SCIENTIFIC DATA

OPEN

Data Descriptor: A hybrid organic-inorganic perovskite dataset

Chiho Kim[1], Tran Doan Huan[1], Sridevi Krishnan[1] & Rampi Ramprasad[1]

Received: 2 February 2017
Accepted: 21 March 2017
Published: 9 May 2017

Hybrid organic-inorganic perovskites (HOIPs) have been attracting a great deal of attention due to their versatility of electronic properties and fabrication methods. We prepare a dataset of 1,346 HOIPs, which features 16 organic cations, 3 group-IV cations and 4 halide anions. Using a combination of an atomic structure search method and density functional theory calculations, the optimized structures, the bandgap, the dielectric constant, and the relative energies of the HOIPs are uniformly prepared and validated by comparing with relevant experimental and/or theoretical data. We make the dataset available at Dryad Digital Repository, NoMaD Repository, and Khazana Repository (http://khazana.uconn.edu/), hoping that it could be useful for future data-mining efforts that can explore possible structure-property relationships and phenomenological models. Progressive extension of the dataset is expected as new organic cations become appropriate within the HOIP framework, and as additional properties are calculated for the new compounds found.

Design Type(s)	data integration objective • database creation objective
Measurement Type(s)	material properties
Technology Type(s)	computational modeling technique
Factor Type(s)	cation
Sample Characteristic(s)	

[1] Institute of Materials Science, University of Connecticut, 97 North Eagleville Rd., Unit 3136, Storrs, Connecticut 06269, USA. Correspondence and requests for materials should be addressed to R.R. (email: rampi.ramprasad@uconn.edu).

Background and Summary

Perovskites belong to a class of inorganic crystals with chemical formula ABX_3, sharing the same structure with calcium titanate $CaTiO_3$. In such a perovskite structure, the inorganic cations A and B are coordinated by 12 and 6 anions X, respectively. By substituting an organic cation for A, the first hybrid organic-inorganic perovskites (HOIPs), namely $CH_3NH_3PbX_3$ (X = Cl, Br, I), were synthesized and characterized in 1978 (ref. 1). HOIPs remained largely unnoticed until the first successful application of $CH_3NH_3PbX_3$ (X = Cl, Br) as photovoltaic absorbers with a power conversion efficiency of 3.8% in 2009 (ref. 2). An enormous number of experimental and computational efforts have then been devoted to optimizing some halide-based HOIPs, e.g., $CH_3NH_3PbI_3$, $HC(NH_2)_2PbI_3$, and $CH_3NH_3SnI_3$, for photovoltaic applications[3–6]. Currently, $CH_3NH_3PbI_3$ and $HC(NH_2)_2PbI_3$ have taken a leading position in providing high performance (reaching 20.1% in the conversion efficiency)[7] and low fabrication cost[3–6].

In fact, there are plenty of choices for the sites A, B, and X in a HOIP. At the site A, methylammonium CH_3NH_3[3–5,8], formamidnium $HC(NH_2)_2$[7,9], and many more[6], have been realized. Cations B can be Pb or Sn while the halogens Br, I, and Cl can be used for X[1,2]. Moreover, the introduction of an organic cation A into the perovskite structure can give raise of many different structural motifs[6,10–12], making the class of halide-based HOIPs highly diverse. Rapidly and thoroughly screening this un-explored domain of the chemical space, for instance, with the emerging data-driven approaches[13–25], may reveal new promising compounds potentially meeting the pressing need for lead-free perovskite solar cell materials[26].

This contribution aims at taking an initial step towards the creation of a comprehensive database of HOIPs, which may be useful for this goal. In fact, this idea has recently been emerging with some datasets of hybrid organic/inorganic perovskites, prepared at some level of computations[27,28]. Our dataset, which contains 1,346 HOIPs, is prepared uniformly at the level of density functional theory (DFT)[29,30] from the initial structures predicted by the minima-hoping method[31,32]. For each material, the equilibrium structure, the relative energies ($\varepsilon_{\mathrm{rel}_1}$ and $\varepsilon_{\mathrm{rel}_2}$, computed with respect to different energy references as described in **Numerical calculations** Section), the atomization energy ($\varepsilon_{\mathrm{at}}$), the dielectric constant (ε), and the direct or indirect energy bandgap (E_{g}) are reported. This dataset, which is available at Dryad Digital Repository, NoMaD Repository, and Khazana Repository, can readily be expanded in multiple ways, i.e., new properties can be calculated from the provided structures, and new HOIPs can also be progressively added. We expect that this dataset can supply a playground for future machine learning based work in this active research area.

Methods

Workflow

Figure 1 summarizes the workflow of the dataset preparation. This procedure starts by collecting 16 organic (molecular) cations A^{+1}, all of which have been considered in the literature[1,6,7,12]. Each of these 16 cations, shown in Fig. 2, is placed at the site A of the $ASnI_3$-based perovskites. This is the starting point for various structure prediction simulations, performed with the minima-hopping method[31,32]. The low-energy structures predicted for $ASnI_3$ are subjected to a preliminary filtering step, keeping 135 prototype structures that are different in the DFT energy and the volume (these quantities are estimated on a not-so-high accuracy level used for the searches). Next, we expand the set of 135 structures by substituting either Ge or Pb for Sn, and, similarly, by substituting either F, Cl, or Br for I. The resulted 1,620 (initial) structures were optimized by DFT at the desired level of accuracy (described in **Numerical calculations**

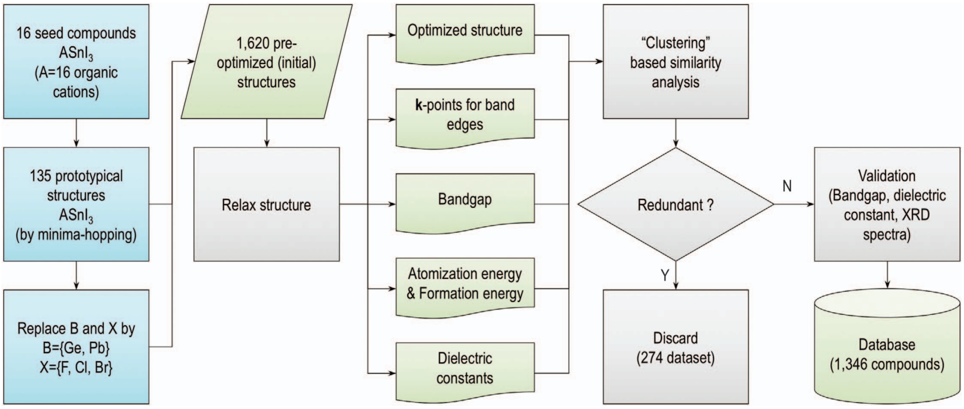

16 seed compounds ASnI3 (A=16 organic cations)
135 prototypical structures ASnI3 (by minima-hopping)
Replace B and X by B={Ge, Pb} X={F, Cl, Br}
1,620 pre-optimized (initial) structures
Relax structure
Optimized structure
k-points for band edges
Bandgap
Atomization energy & Formation energy
Dielectric constants
"Clustering" based similarity analysis
Redundant ?
N
Y
Validation (Bandgap, dielectric constant, XRD spectra)
Discard (274 dataset)
Database (1,346 compounds)

Figure 1. Scheme for preparing the dataset of hybrid organic-inorganic perovskites. Minima-hopping is a structure prediction method that was used for generating an initial set of 135 $ASnI_3$ prototypical structures (where A stands for 16 organic cations), which were used as seeds for the creation of the remaining compounds.

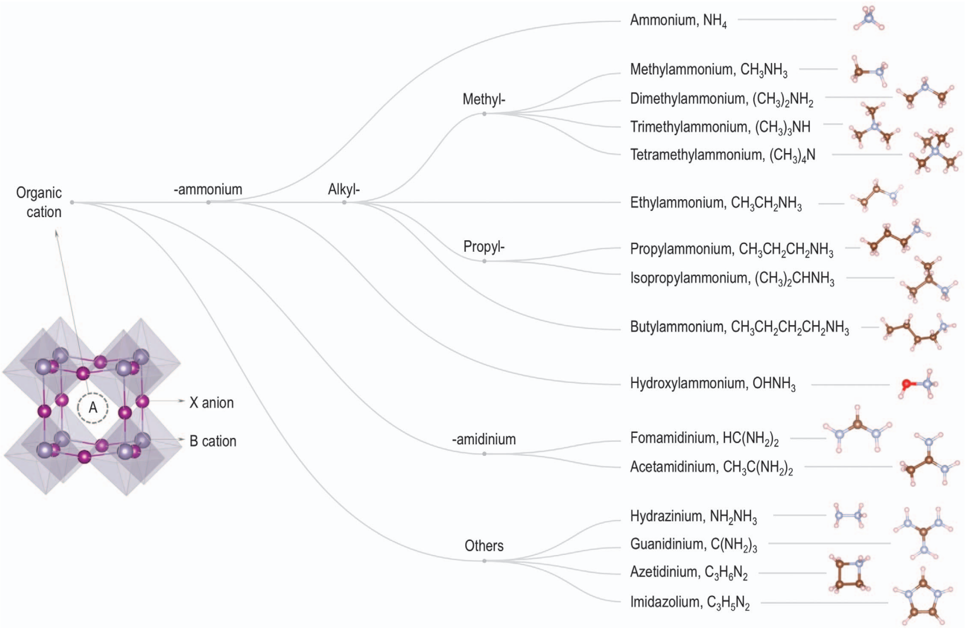

Organic cation
-ammonium
Alkyl-
Methyl-
Propyl-
-amidinium
Others
Ammonium, NH$_4$
Methylammonium, CH$_3$NH$_3$
Dimethylammonium, (CH$_3$)$_2$NH$_2$
Trimethylammonium, (CH$_3$)$_3$NH
Tetramethylammonium, (CH$_3$)$_4$N
Ethylammonium, CH$_3$CH$_2$NH$_3$
Propylammonium, CH$_3$CH$_2$CH$_2$NH$_3$
Isopropylammonium, (CH$_3$)$_2$CHNH$_3$
Butylammonium, CH$_3$CH$_2$CH$_2$CH$_2$NH$_3$
Hydroxylammonium, OHNH$_3$
Fomamidinium, HC(NH$_2$)$_2$
Acetamidinium, CH$_3$C(NH$_2$)$_2$
Hydrazinium, NH$_2$NH$_3$
Guanidinium, C(NH$_2$)$_3$
Azetidinium, C$_3$H$_6$N$_2$
Imidazolium, C$_3$H$_5$N$_2$
A
X anion
B cation

Figure 2. Ball and stick representations of 16 organic cations considered in the HOIP dataset. Carbon, hydrogen, oxygen and nitrogen atoms are shown in dark brown, light pink, red, and gray, respectively.

Section), yielding the relative energies and the atomization energies. Then, the band edge positions in the **k** space, the energy bandgap, and the dielectric constant were calculated for the optimized structures. A post-filtering step is finally performed on the whole dataset, removing redundancy (this time, redundancy is identified at the desired accuracy level of DFT computations), keeping 1,346 distinct data points (summarized in Table 1). Whenever possible, our calculated results are compared with those computed and/or measured data. Relaxed structures of all the materials are finally converted into the crystallographic information format (cif) using the `pymatgen` library[33].

Initial structure accumulation

As briefly demonstrated in the **Workflow** section, our dataset is built up from 135 prototype structures obtained by searching for low-energy structures of 16 HOIPs with chemical formulae ASnI$_3$ (in fact, prototype structures of any material can be searched). In the minima-hopping structure prediction simulations, the DFT-level evergy is used to construct the potential energy surface (PES) of the composition[31,32]. Starting from an initial structure, low-energy minima of the PES are then searched by alternatively performing DFT-based local optimization runs (to locate the nearby minima) and molecular dynamics runs (to escape the identified minima). Thanks to some feedback mechanisms implemented, structure searches using this method is biased, giving some preference to the low-energy domains of the PES. Because of the large number of minima, the searches were performed at a given not-so-high accuracy level of DFT energy, and the minima identified in this step were then refined at the desired level. The power of the minima-hopping has been demonstrated over several classes of crystalline solids[34–36], including three SnI$_3$-based HOIPs[12].

For each of 16 ASnI$_3$ HOIPs, numerous low-energy structures identified are subjected to a filtering step, keeping only those that are different by at least 5 meV/atom in the DFT energy and at least 0.1 Å^3/atom in the structure volume. After the filtering step, 135 prototypical structures of 16 HOIPs were selected, three of which are shown in Fig. 3. In case of isotropic organic cations such as tetramethylammonium, a cubic-like cage formed by the network of Sn and I ions is stabilized in a three-dimensional structure. For the case of anisotropic or polar organic cations, the framework deforms into the two-dimensional planar or pillar motif. More structural variation is possible to be found from further structure searching using different organic cations and/or slightly nonstoichiometric composition in the HOIP system[6,37]. By substituting either Ge or Pb for Sn, and substituting either Cl, F, or Br for I, 1,620 structures of 192 *chemically distinct* HOIPs were obtained. They are the initial structures used to build up the HOIP dataset.

Numerical calculations

General scheme. Our calculations are performed within the DFT[29,30] formalism, using the projector augmented-wave (PAW) method[38] as implemented in the Vienna *Ab initio* Simulation Package (vasp)[39–42]. The default accuracy level of our calculations is 'Accurate', specified by setting PREC = Accurate in all the runs with vasp. The basis set includes plane waves with kinetic energies up to 400 eV, as recommended by

Organic cation A	Cation B and anion X												Total
	Ge				Sn				Pb				
	F	Cl	Br	I	F	Cl	Br	I	F	Cl	Br	I	
Ammonium	2	2	4	3	3	4	3	4	2	2	3	3	35
Methylammonium	6	5	6	6	6	6	6	6	6	5	5	6	69
Dimethylammonium	5	7	9	8	9	8	7	8	5	7	7	7	87
Trimethylammonium	7	8	9	9	7	10	11	11	9	11	9	12	113
Tetramethylammonium	2	3	3	2	1	3	3	2	1	3	3	3	29
Ethylammonium	9	10	11	12	11	11	12	12	12	10	10	11	131
Propylammonium	8	11	13	12	10	13	13	12	11	10	11	13	137
Isopropylammonium	9	8	9	10	9	9	11	9	12	11	8	10	115
Butylammonium	4	3	3	4	4	3	3	3	2	4	4	4	41
Hydroxylammonium	7	7	7	7	5	6	7	7	7	6	7	7	80
Formamidinium	2	3	3	3	2	3	3	3	2	3	3	3	33
Acetamidinium	6	5	6	6	6	6	7	6	5	6	6	6	71
Hydrazinium	8	10	11	11	8	11	12	11	8	9	8	9	116
Guanidinium	3	3	3	3	2	2	3	3	2	3	3	3	33
Azetidinium	13	14	16	15	14	17	16	16	9	13	13	15	171
Imidazolium	6	6	8	8	9	8	7	9	6	7	5	6	85
Total	97	105	121	119	106	120	124	122	99	110	105	118	1,346

Table 1. Summary of the data subclasses in the hybrid organic-inorganic perovskites dataset.

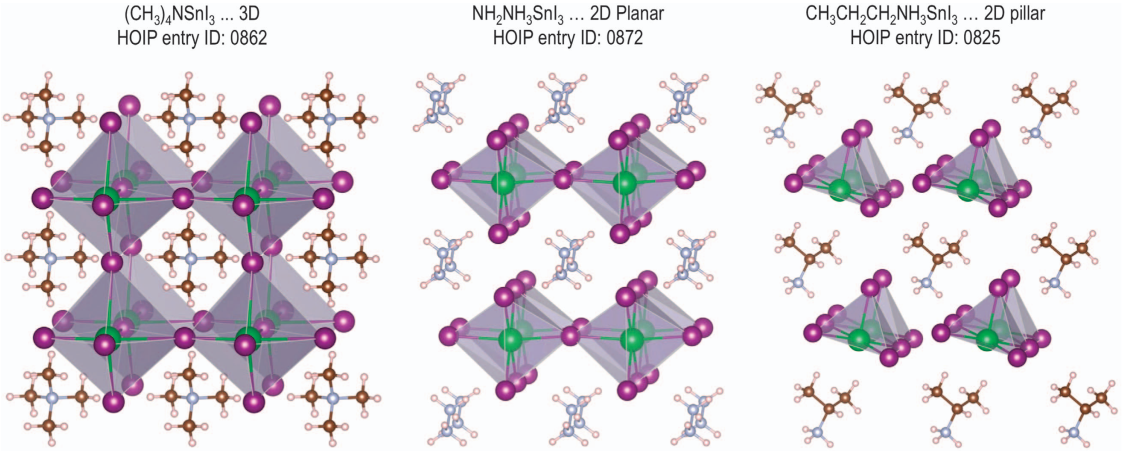


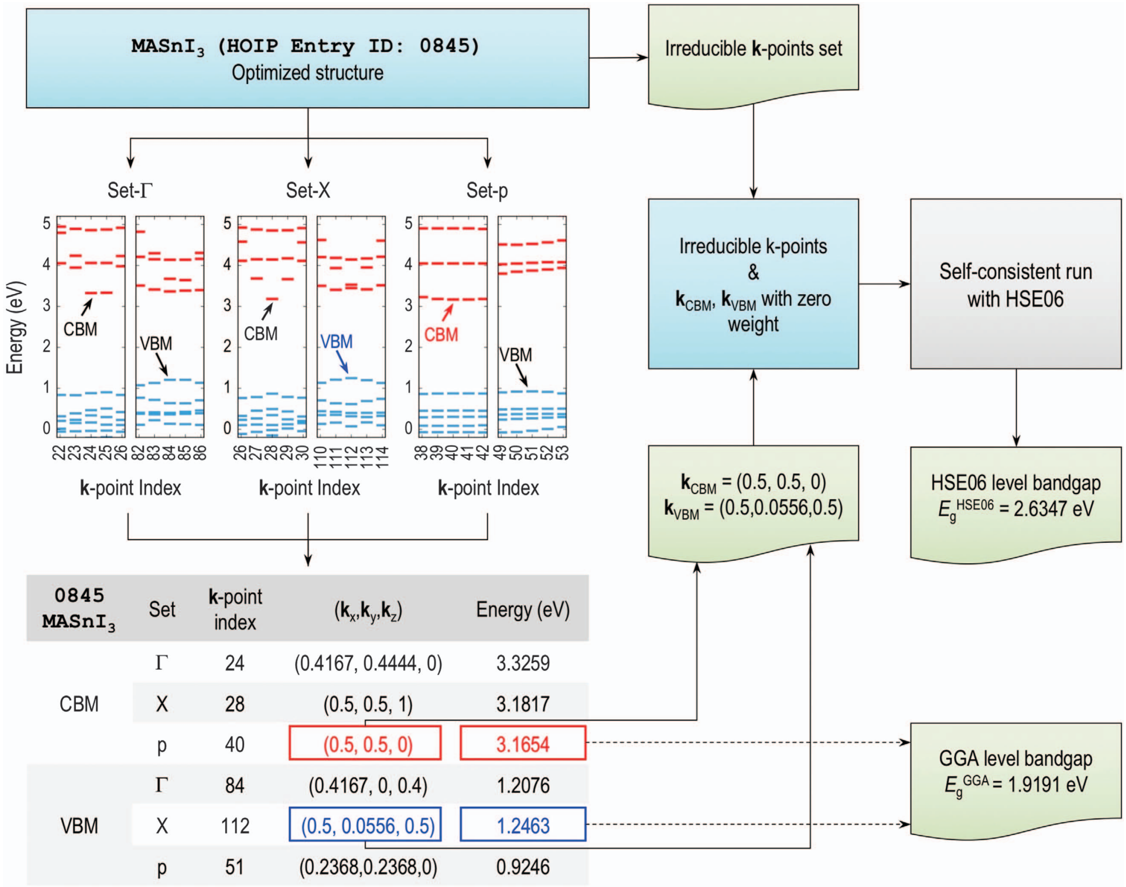
Figure 3. Lowest energy structures of tetramethylammonium, hydrazinium, and propylammonium tin iodide showing three prototypical conformations of organic-inorganic hybrid perovskites. Carbon, hydrogen, nitrogen, tin and iodine atoms are shown in dark brown, light pink, gray, green and purple, respectively.

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 2. The van der Waals dispersion interactions are estimated with the non-local density functional vdW-DF2 (ref. 43). The generalized gradient approximation (GGA) functional associated with vdW-DF2, i.e., refitted Perdew-Wang 86 (rPW86)[44], was used for the exchange-correlation (XC) energies. For all the calculations, except bandgap determination, we sample the Brillouin zones, which are significantly different in shape for the different compounds, by an equispaced (with the spacing of $h_k = 0.20\ \text{Å}^{-1}$), Γ-centered Monkhorst-Pack[45] **k**-points mesh. The equilibration of the examined structures is assumed when the atomic forces are below 0.01 eV/Å. This numerical scheme is consistent with that we used for preparing the polymer dataset reported in ref. 35.

Bandgap determination. The bandgap E_g is perhaps the most desired physical property of HOIPs. Within DFT, E_g is determined as the energy difference between the conduction band minimum (CBM)

Element	POTCAR	Element	POTCAR	Element	POTCAR
Bromine	Br	Carbon	C	Chlorine	Cl
Fluorine	F	Germanium	Ge_d	Hydrogen	H
Iodine	I	Nitrogen	N	Oxygen	O
Lead	Pb_d	Tin	Sn_d		

Table 2. VASP PAW potentials of the elements used for calculations in this work.



0845 MASnI$_3$	Set	k-point index	(k_x,k_y,k_z)	Energy (eV)
	Γ	24	(0.4167, 0.4444, 0)	3.3259
CBM	X	28	(0.5, 0.5, 1)	3.1817
	p	40	(0.5, 0.5, 0)	3.1654
	Γ	84	(0.4167, 0, 0.4)	1.2076
VBM	X	112	(0.5, 0.0556, 0.5)	1.2463
	p	51	(0.2368,0.2368,0)	0.9246

Figure 4. Scheme for calculation of the bandgap of hybrid organic-inorganic perovskites at GGA and HSE06 level of theories. Data entry 0,845 (MASnI$_3$; CH$_3$NH$_3$SnI$_3$, Khazana ID: 2,695) is used for demonstration. Set-Γ, Set-X and Set-p correspond to the **k**-points sets generated within Γ-centered mesh, X-centered mesh, and high symmetry path for P_1 group.

and the valence band maximum (VBM), identified on a given **k**-point mesh. For a solid with an arbitrary primitive cell, the locations of VBM and CBM are generally not known beforehand, and the **k**-point mesh should be very dense in order to locate the band edges accurately. With a mesh of this type, the computation of E_g using the Heyd-Scuseria-Ernzerhof (HSE06)[46,47] exchange-correlation functional, the level of DFT at which the calculated bandgap is expected to be close to the real bandgap, is computationally prohibitive. Although such a computation at the GGA level of DFT is feasible, E_g is generally underestimated by 30% or more[48].

The conduction bands and the valence bands computed at the GGA and HSE06 levels of DFT are essentially similar in the shape. However, they are shifted as a whole with respect to each other and to the true electronic structrures (see, for example, ref. 49). Therefore, our bandgap determination procedure, shown in Fig. 4, includes two steps. First, the locations of VBM and CBM are searched at the GGA level on three different dense **k**-point meshes. The first two meshes (one centered at Γ = (0,0,0) and the other centered at X = (0.5, 0.5, 0.5)) are equispaced with $h_k = 0.15$ Å^{-1}, while the third mesh contains **k**-points distributed along Γ-X-M-Γ-R-M-X-R, the path that has widely been used to represent the electronic band structrure of HOIPs[12,50]. In the second step, the positions of VBM and CBM identified in the first step are used with zero weight for sampling the Brillouin zones using a Monkhorst-Pack **k**-point mesh with $h_k = 0.20$ Å^{-1}, hereby determining the energy difference between CBM and VBM at the HSE06 level of DFT. Although this procedure needs some extra work, we expect that the bandgap computed for HOIPs with an arbitrary primitive cell is reliable.

Atomization and relative energies definitions. The atomization energy of each of these compounds are calculated as

$$\varepsilon_{\mathrm{at}} = E_{\mathrm{ABX_3}} - \sum_i n_i E_i \tag{1}$$

where $E_{\mathrm{ABX_3}}$ is the energy of the HOIP and n_i and E_i are the number and the energy of an isolated atom of the element i respectively. We also report two kinds of relative energies with respect to the atomic constituents and solid constituents.

$$\varepsilon_{\mathrm{rel_2}} = E_{\mathrm{ABX_3}} - E_{\mathrm{A'}} - E_{\mathrm{B}} - \frac{3}{2}E_{\mathrm{X_2}} - \frac{1}{2}E_{\mathrm{H_2}} \tag{2}$$

$$\varepsilon_{\mathrm{rel_2}} = E_{\mathrm{ABX_3}} - E_{\mathrm{A'}} - E_{\mathrm{BX_2}} - E_{\mathrm{HX}} \tag{3}$$

where $E_{\mathrm{A'}}$, E_{B}, $E_{\mathrm{X_2}}$, and $E_{\mathrm{H_2}}$ are the energies of isolated neutral organic molecule A, metallic crystals B, isolated X_2, and H_2 molecules respectively. $E_{\mathrm{BX_2}}$ and E_{HX} are the energies of the metallic halides (BX_2) and hydrogen halides (HX), respectively. For the case of tetramethylammonium cation ($C_4H_{12}N^+$), the energy of neutral trimethylamine (C_3H_9N) was used for $E_{\mathrm{A'}}$, and the energies of the molecules C_2H_6 and CH_3X are used instead of $E_{\mathrm{H_2}}$ and E_{HX} in equations (2) and (3), respectively.

Post-filtering

The preliminary filtering step is performed only on prototypical structures ($ASnI_3$) based on their DFT energy and bandgap estimated during the structure prediction runs with a limited accuracy. Therefore, an additional filtering step is performed on the whole relaxed structures from 1,620 initial structures to remove any possible redundancy. Within this step, all cases with the same chemical composition but different by less than 2% in volume of unit cell Ω, E_g, $\varepsilon_{\mathrm{at}}$, ϵ_{elec} and ϵ_{ion}, are clustered. All the clustered points were inspected visually, keeping only those materials that are distinct. At the end of this step, we are left with 1,346 distinct compounds (also summarized in Table 1). These compounds constitute our final dataset.

Data Records

The complete dataset of HOIP materials can be downloaded as a tarball or can be accessed via Dryad Digital Repository (Data Citation 1) and Khazana Repository (`http://khazana.uconn.edu/`). 1,346 compounds in our final dataset are recorded in Khazana ID from 1,851 to 3,197. All 8,076 ($=1{,}346 \times 6$) DFT runs of the whole dataset (for each structure, there are 6 runs, including relax, dielectric, GGA bandgap with Γ-centered mesh, GGA bandgap with X-centered mesh, GGA bandgap with **k**-points distributed along Γ-X-M-Γ-R-M-X-R, and HSE06 bandgap) are hosted by NoMaD Repository (Data Citation 2).

File format

The information reported in the dataset for a given material is stored in a file, named as N.cif, where N is a cardinal number 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 many visualization software. Other information, including the calculated properties, is provided as the comments lines in the second part of the file as follows (for the example of N = 845).

```
# HOIP entry ID:                    0845
# Khazana ID:                       2695
# Organic cation source:            T.D.Huan et al., Phys. Rev. B 93,094105(2016)
# Label:                            Methylammonium Tin Iodide
# Material class:                   Hybrid organic-inorganic perovskite (MC_ino)
# Geometry class:                   Bulk crystalline materials (GC_cry)
# Organic cation chemical formula: CH3NH3
# Number of atom types:             5
# Total number of atoms:            12
# Atom types:                       C H N Sn I
# Number of each atom:              1 6 1 1 3
# Bandgap, HSE06 (eV):              2.6347
# Bandgap, GGA (eV):                1.9191
# Kpoint for VBM:                   0.5, 0.0556, 0.5
# Kpoint for CBM:                   0.5, 0.5, 0
# Dielectric constant, electronic:4.8562
# Dielectric constant, ionic:       13.0716
# Dielectric constant, total:       17.9278
# Refractive index:                 2.2037
# Atomization energy (eV/atom):     -3.9099
```

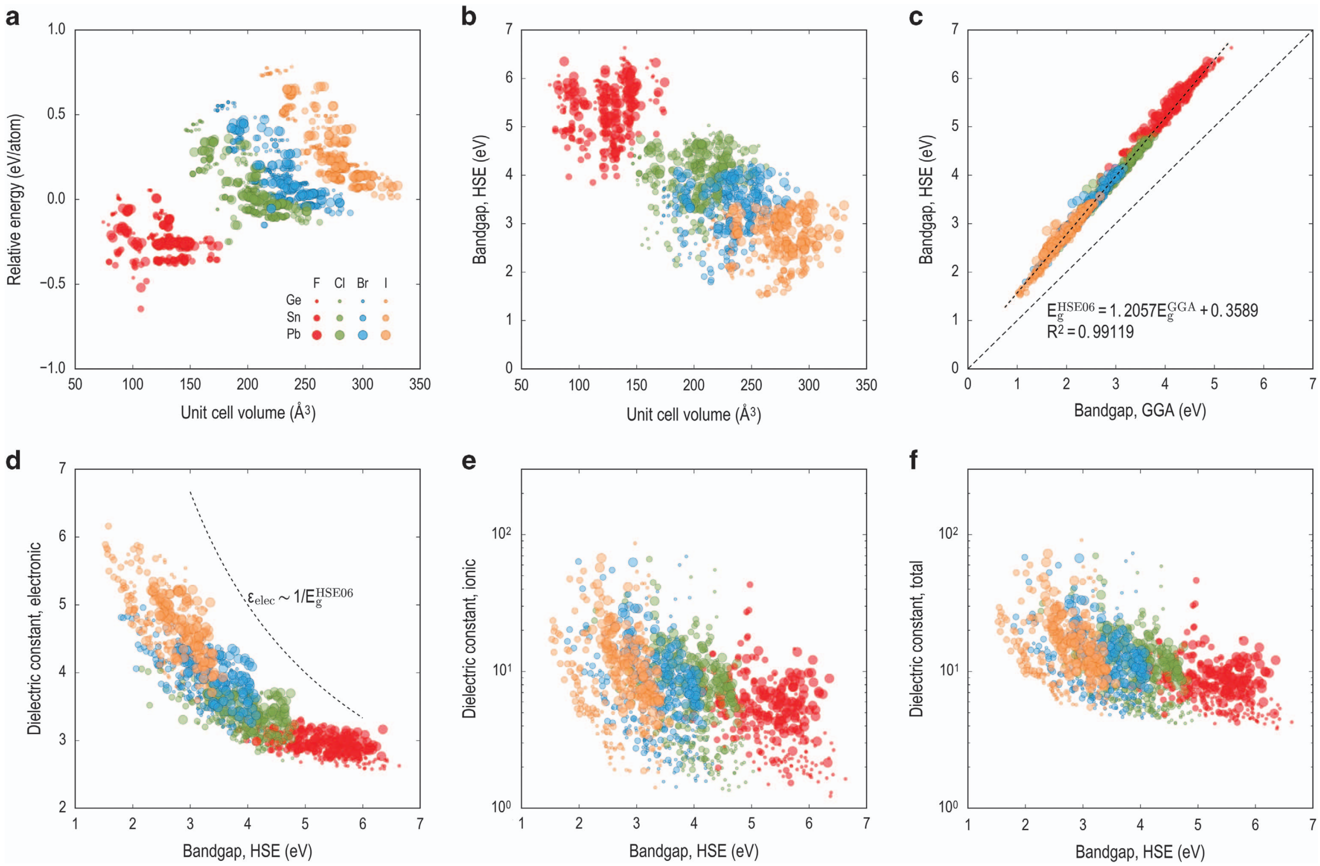


Figure 5. A summary of the HOIP dataset based on the calculated volume of unit cell Ω, relative energy ε_{rel_1}, GGA level bandgap E_g^{GGA}, HSE level bandgap E_g^{HSE06}, and the dielectric constants ϵ_{elec}, ϵ_{ion}, and $\epsilon = \epsilon_{elec} + \epsilon_{ion}$. The panels show (**a**) unit cell volume vs relative energy, (**b**) unit cell volume vs HSE bandgap, (**c**) GGA bandgap vs HSE bandgap, (**d**) HSE bandgap vs electronic dielectric constant, (**e**) HSE bandgap vs ionic dielectric constant, and (**f**) HSE bandgap vs total dielectric constant. In each plot, the color and size of the symbols are coded following the figure keys shown in plot (**a**).

```
# Relative energy1 (eV/atom):      0.2785
# Relative energy2 (eV/atom):      0.4387
# Volume of the unit cell (A^3):   251.45
# Density (g/cm^3):                3.51.
```

While most of the keywords are clear, we used keyword `Label` to provide more detail information of the HOIP compounds, which includes the common name of A organic cation, B cation and X anion. The origin of the formula and structure of organic cations is provided in the keyword `Organic cation source`. Keywords `Material class` and `Geometry class` are set to be 'Hybrid organic-inorganic perovskite' and 'Bulk crystalline materials', respectively.

Graphical summary of the dataset

We visualize the calculated quantities in the property space as shown in Fig. 5. Because the relative energy, unit cell volume of the compound, bandgap and dielectric constant are the primary properties reported by this dataset, six plots, namely $\Omega - \varepsilon_{rel_1}$, $\Omega - E_E^{gHSE06}$, $E_g^{GGA} - E_g^{HSE06}$, $E_g^{HSE06} - \epsilon_{elec}$, $E_g^{HSE06} - \epsilon_{ion}$, and $E_g^{HSE06} - \epsilon$, were shown. Compounds containing different A cations and X anions are represented using different colors and size of the symbols to clarify the role of the chemical contents in controlling the properties of the HOIP.

It can be clearly seen that the dataset is clustered based on the X anions, showing the sequence of F, Cl, Br and I. As shown in Fig. 5a most of F containing HOIP compounds are more favorable to be formed as measured by the relative energy regardless of the A cation contents. Bandgap and unit cell volume are strongly correlated mainly because the electronegativity and the ionic radii of X anions significantly differ for F, Cl, Br and I. Simple and strong correlation between GGA and HSE level bandgap is found as a linear function with scale factor of ~1.2 as shown in Fig. 5c. Small bandgap values varying from 1.5 eV

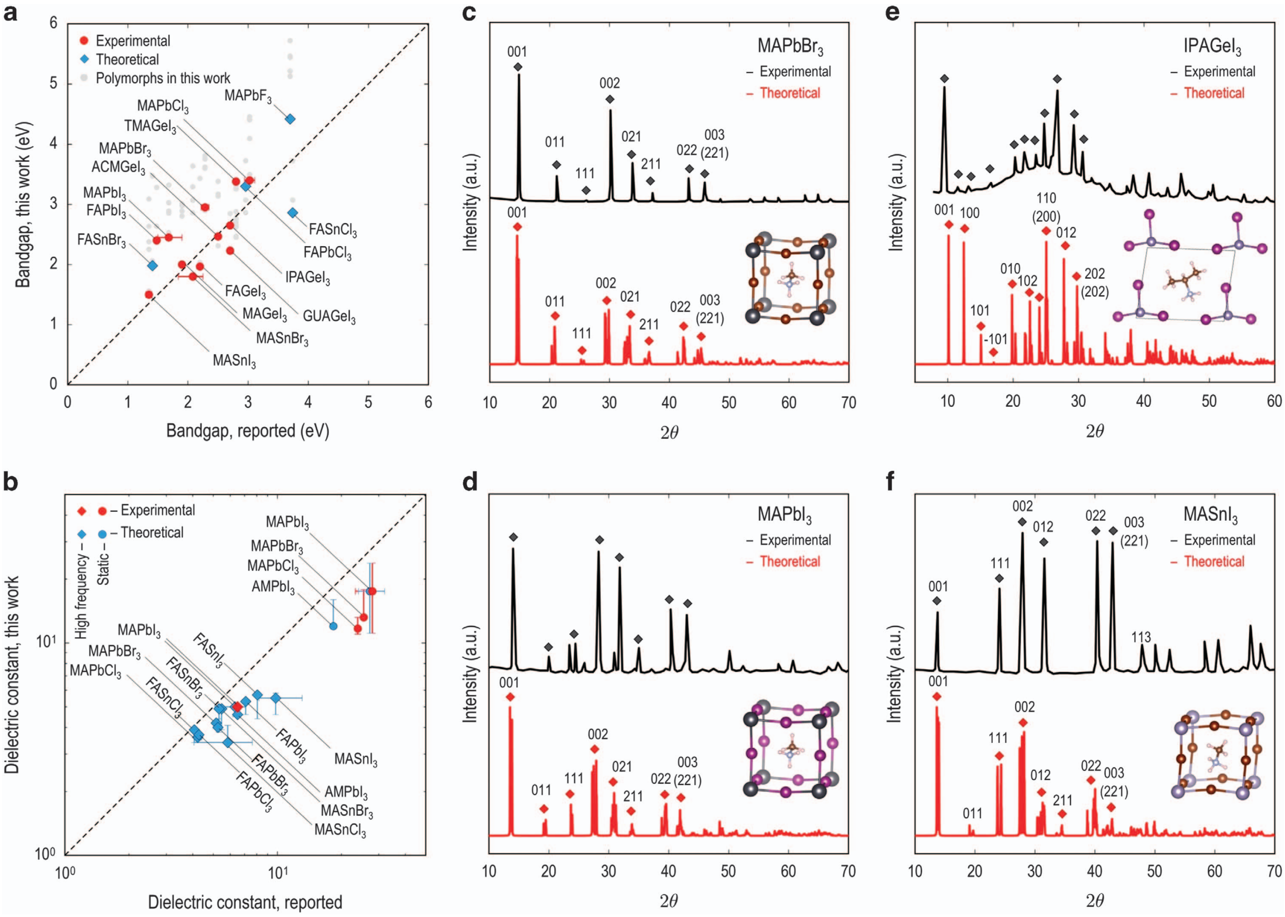


Figure 6. Validation of data computed for some HOIPs by comparing it with the measured data available. Bandgap and dielectric constants computed for the low-energy structures of these compounds are plotted in (**a**,**b**) vs. those experimentally measured, respectively. In these panels, the lowest-energy structure of each HOIP is indicated by a colored symbol while data from the energetically competing structures are shown in gray (**a**) or given within an error bar (**b**). Experimental data of bandgap and dielectric constants of these HOIPs is obtained from refs 8,64–73 and refs 74–83, respectively. In (**c**–**f**), the simulated and measured XRD spectra for $MAPbBr_3$[65], $MAPbI_3$[84], $IPAGeI_3$[73], and $MASnI_3$[5,85], are shown. The reported index of reflection orientation is given on top of each significant peak.

to 1.6 eV, favorable for photovoltaic application, was found for SnI_3 containing HOIP compounds including $CH_3NH_3SnI_3$, $NH_3NH_2I_3SnI_3$, $C_3H_8NSnI_3$. A limit of the form $\epsilon_{elec}{\sim}1/E_g^{HSE06}$ shown in Fig. 5 (d) has also been demonstrated for other classes of materials in the literature[13,35,36,51–62].

Technical Validation

The relative energy computed via equation (2) is physically relevant to examine the relative stability useful for future studies of new HOIPs. As the dataset contains theoretically stable structures, we used the bandgap, dielectric constant, and XRD pattern with Cu Kα (1.54056 Å) for the validation of the calculations. Since available experimental studies for HOIPs seem to be limited to a small subset of the combinatorial possibilities, a small number of experimental bandgap could only be collected from available resources. These correspond to compounds containing acetamidinium (ACM, $C_2H_7N_2$), formamidinium (FA, CH_5N_2), guanidinium (GUA, CH_6N_3), isopropylammonium (IPA, $C_3H_{10}N$), methylammonium (MA, CH_3NH_3), and tetramethylammonium (TMA, $C_4H_{12}N$). Four computed bandgaps are also included in the comparison set. As shown in Fig. 6a, the calculated bandgap for the most stable structure of each case (marked as color coded symbols) agrees well with the data from previous studies. (gray symbols correspond to less stable polymorphs).

In order to further validate the HOIP dataset, experimentally measured and theoretically calculated dielectric constants for both high frequency and static regime are collected and compared with computed dielectric constants. The information is available for a limited number of HOIPs with MA and FA organic cations. Since the computation of dielectric constant using DFPT is highly sensitive to the numerical accuracy of the vibration frequency we used rather tight convergence criterion for the change of total energy by 10^{-8} eV. Figure 6b shows the excellent agreement between previously reported and computed dielectric constants for the selected HOIPs. Finally, we show the XRD spectra calculated for four HOIPs, including $MAPbBr_3$, $MAPbI_3$, $IPAGeI_3$ and $MASnI_3$ in Fig. 6c–f. Each of them is compared with the corresponding measured XRD patterns showing comparable agreement that can be regarded as supportive validation of computational schemes.

Usage Notes

This dataset, which includes 1,346 HOIPs, has been consistently prepared using first-principles calculations. While the HSE06 bandgap E_g^{HSE06} is believed to be fairly close to the true bandgap of the materials, the GGA-rPW86 bandgap is also reported for completeness and for further possible analysis. The reported atomization energy and the dielectric constants are also expected to be accurate.

References

1. Weber, D. $CH_3NH_3PbX_3$, a Pb(II)-System with Cubic Perovskite Structure. *Z. Naturforsch., B: J. Chem. Sci.* **33,** 1443–1445 (1978).
2. Kojima, A., Teshima, K., Shirai, Y. & Miyasaka, T. Organometal Halide Perovskites as Visible-Light Sensitizers for Photovoltaic Cells. *J. Am. Chem. Soc.* **131,** 6050 (2009).
3. Burschka, J. *et al.* Sequential deposition as a route to high-performance perovskite-sensitized solar cells. *Nature* **499,** 316–319 (2013).
4. Liu, M., Johnston, M. B. & Snaith, H. J. Efficient planar heterojunction perovskite solar cells by vapour deposition. *Nature* **501,** 395–398 (2013).
5. Hao, F., Stoumpos, C. C., Cao, D. H., Chang, R. P. H. & Kanatzidis, M. G. Lead-free solid-state organic-inorganic halide perovskite solar cells. *Nature Photon* **8,** 489–494 (2014).
6. Saparov, B. & Mitzi, D. B. Organic-Inorganic Perovskites: Structural Versatility for Functional Materials Design. *Chem. Rev.* **116,** 4558–4596 (2016).
7. Yang, W. S. *et al.* High-performance photovoltaic perovskite layers fabricated through intramolecular exchange. *Science* **348,** 1234–1237 (2015).
8. Baikie, T. *et al.* Synthesis and crystal chemistry of the hybrid perovskite $(CH_3NH_3)PbI_3$ for solid-state sensitised solar cell applications. *J. Mater. Chem. A* **1,** 5628–5641 (2013).
9. Mitzi, D. B. & Liang, K. Synthesis, Resistivity, and Thermal Properties of the Cubic Perovskite $NH_2CH=NH_2SnI_3$ and Related Systems. *J. Solid State Chem.* **134,** 376–381 (1997).
10. Xu, Z. & Mitzi, D. B. $[CH_3(CH_2)_{11}NH_3]SnI_3$:- A Hybrid Semiconductor with MoO3-type Tin(II) Iodide Layers. *Inorg. Chem.* **42,** 6589–6591 (2003).
11. Xu, Z., Mitzi, D. B. & Medeiros, D. R. $[(CH_3)_3NCH_2CH_2NH_3]SnI_4$:-A Layered Perovskite with Quaternary/Primary Ammonium Dications and Short Interlayer Iodine-Iodine Contacts. *Inorg. Chem.* **42,** 1400–1402 (2003).
12. Huan, T. D., Tuoc, V. N. & Minh, N. V. Layered structures of organic/inorganic hybrid halide perovskites. *Phys. Rev. B* **93,** 094105 (2016).
13. Mannodi-Kanakkithodi, A. *et al.* Rational co-design of polymer dielectrics for energy storage. *Adv. Mater.* **28,** 6277–6291 (2016).
14. Huan, T. D., Mannodi-Kanakkithodi, A. & Ramprasad, R. Accelerated materials property predictions and design using motif-based fingerprints. *Phys. Rev. B* **92,** 014106 (2015).
15. Mueller, T., Kusne, A. G. & Ramprasad, R. in *Reviews in Computational Chemistry* Vol. 29 (ed. Parrill A. L. & Lipkowitz K. B.) Ch. 4 (John Wiley & Sons, Inc., 2016).
16. Mannodi-Kanakkithodi, A., Pilania, G., Ramprasad, R., Lookman, T. & Gubernatis, J. E. Multi-objective optimization techniques to design the Pareto front of organic dielectric polymers. *Comput. Mater. Sci.* **125,** 92–99 (2016).
17. Botu, V., Mhadeshwar, A. B., Suib, S. L. & Ramprasad, R. in *Springer Series in Materials Science* Vol. 225 (eds Lookman T., Alexander F. J. & Rajan K.) Ch. 8. (Springer International Publishing, 2016).
18. Kim, C., Pilania, G. & Ramprasad, R. From organized high-throughput data to phenomenological theory: the example of dielectric breakdown. *Chem. Mater.* **28,** 1304–1311 (2016).
19. Pilania, G. *et al.* Machine learning bandgaps of double perovskites. *Sci. Rep* **6,** 19375 (2016).
20. Botu, V., Batra, R., Chapman, J. & Ramprasad, R. Machine learning force fields: construction, validation, and outlook. *J. Phys. Chem. C* **121,** 511–522 (2017).
21. Kim, C., Pilania, G. & Ramprasad, R. Machine learning assisted predictions of intrinsic dielectric breakdown strength of ABX_3 perovskites. *J. Phys. Chem. C* **120,** 14575–1458 (2016).
22. Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C. & Scheffler, M. Big data of materials science: critical role of the descriptor. *Phys. Rev. Lett.* **114,** 105503 (2015).
23. Botu, V., Chapman, J. & Ramprasad, R. A study of adatom ripening on an Al(111) surface with machine learning force fields. *Comput. Mater. Sci.* **129,** 332–335 (2016).
24. Botu, V. & Ramprasad, R. Learning scheme to predict atomic forces and accelerate materials simulations. *Phys. Rev. B.* **92,** 094306 (2015).
25. Botu, V. & Ramprasad, R. Adaptive machine learning framework to accelerate ab initio molecular dynamics. *Int J Quantum Chem* **115,** 1074–1083 (2015).
26. Giustino, F. & Snaith, H. J. Toward Lead-Free Perovskite Solar Cells. *ACS Energy Lett* **1,** 1233–1240 (2016).
27. Castelli, I. E., GarcÌa-Lastra, J. M., Thygesen, K. S. & Jacobsen, K. W. Bandgap calculations and trends of organometal halide perovskites. *APL Materials* **2,** 081514 (2014).
28. Becker, M., Kluner, T. & Wark, M. Formation of hybrid ABX_3 perovskite compounds for solar cell application: First-principles calculations of effective ionic radii and determination of tolerance factors. *Dalton Trans.* **46,** 3500–3509 (2017).
29. Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **136,** B864–B871 (1964).
30. Kohn, W. & Sham, L. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140,** A1133–A1138 (1965).
31. 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,** 9911–9917 (2004).
32. Amsler, M. & Goedecker, S. Crystal structure prediction using the minima hopping method. *J. Chem. Phys.* **133,** 224104 (2010).
33. Ong, S. P. *et al.* Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68,** 314–319 (2013).
34. 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).
35. Huan, T. D. *et al.* A polymer dataset for accelerated property prediction and design. *Sci. Data* **3,** 160012 (2016).

36. Baldwin, A. F. *et al.* Rational design of organotin polyesters. *Macromolecules* **48,** 2422–2428 (2015).
37. Albero, J., Asiri, A. M. & Garcia, H. Influence of the composition of hybrid perovskites on their performance in solar cells. *J. Mater. Chem. A* **4,** 4353–4364 (2016).
38. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50,** 17953–17979 (1994).
39. Kresse, G. & Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* **47,** 558–561 (1993).
40. Kresse, G. Ab initio Molekular Dynamik für flüssige Metalle. Ph.D. thesis Technische Universität Wien, (1993).
41. 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,** 15–50 (1996).
42. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54,** 11169–11186 (1996).
43. 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).
44. Murray, E. D., Lee, K. & Langreth, D. C. Investigation of exchange energy density functional accuracy for interacting molecules. *J. Chem. Theor. Comput* **5,** 2754–2762 (2009).
45. Monkhorst, H. J. & Pack, J. D. Special points for Brillouin-zone integrations. *Phys. Rev. B* **13,** 5188–5192 (1976).
46. Heyd, J., Scuseria, G. E. & Ernzerhof, M. Hybrid functionals based on a screened Coulomb potential. *J. Chem. Phys.* **118,** 8207–8215 (2003).
47. 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).
48. Perdew, J. P. Density functional theory and the band gap problem. *Int. J. Quant. Chem* **28,** 497–523 (1985).
49. Ramprasad, R., Glassford, K. M., Adams, J. B. & Masel, R. I. CO on Pd(110): determination of the optimal adsorption site. *Surf. Sci* **360,** 31–42 (1996).
50. He, Y. & Galli, G. Perovskites for Solar Thermoelectric Applications: A First Principle Study of $CH_3NH_3AI_3$ (A = Pb and Sn). *Chem. Matter* **26,** 5394–5400 (2014).
51. Wang, C. *et al.* Computational strategies for polymer dielectrics design. *Polymer* **55,** 979–988 (2014).
52. Wang, C. & Ramprasad, R. Novel hybrid polymer dielectrics based on group 14 chemical motifs. *Int. J. Hi. Spe. Ele. Syst* **23,** 1420002 (2014).
53. Baldwin, A. F. *et al.* Poly(dimethyltin glutarate) as a prospective material for high dielectric applications. *Adv. Mater.* **27,** 346–351 (2015).
54. Baldwin, A. F. *et al.* Effect of incorporating aromatic and chiral groups on the dielectric properties of poly(dimethyltin esters). *Macromol. Rapid Commun.* **35,** 2082–2088 (2014).
55. Ma, R. *et al.* Rationally designed polyimides for high-energy density capacitor applications. *ACS Appl. Mater. Interfaces* **6,** 10445–10451 (2014).
56. Sharma, V. *et al.* Rational design of all organic polymer dielectrics. *Nat. Comm.* **5,** 4845 (2014).
57. Mannodi-Kanakkithodi, A., Wang, C. C. & Ramprasad, R. Compounds based on Group 14 elements: building blocks for advanced insulator dielectrics design. *J. Mater. Sci.* **50,** 801–807 (2015).
58. Ma, R. *et al.* Rational design and synthesis of polythioureas as capacitor dielectrics. *J. Mater. Chem. A* **3,** 14845–14852 (2015).
59. Mannodi-Kanakkithodi, A., Pilania, G., Huan, T. D., Lookman, T. & Ramprasad, R. Machine learning strategy for the accelerated design of polymer dielectrics. *Sci. Rep.* **6,** 20952 (2016).
60. Huan, T. D. *et al.* Advanced polymeric dielectrics for high energy density applications. *Prog. Mater. Sci.* **83,** 236–269 (2016).
61. Mannodi-Kanakkithodi, A., Pilania, G. & Ramprasad, R. Critical assessment of regression-based machine learning methods for polymer dielectrics. *Comput. Mater. Sci.* **125,** 123–135 (2016).
62. Zhu, H., Tang, C., Fonseca, L. R. C. & Ramprasad, R. Recent progress in ab initio simulations of hafnia-based gate stacks. *J. Mater. Sci* **47,** 7399–7416 (2012).
63. Towns, J. *et al.* XSEDE: accelerating scientific discovery. *Comput. Sci. Engin* **16,** 62–74 (2014).
64. Hirasawa, M., Ishihara, T. & Goto, T. Exciton features in 0-, 2-, and 3-dimensional networks of $[PbI_6]_4$- octahedra. *J. Phys. Soc. Jpn* **63,** 3870–3879 (1994).
65. Baikie, T. *et al.* A combined single crystal neutron/X-ray diffraction and solid-state nuclear magnetic resonance study of the hybrid perovskites $CH_3NH_3PbX_3$ (X = I, Br and Cl). *J. Mater. Chem. A* **3,** 9298–9307 (2015).
66. Geng, W., Zhang, L., Zhang, Y.-N., Lau, W.-M. & Liu, L.-M. First-principles study of lead iodide perovskite tetragonal and orthorhombic phases for photovoltaics. *J. Phys. Chem. C* **118,** 19565–19571 (2014).
67. Kitazawa, N., Watanabe, Y. & Nakamura, Y. Optical properties of $CH_3NH_3PbX_3$ (X = halogen) and their mixed-halide crystals. *J. Mater. Sci.* **37,** 3585–3587 (2002).
68. El-Mellouhi, F. *et al.* Hydrogen bonding and stability of hybrid organic-inorganic perovskites. *ChemSusChem* **9,** 2648–2655 (2016).
69. Bernal, C. & Yang, K. First-principles hybrid functional study of the organic-inorganic perovskites $CH_3NH_3SnBr_3$ and $CH_3NH_3SnI_3$. *J. Phys. Chem. C* **118,** 24383–24388 (2014).
70. Papavassiliou, G. & Koutselas, I. Structural, optical and related properties of some natural three- and lower-dimensional semiconductor systems. *Synth. Met* **71,** 1713–1714 (1995).
71. Eperon, G. E. *et al.* Formamidinium lead trihalide: a broadly tunable perovskite for efficient planar heterojunction solar cells. *Energy Environ. Sci* **7,** 982–988 (2014).
72. Ma, Z.-Q., Pan, H. & Wong, P. K. A first-principles study on the structural and electronic properties of Sn-based organic-inorganic halide perovskites. *J. Electron. Mater.* **45,** 5956–5966 (2016).
73. Stoumpos, C. C. *et al.* Hybrid germanium iodide perovskite semiconductors: active lone pairs, structural distortions, direct and indirect energy gaps, and strong nonlinear optical oroperties. *J. Am. Chem. Soc.* **137,** 6804–6819 (2015).
74. Feng, J. & Xiao, B. Effective Masses and Electronic and Optical Properties of Nontoxic $MASnX_3$ (X = Cl, Br, and I) Perovskite Structures as Solar Cell Absorber: A Theoretical Study Using HSE06. *J. Phys. Chem. C* **118,** 19655–19660 (2014).
75. Ju, M.-G., Sun, G., Zhao, Y. & Liang, W. A computational view of the change in the geometric and electronic properties of perovskites caused by the partial substitution of Pb by Sn. *Phys. Chem. Chem. Phys.* **17,** 17679–17687 (2015).
76. Hirasawa, M., Ishihara, T., Goto, T., Uchida, K. & Miura, N. Magnetoabsorption of the lowest exciton in perovskite-type compound $(CH_3NH_3)PbI_3$. *Phys. B* **201,** 427–430 (1994).
77. Frost, J. M., Butler, K. T. & Walsh, A. Molecular ferroelectric contributions to anomalous hysteresis in hybrid perovskite solar cells. *APL Mater* **2,** 081506 (2014).
78. Onoda-Yamamuro, N., Matsuo, T. & Suga, H. Dielectric study of $\{CH_3NH_3PbX_3\}$ (X = Cl, Br, I). *J. Phys. Chem. Solids* **53,** 935–939 (1992).
79. Dong, Q. *et al.* Electron-hole diffusion lengths>175 μm in solution-grown $CH_3NH_3PbI_3$ single crystals. *Science* **347,** 967–970 (2015).

80. Poglitsch, A. & Weber, D. Dynamic disorder in methylammoniumtrihalogenoplumbates (II) observed by millimeter-wave spectroscopy. *J. Chem. Phys.* **87,** 6373–6378 (1987).
81. Umari, P. & Mosconi, E. Relativistic GW calculations on $CH_3NH_3PbI_3$ and $CH_3NH_3SnI_3$ Perovskites for Solar Cell Applications. *Sci. Rep.* **4,** 4467 (2014).
82. Brivio, F., Walker, A. B. & Walsh, A. Structural and electronic properties of hybrid perovskites for high-efficiency thin-film photovoltaics from first-principles. *APL Mater* **1,** 042111 (2013).
83. Bokdam, M. *et al.* Role of Polar Phonons in the Photo Excited State of Metal Halide Perovskites. *Sci. Rep.* **6,** 28618 (2016).
84. Koh, T. M. *et al.* Formamidinium-containing metal-halide: an alternative material for near-IR absorption perovskite solar cells. *J. Phys. Chem. C* **118,** 16458–16462 (2014).
85. Dang, Y. *et al.* Formation of hybrid perovskite tin iodide single crystals by top-seeded solution growth. *Angew. Chem. Int. Ed.* **55,** 3447–3450 (2016).

Data Citations

1. Kim, C., Huan, T. D., Krishnan, S. & Ramprasad, R. *Dryad Digital Repository*, http://dx.doi.org/10.5061/dryad.gq3rg (2017).
2. Kim, C., Huan, T. D., Krishnan, S. & Ramprasad, R. *NoMaD Repository*, http://dx.doi.org/10.17172/NOMAD/2017.03.15-1 (2017).

Acknowledgements

Computational work was made possible through XSEDE[63] computational resource allocation number TG-DMR080058N. The authors thank Stefan Goedecker and Max Amsler for the minima-hopping code.

Author Contributions

C.K. and T.D.H. contributed equally to the work and manuscript. R.R. designed and supervised this project. All authors discussed the results, wrote, and shaped the manuscript. The DFT computations were performed by C.K., S.K., and T.D.H. Data repository (Khazana) was designed and maintained by C.K.

Additional Information

Competing interests: The authors declare no competing financial interests.

How to cite this article: Kim, C. *et al.* A hybrid organic-inorganic perovskite dataset. *Sci. Data* 4:170057 doi: 10.1038/sdata.2017.57 (2017).

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