www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES

» Electronic properties and materials

» Computational chemistry

» Density functional theory

» Atomistic models

Received: 25 September 2015

Accepted: 26 January 2016

Published: 01 March 2016

A polymer dataset for accelerated property prediction and design

Tran Doan Huan[1], Arun Mannodi-Kanakkithodi[1], Chiho Kim[1], Vinit Sharma[1,]†, Ghanshyam Pilania[2] & Rampi Ramprasad[1]

Emerging computation- and data-driven approaches are particularly useful for rationally designing materials with targeted properties. Generally, these approaches rely on identifying structure-property relationships by learning from a dataset of sufficiently large number of relevant materials. The learned information can then be used to predict the properties of materials not already in the dataset, thus accelerating the materials design. Herein, we develop a dataset of 1,073 polymers and related materials and make it available at http://khazana.uconn.edu/. This dataset is uniformly prepared using first-principles calculations with structures obtained either from other sources or by using structure search methods. Because the immediate target of this work is to assist the design of high dielectric constant polymers, it is initially designed to include the optimized structures, atomization energies, band gaps, and dielectric constants. It will be progressively expanded by accumulating new materials and including additional properties calculated for the optimized structures provided.

Design Type(s)	database creation objective • Polymer Chemistry • 3D structure prediction
Measurement Type(s)	material properties
Technology Type(s)	computational modeling technique
Factor Type(s)	chemical compound
Sample Characteristic(s)	

[1] Institute of Materials Science, University of Connecticut, 97 North Eagleville Rd., Unit 3136, Storrs, Connecticut 06269, USA. [2] Materials Science and Technology Division, Los Alamos National Laboratory, Los Alamos, 87545 New Mexico, USA. †Present address: Materials Science and Technology Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA. Correspondence and requests for materials should be addressed to R.R. (email: rampi.ramprasad@uconn.edu).

Background & Summary

A central tenet of data-driven materials discovery is that if the volume of accumulated or available data is sufficiently large, and if it can be mined properly with suitable data-driven techniques, the process of designing a new material could be more efficient and rational[1–11]. This notion has lead to the development of many useful materials databases[12–18]. The present contribution deals with polymeric materials. Given the complexity of the chemical and configurational/morphological space of polymeric materials, the creation of a database focusing on this materials class is challenging. Nevertheless, if systematic steps can be taken in this direction, consistent with the charter of the Materials Genome Initiative, we will progressively get closer to the rational design and discovery of application-specific polymers.

Within this context, it is worth noting that the recent rational development of nearly a hundred novel polymeric dielectrics for capacitive or electrostatic energy storage[19–26] has benefitted from the synergy between experimental and computational efforts, of which computations at various levels, including force fields[27–30] and density functional theory (DFT)[31,32], have provided critical guidance. Given a polymer chemical composition, the computational step mainly involves predicting the lowest-energy structures and computing the associated dielectric constant ε and band gap E_g. Those with high ε and high E_g were then identified, leading to the experimental realizations of polymers with desired performances such as high energy density, low loss, etc., refs 19–26.

This contribution describes a dataset of 1,073 polymers and related materials as the first step aiming at the rational design of polymers by data-driven approaches. The dataset reported herein, referred to as “polymer dataset” for convenience, was prepared at a uniform and consistent level of first-principles DFT computations. Since our initial goal is to assist the design of high dielectric constant polymers for energy storage, the polymer dataset supplies the equilibrium (relaxed) structures of the materials associated with relevant calculated properties, including the atomization energy $\mathcal{E}_{at}$, the dielectric constant ε and the energy band gap E_g. The initial structures used for the preparation were collected either from other available sources or, quite often, from computational structure searches. This dataset, which is available at http://khazana.uconn.edu/, can readily be expanded in multiple ways, i.e., new properties can be calculated from the provided equilibrium structures, and new materials with relevant calculated properties can also be progressively added. Furthermore, it may also serve as a playground for data-mining.

Methods

Workflow

The workflow in Fig. 1 summarizes the preparation of the polymer dataset. In the first step, crystal structures of polymers and related compounds were collected from various available sources, including the reported literature, the *Crystallography Open Database* (COD)[15], and our structure prediction works[20–26]. Those obtained from structure prediction runs were subjected to a preliminary filter (described below), removing any obvious redundancy of identical structures. Then, the selected structures were optimized by DFT calculations, yielding the equilibrium structures and their atomization energies $\mathcal{E}_{at}$. The energy band gap E_g was then calculated on a dense grid of **k** points while their dielectric constant

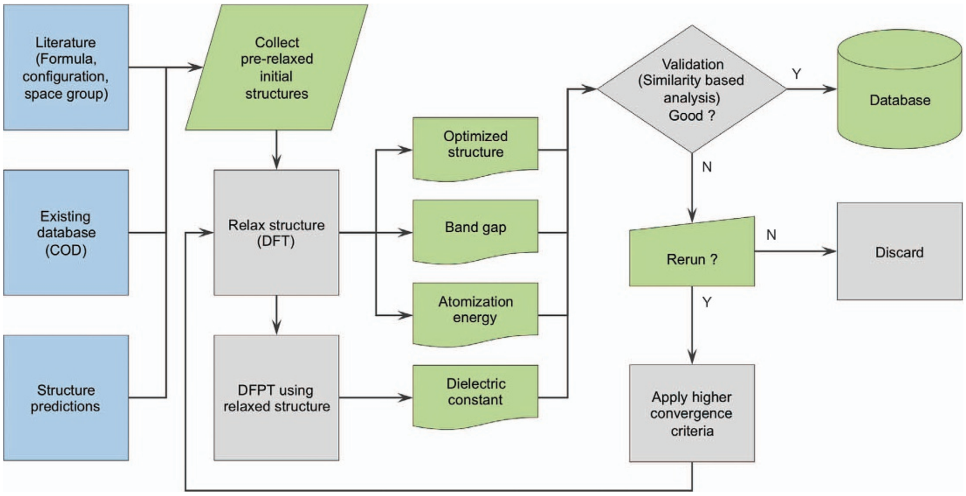
Literature (Formula, configuration, space group)
Existing database (COD)
Structure predictions
Collect pre-relaxed initial structures
Relax structure (DFT)
DFPT using relaxed structure
Optimized structure
Band gap
Atomization energy
Dielectric constant
Validation (Similarity based analysis) Good ?
Y
N
Database
Rerun ?
N
Y
Discard
Apply higher convergence criteria

Figure 1. Scheme for preparing the dataset of polymers and related materials. USPEX and minima-hopping are two structure prediction methods that were used for generating a majority of the dataset.

ε, which is composed of an electronic part $\varepsilon_{\text{elec}}$ and an ionic part ε_{ion}, was computed within the framework of density functional perturbation theory (DFPT)[33]. In the next step, the computational scheme and the calculated results were validated with available measured data, including the measured band gap E_g, the dielectric constant ε and/or the infrared spectroscopy (IR) measurements. Those which do not agree with the available experimental data are subjected to further calculations at tighter convergence criteria of residual atomic force (see Technical Validation for more details), and if better agreement is not reached, these points are removed from the dataset. A post-filtering step was finally performed on the whole dataset, keeping only distinct data points. Relaxed structures of all the materials are finally converted into the crystallographic information format (cif) using the pymatgen library[34]. A note was also provided together with the dataset, indicating the convergence criteria of the datapoints reported herein.

Structure accumulation

Our dataset includes three primary subsets, each of them originating from a distinct source. Subset 1 consists of *common polymers* which have already been synthesized, resolved, and reported elsewhere. This set contains 34 polymers, listed in Table 1. Collecting polymer structures of this class is challenging because the reported data is widely scattered, and in case the information obtained is sufficient to reconstruct structures, this work has to be done manually and hence, substantially laborious. We further note that only for a few of them, measurement for band gap, dielectric constant, and/or infrared (IR) spectrum have been performed. This data was used for the validation step.

Subset 2 includes 314 new organic polymers (284 of them have been used in ref. 11) and 472 new organometallic polymers. Their structures were generated from a computation-driven strategy[19,20] which has been used to rationally design various classes of polymeric dielectrics[11,20–26]. The starting point of this strategy is a pool of common polymer building blocks, which are either organic, e.g., –CH_2–, –NH–, –CO–, –O–, –CS–, –C_6H_4–, and –C_4H_2S–, or inorganic (metal-containing) like –COO–$Sn(CH_3)_2$–OCC–, –SnF_2–, and –$SnCl_2$–. The repeat unit of an organic polymer is then created by concatenating a given number of organic building blocks while that of an organometallic polymer contains at least one inorganic block linked with a chain of several CH_2 groups. Next, chains of the repeat units (illustrated in Fig. 2) are packed in low-energy crystal structures which are determined by Universal Structure Predictor: Evolutionary Xtallography (USPEX)[23,35] or minima-hopping (MH)[36,37], two of the currently most powerful structure prediction methods. In brief, these methods allow for predicting the low-energy structures of a material as the local minima of the potential energy surface, constructed from DFT energy. The efficiency of these methods have been successfully demonstrated for many different materials classes[38–41], including a large number of organic[20,23] and organometallic polymers[24–26].

For each structure prediction run, the lowest-energy structure and those within 200 meV per atom above it were collected. The number of structures within this energy window is material-dependent, ranging from several to several dozens. Because many of them are just slightly different by small perturbations in the atomic arrangement, a preliminary filtering step was used to remove this

Polymer	Ref.	Polymer	Ref.
Polyethylene	61	Isotactic polypropylene	62
Polyethylene oxide	23	Polyglutamic acid	23
Cellulose	23	Poly(1,1,2-trifluoroethene)	63
Clathrate syndiotactic polystyrene	64	Poly(2,5-dihydrothiophene-2,5-diyl)	65
Poly ε-caprolactone	66	Poly(2,6-benzothiazole)	67
Poly(3,3,3-trifluoro-2-methyloxirane)	68	Poly(ethene-alt-hexafluoroacetone)	69
Polyethylene adipate	70	Polyethylene suberate	70
Polyoxymethylene	23	Poly(p-phenylene oxide)	23
Poly(p-phenylene sulfide)	71	Poly(propylene sulfide)	72
Poly(p-xylylene)	23	Poly-tetrafluoroethylene-alt-ethylene	73
Poly(tetramethylene terephthalate)	74	Poly(trimethylene sebacate)	75
Poly(vinyl fluoride)	76	Polyethylene terephthalate	77
Polyvinylidene fluoride (delta)	78	Polyvinylidene fluoride (beta)	78
Syndiotactic polypropylene	79	Poly(2,5-benzoxazole)	67
Poly(2-vinylpyridine)	80	Polyacrylonitrile	81
Polyglycine	82	Poly (m-phenylene isophthalamide)	83
Poly(m-pyridine)	84	Poly(p-phenylene benzobisoxazole)	85

Table 1. List of the common polymers summarized in this dataset and the corresponding references.

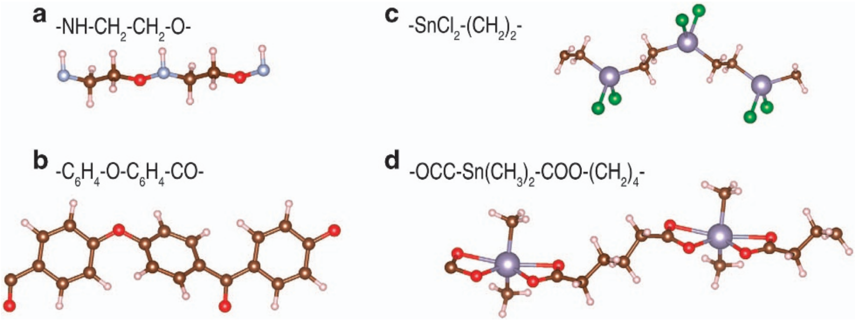
a -NH-CH$_2$-CH$_2$-O-
b -C$_6$H$_4$-O-C$_6$H$_4$-CO-
c -SnCl$_2$-(CH$_2$)$_2$-
d -OCC-Sn(CH$_3$)$_2$-COO-(CH$_2$)$_4$-

Figure 2. Organic polymer chains with repeat units of $–NH–CH_2–CH_2–O–$ (**a**) and $–C_6H_4–O–C_6H_4–CO–$ (**b**) and organometallic polymer chains with repeat units of $–SnCl_2–(CH_2)_2–$ (**c**) and $–OOC–Sn–(CH_3)_2–COO–(CH_2)_4–$ (**d**). Carbon, hydrogen, oxygen, nitrogen, chlorine, and tin atoms are shown in dark brown, light pink, red, light cyan, green, and dark cyan, respectively.

redundancy. In particular, we used a clustering algorithm (hierarchical) to group those which are different by less than 5 meV per atom in $\mathcal{E}_{\rm at}$ and less than 0.1 eV in $E_{\rm g}$, keeping the representative structures. Only those with polymeric motifs, when visually confirmed, are selected for the next steps. In the predicted polymer structures, especially for those of organometallic polymers, these polymeric chains are not necessarily isolated, i.e., inter-chain bonds may occur in various fashions[24–26].

The material structures used to prepare subset 3 were collected from COD. Generally, materials provided by COD are not polymers, but a number of them are collected in this dataset as they are closely related to the examined polymers. Although collecting materials structures from this database is straightforward, we limited ourselves to only those whose cell volumes are not too large, i.e., roughly 1,500 Å^3 and below. This subset contains 253 molecular organic and organometallic crystals, 178 of them have recently been used in ref. 10 by some of us.

Table 2 summarizes the contents of the polymer dataset, which contains both polymers (subset 1 and 2) and non-polymers (subset 3). In terms of chemistry, the included materials can be classified as either organic or organometallic, incorporating different metals in their backbone. The complete list of chemical elements that appear in this dataset is given in Table 3.

Numerical calculations

The computed data reported in our dataset was prepared with density functional theory (DFT)[31,32] calculations, using the projector augmented-wave (PAW) formalism[42] as implemented in Vienna *Ab initio* Simulation Package (vasp)[43–46]. The default accuracy level of our calculations is ``Accurate'', specified by setting PREC = Accurate in all the runs with vasp. The basis set includes all the plane waves with kinetic energies up to 400 eV, as recommended by vasp manual for this level of accuracy. PAW datasets of version 5.2, which were used to describe the ion-electron interactions, are also summarized in Table 3. The van der Waals dispersion interactions, known[47] to be important in stabilizing soft materials dominated by non-bonding interactions like polymers[48], were estimated with the non-local density functional vdW-DF2 (ref. 49). The generalized gradient approximation (GGA) functional associated with vdW-DF2, i.e., refitted Perdew-Wang 86 (rPW86)[50], was used for the exchange-correlation (XC) energies.

Because the examined material structures are significantly different in terms of the cell shape, the sampling procedure of their Brillouin zones must be handled appropriately. For each structure, a Monkhorst-Pack **k**-point mesh[51] of a given spacing parameter $h_{\rm k}$ in the reciprocal space was used. For the geometry optimization and dielectric constant calculations, $h_{\rm k} = 0.25$ Å^{-1} while the band gap calculations have been performed on a finer Γ-centered mesh with $h_{\rm k} = 0.20$ Å^{-1}. We further set the lower limit for the Monkhorst-Pack mesh dimensionality, that is, the number of grid points along any reciprocal axis is no less than 3, regardless of how short the reciprocal lattice dimension along this axis is.

During the relaxation step, we optimized both the cell and the atomic degrees of freedom of the materials structures until atomic forces are smaller than 0.01 eV Å^{-1}. Calculations for band gap $E_{\rm g}$ was then carried out on top of the equilibrium structures. Because $E_{\rm g}$ is typically underestimated with a GGA XC functional like rPW86 (ref. 52), this important physical property has also been calculated with the hybrid Heyd-Scuseria-Ernzerhof (HSE06) XC functional[53,54] with an expectation that the calculated result would become much closer to the true material band gap. Both $E_{\rm g}^{\rm GGA}$ and $E_{\rm g}^{\rm HSE06}$, the band gap calculated at the GGA-rPW86 and HSE06 levels of theory, are provided in all the entries of the dataset (see File format for more details). Finally, the dielectric constant ε of these structures was calculated within the DFPT formalism as implemented in vasp package. Calculations of this type involve the determination of the lattice vibrational spectra at Γ, the center of the Brillouin zone. This information is also used to compute the IR spectra of some structures for the purpose of validation.

ID	No. of points	Descriptions	Reference
0001–0034	34	Common polymers	[23,61–85]
0035–0348	314	New organic polymers	[11,20,21,23]
0349–0410	62	Poly(tin ester)	[24–26]
0411–0447	37	Titanium containing polymers	
0448–0460	13	Calcium containing polymers	
0461–0470	10	Aluminum containing polymers	
0471–0572	102	Zinc containing polymers	
0573–0588	16	Magnesium containing polymers	
0589–0610	22	Zirconium containing polymers	
0611–0630	20	Hafnium containing polymers	
0631–0741	111	Cadmium containing polymers	
0742–0763	22	$SnCl_2$ containing polymers	
0764–0796	33	SnF_2 containing polymers	
0797–0820	24	Lead containing polymers	
0821–0854	34	Molecular crystals of C and H	[10,15]
0855–0998	144	Molecular crystals of C, H, & O	[10,15]
0999–1050	52	Molecular crystals of C, H, N, & O	[15]
1051–1073	23	Molecular crystals of C, H, O, & Sn	[15]

Table 2. Summary of the data subclasses in the polymer dataset.

Element	POTCAR	Element	POTCAR	Element	POTCAR
Aluminum	Al	Bromine	Br	Carbon	C
Calcium	Ca_sv	Cadmium	Cd	Chlorine	Cl
Fluorine	F	Hydrogen	H	Hafnium	Hf_sv
Magnesium	Mg_sv	Nitrogen	N	Oxygen	O
Phosphorus	P	Lead	Pb_d	Sulfur	S
Tin	Sn_d	Titanium	Ti_sv	Zinc	Zn
Zirconium	Zr_sv				

Table 3. VASP PAW potentials of the elements used for calculations in this work.

Post-filtering

Given that the sources of the polymer dataset reported herein are diversified, any clear duplicate and/or redundancy should be identified and removed. Because the preliminary filtering step was performed only on subset 2 based on their DFT energy and band gap estimated during the structure prediction runs with a limited accuracy, an additional filtering step was performed on the whole dataset. Within this step, all cases with the same chemical composition but different by less than 0.1 eV in E_g, less than 5 meV per atom in $\mathcal{E}_{at}$, and less than 0.1 in both ε_{elec} and ε_{ion}, are clustered. At this point, the number of clustered points is not large, and all of them were inspected visually, keeping only distinct materials.

Data Records

The complete dataset of 1,073 polymers and related materials can be downloaded as a tarball from Dryad Repository (Data Citation 1) or can be accessed via http://khazana.uconn.edu/ (all the records with ID from 0001 to 1073). All 4,292 DFT runs of the entire dataset (for each structure, there are 4 runs, including relax, dielectric, GGA band gap, and HSE06 band gap) are hosted by NoMaD Repository (Data Citation 2).

File format

All the information reported in the dataset for a given material is stored in a file, named as 0001.cif, where a cardinal number (0001 in this example) is used for the identification of the entry in the dataset. The first part of a file of this type is devoted to the optimized structure in the standard cif format which is compatible with majority of visualization software. Other information, including the calculated properties, is provided as the comments lines in the second part of the file as follow

```
# Source: VSharma_etal:NatCommun.5.4845(2014)
# Class: organic_polymer_crystal
# Label: Polyimide
# Structure prediction method used: USPEX
# Number of atoms: 32
# Number of atom types: 4
# Atom types: C H O N
# Dielectric constant, electronic: 3.71475E+00
# Dielectric constant, ionic: 1.54812E+00
# Dielectric constant, total: 5.26287E+00
# Band gap at the GGA level (eV): 2.05350E+00
# Band gap at the HSE06 level (eV): 3.30140E+00
# Atomization energy (eV/atom): -6.46371E+00
# Volume of the unit cell (A^3): 2.79303E+02
```

While most of the keywords are clear, we used Source to provide the origin of the material structure and Class to refer to the class of materials which can either be “organic polymer crystal”, “organometallic polymer crystal”, “organic molecular crystal”, or “organometallic molecular crystal”. Keyword Label was used to provide more detailed information on the material, which can be the common name of the material if it is available, the ID of the record obtained from COD, or the repeat unit of the polymer structure predicted.

Graphical summary of the dataset

To graphically summarize the polymer dataset, we visualize it in the property space. Because the band gap and the dielectric constant are the primary properties reported by this dataset, three plots, namely $E_g^{\rm HSE06}-\varepsilon_{\rm elec}$, $E_g^{\rm HSE06}-\varepsilon_{\rm ion}$, and $E_g^{\rm HSE06}-\varepsilon$, were compiled and shown in Fig. 3. Materials from different classes are shown in different colors to clarify the role of the polymer chemical composition in controlling E_g and ε. Within the recent effort of developing polymers for high-energy-density applications[19–26], such plots are useful for identifying promising candidates, i.e., those which have high dielectric constant while maintaining sufficient band gap ($E_g \geq 3$ eV).

Figure 3a clearly indicates a limit of the form $\varepsilon_{\rm elec} \sim 1/E_g$ between $\varepsilon_{\rm elec}$ and E_g, which is applicable for both organic and organometallic classes of materials. We note that this behavior has also been reported elsewhere[10,19]. Figure 3c, on the other hand, demonstrates that the classes of organic and organometalic polymers and molecular crystals occupy different regions in the property space. At a given value of band gap, the organometallic polymers are generally much higher than the organic polymers in terms of the dielectric constant. While a fairly large number of organometallic polymers were already developed[24–26], this observation suggests that there remains significant room for manipulating the dielectric constant of the organometallic polymers.

Technical Validation

Among the materials properties reported in the present dataset, the atomization energy $\mathcal{E}_{\rm at}$ is physically relevant and has always been used as a standard method for examining the thermodynamic stability of various classes of materials, including inorganic crystals[38–41] and polymers[19–26]. While the band gap $E_g^{\rm GGA}$ calculated at the GGA level of DFT is not ready to be compared with the measured data due to the aforementioned well-known underestimation[52], $E_g^{\rm HSE06}$ (the band gap calculated with the HSE06 XC

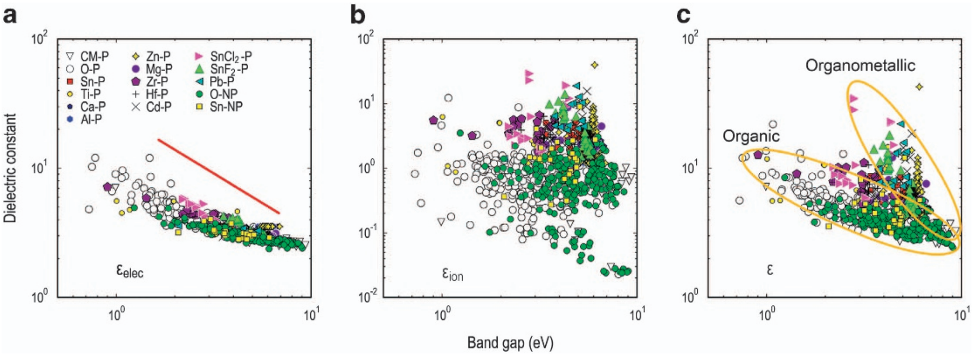

Figure 3. A summary of the polymer dataset based on the calculated band gap $E_g^{\rm HSE06}$ and the dielectric constants $\varepsilon_{\rm elec}$ (**a**), $\varepsilon_{\rm ion}$ (**b**), and $\varepsilon = \varepsilon_{\rm elec} + \varepsilon_{\rm ion}$ (**c**). In the figure keys, “CM”, “P”, “NP”, and “O” refer to “Common”, “Polymer”, “Non-Polymer”, and “Organic”, respectively. For organometallic polymers, the identity of the metal element included is used. The polymers developed by the structure prediction based pathway in refs 19–26 are labeled as “Dev-P”.

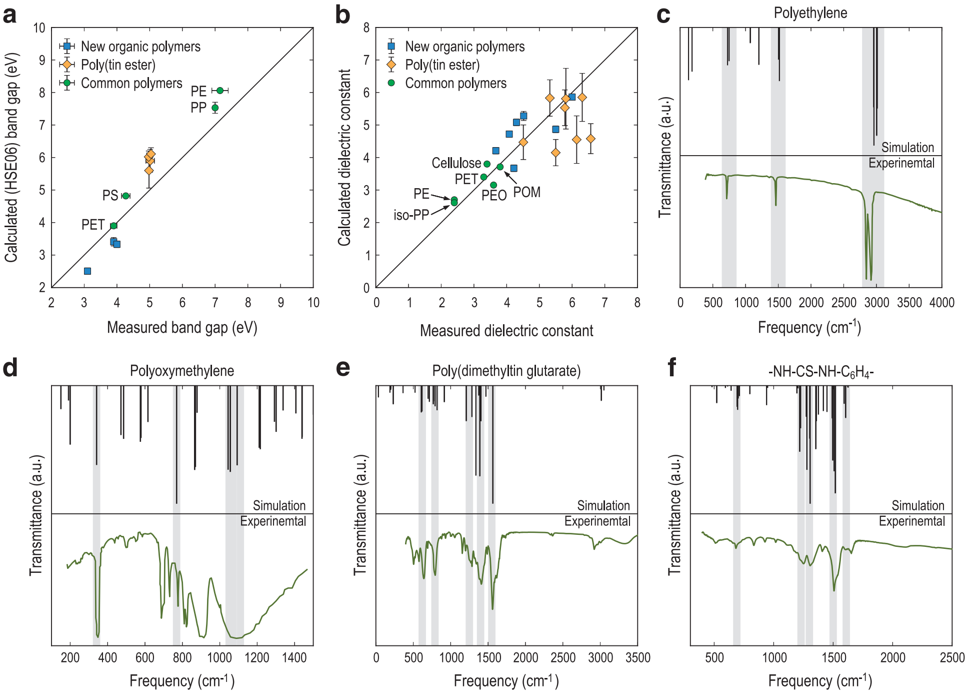

Figure 4. Calculated and measured dielectric constants of (**a**) several inorganic compounds, and (**b**) the polymers reported in refs 20–22 (new organic polymers) and refs 24–26 (poly(tin ester)). The error bars originated from different (energetically competing) structures predicted for a given polymer. For organometallic polymers, the error bars are significant due to the diversity of structural motifs involving the aforementioned inter-chain bonds, which are not present in organic polymers. In (**c**), (**d**), (**e**), and (**f**), the simulated and measured infrared spectra of orthorhombic polyethylene, orthorhombic polyoxymethylene, poly (dimethyltin glutarare), and polythiourea are shown. The experimental data of these three polymers was taken from refs 60,55,24,20, respectively. Shadow areas are given to indicate the agreement between simulated and measured transmitance peaks.

functional) is expected to be rather close to the true band. We show in Fig. 4a $E_{\rm g}^{\rm HSE06}$ of 11 polymers for which the band gap has been measured experimentally. The calculated band gap seems to agree pretty well with the measured data with a numerical discrepancy of about 20% and below.

We now consider the calculations of the dielectric constants, namely $\varepsilon_{\rm elec}$ and $\varepsilon_{\rm ion}$. Overall, the theoretical foundations and the implementations for calculating $\varepsilon_{\rm elec}$ and $\varepsilon_{\rm ion}$ are well developed and tested, leading to rather accurate results. Within the DFT-based perturbative approach, $\varepsilon_{\rm elec}$ is computed via the response to the external field perturbations while $\varepsilon_{\rm ion}$ is evaluated through the phonon frequencies at the Γ point of the Brillouin zone. To be precise, the dielectric response of a crystalline insulator to an external electric field **E** is given in terms of a frequency-dependent tensor $\varepsilon^{\alpha\beta}(\omega)$. To linear order, the electronic contribution of the dielectric tensor is given by

$$\varepsilon^{\alpha\beta}_{\rm elec}(\omega) = 1 + 4\pi\frac{\partial P_\alpha}{\partial E_\beta}, \tag{1}$$

where P_α is the component along the α direction of the induced polarization **P**. On the other hand, the ionic part of the dielectric tensor is determined as

$$\varepsilon^{\alpha\beta}_{\rm ion}(\omega) = \frac{4\pi}{\Omega}\sum_m \frac{S_{m\alpha\beta}}{\omega^2_{m,\mathbf{q}=0} - \omega^2}. \tag{2}$$

In this expression, Ω is the volume of the simulation cell, appearing as a normalization factor. The sum is taken over the index m of the phonon normal modes, which assumes the frequency $\omega_{m,\mathbf{q}=0}$ at the Brillouin zone center ($\mathbf{q}=0$) while the mode oscillator strength $S_{m\alpha\beta}$ is determined through the Born effective charge $Z_{s,\alpha\beta}{}^{\star}$ of the atom s. For an isotropic material, the dielectric constant of the practical interest is taken to be the average value of its diagonal elements at the static limit, i.e., $\varepsilon = \frac{1}{3}\sum_\alpha[\varepsilon^{\alpha\alpha}(\omega\to 0)]$.

Equation 2 implies that at the limit of $\omega \to 0$, $\varepsilon_{\text{ion}}^{\alpha\beta}(\omega)$ is rather sensitive to the numerical accuracy of $\omega_{m,\mathbf{q}=0}$, which, in turn, suggests highly equilibrated materials structures for the DFPT calculations. As mentioned in the Workflow Section, if the calculated dielectric constant ε of a polymer is different from its measured data (this information is available for just a limited number polymers in subset 1 and 2) by more than 20%, the structures are further optimized until the residual atomic forces are smaller than 0.001 eV Å^{-1}. Only those with calculated dielectric constant within 20% of the experimental data [shown in Fig. 4b] are kept.

Within our dataset, the IR spectrum was measured for some materials. From the computational side, this material characteristic can also be calculated rather accurately from the byproducts of the dielectric constant calculations with DFPT. In particular, the intensities of the infrared-active modes are given by[56]

$$I_m \propto \sum_{\alpha} \left| \sum_{s\beta} Z^*_{s,\alpha\beta} e_{m,s\beta} \right|^2, \qquad (3)$$

where $e_{m,s\beta}$ is the β component of the normalized vibrational eigenvector of the mode m at the atom s. Obviously, all of the necessary quantities needed to calculate I_m according to Equation 3 can be obtained within the DFPT-based computational scheme of ε, thus requiring essentially no computational overhead. This approach has widely been used in characterizing various classes of materials[57,58]. We show in Figure 4c–f the IR spectra calculated for four polymers, including orthohombic polyethylene, orthohombic polyoxymethylene, poly(dimethyltin glutarate)[24], and polythiourea[20], each of them is compared with the corresponding measured IR spectrum. The excellent agreement between the calculated and the measured IR spectra can be regarded as a supportive validation of the computational scheme based on DFT calculations used for this polymer dataset.

Usage Notes

This dataset, which includes a variety of known and new organic and organometallic polymers and related materials, has been consistently prepared using first-principles calculations. While the HSE06 band gap $E_{\text{g}}^{\text{HSE06}}$ is believed to be fairly close to the true band gap of the materials, the GGA-rPW86 band gap is also reported for completeness and for further possible analysis. The reported atomization energy and the dielectric constants are also expected to be accurate.

The polymer dataset is one among many recently developed datasets which can be used for designing materials by various data-driven approaches. To be specific, this dataset is expected to be useful in the development of polymers for energy storage and electronics applications. Moving forward, the development of this dataset will be continuously validated and updated, and the most recent version can be accessed at repository http://khazana.uconn.edu/.

References

1. Curtarolo, S. *et al.* The high-throughput highway to computational materials design. *Nat. Matter* **12,** 191 (2013).
2. Hautier, G., Jain, A. & Ong, S. From the computer to the laboratory: materials discovery and design using first-principles calculations. *J. Mater. Sci.* **47,** 7317 (2012).
3. Rajan, K. Materials informatics. *Mater. Today* **8,** 38 (2005).
4. Schön, J. C. How can databases assist with the prediction of chemical compounds? *Z. Anorg. Allg. Chem.* **640,** 2717 (2014).
5. Curtarolo, S., Morgan, D., Persson, K., Rodgers, J. & Ceder, G. Predicting crystal structures with data mining of quantum calculations. *Phys. Rev. Lett.* **91,** 135503 (2003).
6. Hansen, K. *et al.* Assessment and validation of machine learning methods for predicting molecular atomization energies. *J. Chem. Theory Comput.* **9,** 3404 (2013).
7. Meredig, B. *et al.* Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B* **89,** 094104 (2014).
8. Bhat, T. N., Bartolo, L. M., Kattner, U. R., Campbell, C. E. & Elliott, J. T. Strategy for extensible, evolving terminology for the Materials Genome Initiative efforts. *JOM* **67,** 1866 (2015).
9. Pilania, G., Wang, C., Jiang, X., Rajasekaran, S. & Ramprasad, R. Accelerating materials property predictions using machine learning. *Sci. Rep* **3,** 2810 (2013).
10. Huan, T. D., Mannodi-Kanakkithodi, A. & Ramprasad, R. Accelerated materials property predictions and design using motif-based fingerprints. *Phys. Rev. B* **92,** 014106 (2015).
11. Mannodi-Kanakkithodi, A., Pilania, G., Huan, T. D., Lookman, T. & Ramprasad, R. Machine learning strategy for the accelerated design of polymer dielectrics. *Sci. Rep.* doi:10.1038/srep20952 (2016).
12. Jain, A. *et al.* Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* **1,** 011002 (2013).
13. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). *JOM* **65,** 1501 (2013).
14. Taylor, R. H. *et al.* A RESTful API for exchanging materials data in the AFLOWLIB.org consortium. *Comput. Mater. Sci.* **93,** 178 (2014).
15. Gražulis, S. *et al.* Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic Acids Res.* **40,** D420 (2012).
16. Otsuka, S., Kuwajima, I., Hosoya, J., Xu, Y. & Yamazaki, M. in *International Conference on Emerging Intelligent Data and Web Technologies (EIDWT)* pp 22–29 (IEEE, Tirana, 2011).
17. Reymond, J.-L. The chemical space project. *Acc. Chem. Res.* **48,** 722 (2015).
18. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Dat* **1,** 140022 (2014).
19. Wang, C. C. *et al.* Computational strategies for polymer dielectric design. *Polymer* **55,** 979 (2014).
20. Sharma, V. *et al.* Rational design of all organic polymer dielectrics. *Nat. Commun* **5,** 4845 (2014).

21. Lorenzini, R., Kline, W., Wang, C., Ramprasad, R. & Sotzing, G. The rational design of polyurea & polyurethane dielectric materials. *Polymer* **54,** 3529 (2013).
22. Ma, R. *et al.* Rational design and synthesis of polythioureas as capacitor dielectrics. *J. Mater. Chem. A* **3,** 14845 (2015).
23. Zhu, Q., Sharma, V., Oganov, A. R. & Ramprasad, R. Predicting polymeric crystal structures by evolutionary algorithms. *J. Chem. Phys.* **141,** 154102 (2014).
24. Baldwin, A. F. *et al.* Poly(dimethyltin glutarate) as a prospective material for high dielectric applications. *Adv. Matter* **27,** 346 (2015).
25. Baldwin, A. F. *et al.* Rational design of organotin polyesters. *Macromolecules* **48,** 2422 (2015).
26. Baldwin, A. F. *et al.* Effect of incorporating aromatic and chiral groups on the dielectric properties of poly(dimethyltin esters). *Macromol. Rapid Commun.* **35,** 2082 (2014).
27. Banks, J. L. *et al.* Integrated modeling program, applied chemical theory (IMPACT). *J. Comput. Chem.* **26,** 1752 (2005).
28. Jorgensen, W. L., Ulmschneider, J. P. & Tirado-Rives, J. Free Energies of Hydration from a Generalized Born Model and an All-Atom Force Field. *J. Phys. Chem. B* **108,** 16264 (2004).
29. Vanommeslaeghe, K. & MacKerell, Jr., A. D. Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. *J. Chem. Infor. Model.* **52,** 3144 (2012).
30. Vanommeslaeghe, K., Raman, E. P. & MacKerell, Jr., A. D. Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges. *J. Chem. Infor. Model.* **52,** 3155 (2012).
31. Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **136,** B864 (1964).
32. Kohn, W. & Sham, L. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140,** A1133 (1965).
33. Baroni, S., de Gironcoli, S. & Dal Corso, A. Phonons and related crystal properties from density-functional perturbation theory. *Rev. Mod. Phys.* **73,** 515 (2001).
34. Ong, S. P. *et al.* Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68,** 314 (2013).
35. Glass, C. W., Oganov, A. R. & Hansen, N. USPEX-Evolutionary crystal structure prediction. *Comput. Phys. Commun.* **175,** 713 (2006).
36. Goedecker, S. Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems. *J. Chem. Phys.* **120,** 9911 (2004).
37. Amsler, M. & Goedecker, S. Crystal structure prediction using the minima hopping method. *J. Chem. Phys.* **133,** 224104 (2010).
38. Huan, T. D., Amsler, M., Tuoc, V. N., Willand, A. & Goedecker, S. Low-energy structures of zinc borohydride $Zn(BH_4)_2$. *Phys. Rev. B* **86,** 224110 (2012).
39. Huan, T. D. *et al.* Thermodynamic stability of alkali metal/zinc double-cation borohydrides at low temperatures. *Phys. Rev. B* **88,** 024108 (2013).
40. Huan, T. D., Sharma, V., Rossetti, G. A. & Ramprasad, R. Pathways towards ferroelectricity in hafnia. *Phys. Rev. B* **90,** 064111 (2014).
41. Sharma, H., Sharma, V. & Huan, T. D. Exploring $PtSO_4$ and $PdSO_4$ phases: an evolutionary algorithm based investigation. *Phys. Chem. Chem. Phys.* **17,** 18146 (2015).
42. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50,** 17953 (1994).
43. Kresse, G. & Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* **47,** 558 (1993).
44. Kresse, G. (Ph.D. thesis), Ab initio Molekular Dynamik für flüssige Metalle, Technische Universität Wien, (1993).
45. Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **6,** 15 (1996).
46. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54,** 11169 (1996).
47. Woods, L. M. *et al.* Preprint, arXiv:1509.03338.
48. Liu, C.-S., Pilania, G., Wang, C. & Ramprasad, R. How Critical Are the van der Waals Interactions in Polymer Crystals? *J. Phys. Chem. A* **116,** 9347 (2012).
49. Lee, K., Murray, É. D., Kong, L., Lundqvist, B. I. & Langreth, D. C. Higher-accuracy van der Waals density functional. *Phys. Rev. B* **82,** 081101(R) (2010).
50. Murray, E. D., Lee, K. & Langreth, D. C. Investigation of exchange energy density functional accuracy for interacting molecules. *J. Chem. Theor. Comput* **5,** 2754 (2009).
51. Monkhorst, H. J. & Pack, J. D. Special points for Brillouin-zone integrations. *Phys. Rev. B* **13,** 5188 (1976).
52. Perdew, J. P. Density functional theory and the band gap problem. *Int. J. Quant. Chem.* **28,** 497 (1985).
53. Heyd, J., Scuseria, G. E. & Ernzerhof, M. Hybrid functionals based on a screened Coulomb potential. *J. Chem. Phys.* **118,** 8207 (2003).
54. Krukau, A. V., Vydrov, O. A., Izmaylov, A. F. & Scuseria, G. E. Influence of the exchange screening parameter on the performance of screened hybrid functionals. *J. Chem. Phys.* **125,** 224106 (2006).
55. Carazzolo, G. & Mammi, M. Crystal structure of a new form of polyoxymethylene. *J. Polym. Sci. A* **1,** 965 (1963).
56. Brüesch, P. *Phonons: Theory and Experiments II, of Springer Series in Solid-State Sciences* Vol. 65, Chap. 2, pp 8–64 (Springer: Berlin, 1986).
57. Giannozzi, P. & Baroni, S. Vibrational and dielectric properties of C60 from density-functional perturbation theory. *J. Chem. Phys.* **100,** 8537 (1994).
58. Wang, C. C., Pilania, G. & Ramprasad, R. Dielectric properties of carbon-, silicon-, and germanium-based polymers: A first-principles study. *Phys. Rev. B* **87,** 035103 (2013).
59. Towns, J. *et al.* XSEDE: accelerating scientific discovery. *Comput. Sci. Engin* **16,** 62 (2014).
60. Wigman, L. S., Hart, E. E. & Gombatz, C. IR spectroscopy using disposable polyethylene cards: a replacement for KBr pellets and mulls. *J. Chem. Educ.* **73,** 677 (1996).
61. Peacock, A. *Handbook of Polyethylene: Structures: Properties, and Applications*. 1 ed. (CRC Press: New York, US, 2000).
62. Hikosaka, M. & Seto, T. The order of the molecular chains in isotactic polypropylene crystals. *Polym. J.* **5,** 111 (1973).
63. Kolda, R. R. & Lando, J. B. The effect of hydrogen-fluorine defects on the conformational energy of polytrifluoroethylene chains. *J. Macromol. Sci. Phys.* **11,** 21 (1975).
64. De Rosa, C., Guerra, G., Petraccone, V. & Pirozzi, B. Crystal structure of the emptied clathrate form (δ_e form) of syndiotactic polystyrene. *Macromolecules* **30,** 4147 (1997).
65. Kobayashi, M. *et al.* Synthesis and properties of chemically coupled poly(thiophene). *Synth. Met.* **9,** 77 (1984).
66. Dorset, D. L. Direct determination of polymer crystal structures from fibre and powder X-ray data. *Polymer* **38,** 247 (1997).
67. Fratini, A. V., Cross, E. M., O'brien, J. F. & Adams, W. W. The structure of poly-2,5-benzoxazole (ABPBO) and poly-2,6-benzothiazole (ABPBT) fibers by X-ray diffraction. *J. Macromol. Sci. Phys.* **24,** 159 (1985).
68. Kumpanenko, I. V., Kazaskii, K. S., Ptitsyna, N. V. & Kushnerev, M. Y. Structural study of polymeric 3,3,3-trifluoro-1,2-epoxypropane. *Polym. Sci. USSR* **12,** 930 (1970).

69. Matsubayashi, H., Chatani, Y., Tadokoro, H., Tabata, Y. & Ito, W. Molecular and crystal structure of hexafluoroacetone-ethylene alternating copolymer. *Polym. J.* **9,** 145 (1977).
70. Turner-Jones, A. & Bunn, C. W. The crystal structure of polyethylene adipate and polyethylene suberate. *Acta Cryst* **15,** 105 (1962).
71. Tabor, B. J., Magré, E. P. & Boon, J. The crystal structure of poly-p-phenylene sulphide. *Eur. Polym. J.* **7,** 1127 (1971).
72. Sakakihara, H., Takahashi, Y., H., T., Sigwalt, P. & Spassky, N. Structural studies of the optically active and racemic poly (propylene sulfides). *Macromolecules* **2,** 515 (1969).
73. Tanigami, T. *et al.* Structural studies on ethylene-tetrafluoroethylene copolymer 1. Crystal structure. *Polymer* **27,** 999 (1986).
74. Mencik, Z. The crystal structure of poly(tetramethylene terephthalate). *J. Polym. Sci.: Polym. Phys. Ed.* **13,** 2173 (1975).
75. Jourdan, N., Deguire, S. & Brisse, F. Structural study of linear polyesters. 1. crystal structure of poly(trimethylene sebacate), established from X-ray and electron diffraction data. *Macromolecules* **28,** 8086 (1995).
76. Lando, J. B. & Hanes, M. D. X-ray Analysis of Poly(vinyl fluoride). *Macromolecules* **28,** 1142 (1995).
77. de, P., Daubeny, R. & Bunn, C. W. The crystal structure of polyethylene terephthalate. *Proc. R. Soc. A* **226,** 531 (1954).
78. Hasegawa, R., Kobayashi, M. & Tadokoro, H. Molecular conformation and packing of poly(vinylidene fluoride). Stability of three crystalline forms and the effect of high pressure. *Polym. J.* **591** (1972).
79. De Rosa, C. & Corradini, P. Crystal structure of syndiotactic polypropylene. *Macromolecules* **26,** 5711 (1993).
80. Puterman, M., Kolpak, F. J., Blackwell, J. & Lando, J. B. X-ray structure determination of isotactic poly(2-vinylpyridine). *J. Pol. Sci.: Polym. Phys. Ed.* **15,** 805 (1977).
81. Hobson, R. J. & Windle, A. H. Crystalline structure of atactic polyacrylonitrile. *Macromolecules* **26,** 6903 (1993).
82. Lotz, B. Crystal structure of polyglycine I. *J. Mol. Bio.* **87,** 169 (1974).
83. Kakida, H., Chatani, Y. & Tadokoro, H. Crystal structure of poly(m-phenylene isophthalamide). *J. Polym. Sci.: Polym. Phys. Ed.* **14,** 427 (1976).
84. Kobayashi, N. *et al.* Chain Distortion of m-Linked Aromatic Polymers: Poly(m-phenylene) and Poly(m-pyridine). *Macromolecules* **37,** 7986 (2004).
85. Tashiro, K. *et al.* Confirmation of the crystal structure of poly(p-phenylene benzobisoxazole) by the X-ray structure analysis of model compounds and the energy calculation. *J. Polym. Sci. Part B: Polym. Phys.* **39,** 1296 (2001).

Data Citations

1. Huan, T. D. *et al. Dryad Digital Repository*. http://dx.doi.org/10.5061/dryad.5ht3n (2015).
2. Huan, T. D. *et al. NoMaD Repository*. http://dx.doi.org/10.17172/NOMAD/2016.01.27-1 (2016).

Acknowledgements

The present work was supported by Multidisciplinary University Research Initiative (MURI) grant from the Office of Naval Research under Award No. N00014-10-1-0944. G.P. acknowledges the support provided by U.S. Department of Energy through the LANL/LDRD Program's Director's postdoctoral fellowship. A.M.K. would like to thank Turab Lookman at Los Alamos National Laboratory for providing access to computational resources. Computational work was made possible through XSEDE computational resource allocation number TG-DMR080058N[59].

Author Contributions

T.D.H. wrote the paper with inputs and critique from all authors. Data accumulations and first-principles calculations were performed by T.D.H., A.M.K., C.K., V.S., and G.P. Dataset was refined, validated, and finalized by T.D.H. Data repository (Khazana) was designed and maintained by C.K. This project was initiated, designed and supervised by R.R. Contributions from T.D.H. and A.M.K. are equal.

Additional Information

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

How to cite this article: Huan, T. D. *et al.* A polymer dataset for accelerated property prediction and design. *Sci. Data* 1:140003 doi: 10.1038/sdata.2016.12 (2016).