

Machine-learning predictions of polymer properties with Polymer Genome ^F

Cite as: J. Appl. Phys. **128**, 171104 (2020); <https://doi.org/10.1063/5.0023759>

Submitted: 06 August 2020 . Accepted: 10 October 2020 . Published Online: 05 November 2020

^{id} Huan Doan Tran, ^{id} Chiho Kim, ^{id} Lihua Chen, Anand Chandrasekaran, ^{id} Rohit Batra, Shruti Venkatram, Deepak Kamal, ^{id} Jordan P. Lightstone, Rishi Gurnani, ^{id} Pranav Shetty, Manav Ramprasad, ^{id} Julia Laws, Madeline Shelton, and ^{id} Rampi Ramprasad

COLLECTIONS

^F This paper was selected as Featured



View Online



Export Citation



CrossMark

Meet the Next Generation
of Quantum Analyzers

And Join the Launch
Event on November 17th



Register now



Zurich
Instruments



Machine-learning predictions of polymer properties with Polymer Genome

Cite as: J. Appl. Phys. **128**, 171104 (2020); doi: [10.1063/5.0023759](https://doi.org/10.1063/5.0023759)

Submitted: 6 August 2020 · Accepted: 10 October 2020 ·

Published Online: 5 November 2020



Huan Doan Tran,  Chiho Kim,  Lihua Chen,  Anand Chandrasekaran, Rohit Batra,  Shruti Venkatram, Deepak Kamal, Jordan P. Lightstone,  Rishi Gurnani, Pranav Shetty,  Manav Ramprasad, Julia Laws,  Madeline Shelton, and Rampi Ramprasad^{a)} 

AFFILIATIONS

School of Materials Science and Engineering, Georgia Institute of Technology, 771 Ferst Drive NW, Atlanta, Georgia 30332, USA

^{a)}Author to whom correspondence should be addressed: rampi.ramprasad@mse.gatech.edu

ABSTRACT

Polymer Genome is a web-based machine-learning capability to perform near-instantaneous predictions of a variety of polymer properties. The prediction models are trained on (and interpolate between) an underlying database of polymers and their properties obtained from first principles computations and experimental measurements. In this contribution, we first provide an overview of some of the critical technical aspects of Polymer Genome, including polymer data curation, representation, learning algorithms, and prediction model usage. Then, we provide a series of pedagogical examples to demonstrate how Polymer Genome can be used to predict dozens of polymer properties, appropriate for a range of applications. This contribution is closed with a discussion on the remaining challenges and possible future directions.

Published under license by AIP Publishing. <https://doi.org/10.1063/5.0023759>

I. INTRODUCTION

In the last decade, interest in utilizing data-driven informatics approaches has intensified within materials science and engineering.^{1–3} One sub-domain of materials' research which appears to be ripe for informatics-based forays is polymer science and engineering.^{4–9} Polymeric materials are simple and complex at the same time. Typically composed of the smallest atoms of the periodic table, polymers can display extraordinary diversity at very small and large scales, ranging from an immense array of possibilities for atomic-level connectivity, chain packing, and morphology (the last being a catch-all expression to capture crystallinity, phase separation, porosity, and microstructure). This diversity of structure leads to a plethora of attractive properties as reflected by the ubiquity of polymers in everyday life and high-technology.^{10,11}

The vast chemo-structural space of polymer possibilities leads to enormous challenges with respect to studying them (either using experimental or computational methods), especially when one is interested in searching this space for attractive candidates for a given application.¹² Furthermore, it is also non-trivial to effectively harness the existing (and exponentially growing) knowledge base of past studies toward further developments and discoveries. Recent developments in the polymer informatics arena are attempting to fill the above gap by effectively exploiting available data (or using

intentionally created data) and advanced machine-learning (ML) algorithms.^{13,14} These methods may be used to rapidly estimate properties of new materials.^{1,8,9,15–19} Moreover, opportunities exist for inverting the property prediction pipeline to efficiently identify materials that satisfy target property or performance objectives.^{18–21}

One such development, which we call the Polymer Genome project,^{8,9} is discussed here and portrayed schematically in Fig. 1. The essential ingredients of the Polymer Genome project (or any such informatics effort) are the following. Systematic and continuous accumulation of (experimental and computational) polymer data is the necessary first ingredient. In Polymer Genome, such data are either being acquired from a variety of literature sources^{29–33} or being generated using computations in a high-throughput and consistent manner.³⁴ ML algorithms then convert these data to knowledge (and predictive models) in a step-by-step manner. The first step within the ML pipeline is converting the data into machine readable form via a fingerprinting step that encodes features of the polymer at a variety of hierarchical length scales in a numerical fingerprint vector. The next step in the ML pipeline is the learning step, during which the polymer fingerprint vectors are mapped onto the corresponding polymer property values, using one of many algorithms; this step is essentially a function finding exercise, i.e., the best hypothesis function that links the

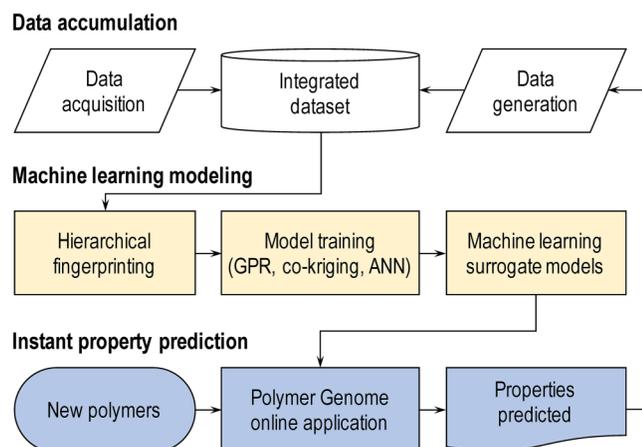


FIG. 1. An overview of the architecture of Polymer Genome.

fingerprint vectors and the property values is identified following robust statistical practices. The hypothesis function, one for each property for which data are available, constitutes a “surrogate” model for the instantaneous prediction of the corresponding polymer property.

The above workflow has been utilized to develop numerous surrogate models, which are deployed and may be used at www.polymergenome.org. Predictions for several dozen attributes of polymers may be made using this platform at the present time. A user-friendly graphical user interface (GUI) is provided to easily build and query polymers of interest. In addition to providing an ultrafast capability to estimate the properties of new polymers, these prediction models are also used to guide further data generation (e.g., based on whether a polymer has an attractive property value or if the uncertainty of the prediction is too high) either through computational or empirical work (see Fig. 1).

The goal of the present Tutorial is to introduce Polymer Genome as a practical tool for the polymer community. We will outline the critical scientific and technical aspects of Polymer Genome, including polymer data accumulation and generation, the fingerprinting scheme using which polymers can be represented numerically in a machine readable form, the learning algorithms used for developing surrogate models for the prediction of polymer properties, and the online platform for handling the interactions with end-users (see Fig. 1). Then, a set of tutorials will be provided, illustrating the applications of Polymer Genome to solve some practical problems involving property predictions and design. This Tutorial is closed by a discussion on the remaining challenges and the future development plan of Polymer Genome.

II. POLYMER GENOME PIPELINE

A. Polymer data

A comprehensive summary of the data sets, the predictive (surrogate) models, and the polymer properties supported by Polymer Genome is given in Table I. Overall, data sets corresponding to several dozen polymer attributes were utilized to build

surrogate models for dozens of properties. In fact, some polymer properties pose extra dimensions, requiring additional data sets for the model training. For example, the model that predicts if a polymer can (or cannot) be dissolved by each of 24 regular solvents (the completed list can be found in Ref. 25) was trained on 24 corresponding data sets. The capability of predicting the permeability of a polymer to six gases (CH_4 , CO_2 , He, N_2 , O_2 , and H_2) was developed from six distinct data sets.²⁷ The dielectric constant measured at nine frequencies ranging from 60 Hz to 10^{15} Hz was utilized to allow Polymer Genome to predict the frequency-dependent dielectric constant of polymers.²³ These data sets are also structurally diverse, containing both linear and ladder polymers (see Sec. II B and Fig. 3 for more information). This significant complexity introduces both challenges and opportunities for encoding the chemical structure of polymers and developing the surrogate property prediction models.

The majority of the property prediction models in Polymer Genome utilizes experimentally measured data. Within this polymer data class, the biggest entities, i.e., the solvent/non-solvent data set (6721 polymers) and the glass transition temperature data set (5076 polymers), are far bigger than those involving computational data, except the bandgap data set. While experimental data are enormously important for Polymer Genome, collecting such information from published journals, printed handbooks, and online repositories^{29–33} is challenging because of both technical and non-technical reasons, requiring laborious manual data extraction and validation.

Data from computational sources were generated^{34,35} using density functional theory (DFT) as implemented in VASP software.^{36–39} Within this high-throughput computational workflow, polymer models of increasing sophistication, i.e., polymer chains and crystals, were constructed before relevant properties can be computed. Currently, seven data sets have their origins in computations, including polymer chain and crystal bandgap (computed using the HSE06 exchange-correlation functional²²), atomization energy, ionization energy, electron affinity, static dielectric constant, and refractive index. We note that a separate refractive index model that utilizes primarily experimentally measured data are also available on Polymer Genome; the model based on computational data tends to over-estimate the refractive index as this data set corresponds to polymer crystals that tend to be denser than real polymers. The primary challenge of this workflow is that predicting polymer crystal structure is computationally intensive, specifically when established methods, e.g., minima-hopping^{40,41} and USPEX,^{42,43} are used.^{34,44,45} A new efficient method, referred to as polymer structure predictor, was recently developed, strictly enforcing pre-defined atomic connectivity and known modes of chain packing.³⁵ In the near future, this method will be used to autonomously explore the polymer space for those satisfying targeted properties⁴⁶ and systematically generating/accumulating polymer data.

The curated polymer data sets were unified in a principal data set of 13 347 polymers, nearly all of which have been experimentally synthesized and reported elsewhere. This master data set is visually portrayed in Fig. 2(a). In Figs. 2(b) and 2(c), it is shown as the background on which two representative polymer property data sets, i.e., polymer chain bandgap and glass transition temperature, are overlaid. In the future, computations will be used to maximize

TABLE I. A summary of the curated polymer data sets, the developed models, and the polymer properties supported by Polymer Genome. These properties are arranged in some categories, including “electronic properties” (rows 1–4), “response properties” (5–8), “mechanical properties” (9–10), “thermal properties” (11–13), “solubility properties” (14–15), “permeability properties” (16), “physical and thermodynamic properties” (17–20), and “other properties” (21–22). Here, GPR, CK, and ANN stand for Gaussian process regression, co-Kriging, and artificial neural network, respectively. Model performance is given in terms of either classification accuracy (for the polymer/solvent compatibility) or $RMSE_{CV}$, the averaged cross-validation (CV) test error of the CV models created when the 100%-data model is trained. References and notes are provided when available.

No.	Polymer properties	Data		ML Algo.	$RMSE_{CV}$	Notes	Reference
		Source	Size				
1	Polymer crystal bandgap	Comput.	562	GPR	0.26 eV	Training data produced using using HSE06 XC functional ²²	8
2	Polymer chain bandgap	Comput.	3881	GPR	0.24 eV	Training data produced using using HSE06 XC functional ²²	
3	Ionization energy	Comput.	371	GPR	0.21 eV		
4	Electron affinity	Comput.	371	GPR	0.18 eV		
5	Static dielectric constant (crystal)	Comput.	383	GPR	0.38		8
6	Frequency-dependent dielectric constant	Exper.	1193	GPR	0.16	Training data include measurements at 60, 10 ² , 10 ³ , 10 ⁴ , 10 ⁵ , 10 ⁶ , 10 ⁷ , 10 ⁹ , and 10 ¹⁵ Hz	23
7	Refractive index (bulk resin)	Exper.	516	GPR	0.04		24
8	Refractive index (crystal)	Comput.	383	GPR	0.07		8
9	Tensile strength	Exper.	672	GPR	4.75 MPa		
10	Young's modulus	Exper.	629	GPR	120 MPa		
11	Glass transition temperature	Exper.	5076	GPR	18.8 K		8
12	Melting temperature	Exper.	2084	GPR	27.1 K		
13	Thermal decomposition temperature	Exper.	3545	GPR	28.03 K		
14	Polymer/solvent (in) compatibility	Exper.	6721	ANN	93% accurate classification	The compatibility with 24 solvents is predicted	25
15	Solubility parameter	Exper.	112	GPR	0.47 MPa ^{1/2}		26
16	Gas permeability	Exper.	1779	GPR	1.2 Barrer	The permeability to CH ₄ , CO ₂ , He, N ₂ , O ₂ , and H ₂ is predicted	27
17	Polymer density	Exper.	890	GPR	0.03 g/cc		8
18	Atomization energy	Comput.	391	GPR	0.01 eV/atom		8
19	Specific heat	Exper.	80	GPR	0.07 J/gK		
20	Fractional free volume	Exper.	133	GPR	0.01		
21	Limiting oxygen index	Exper.	101	GPR	3.73%		
22	Tendency to crystallize	Exper.	429/107	CK	8.38%	Training data include low- and high-fidelity data	28

the coverage within the principal data set, which will also be gradually expanded.

B. Polymer fingerprints

Materials' data are generally very diverse in nature and format and, thus, not directly readable/ready for computer learning. In practice, the various cases under study must be represented numerically by a fingerprint or descriptor in the machine-learning process.^{15,16,47–50} Good fingerprints should be closely related to the nature of materials and properties, adequately capturing enough chemo-structural information of the materials, and satisfying certain requirements, e.g., being invariant with respect to transformations that do not change the materials in any physical way. A good review of materials' fingerprints can be found in Ref. 15.

The chemical structure of the repeat unit of a polymer may be represented by a string of characters called SMILES, which stands for the *simplified molecular-input line-entry system*.⁵¹ SMILES was initially defined for molecules and has been extended to polymers by explicitly specifying the connecting points of polymer repeat units.⁸ At the present time, Polymer Genome supports two main classes of polymers, i.e., linear polymers and ladder polymers, the former has two connecting points and the latter has four connecting points in each repeat unit. For an illustration of the polymer SMILES concept, Fig. 3 sketches the chain, the repeat unit, and the SMILES string of poly(isobutylene), i.e., a linear polymer, and poly(naphthalene-2,3:6,7-tetrayl-6,7-dimethylene), i.e., a ladder polymer. Generally, writing a SMILES string of a complex polymer is cumbersome, thus a detailed guideline of polymer SMILES and a GUI-based polymer draw tool are provided at www.polymergenome.org.

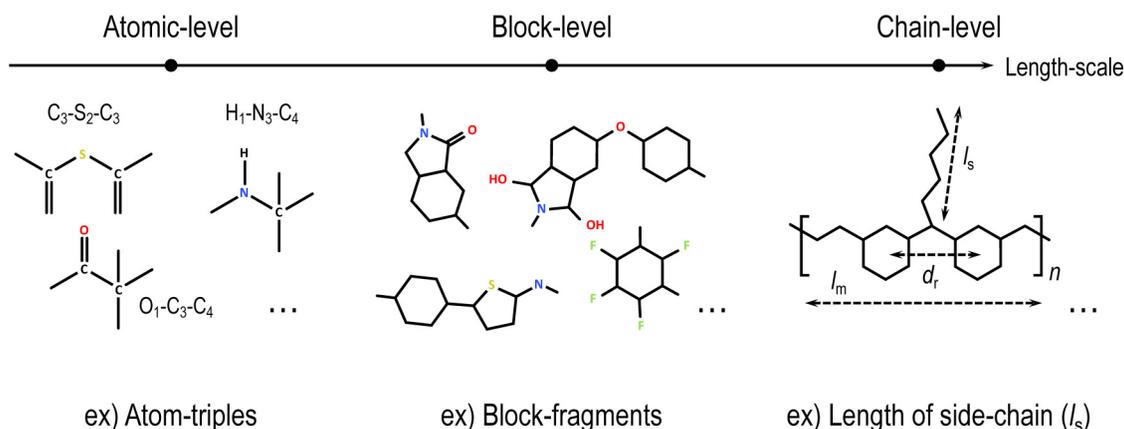


FIG. 4. Hierarchical fingerprints used to represent polymers in the Polymer Genome pipeline.

For each polymer property, the entire list of ≈ 3000 fingerprint components were down selected significantly and separately using the Recursive Feature Elimination (RFE) or the Least Absolute Shrinkage and Selection Operator (LASSO) algorithms.⁵² Then, the surviving fingerprint components were checked for possible correlations and those that are correlated are simplified. During this process, least important fingerprint components are pruned, keeping only the most relevant components in order to form the optimal fingerprint. As discussed above, the final version of the fingerprint obtained for each property contains the components that capture the most important underlying characteristic processes. These algorithms were found⁸ to be critical for eliminating information redundancy, which likely adds unnecessary noise to the polymer data and leads to the development of ML models that are faster and more accurate. However, we also note that the fingerprint component reduction could potentially reduce the generalizability of the models when they encounter the cases that do not clearly exhibit the correlations—strictly speaking, this possibility is rooted at the finiteness of the data. Therefore, for a few models whose performance was not improved significantly during the feature reduction step, we kept the original version of fingerprint in order to maintain their generalizability.

C. Machine-learning algorithms

Learning algorithms are needed next to establish mappings between polymer fingerprints and properties. Among the models supplied by Polymer Genome (see Table I), the vast majority were developed using Gaussian process regression (GPR or Kriging)^{13,14} with a radial basic function kernel. There are several reasons for the preference given to this elegant non-parametric Bayesian method. First, GPR is explicitly similarity-based and, therefore, intuitive. Second, by assuming the output is a realization of a Gaussian process, GPR provides a built-in measure of the prediction uncertainty. Finally, the current polymer data sets are not too big, thus training a GPR model and using it to make predictions is not computationally intensive.

Co-Kriging (CK)⁵³ is an information-fusion approach that is ideal when multiple sources of data (perhaps with different levels of fidelity) are available for the same property.^{54–56} CK is used in Polymer Genome to predict the tendency of a polymer to crystallize, which can be quantified based on two measures (with different levels of fidelity). These two measures allowed us to create two separate data sets that quantify the tendency of crystallization.²⁸ The first set contains 107 “high-fidelity” data points, measured either directly using methods like nuclear magnetic resonance, x-ray diffraction, and infrared spectroscopy or indirectly from the experimental data of extensive properties like heat fusion and density. In the second set, 429 “low-fidelity” data points were obtained computationally using the group contribution method.³¹ These data sets are considered as two Gaussian processes, the former is the sum of the latter (scaled by a factor) and another independent process.^{54,55} In other words, the CK formalism fuses high- and low-fidelity data sets into a model whose prediction accuracy is significantly improved.⁵⁶

An artificial neural network (ANN) consists of numerous nodes or neurons, arranged in a series of layers, starting from the input layer, going through hidden layers before ending at the (last) output layer. Each neuron receives signals from all the neurons of the prior layer (or import directly from the input if it is in the input layer), processes the data, and transmits the activated outputs to all the neurons of the next layer (or export directly as the output if it is in the output layer). This architecture can capture very well the highly non-linear hidden relationships between materials’ structures and their properties and has been widely used in materials’ research during the last decade.^{57–62} In Polymer Genome, the architecture of ANN is particularly suitable for the data structure and the learning problem of the solvent/non-solvent prediction model.²⁵ In the near future, the powerful ANN algorithm may be considered for other models.

D. Machine-learning models

The demonstrated developments of data generation and curation, polymer fingerprinting, and learning algorithms set the stage for

learning the fingerprinted data and creating ML prediction models. During this data-learning process, the models were created using five-fold cross validation.^{15,16,63} In this multi-step standard procedure in supervised learning for minimizing the risk of overfitting, the training data set is split into k bins first. Then, each of these bins is left out for testing the model that is trained on the union of the remaining $k - 1$ bins. This step involves creating k models, and the model that performs the best on the designated test set is selected.

For each of the property/performance of polymers for which data were curated, a ML model was developed and implemented. Essential information of these models, including details of the training data, the algorithm, the cross-validation root mean square errors of the models, and the available references, is summarized in Table I. Figure 5 visualizes the performance of a subset of 12 models developed and implemented in Polymer Genome, including polymer crystal and chain bandgap, frequency-dependent dielectric constant, gas permeability, specific heat, tendency to crystallize, tensile strength, Young's modulus, glass transition temperature, melting temperature, thermal decomposition temperature, and polymer density. Some essential information of these models is also given in Fig. 5.

E. Polymer Genome online platform

The Polymer Genome online platform was created and made freely accessible at www.polymergenome.org, offering end-users a convenient toolkit to access the ML models for polymer property predictions. Using a GUI, users can easily specify and query the polymers of interest. Working under this interface layer, the Polymer Genome platform obtains the polymer SMILES string, converts it into fingerprints, predicts its properties using the implemented models, and finally returns the results. The whole process will take up to a minute if not seconds. Polymer Genome platform was developed using Python and standard web programming languages such as Hypertext Preprocessor (PHP) and Hypertext Markup Language (HTML).

Polymer Genome offers various options to query a polymer, i.e., by using its name, common abbreviation, the building block representation of its repeat unit,^{64,65} its class, its SMILES string, and especially by drawing it using the implemented GUI-based polymer draw tool, which is shown in Fig. 6. Because writing a SMILES string for a complex polymer is generally not straightforward and often very time-consuming, the polymer draw tool offers a very

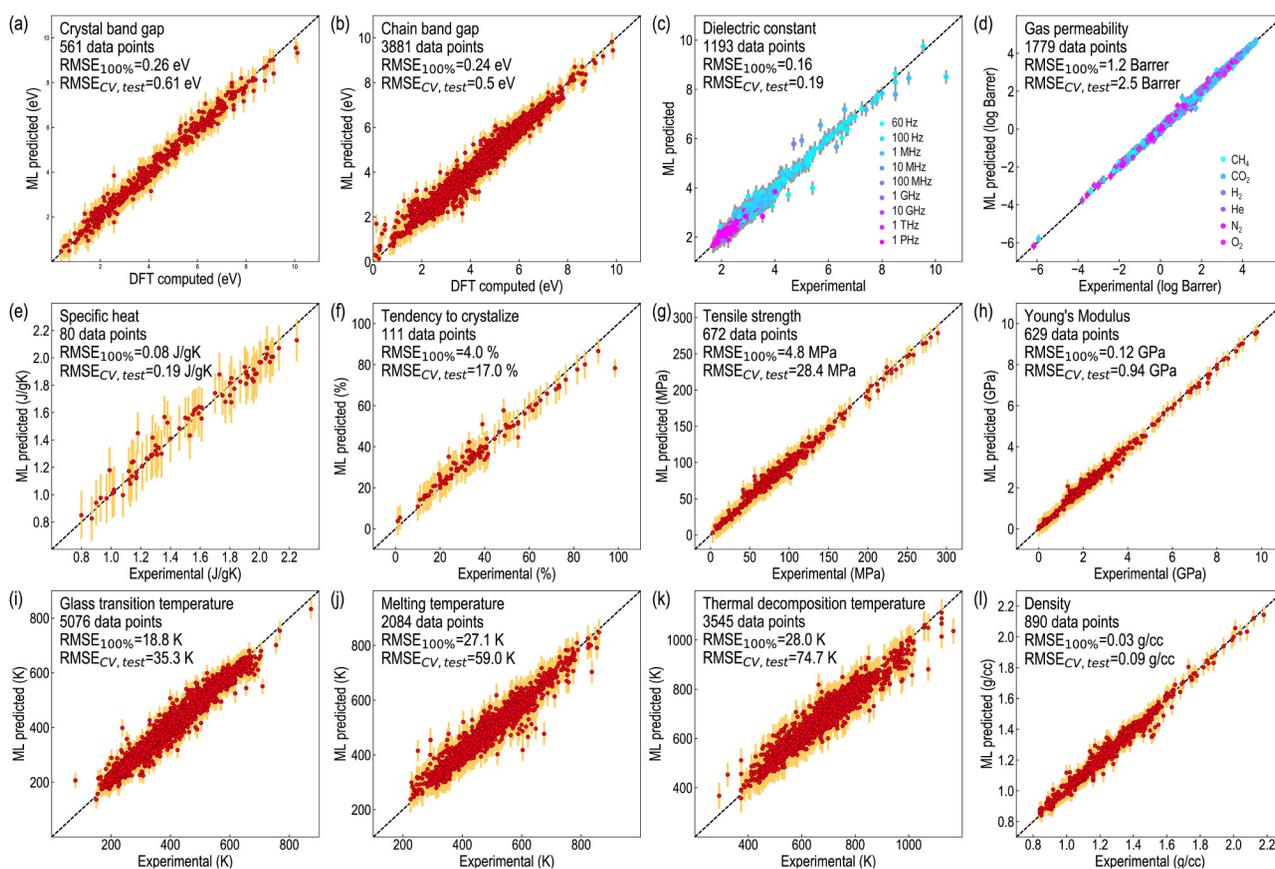


FIG. 5. Visualized performance of 12 representative (out of more than 20) surrogate models developed in the Polymer Genome project, given in sub-panels (a)–(l). Essential information of these models, in which $RMSE_{100\%}$ is the root mean square error of the model trained on 100% (the entire) of the data and $RMSE_{CV, test}$ is the average of the cross-validation test error of the models created when the 100%-data model is trained, is also given.

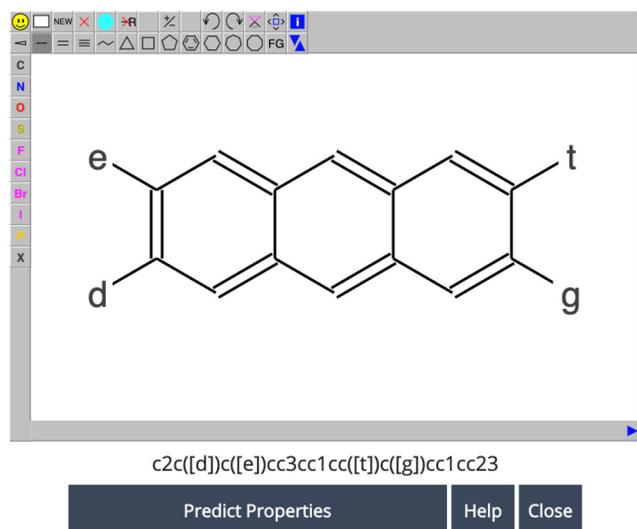


FIG. 7. An overview of Polymer Genome online platform available at www.polymergenome.org. Keyword kevlar is used as an example user input to show resulting Polymer details page.

convenient and powerful method for graphically drawing the queried polymers. This tool handles linear and ladder polymers, the latter is a specific class of cross-linked polymers, involving cross-links between pairs of polymer chains. Comprehensive user guidelines are provided at www.polymergenome.org and by some YouTube videos, which can be found by searching for “Polymer Genome.”

Accepting the query for a polymer from users, Polymer Genome returns its class, abbreviation, synonyms, and similar polymers, the 3D visualization of the repeat unit with atomic coordinates, and its predicted properties. Dozens of properties predicted are categorized into multiple groups, including electronic properties, dielectric and optical properties, mechanical properties, thermal properties, physical and thermodynamic properties, and solubility properties. An example of the search result page is given in Fig. 7.

III. TUTORIALS: POLYMER GENOME FOR POLYMER PROPERTIES PREDICTIONS

This section is devoted to a set of eight tutorial problems, designed to provide end-users systematic and pedagogical guidelines in the usage of Polymer Genome.

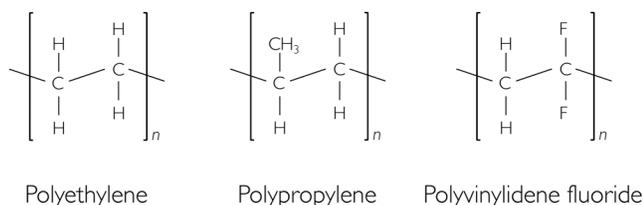


FIG. 8. The chemical structure of polyethylene, polypropylene, and polyvinylidene.

A. Polymer SMILES

Description. Please write the SMILES string of polyethylene, polypropylene, and polyvinylidene fluoride whose repeat units are shown in Fig. 8.

Solution. By referring to the guidelines available at www.polymergenome.org, the SMILES strings of these polymers can be written as [*]CC[*] for polyethylene, [*]C(C)C[*] for polypropylene, and [*]CC(F)(F)[*] for polyvinylidene fluoride.

B. GUI polymer draw tool for complex polymers

Description. It is not easy to directly or manually construct a SMILES string for a complex polymer such as polyvinylpyrrolidone⁶⁶ and PIM-1 (polymers of intrinsic microporosity),⁶⁷ as shown in Fig. 9. Please use the polymer draw tool to sketch these polymers and obtain the SMILES strings.

Solution. In this Tutorial problem, users will find using the GUI-based polymer draw tool is much more convenient. After completing the drawing, users will obtain the SMILES strings of these polymers as [*]CC([*])N1CCCC1=O and C12=CC4=C(C=C1CC3=C(C2)C(C#N)=C([g])C([t])=C3C#N)C5(C(C4(C)C)C6=C(CC5)C=C(O[d])C(O[e])=C6. We note that the SMILES strings are not unique for a polymer, i.e., the same polymer may be represented using different SMILES strings. Thus, it is possible that users may arrive at different SMILES strings for these polymers under discussion.

C. Save drawn polymers for later use

Description. Polymers in the same family may share some common substructures, and drawing them separately is quite inconvenient/inefficient. In this Tutorial, users are asked to (1) draw polymer (a) in Fig. 10, (2) right click (on a Windows or Linux computer) or hold Ctrl and click the mouse (on a Mac) on the draw tool, select “Copy as MOL,” and paste the copied text into a file with extension .mol, (3) open a blank draw windows, right click (or holding Ctrl while click the mouse) again, select “Paste MOL or SDF or SMILES,” either paste the text from the saved file or upload it, click “Accept” to import the saved information, and (4) continue editing the imported polymer to make polymer (b) in Fig. 10.

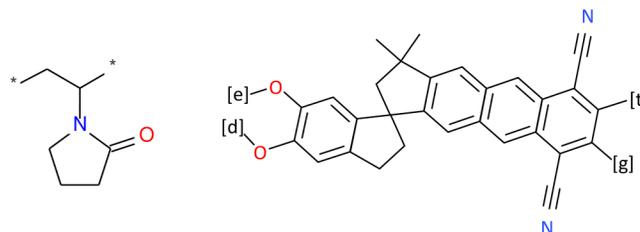


FIG. 9. The chemical structure of polyvinylpyrrolidone, a linear polymer (left) and a PIM-1 (polymers of intrinsic microporosity), a ladder polymer (right).

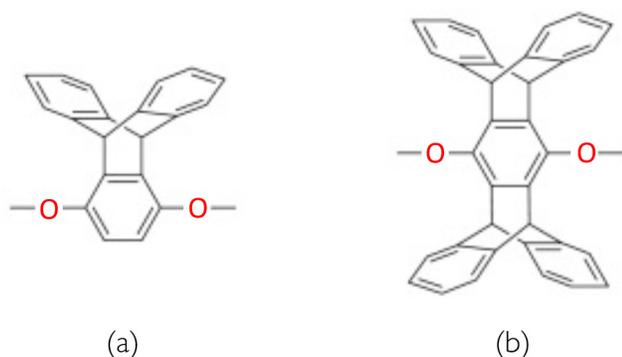


FIG. 10. Two complex polymers, given in (a) and (b), used in Tutorial "Save drawn polymers for later use."

Solution. After completing the aforementioned steps, users will find saving and reusing the drawn polymers a very convenient practice that could save lots of time when using Polymer Genome.

D. Query polymers with block notations

Description. Because of some historical reasons, Polymer Genome supports the use of some pre-defined polymer building blocks, some of them are $-\text{CH}_2-$, $-\text{NH}-$, $-\text{C}_6\text{H}_4-$, and $-\text{C}_4\text{H}_2\text{S}-$, for defining a class of linear polymers.^{64,65} A full list of these blocks and their SMILES string can be found in the guidelines at www.polymergenome.org. Users are now asked to use Polymer Genome in order to (1) predict the crystal bandgap and the total dielectric constant of three polymers whose repeat unit are $-\text{NH}-\text{CO}-\text{NH}-\text{C}_6\text{H}_4-$, $-\text{CO}-\text{NH}-\text{CO}-\text{C}_6\text{H}_4-$, and $-\text{NH}-\text{CS}-\text{NH}-\text{C}_6\text{H}_4-$ in the block notation and then (2) compare the predictions with the results obtained by DFT computations which can be found in Fig. 3 of Ref. 64.

Solution. Users will find the predicted values at about $\approx 5\%$ from the DFT computed values within less than a minute. The primary advantage of using Polymer Genome is the speed with which the results can be obtained. For reference, the DFT computations reported in Ref. 64, which include (1) polymer crystal structure prediction and (2) bandgap and dielectric constant calculations for the predicted structures, need days to weeks to complete.

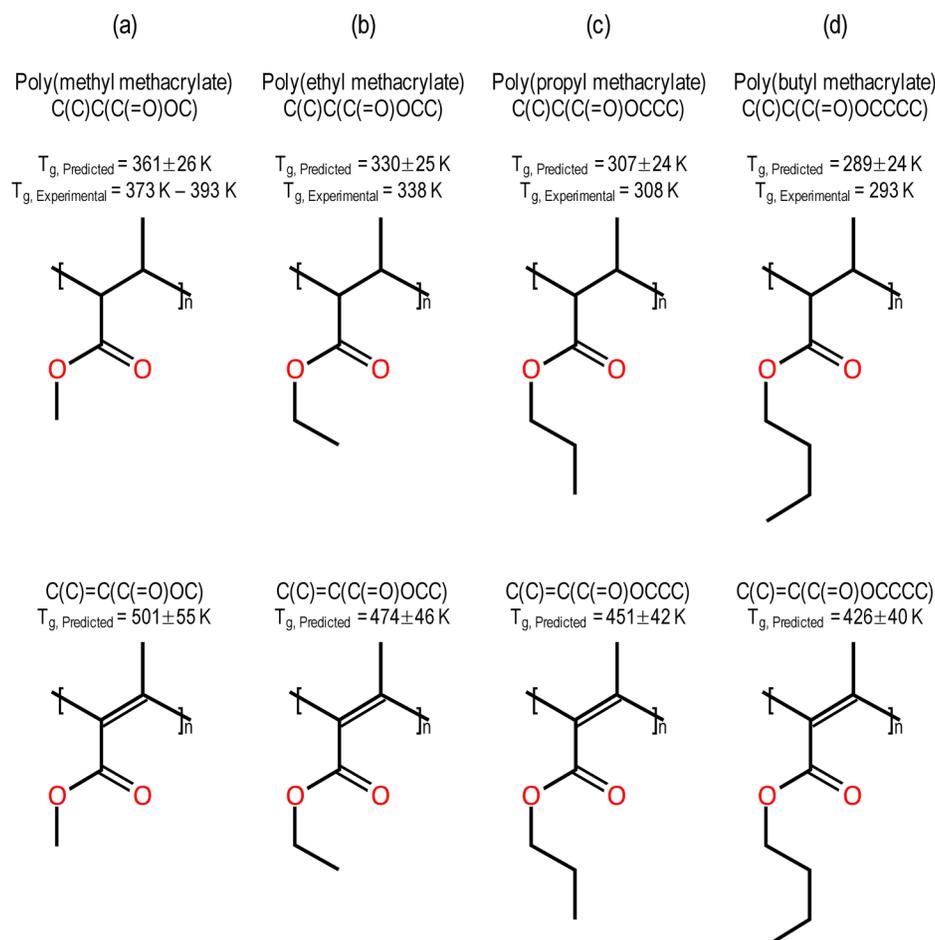


FIG. 11. Predicted and measured T_g of (a) poly(methyl methacrylate), (b) poly(ethyl methacrylate), (c) poly(propyl methacrylate), and (d) poly(butyl methacrylate) (top panel). The clear trend is that the longer side chain, the lower T_g . In the bottom panel, a double bond is introduced in the main chain of these polymers, raising its stiffness, and resulting in consistently higher predicted T_g .

E. Trends in the glass transition temperature

Description: There are several guidelines for the important features affecting the glass transition temperature T_g of polymers, two of them will be tested in this Tutorial. First, polymers with longer side chain length could generally have smaller T_g . Second, the higher the stiffness of the main chain (polymer backbone), the higher the T_g . Please use the Polymer Genome to explore the aforementioned effects and confirm the expected trends.

Solution: Poly(methacrylate) is a prototypical polymer that can be used to verify both guidelines. Figures 11(a)–11(d) show the structure of poly(methyl methacrylate), poly(ethyl methacrylate), poly(propyl methacrylate), and poly(butyl methacrylate), which are formed by attaching a methyl group, an ethyl group, a propyl group, and a butyl group, respectively, at the end of the prototypical polymer main chain. By examining the predicted glass transition temperature, the first trend can be validated. Now, a double bond is introduced in the main chain of these four polymers, elevating the main chain stiffness. Consequently, user will find the predicted glass transition temperature is significantly increased.

F. High refractive index polymers

Description: High refractive index polymers are particularly useful for advanced photonic devices. In addition to a high refractive index n , suitable candidates should have high thermal stability, of which a measure is the glass transition temperature T_g . As a typical solution for designing intrinsic high refractive index polymers, aromatic rings and/or sulfur-containing groups are used.⁶⁸ Interestingly, as learned in the “trends in the glass transition temperature” tutorial, aromatic rings could also enhance the stiffness of the polymer backbone, thus raising T_g . Please use Polymer Genome to (1) predict the refractive index and the glass transition temperature of ten sulfur-containing polyimides given in Table II of Ref. 68, (2) compare the predicted refractive index with that given in the same reference, and (3) select the promising candidates with $n \geq 1.75$ and $T_g \geq 450$ K.

Solution: Predictions were made for four out of ten polymers requested, and the results are shown in Table II. Users can easily extend this table and find the promising candidates. One of them was already identified and highlighted in bold.

TABLE II. Sulfur-containing polyimides that have high refractive index n and, thus, could be useful for photonic devices if having high T_g . The reference refractive index n^{ref} was taken from Ref. 68 while n^{pred} and T_g^{pred} were obtained using Polymer Genome. The entry highlighted in bold is a promising candidate.

SMILES	n^{ref}	n^{pred}	T_g^{pred}
[*]c3ccc(Sc2ccc(Sc1ccc([*])cc1)cc2)cc3	1.75	1.72 ± 0.06	405 ± 50
[*]c3ccc(Sc2ccc(Sc1ccc([*])cc1)s2)cc3	1.75	1.77 ± 0.06	347 ± 89
[*]c3ccc(Sc2nnc(Sc1ccc([*])cc1)s2)cc3	1.75	1.71 ± 0.09	312 ± 154
[*]c5ccc(Sc4c1SCCSc1c(Sc2ccc([*])cc2)c3SCCSc34)cc5	1.77	1.80 ± 0.07	462 ± 68

G. Polymeric membranes for gas separation

Description: A polymer that is good for separating O_2 and N_2 gases should have (1) high O_2 permeability to allow O_2 pass through and (2) high O_2/N_2 selectivity to pass only a smaller amount of N_2 compared to O_2 . The O_2/N_2 selectivity is defined as the ratio between the O_2 permeability and the N_2 permeability. In this tutorial problem, the targeted permeability of O_2 is 2000 Barrer and above, the O_2/N_2 selectivity is greater than 2, and the N_2 permeability is not higher than 1800 Barrer.

Solution: We will start from a template polymer, i.e., fluorenyl-poly(diarylacetylene) with the SMILES [*]C(=C([*])c1cc(F)cc(F)c1)c4ccc3c2cccc2C(C)(C)c3c4. This polymer has sufficiently large O_2 permeability (2727 Barrer) but the N_2 permeability (1689 Barrer) is still not small enough to have the O_2/N_2 selectivity of 2 and above. Now, users are asked to modify this polymer by replacing the left-side pendant c1cc(F)cc(F)c1 of this template polymer by one of the following options: -O, -N, -COOH, -c1ccc1, c1c(N)cc1, -c1cc(N)ccc1, -c1cc(Cl)cc(Cl)c1, -c1cc(Br)cc(Br)c1, -c1cc(I)cc(I)c1, and -c1cc5cccc5cc1, and tabulate the results. Note that the GUI-based polymer draw tool is useful for this work. The results are summarized in Table III, showing two candidates meeting all three aforementioned requirements.

H. Finding solvents for multiple polymers

Description: Users are asked to prepare a solution-deposited coating that is a combination of four different polymers, including (1) poly(dioctyloxyphosphazene), (2) poly[1-(2,3,4,5,6-

TABLE III. Predicted O_2 permeability, N_2 permeability, and O_2/N_2 selectivity of ten polymers obtained in Sec. III G. Entries highlighted in bold are candidates that meet the required criteria.

SMILES	O_2 permeability	N_2 permeability	O_2/N_2 selectivity
[*]C(=C([*])O)c3ccc2c1cccc1C(C)(C)c2c3	1112	711	6
[*]C(=C([*])N)c3ccc2c1cccc1C(C)(C)c2c3	1187	567	2.1
[*]C(=C([*])COO)c4ccc3c2cccc2C(C)(C)c3c4	525	299	1.8
[*]C(=C([*])c1ccc1)c4ccc3c2cccc2C(C)(C)c3c4	6642	3384	2.0
[*]C(=C([*])c1c(N)cc1)c4ccc3c2cccc2C(C)(C)c3c4	3683	1622	2.3
[*]C(=C([*])c1cc(N)ccc1)c4ccc3c2cccc2C(C)(C)c3c4	1638	797	2.1
[*]C(=C([*])c1cc(Cl)cc(Cl)c1)c4ccc3c2cccc2C(C)(C)c3c4	2521	1397	1.8
[*]C(=C([*])c1cc(Br)cc(Br)c1)c4ccc3c2cccc2C(C)(C)c3c4	2917	1412	2.1
[*]C(=C([*])c1cc(I)cc(I)c1)c4ccc3c2cccc2C(C)(C)c3c4	2597	1400	1.9
[*]C(=C([*])c1cc5cccc5cc1)c4ccc3c2cccc2C(C)(C)c3c4	4194	3292	1.3

TABLE IV. Predicted solvents for four polymers considered, whose SMILES strings are also provided for convenience. The predicted solvents appearing in the predictions for all the polymers are highlighted in bold font.

Polymer	poly(dioctyloxyphosphazene) <chem>[*]N=P([*])(OCCCCCCCC)OCCCCCCCC</chem>
Solvents	M-cresol, dichloromethane, acetic acid , NMP, chlorobenzene , nitrobenzene, THF, chloroform, benzene, toluene
Polymer	poly[1-(2,3,4,5,6-pentafluorophenyl)ethylene] <chem>[*]CC([*])c1c(F)c(F)c(F)c(F)c1F</chem>
Solvents	Chlorobenzene , NMP, DMAc, nitrobenzene, acetic acid , M-cresol, N-butanol, acetonitrile, DMF, 1,4-dioxane
Polymer	poly(1-phenylethene-1,2-diyl) <chem>[*]C=C([*])c1ccccc1</chem>
Solvents	Chlorobenzene , NMP, DMAc, nitrobenzene, acetic acid , M-cresol, N-butanol, acetonitrile, DMF, 1,4-dioxane
Polymer	poly(oxydecanedioyl) <chem>[*]OC(=O)CCCCCCCC([*])=O</chem>
Solvents	Dichloromethane, NMP, Acetic acid , chlorobenzene , benzene, toluene, chloroform, THF, DMAc, 1,4-dioxane

pentafluorophenyl)ethylene], (3) poly(1-phenylethene-1,2-diyl), and (4) poly(oxydecanedioyl). For convenience, the SMILES strings of these polymers can be found in Table IV. In order to perform this task, the polymers must all be soluble in the same solvent. Among the limited types of solvents in the inventory (the full list can be found in Ref. 25), select two solvents that can be used to dissolve all the polymers on the list.

Solution: The list of solvents predicted by Polymer Genome for all four polymers in consideration is shown in Table IV. Based on the obtained results, chlorobenzene and acetic acid, highlighted in bold font in the table, can be used to dissolve all the polymers.

IV. GOING FORWARD

The emergence of polymer informatics has opened up a pathway to instantly estimate the properties of new polymers and efficiently explore the staggering polymer space. Polymer Genome is a recent development in this sub-domain of materials research. By harnessing the existing knowledge base of past studies, an ecosystem of new machine-learning based tools has been systematically created, implemented, and deployed, serving the growing needs of polymer scientists from both academic and industrial domains. Needless to say, there are multiple open problems that need to be addressed in the future.

The current polymer data sets and predictive models of Polymer Genome do not handle network polymers, polymer blends, copolymers, and those with species other than C, H, N, O, B, F, Si, P, S, Cl, Br, and I. Polymers that have metal atoms in the backbones, also referred to as organometallic polymers, may host novel functionalities due to the nature of the carbon-metal bonds.^{34,45,69,70} The first step to closing this gap is to collect and curate literature data on these polymer subclasses, either

manually or using more sustainable natural language processing based methods. Computational data can also be generated when the polymer space is explored in an efficient manner using not just high-throughput but also autonomous computational workflows.⁴⁶

Further innovations in fingerprint developments can also be foreseen. First, when the polymer data are expanded to the new chemical, morphology, and processing condition domains, new fingerprint components are required. Second, the current polymer fingerprint scheme does not capture conformational and chiral degrees of freedom, and this deficiency should be solved in some ways. Finally, when the number of fingerprint components increases, more advanced feature engineering techniques should be explored for identifying the most relevant information for model development.

Going further, computer algorithms may also be used to discover the data representations (fingerprints), e.g., using variational auto-encoders⁷¹⁻⁷³ or by learning the SMILES of polymers. As the data set size and diversity increase, deep learning approaches^{72,73} that can simultaneously ingest the entire data set for all properties of interest and predict these properties at the same time, e.g., using multi-task learning, is expected to play increasing roles in polymer informatics.

Perhaps one of the most important motivations of the development of ML prediction models is that they can be used to design polymers with targeted properties for targeted applications. Because polymer properties can be predicted almost instantly, an intelligent enough algorithm could drive the polymer space exploration toward a designated target efficiently within a reasonable time scale. While some proofs-of-concept of this vision have been reported,¹⁸⁻²¹ more sophisticated, efficient, and robust methods should be further developed for making this goal become practical.

AUTHORS' CONTRIBUTIONS

H.D.T. and C.K. contributed equally to the creation of the manuscript.

ACKNOWLEDGMENTS

The authors are grateful for the financial support of various aspects of this work by the Office of Naval Research, Department of Energy, Toyota Research Institute, and the Kolon Center for Lifecycle Innovation. Computational support from XSEDE is also acknowledged. The authors are thankful to Kenny Lipkowitz, Blair Brettmann, and Ryan Lively for fruitful discussions. A couple of the tutorial examples were inspired by classroom activities created by Blair Brettmann.

DATA AVAILABILITY

The DFT data that support the findings of this study are openly available in khazana.gatech.edu, Ref. 74.

REFERENCES

- ¹Information Science for Materials Discovery and Design, edited by T. Lookman, F. J. Alexander, and K. Rajan (Springer International Publishing, 2016).
- ²K. Rajan, *Mater. Today* **8**, 38 (2005).
- ³A. Agrawal and A. Choudhary, *Appl. Phys. Lett. Mater.* **4**, 053208 (2016).
- ⁴D. J. Audus and J. J. de Pablo, *ACS Macro Lett.* **6**, 1078 (2017).
- ⁵G. Chen, Z. Shen, A. Iyer, U. F. Ghumman, S. Tang, J. Bi, W. Chen, and Y. Li, *Polymers* **12**, 163 (2020).
- ⁶N. E. Jackson, M. A. Webb, and J. J. de Pablo, *Curr. Opin. Chem. Eng.* **23**, 106 (2019).
- ⁷J. S. Peerless, N. J. Milliken, T. J. Oweida, M. D. Manning, and Y. G. Yingling, *Adv. Theor. Simul.* **2**, 1800129 (2019).
- ⁸C. Kim, A. Chandrasekaran, T. D. Huan, D. Das, and R. Ramprasad, *J. Phys. Chem. C* **122**, 17575 (2018).
- ⁹A. Mannodi-Kanakkithodi, A. Chandrasekaran, C. Kim, T. D. Huan, G. Pilania, V. Botu, and R. Ramprasad, *Mater. Today* **21**, 785 (2018).
- ¹⁰G. W. Ehrenstein, *Polymeric Materials: Structure, Properties, Applications* (Carl Hanser Verlag GmbH Co KG, 2012).
- ¹¹T. D. Huan, S. Boggs, G. Teysedre, C. Laurent, M. Cakmak, S. Kumar, and R. Ramprasad, *Prog. Mater. Sci.* **83**, 236 (2016).
- ¹²T. E. Gartner III and A. Jayaraman, *Macromolecules* **52**, 755 (2019).
- ¹³*Gaussian Processes for Machine Learning*, edited by C. E. Rasmussen and C. K. I. Williams (The MIT Press, Cambridge, MA, 2006).
- ¹⁴C. K. I. Williams and C. E. Rasmussen, in *Advances in Neural Information Processing Systems 8*, edited by D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo (MIT Press, 1995).
- ¹⁵R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim, *npj Comput. Mater.* **3**, 54 (2017).
- ¹⁶T. Mueller, A. G. Kusne, and R. Ramprasad, *Reviews in Computational Chemistry* (Wiley, New York, 2016), Chap. 4.
- ¹⁷G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, and R. Ramprasad, *Sci. Rep.* **3**, 2810 (2013).
- ¹⁸T. D. Huan, A. Mannodi-Kanakkithodi, and R. Ramprasad, *Phys. Rev. B* **92**, 014106 (2015).
- ¹⁹A. Mannodi-Kanakkithodi, G. Pilania, T. D. Huan, T. Lookman, and R. Ramprasad, *Sci. Rep.* **6**, 20952 (2016).
- ²⁰R. Batra, H. Dai, T. D. Huan, L. Chen, C. Kim, W. R. Gutekunst, L. Song, and R. Ramprasad, submitted (2020).
- ²¹C. Kim, R. Batra, L. Chen, H. Tran, and R. Ramprasad, *Comput. Mater. Sci.* **186**, 110067 (2020).
- ²²J. Heyd, G. E. Scuseria, and M. Ernzerhof, *J. Chem. Phys.* **118**, 8207 (2003).
- ²³L. Chen, C. Kim, R. Batra, J. P. Lightstone, C. Wu, Z. Li, A. A. Deshmukh, Y. Wang, H. D. Tran, P. Vashishta *et al.*, *npj Comput. Mater.* **6**, 1 (2020).
- ²⁴J. P. Lightstone, L. Chen, C. Kim, R. Batra, and R. Ramprasad, *J. Appl. Phys.* **127**, 215105 (2020).
- ²⁵A. Chandrasekaran, C. Kim, S. Venkatraman, and R. Ramprasad, *Macromolecules* **53**, 4764 (2020).
- ²⁶S. Venkatram, C. Kim, A. Chandrasekaran, and R. Ramprasad, *J. Chem. Inf. Model.* **59**, 4188 (2019).
- ²⁷G. Zhu, C. Kim, A. Chandrasekaran, J. D. Everett, R. Ramprasad, and R. P. Lively, *J. Polymer Eng.* **40**, 451 (2020).
- ²⁸S. Venkatram, R. Batra, L. Chen, C. Kim, M. Shelton, and R. Ramprasad, *J. Phys. Chem. B* **124**, 6046 (2020).
- ²⁹*Polymer Handbook*, 4th ed., edited by J. Brandup, E. H. Immergut, and E. A. Grulke (John Wiley & Sons, New York, 1999).
- ³⁰*Handbook of Polymers*, 2nd ed., edited by G. Wypych (ChemTec Publishing, Toronto, 2016).
- ³¹D. W. Van Krevelen and K. Te Nijenhuis, *Properties of Polymers: Their Correlation with Chemical Structure; Their Numerical Estimation and Prediction from Additive Group Contributions* (Elsevier, 2009).
- ³²*Polymer Data Handbook*, 2nd ed., edited by J. E. Mark (Oxford University Press, New York, 2009).
- ³³S. Otsuka, I. Kuwajima, J. Hosoya, Y. Xu, and M. Yamazaki, in *2011 International Conference on Emerging Intelligent Data and Web Technologies (EIDWT)* (IEEE, Tirana, 2011), pp. 22–29.
- ³⁴T. D. Huan, A. Mannodi-Kanakkithodi, C. Kim, V. Sharma, G. Pilania, and R. Ramprasad, *Sci. Data* **3**, 160012 (2016).
- ³⁵T. D. Huan and R. Ramprasad, *J. Phys. Chem. Lett.* **11**, 5823 (2020).
- ³⁶G. Kresse and J. Hafner, *Phys. Rev. B* **47**, 558 (1993).
- ³⁷G. Kresse, “Ab initio molekular dynamik für flüssige metalle,” Ph.D. thesis (Technische Universität Wien, 1993).
- ³⁸G. Kresse and J. Furthmüller, *Comput. Mater. Sci.* **6**, 15 (1996).
- ³⁹G. Kresse and J. Furthmüller, *Phys. Rev. B* **54**, 11169 (1996).
- ⁴⁰S. Goedecker, *J. Chem. Phys.* **120**, 9911 (2004).
- ⁴¹M. Amsler and S. Goedecker, *J. Chem. Phys.* **133**, 224104 (2010).
- ⁴²C. W. Glass, A. R. Oganov, and N. Hansen, *Comput. Phys. Commun.* **175**, 713 (2006).
- ⁴³A. R. Oganov and C. W. Glass, *J. Chem. Phys.* **124**, 244704 (2006).
- ⁴⁴Q. Zhu, V. Sharma, A. R. Oganov, and R. Ramprasad, *J. Chem. Phys.* **141**, 154102 (2014).
- ⁴⁵A. F. Baldwin, T. D. Huan, R. Ma, A. Mannodi-Kanakkithodi, M. Tefferi, N. Katz, Y. Cao, R. Ramprasad, and G. A. Sotzing, *Macromolecules* **48**, 2422 (2015).
- ⁴⁶J. H. Montoya, K. T. Winther, R. A. Flores, T. Bligaard, J. H. Hummelshøj, and M. Aykol, *Chem. Sci.* **11**, 8517 (2020).
- ⁴⁷M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, *Phys. Rev. Lett.* **108**, 058301 (2012).
- ⁴⁸B. Huang and O. A. von Lilienfeld, *J. Chem. Phys.* **145**, 161102 (2016).
- ⁴⁹K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller, and E. K. U. Gross, *Phys. Rev. B* **89**, 205118 (2014).
- ⁵⁰T. D. Huan, R. Batra, J. Chapman, S. Krishnan, L. Chen, and R. Ramprasad, *npj Comput. Mater.* **3**, 37 (2017).
- ⁵¹D. Weininger, *J. Chem. Inf. Comput. Sci.* **28**, 31 (1988).
- ⁵²I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, *Mach. Learn.* **46**, 389 (2002).
- ⁵³M. C. Kennedy and A. O’Hagan, *Biometrika* **87**, 1 (2000).
- ⁵⁴G. Pilania, J. Gubernatis, and T. Lookman, *Comput. Mater. Sci.* **129**, 156 (2017).
- ⁵⁵R. Batra, G. Pilania, B. P. Uberuaga, and R. Ramprasad, *ACS Appl. Mater. Interfaces* **11**, 24906 (2019).
- ⁵⁶A. Patra, R. Batra, A. Chandrasekaran, C. Kim, T. D. Huan, and R. Ramprasad, *Comput. Mater. Sci.* **172**, 109286 (2020).
- ⁵⁷J. Gasteiger and J. Zupan, *Angew. Chem. Int. Ed.* **32**, 503 (1993).
- ⁵⁸J. Behler and M. Parrinello, *Phys. Rev. Lett.* **98**, 146401 (2007).
- ⁵⁹T. B. Blank, S. D. Brown, A. W. Calhoun, and D. J. Doren, *J. Chem. Phys.* **103**, 4129 (1995).
- ⁶⁰N. Kuritz, G. Gordon, and A. Natan, *Phys. Rev. B* **98**, 094109 (2018).
- ⁶¹A. Chandrasekaran, D. Kamal, R. Batra, C. Kim, L. Chen, and R. Ramprasad, *npj Comput. Mater.* **5**, 22 (2019).
- ⁶²D. Kamal, A. Chandrasekaran, R. Batra, and R. Ramprasad, *Mach. Learn. Sci. Technol.* **1**, 025003 (2020).
- ⁶³T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. (Springer, New York, 2009).
- ⁶⁴V. Sharma, C. C. Wang, R. G. Lorenzini, R. Ma, Q. Zhu, D. W. Sinkovits, G. Pilania, A. R. Oganov, S. Kumar, G. A. Sotzing, S. A. Boggs, and R. Ramprasad, *Nat. Commun.* **5**, 4845 (2014).
- ⁶⁵A. Mannodi-Kanakkithodi, G. Treich, T. D. Huan, R. Ma, M. Tefferi, Y. Cao, G. Sotzing, and R. Ramprasad, *Adv. Mater.* **28**, 6277 (2016).
- ⁶⁶F. Haaf, A. Sanner, and F. Straub, *Polymer J.* **17**, 143 (1985).
- ⁶⁷N. B. McKeown and P. M. Budd, *Chem. Soc. Rev.* **35**, 675 (2006).

⁶⁸J.-G. Liu and M. Ueda, *J. Mater. Chem.* **19**, 8907 (2009).

⁶⁹G. M. Treich, S. Nasreen, A. Mannodi Kanakkithodi, R. Ma, M. Tefferi, J. Flynn, Y. Cao, R. Ramprasad, and G. A. Sotzing, *ACS Appl. Mater. Interfaces* **8**, 21270 (2016).

⁷⁰S. Nasreen, M. L. Baczkowski, G. M. Treich, M. Tefferi, C. Anastasia, R. Ramprasad, Y. Cao, and G. A. Sotzing, *Macromol. Rapid Commun.* **40**, 1800679 (2019).

⁷¹D. P. Kingma and M. Welling, *arXiv:1312.6114* (2013).

⁷²I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, 2016), see <http://www.deeplearningbook.org>.

⁷³D. Foster, *Generative Deep Learning: Teaching Machines to Paint, Write, Compose, and Play* (O'Reilly Media, 2019).

⁷⁴Khazana, Materials data and tools from the Ramprasad Group, <https://khazana.gatech.edu/>.