Chem. Mater.* **2010**, *22*, 3762–3767.
- (9) Oliynyk, A. O.; Antono, E.; Sparks, T. D.; Ghadbeigi, L.; Gaultois, M. W.; Meredig, B.; Mar, A. High-Throughput Machine-Learning-Driven Synthesis of Full-Heusler Compounds. *Chem. Mater.* **2016**, *28*, 7324–7331.
- (10) Pankajakshan, P.; Sanyal, S.; de Noord, O. E.; Bhattacharya, I.; Bhattacharyya, A.; Waghmare, U. Machine Learning and Statistical Analysis for Materials Science: Stability and Transferability of Fingerprint Descriptors and Chemical Insights. *Chem. Mater.* **2017**, *29*, 4190–4201.
- (11) Kim, C.; Pilania, G.; Ramprasad, R. From Organized High-Throughput Data to Phenomenological Theory Using Machine Learning: The Example of Dielectric Breakdown. *Chem. Mater.* **2016**, *28*, 1304–1311.
- (12) Jain, A.; Shin, Y.; Persson, K. A. Computational Predictions of Energy Materials Using Density Functional Theory. *Nat. Rev. Mater.* **2016**, *1*, 15004.
- (13) Mannodi-Kanakkithodi, A.; Pilania, G.; Huan, T. D.; Lookman, T.; Ramprasad, R. Machine Learning Strategy for Accelerated Design of Polymer Dielectrics. *Sci. Rep.* **2016**, *6*, 20952.
- (14) Ghadbeigi, L.; Harada, J. K.; Lettiere, B. R.; Sparks, T. D. Performance and Resource Considerations of Li-Ion Battery Electrode Materials. *Energy Environ. Sci.* **2015**, *8*, 1640–1650.
- (15) Hattrick-Simpers, J.; Wen, C.; Lauterbach, J. The Materials Super Highway: Integrating High-Throughput Experimentation Into Mapping the Catalysis Materials Genome. *Catal. Lett.* **2015**, *145*, 290–298.
- (16) Hill, J.; Mannodi-Kanakkithodi, A.; Ramprasad, R.; Meredig, B. In *Computational Materials System Design*; Shin, D., Saal, J., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp 193–225.
- (17) Mannodi-Kanakkithodi, A.; Huan, T. D.; Ramprasad, R. Mining Materials Design Rules From Data: The Example of Polymer Dielectrics. *Chem. Mater.* **2017**, *29*, 9001–9010.
- (18) Huan, T. D.; Batra, R.; Chapman, J.; Krishnan, S.; Chen, L.; Ramprasad, R. A Universal Strategy for the Creation of Machine Learning-Based Atomistic Force Fields. *npj Comput. Mater.* **2017**, *3*, 37.
- (19) Botu, V.; Batra, R.; Chapman, J.; Ramprasad, R. Machine Learning Force Fields: Construction, Validation, and Outlook. *J. Phys. Chem. C* **2017**, *121*, 511–522.
- (20) Botu, V.; Chapman, J.; Ramprasad, R. A Study of Adatom Ripening on an Al (111) Surface with Machine Learning Force Fields. *Comput. Mater. Sci.* **2017**, *129*, 332–335.
- (21) Kim, C.; Huan, T. D.; Krishnan, S.; Ramprasad, R. A Hybrid Organic-Inorganic Perovskite Dataset. *Sci. Data* **2017**, *4*, 170057.
- (22) Materials Genome Initiative. <https://www.mgi.gov/> (accessed June 19, 2018).
- (23) The Novel Materials Discovery (NOMAD) Laboratory. <https://nomad-coe.eu/> (accessed June 19, 2018).
- (24) Tolle, K. M.; Tansley, D. S. W.; Hey, A. J. G. The Fourth Paradigm: Data-Intensive Scientific Discovery Point of View. *Proc. IEEE* **2011**, *99*, 1334–1337.
- (25) Chen, L.; Huan, T. D.; Ramprasad, R. Electronic Structure of Polyethylene: Role of Chemical, Morphological and Interfacial Complexity. *Sci. Rep.* **2017**, *7*, 6128.
- (26) Mannodi-Kanakkithodi, A.; Treich, G. M.; Huan, T. D.; Ma, R.; Tefferi, M.; Cao, Y.; Sotzing, G. A.; Ramprasad, R. Rational Co-Design of Polymer Dielectrics for Energy Storage. *Adv. Mater.* **2016**, *28*, 6277–6291.
- (27) Treich, G. M.; Tefferi, M.; Nasreen, S.; Mannodi-Kanakkithodi, A.; Li, Z.; Ramprasad, R.; Sotzing, G. A.; Cao, Y. A Rational Co-Design Approach to the Creation of New Dielectric Polymers with High Energy Density. *IEEE Trans. Dielectr. Electr. Insul.* **2017**, *24*, 732–743.
- (28) Baldwin, A. F.; Huan, T. D.; Ma, R.; Mannodi-Kanakkithodi, A.; Tefferi, M.; Katz, N.; Cao, Y.; Ramprasad, R.; Sotzing, G. A. Rational Design of Organotin Polyesters. *Macromolecules* **2015**, *48*, 2422.
- (29) Zhu, Q.; Sharma, V.; Oganov, A. R.; Ramprasad, R. Predicting Polymeric Crystal Structures by Evolutionary Algorithms. *J. Chem. Phys.* **2014**, *141*, 154102.
- (30) Lorenzini, R.; Kline, W.; Wang, C.; Ramprasad, R.; Sotzing, G. The Rational Design of Polyurea & Polyurethane Dielectric Materials. *Polymer* **2013**, *54*, 3529.
- (31) Baldwin, A. F.; Ma, R.; Huan, T. D.; Cao, Y.; Ramprasad, R.; Sotzing, G. A. Effect of Incorporating Aromatic and Chiral Groups on the Dielectric Properties of Poly(dimethyltin Esters). *Macromol. Rapid Commun.* **2014**, *35*, 2082.

- (32) Mannodi-Kanakthodi, A.; Pilania, G.; Ramprasad, R.; Lookman, T.; Gubernatis, J. E. Multi-Objective Optimization Techniques to Design the Pareto Front of Organic Dielectric Polymers. *Comput. Mater. Sci.* **2016**, *125*, 92–99.
- (33) Huan, T. D.; Boggs, S.; Teyssedre, G.; Laurent, C.; Cakmak, M.; Kumar, S.; Ramprasad, R. Advanced Polymeric Dielectrics for High Energy Density Applications. *Prog. Mater. Sci.* **2016**, *83*, 236–269.
- (34) Audus, D. J.; de Pablo, J. J. Polymer Informatics: Opportunities and Challenges. *ACS Macro Lett.* **2017**, *6*, 1078–1082.
- (35) Williams, C. K. I.; Rasmussen, C. E. Gaussian Processes for Regression. *Advances in Neural Information Processing Systems 8*. MIT Press: Cambridge, MA, 1996; pp 514–520.
- (36) Bicerano, J. *Prediction of Polymer Properties*; Marcel Dekker, Inc.: New York, 2002.
- (37) Barton, A. F. M. *Handbook of Solubility Parameters and Other Cohesion Parameters*; CRC Press, Inc.: Boca Raton, FL, 1983.
- (38) Huan, T. D.; Mannodi-Kanakthodi, A.; Kim, C.; Sharma, V.; Pilania, G.; Ramprasad, R. A Polymer Dataset for Accelerated Property Prediction and Design. *Sci. Data* **2016**, *3*, 160012.
- (39) Sharma, V.; Wang, C. C.; Lorenzini, R. G.; Ma, R.; Zhu, Q.; Sinkovits, D. W.; Pilania, G.; Oganov, A. R.; Kumar, S.; Sotzing, G. A.; et al. Rational Design of All Organic Polymer Dielectrics. *Nat. Commun.* **2014**, *5*, 4845.
- (40) Wang, C. C.; Pilania, G.; Boggs, S. A.; Kumar, S.; Breneman, C.; Ramprasad, R. Computational Strategies for Polymer Dielectrics Design. *Polymer* **2014**, *55*, 979.
- (41) Heyd, J.; Scuseria, G. E.; Ernzerhof, M. Hybrid Functionals Based on a Screened Coulomb Potential. *J. Chem. Phys.* **2003**, *118*, 8207–8215.
- (42) Baroni, S.; de Gironcoli, S.; Dal Corso, A.; Giannozzi, P. Phonons and Related Crystal Properties From Density-Functional Perturbation Theory. *Rev. Mod. Phys.* **2001**, *73*, 515–562.
- (43) Huan, T. D.; Amsler, M.; Tuoc, V. N.; Willand, A.; Goedecker, S. Low-Energy Structures of Zinc Borohydride $\text{Zn}(\text{BH}_4)_2$. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2012**, *86*, 224110.
- (44) Sharma, H.; Sharma, V.; Huan, T. D. Exploring PtSO_4 and PdSO_4 Phases: An Evolutionary Algorithm Based Investigation. *Phys. Chem. Chem. Phys.* **2015**, *17*, 18146.
- (45) Huan, T. D.; Sharma, V.; Rossetti, G. A.; Ramprasad, R. Pathways Towards Ferroelectricity in Hafnia. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *90*, 064111.
- (46) Huan, T. D.; Amsler, M.; Sabatini, R.; Tuoc, V. N.; Le, N. B.; Woods, L. M.; Marzari, N.; Goedecker, S. Thermodynamic Stability of Alkali Metal/Zinc Double-Cation Borohydrides at Low Temperatures. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *88*, 024108.
- (47) Baldwin, A. F.; Ma, R.; Mannodi-Kanakthodi, A.; Huan, T. D.; Wang, C.; Marszalek, J. E.; Cakmak, M.; Cao, Y.; Ramprasad, R.; Sotzing, G. A.; et al. Poly(dimethyltin Glutarate) as a Prospective Material for High Dielectric Applications. *Adv. Mater.* **2015**, *27*, 346.
- (48) Ma, R.; Sharma, V.; Baldwin, A. F.; Tefferi, M.; Offenbach, I.; Cakmak, M.; Weiss, R.; Cao, Y.; Ramprasad, R.; Sotzing, G. A. Rational Design and Synthesis of Polythioureas as Capacitor Dielectrics. *J. Mater. Chem. A* **2015**, *3*, 14845–14852.
- (49) Sperling, L. *Introduction to Physical Polymer Science*; Wiley: 2005.
- (50) Flory, P. J. Thermodynamics of High Polymer Solutions. *J. Chem. Phys.* **1942**, *10*, 51–61.
- (51) Huggins, M. L. Solutions of Long Chain Compounds. *J. Chem. Phys.* **1941**, *9*, 440–440.
- (52) Huan, T. D.; Mannodi-Kanakthodi, A.; Ramprasad, R. Accelerated Materials Property Predictions and Design Using Motif-Based Fingerprints. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2015**, *92*, 14106.
- (53) Nantasenamat, C.; Isarankura-Na-Ayudhya, C.; Naenna, T.; Prachayasittikul, V. A Practical Overview of Quantitative Structure-Activity Relationship. *EXCLI J.* **2009**, *8*, 74–88.
- (54) Nantasenamat, C.; Isarankura-Na-Ayudhya, C.; Prachayasittikul, V. Advances in Computational Methods to Predict the Biological Activity of Compounds. *Expert Opin. Drug Discovery* **2010**, *5*, 633–654.
- (55) RDKit Open Source Toolkit for Cheminformatics. <http://www.rdkit.org/> (accessed June 19, 2018).
- (56) Labute, P. A Widely Applicable Set of Descriptors. *J. Mol. Graphics Modell.* **2000**, *18*, 464–477.
- (57) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43*, 3714–3717.
- (58) Prasanna, S.; Doerksen, R. Topological Polar Surface Area: A Useful Descriptor in 2D-QSAR. *Curr. Med. Chem.* **2009**, *16*, 21–41.
- (59) Sicher, M.; Mohr, S.; Goedecker, S. Efficient Moves for Global Geometry Optimization Methods and Their Application to Binary Systems. *J. Chem. Phys.* **2011**, *134*, 044106.
- (60) Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene Selection for Cancer Classification Using Support Vector Machines. *Mach. Learn.* **2002**, *46*, 389–422.
- (61) Vu, K.; Snyder, J. C.; Li, L.; Rupp, M.; Chen, B. F.; Khelif, T.; Müller, K.-R.; Burke, K. Understanding Kernel Ridge Regression: Common Behaviors From Simple Functions to Density Functionals. *Int. J. Quantum Chem.* **2015**, *115*, 1115–1128.
- (62) Polymer Genome. <http://www.polymergenome.org> (accessed June 19, 2018).
- (63) Wiswesser, W. J. *Line-Formula Chemical Notation*; Crowell: New York, 1954.
- (64) Smith, E. G. *The Wiswesser Line-Formula Chemical Notation*; McGraw-Hill: New York, 1968.
- (65) SMARTS a Language for Describing Molecular Patterns. <http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html> (accessed June 19, 2018).
- (66) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Model.* **1992**, *32*, 244–255.
- (67) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31–36.
- (68) Bienfait, B.; Ertl, P. JSME: A Free Molecule Editor in JavaScript. *J. Cheminf.* **2013**, *5*, 24.