npj | Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

Active learning for accelerated design of layered materials

Lindsay Bassman Oftelie[1,2], Pankaj Rajak[1,3], Rajiv K. Kalia[1,2,3,4], Aiichiro Nakano[1,2,3,4,5], Fei Sha[4,5], Jifeng Sun[6], David J. Singh[6], Muratahan Aykol[7], Patrick Huck[7], Kristin Persson[7] and Priya Vashishta[1,2,3,4]

Hetero-structures made from vertically stacked monolayers of transition metal dichalcogenides hold great potential for optoelectronic and thermoelectric devices. Discovery of the optimal layered material for specific applications necessitates the estimation of key material properties, such as electronic band structure and thermal transport coefficients. However, screening of material properties via brute force ab initio calculations of the entire material structure space exceeds the limits of current computing resources. Moreover, the functional dependence of material properties on the structures is often complicated, making simplistic statistical procedures for prediction difficult to employ without large amounts of data collection. Here, we present a Gaussian process regression model, which predicts material properties of an input hetero-structure, as well as an active learning model based on Bayesian optimization, which can efficiently discover the optimal hetero-structure using a minimal number of ab initio calculations. The electronic band gap, conduction/valence band dispersions, and thermoelectric performance are used as representative material properties for prediction and optimization. The Materials Project platform is used for electronic structure computation, while the BoltzTraP code is used to compute thermoelectric properties. Bayesian optimization is shown to significantly reduce the computational cost of discovering the optimal structure when compared with finding an optimal structure by building a regression model to predict material properties. The models can be used for predictions with respect to any material property and our software, including data preparation code based on the Python Materials Genomics (PyMatGen) library as well as python-based machine learning code, is available open source.

npj Computational Materials (2018)4:74 ; https://doi.org/10.1038/s41524-018-0129-0

INTRODUCTION

Since the advent of single-layer graphene, a wide variety of two-dimensional (2D) materials have been isolated with a suite of interesting properties and applications.[1] In particular, 2D monolayers of semiconducting transition metal dichalcogenides (TMDCs) have proven worthy candidate materials for next-generation optoelectronic and thermoelectric devices due to their tunable band gap, transport and other properties, combined with their mechanical strength and chemical stability.[2–5] An important aspect of these layered materials is the discrete nature of the van der Waals (vdW) forces that bond the layers. This weak vdW bonding facilitates synthesis of 2D monolayers from bulk via mechanical or chemical exfoliation. It also allows for stacking of lattice-mismatched monolayers from different species of TMDCs, thereby enabling the formation of unlimited combination of multilayers. It should be noted, however, that the interlayer interactions are nonetheless sizable enough to strongly affect electronic behavior. Therefore, the electronic properties of multi-layer hetero-structures can vary in a highly nontrivial manner, depending on the total number of layers, the specific layer ordering, and each layer's composition. The possibility to vertically stack heterogeneous TMDC monolayers opens the door to a whole new class of hybrid materials where one can, in principle, engineer electronic, transport, optical, and other properties in well-defined, controllable material structures.[6–10]

As a specific example, we consider hetero-structures formed from monolayers of group VI TMDCs: MoS_2,$MoSe_2$,$MoTe_2$,WS_2, WSe_2,WTe_2. We take this set to be our alphabet $\Sigma = \{MoS_2,MoSe_2, MoTe_2,WS_2,WSe_2,WTe_2\}$, where each species of TMDC becomes a distinct symbol. Hetero-structures are then represented by strings, $w = a_1a_2\ldots a_N$, where the string length $|w|$ corresponds to the number of layers N in the stacked hetero-structure. For example, a three-layer hetero-structure may be written as $w = MoSe_2WTe_2$-MoS_2, where $|w| = 3$. We define the set of all N-layered hetero-structures as $H_N = \{a^N = a_1a_2\ldots a_N | a_i \in \Sigma; i = 1,\ldots,N\}$. As the number of layers N increases, the size of H_N increases exponentially. Furthermore, the computation time required to calculate the electronic properties for each hetero-structure using standard density functional theory (DFT) calculations increases rapidly as $O(n^3)$, where n is the number of electrons. Thus, performing exhaustive exploration of H_N quickly becomes intractable for large N.

Recently, machine learning techniques applied to the material science domain have shown great success in high-throughput screening and material property prediction.[11–13] For property prediction, material properties are computed for some percentage of a class of structures and are divided into a training set and a

[1]Collaboratory for Advanced Computing and Simulations, University of Southern California, Los Angeles, CA 90089, USA; [2]Department of Physics and Astronomy, University of Southern California, Los Angeles, CA 90089, USA; [3]Department of Chemical Engineering and Material Science, University of Southern California, Los Angeles, CA 90089, USA; [4]Department of Computer Science, University of Southern California, Los Angeles, CA 90089, USA; [5]Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089, USA; [6]Department of Physics and Astronomy, University of Missouri, Columbia, MO 65211-7010, USA and [7]Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA
Correspondence: Aiichiro Nakano (anakano@usc.edu)
These authors contributed equally: Lindsay Bassman Oftelie, Pankaj Rajak.

Received: 1 May 2018 Accepted: 13 November 2018
Published online: 10 December 2018

test set. A statistical model is built using the training data set and then used to predict the material properties of the structures in the remaining test data set. This separation of data into training and test sets is essential for quantifying how the model will generalize to new structures. With careful selection of model parameters and penalization of overfitting of the statistical model, the model can be used to make high accuracy predictions of material properties of the remaining materials in the class of structures, bypassing expensive ab initio computations. Success has been shown in predicting material properties such as band gap,[14–17] dielectric breakdown strength,[18,19] melting point,[15] and defects[20] using various machine learning methods, such as support vector regression,[14,15] neural networks,[21,22] and kernel ridge regression.[17,23]

Here, we use Gaussian process regression (GPR) in the space of vertically stacked TMDC's for prediction of two distinct types of properties derived from electronic structure. The first property type includes attributes of the band structure, which are critical parameters for optoelectronic materials. Specifically, we use GPR to predict the band gap value of structures as well as the conduction band minimum (CBM) and valence band maximum (VBM) dispersion curves in momentum space. The second property type is a simplified proxy for the thermoelectric figure of merit, which is a complex combination of electrical and thermal transport properties that is challenging to optimize.[24] Good thermoelectric materials have highly complex band structures, markedly different from standard isotropic parabolic bands, and the thermoelectric performance depends critically on them. Importantly, the concept of reduced dimensionality to achieve high thermoelectric performance has been very influential in thermoelectrics,[25] but has proven difficult to achieve in practical materials. The ability to engineer stacks may provide a unique opportunity to finally realize the benefits of reduced dimensionality to achieve high thermoelectric performance. Here, we focus on the electronic transport component of the thermoelectric figure of merit and use GPR to predict a recently proposed, band structure-dependent, electronic fitness function[26] (EFF) as a function of dopant concentration for both n-type- and p-type-doped hetero-structures. With a very complex structure dependence, this material property provides an excellent test case for our machine learning algorithms.

In many instances, instead of the ability to predict a structure's material properties, we only need to find the structure that has an optimal value for a given property, defined by the specific application. Here, optimality does not necessarily refer to the maximum or minimum value of the property; it can also refer to a property value that is closest to some desired value. For example, finding a structure with optimal band gap could refer to the structure with the maximum band gap, or the structure with a band gap closest to, say, the Shockley–Queisser limit for efficiency of solar cells.[27] Building a regression model to predict the property value of a structure, and then using this model to search for the structure with the optimal property, is neither an efficient nor scalable process for solving this problem. Instead, a machine learning model based on active learning is well suited here. In active learning, each training data point has an associated reward, and the point with the highest reward is computed and appended to the training data set in each iteration during training. The reward is computed using a reward function, which signifies the decrease in uncertainty associated with the model if a particular data point is selected. Here, we use a type of active learning called Bayesian optimization (BO), a method that optimizes black box functions[28–31] with minimal function evaluations, to discover the structure with an optimal property value.

In this work, we use BO to search for either the structure with maximum band gap or a band gap closest to 1.1 eV, the Shockley–Queisser limit for efficiency of solar cells.[27] We also apply BO to find the best thermoelectric hetero-structure. Since every structure has a convex EFF curve versus dopant concentration, each structure will have a peak EFF value at some dopant concentration, which can vary from structure to structure. We use BO to find the structure that has the maximum peak EFF value.

While we focus on band gap and thermoelectric EFF values as representative material properties, these methods can be used to predict or find optimal structures with respect to any other material property for which data are available. Similarly, while the proposed methods were validated on the class of three-layered hetero-structures, they can readily be applied to any class of *N*-layered hetero-structures, where exhaustive structure calculation is prohibitive.

RESULTS AND DISCUSSION

Figure 1 shows the four main steps in our property and optimal-structure prediction: data preparation, data computation, determination of numerical feature vectors to represent the structures, and machine learning algorithm development. Critical aspects of each step are outlined in the subsections below, while more technical details can be found in the Methods section.

Data collection and preparation

While the goal of this machine learning pursuit is to reduce the number of electronic property calculations needed to screen a class of materials for different applications, here we validate the method against DFT calculations for the entire class of three-layered hetero-structures, H_3. First, we create structure files for all unique three-layer hetero-structures and upload them to the Materials Project[32] (MP) database. The MP framework then performs electronic structure computation with DFT to obtain band structures for each material and subsequently performs transport calculations using BoltzTraP code[33] to get the thermoelectric EFF of each structure as a function of carrier-dopant concentration. This function reflects complex band structures as it relates to electronic transport functions relevant to thermoelectric performance. Specifically, the thermoelectric EFF is large for electronic structures that decouple the normally inverse relationship between electrical conductivity and thermopower. The electronic band structure and the thermoelectric EFF for a sample structure are given in Figure S1 in Supplementary Information.

Feature vector

For the machine learning algorithms to learn how the electronic property data relate to their corresponding structures, a numerical representation of the structure is required, as opposed to the character strings we, as humans, use to identify the different structures (e.g., $MoSe_2WTe_2MoS_2$). Since our material design space is limited to TMDCs, many atomic properties often used in feature vectors are either irrelevant or too similar across our materials to be useful. Therefore, we chose three of the most relevant atomic properties for prediction of band gaps found in material informatics literature:[14–16,21,34–36] (i) electronegativity (EN), or the tendency of the atom to attract an electron based on energy, (ii) first ionization potential (IP), or the energy required to remove an electron from the atom, and (iii) atomic radius (AR). Feature vectors may be composed to represent the hetero-structure by combining any subset of these atomic properties for each atomic species in each layer into a vector. Since each hetero-structure has three layers, and each layer contains two atomic species, a feature vector using one of these atomic properties, say IP, would be a six-dimensional vector, while a feature vector using two of these atomic properties would be a 12-dimensional vector. Values for the electronegativities, first IPs, and atomic radii for each of the five atomic species found in the hetero-structures are given in Table S4 in Supplementary Information.

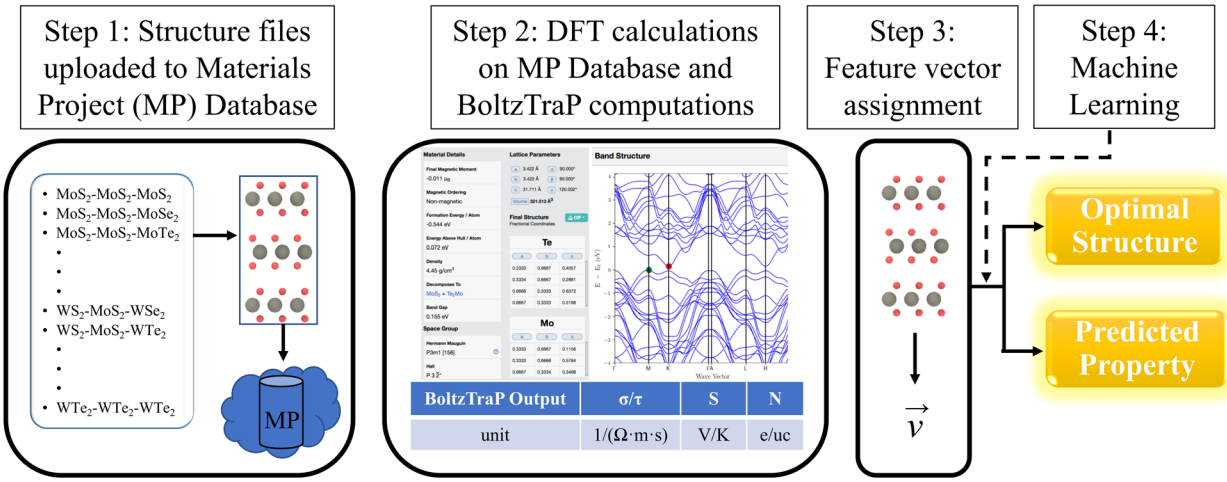

Fig. 1 Workflow for optimal structure and property prediction. First, structure files for a family of *N*-layered materials are created and uploaded to the Materials Project (MP) database. Second, the MP infrastructure performs all DFT calculations, and subsequently, transport calculations using BoltzTraP code are performed. A snapshot of the material property data computed by MP database is pictured, along with the thermoelectric parameters computed by BoltzTraP. Third, a numerical feature vector is assigned to uniquely represent each structure. Fourth, and finally, machine learning techniques are applied to the data to make predictions for either a material property or an optimal structure

We perform an extensive search of various combinations of these three atomic properties to compose the best feature vector for a given target property. Upon examining various combinations, we found that "stack-dependent" feature vectors provide the best prediction accuracy for all target material properties, where the atomic property used to represent a given layer depends on that layer's position in the stack. In the case of three-layered hetero-structures, layers 1 and 3 (i.e., outermost layers) are indistinguishable from each other due to mirror symmetry, whereas layer 2 (inner layer) is distinct from them. Thus, a stack-dependent feature vector can assign one atomic property to represent layers 1 and 3, while a different atomic property to represent layer 2. The best performance was achieved in BO searches for maximum band gap when layers 1 and 3 are represented by IP and layer 2 by AR. Good accuracy was also achieved by models in which each atomic species in each layer is represented using their IP and AR values. In general, models using EN to represent chalcogens (S, Se, and Te) tend to have lower accuracy for band gap.

For thermoelectric EFF in n-type-doped structures, the highest accuracy was achieved in BO searches when layers 1 and 3 are represented by AR and layer 2 by EN, whereas for thermoelectric EFF in p-type-doped structures, the highest accuracy was observed when layers 1 and 3 are represented by EN and layer 2 by IP. Also, in p-type-doped structures, models using EN to represent chalcogens tend to have higher accuracy. Prediction accuracies of various feature vectors for the predictions of different target properties are found in the section entitled "Feature Learning and Model Selection" in the Supplementary Information. The best feature vector for each target property was used in the corresponding prediction models presented below.

Gaussian process regression

We build GPR models to predict (i) the band gap, (ii) the VBM and CBM dispersion curves, and (iii) the thermoelectric EFF curve as a function of carrier-dopant concentration for both n-type- and p-type-doped hetero-structures. In the case of predicting the band gap, the target variable Y is a scalar, whereas in the case of predicting the VBM, CBM, and EFF curves, the target variable $\vec{Y}$ is made into a vector by discretizing a continuous curve into discrete points. A different model is trained for each point in order to create a predicted curve. Training each model involves calculating a percentage (in our case, ranging from 40% to 70%) of the entire class of three-layered hetero-structures, which are randomly selected to serve as each regression model's training set. The remaining structures, which the regression model has never encountered, are used as the test data set. Each model is then tasked with predicting its respective target variable (band gap, discrete points in the VBM, CBM, and EFF curves) for all structures in the test data set. Details of the GPR model may be found in the Methods section.

We found that training data sets with fewer than 60% of structures did not produce reliable predictions, while training sets with more than 60% did not show additional improvement. Since each GPR model is created by randomly selecting a percentage of the structures to act as the initial training data set, we created 100 independent GPR models and collected statistics on their prediction accuracy to average out any effects from the particular initial training data set chosen. Resulting predicted values versus their ground truth values are shown in Fig. 2 for one instance of a GPR model for each target variable, where models were trained with 60% of the structures in their training data sets. The ground truth values are results obtained using DFT and BoltzTraP code. Predictions made with models using smaller and larger percentages for the training data set are shown in Figures S2, S3, and S4 in Supplementary Information. Figure 2a shows predicted and ground truth band gap values of the three-layered hetero-structures in the test data set from one instance of the band gap prediction regression model with a particular, randomly selected training data set. Figure 2b shows the predicted and ground truth values for the VBM and CBM dispersion curves in momentum space for a sample three-layered hetero-structure. Figure 2c, d shows the predicted and ground truth EFF curves versus carrier concentration for a sample n(p)-type-doped three-layered hetero-structures.

As can be seen in Fig. 2, it is possible to build different GPR models to successfully predict a wide variety of material properties. One figure of merit for the accuracy of the predictions is the mean-square error (MSE), which is given by the equation, $\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N_s}\left(Y_{i,\,\mathrm{true}} - Y_{i,\,\mathrm{predicted}}\right)^2$, where N_s is the number of structures for which predictions were made. In the case of band

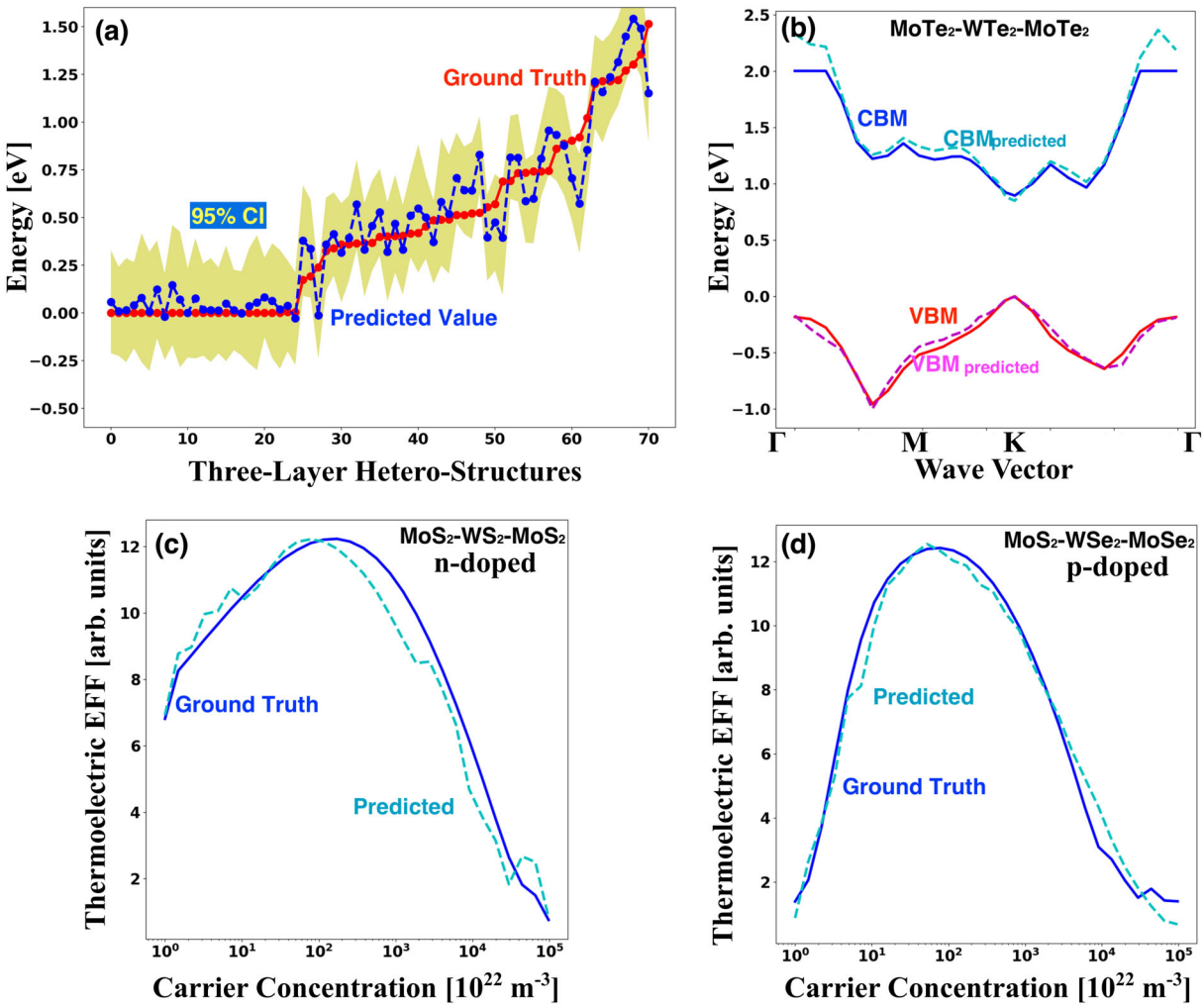

Fig. 2 **a** Predicted (blue) and ground truth (red) band gap values of three-layer hetero-structures in the test data set. Yellow shading represents a 95% confidence interval (CI) of the predicted results. **b** Predicted (dashed lines) and ground truth (solid lines) values for the VBM and CBM dispersion curves in momentum space for a sample three-layered hetero-structure. **c**, **d** Predicted (dashed lines) and ground truth (solid lines) thermoelectric EFF curves versus carrier concentration for a sample n(p)-type-doped three-layered hetero-structure. In all models, 60% of the three-layered structures were randomly selected to comprise the training data set

gap, the target variable Y is a scalar and this form of MSE is applicable. For predicting CBM and VBM dispersion curves and the thermoelectric EFF curve, where the target variable $\vec{Y}$ is a vector, we compute the overall MSE by averaging over the individual MSE's for each point in the target vector. MSE's for the various models with different-sized training data sets are given in the sections entitled "BandGap Prediction", Figures S3 and S4, and Table S2 in Supplementary Information.

Bayesian optimization

For applications where we need only find the structure with a desired value of some property (especially in cases where computation of each structure's material property is expensive), BO can be used for efficient (i.e., with minimal structure calculations) discovery of the optimal structure. In the BO process, a GPR model is first built using a randomly selected, small percentage of structures as the training data set (much smaller than the training data set size in the GPR models presented above). Since the true functional form of the relationship between structure and material property is unknown, and since the GPR model only provides crude predictions due to the small size of the training set, the procedure optimizes a surrogate function, called the acquisition function,[29] to determine which structure to evaluate next. The acquisition function selects the next structure based on a trade-off between exploration (to diversify the search) and exploitation (to follow the trend found by the current best estimates). Among the available acquisition functions, such as probability of improvement, upper confidence bounds, and expected improvement (EI), we used EI, the most widely used acquisition function due to its simple analytical expression (see Table 1 in the Methods section). The EI value for the remaining uncalculated structures is computed, and the material properties of the structure with the maximum EI are calculated next. This completes one iteration of BO. Each newly computed structure is added to the training data set, and the next iteration begins with a new GPR model built from the augmented training data set. The total number of iterations is up to the algorithm designer and constrained by how expensive it is to calculate new structure data. Here, we use 30 iterations for each BO run, as this was sufficiently many to predict the optimal structure with high accuracy, but few enough to remain computationally feasible. Specifics of the BO technique can be found in the Methods section.

In this work, BO models were first created to find either the structure with the maximum band gap or the structure with a band gap value closest to 1.1 eV (Shockley–Queisser limit for

Table 1. Pseudocode for the Bayesian optimization algorithm and outline for computing the acquisition function used in the above algorithm

Bayesian Optimization Algorithm

1. Data set: $D_{1:n} = \{x_i, y_i\} i = 1$ to n
2. Build Gaussian process regression model: $y = f_n(x) \sim GP(m(x), k(x,x'))$
3. Bayesian optimization () {
 for $t = 1$ to t_{max}
 a. Find next x_t by optimizing the acquisition function u over GP
 $x_t = argmax_x u(x|D_{1:n})$
 b. Compute the value y_t for this new x_t
 c. Augment (x_t, y_t) into data set $D_{1:n} = \{x_i, y_i\}$
 d. Update the Gaussian Process Regression model
}

Acquisition function u

1. Find x such that Expected Improvement ($\mathbb{E}$) is maximum
 $X = argmax_x EI(x) = argmax_x \mathbb{E}(\max\{0, f_n(x) - f^{max}\}|D_{1:n})$
 where, $f_n(x)$ is a Gaussian process regression model made from $D_{1:n}$ and f^{max} is the maximum value of this function.
2. Equations to compute Expected Improvement ($\mathbb{E}$|)

$$EI(x) = \begin{cases} (\mu(x) - f^{max})\Phi(Z) + \sigma(x)\phi(z) & if\ \sigma(x) > 0 \\ 0 & 0\ if\ \sigma(x) = 0 \end{cases}$$

$$Z = \frac{\mu(x) - f^{max}}{\sigma(x)}$$

 where $\mu(x)$ and $\sigma(x)$ are predicted mean and standard deviation values for x by Gaussian process regression model $f_n(x)$.
 $\varphi(Z)$ and $\Phi(Z)$are probability density function (PDF) and cumulative density function (CDF) of standard normal distribution
3. Return x

efficiency of solar cells[27]). Since each BO run begins by randomly selecting a small number of structures to compute for the initial training data set, we performed 500 independent BO runs and collected statistics on the optimal structure returned by each of the independent runs to average out any effects from the particular, randomly selected initial training data set. Figure 3a shows a scatter plot of the band gap values for the three-layered hetero-structures, where structures are ordered along the x-axis in increasing band gap value. Structures that were most frequently returned as the optimal structure are highlighted in red (for maximum band gap) and green (for desired band gap). The pie charts in Figs. 3b, c show the percentages of the 500 independent BO runs that the most frequently found optimal structures were returned. For example, in the BO search for the maximum band gap, Fig. 3b shows that 79% of BO runs correctly found the structure with the maximum band gap of 1.7 eV. In 15% of runs, the second-best structure, with a band gap of 1.5 eV was found, and in 5% of runs, the structure with the third-highest band gap value of 1.3 eV was returned.

BO was also applied to the class of four-layered hetero-structures to test the method on a class of layered structures for which it is difficult to compute materials properties for the full set. Since the full set of four-layered materials was not computed, we cannot guarantee that the BO process finds the optimal band gap structure (in this case, the structure with maximum band gap), but we did find that the model was consistently able to find a material with a higher band gap value than the highest band gap value computed in the initial training data set, as shown in Fig. 4. Here, the initial training data set was comprised of 30 randomly selected structures and 30 iterations were performed in each BO run. Since a stack-dependent feature vector, where atoms in outermost layers are represented by IP and those in innermost layers by AR, had the highest accuracy during BO search of maximum band gap, we have used a similar feature vector to represent four-layer structures. Here, atomsic species in layers 1 and 4 are represented by IP and those in layers 2 and 3 by AR.

Finally, BO was used to discover the three-layered hetero-structure with the highest peak EFF value for both n-type and p-type doping. Figures 5a, b show the peak EFF values for p-type- and n-type-doped three-layered hetero-structures, respectively. Here, the structures are sorted in ascending order of EFF value along the x-axis, with structures most frequently returned as the optimal structure in a set of 500 independent BO runs highlighted in different colors. Corresponding pie charts in Figs. 5c, d show the percentages of the 500 independent BO runs in which the most frequently found optimal structures were returned. As shown in Fig. 5, the two materials found as the best candidates for n-type (p-type) thermoelectric devices were $MoSe_2$-WS_2-WS_2 and WSe_2WTe_2-WSe_2 (WTe_2-$MoTe_2$-WTe_2 and $MoSe_2$-WSe_2-WSe_2). A physical explanation of why these materials emerged as optimal candidates is found in the "Discussion of Optimal Thermoelectric Materials" section in Supplementary Information.

Once again, it is seen that BO can successfully find a (nearly) optimal structure with high probability, in this case for a material property that has a far more complex relationship with the hetero-structure's electronic structure than band gap. In order to show the effectiveness and generalizability of BO for optimal property prediction beyond band gap and thermoelectric-EFF values, we tested the algorithm on a separate data set of adsorbate energies for a variety of adsorbate/surface material pairs.[37] After evaluating only 20% of the data set, 82% of 500 independent BO runs successfully identified the adsorbate/surface material pair with the minimum adsorption energy. Details of the data set, feature vector used, and model accuracy are found in the section entitled "Bayesian Optimization for Adsorption Energy" in Supplementary Information. Thus, the BO method can be successfully used for the discovery of a maximum, minimum, or desired value of a range of material properties.

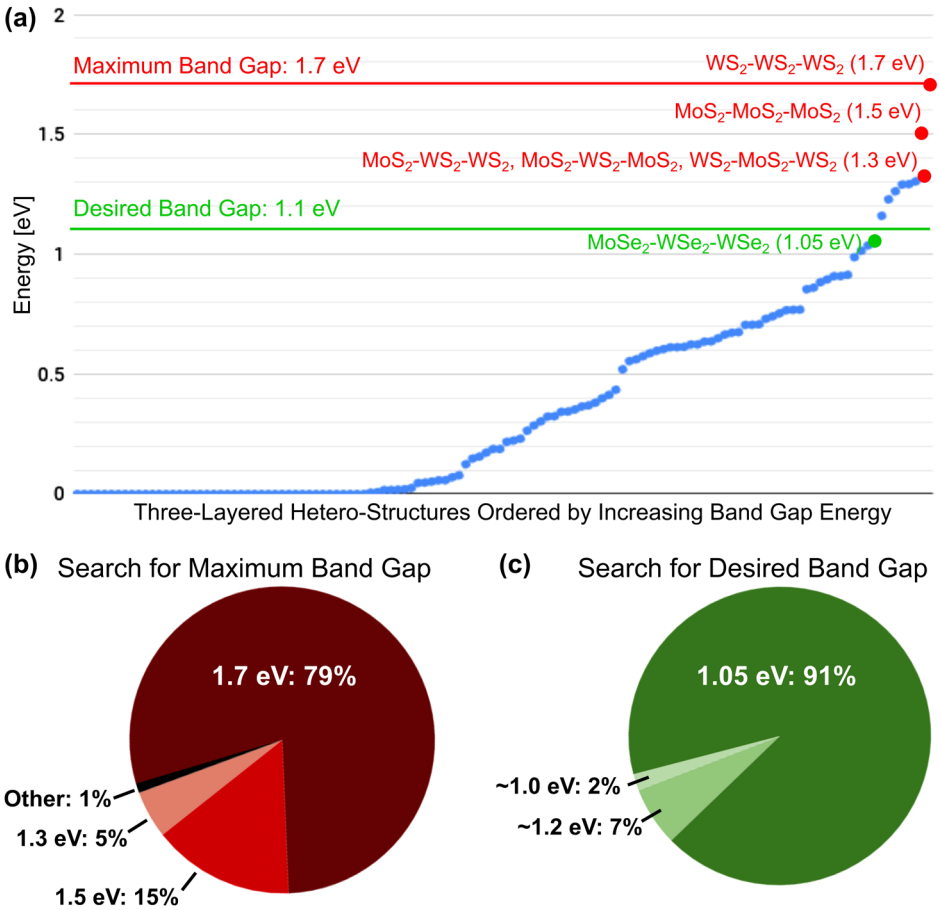

Fig. 3 **a** Band gap values for all three-layered hetero-structures, where hetero-structures on the *x*-axis are ordered by increasing band gap value. A table of structure names and corresponding band gap values is found as Table S5 in Supplementary Information. Highlighted points denote hetero-structures returned most frequently as the optimal structure by a BO model searching for the maximum band gap (red) and a desired value of 1.1 eV (green). **b** Pie chart showing the distribution of optimal band gap values returned in 500 independent BO searches for the maximum band gap. **c** Pie chart showing the distribution of optimal band gap found in 500 independent BO searches for a desired band gap value of 1.1 eV. $MoSe_2$-WSe_2-WSe_2, with a band gap of 1.05 eV, has the closest band gap value to 1.1 eV and was returned as the optimal structure in 91% of the 500 independent BO searches. WSe_2-WSe_2-WSe_2 and MoS_2-MoS_2-WS_2, with band gap values of 1.23 and 1.26 eV, respectively, were returned in 7% of the BO searches, while WSe_2-$MoSe_2$-WSe_2 and $MoSe_2$-WSe_2-$MoSe_2$, with band gap values of 1.01 and 1.04 eV, respectively, were returned in the remaining 2% of BO searches

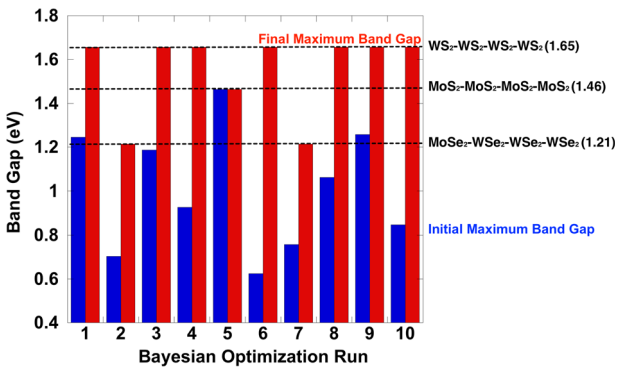

Fig. 4 Initial and final maximum band gap values for ten different BO runs on the class of four-layered hetero-structures. Optimal structures, along with their band gap value are labeled

METHODS

Reference data preparation

Structure files were prepared for all unique three-layered hetero-structures $w \in H_3$. When preparing configurations for three-layer hetero-structures, in which each layer can be one of six compounds, one may naively think that there would be $6^3 = 216$ configurations. However, a structure with stacking sequence ABC (ABB) is the same as the structure with stacking sequence CBA (BBA), as these two structures are simply 180° rotations of one another. For all such pairs, we prepared a structure file only for the first one in lexicographical order but not the other (i.e., $MoSe_2$-WSe_2-WS_2 was kept in the data set, while WS_2-WSe_2-$MoSe_2$ was discarded). This resulted in 126 unique three-layer structures files. In order to construct each three-layered hetero-structure *w*, we used a supercell of WTe_2 (as determined by experiment,[38] found in the Inorganic Crystal Structure Database[39]) as a template for atomic position and substituted atomic species as necessary to create each specific configuration for all unique three-layered hetero-structures. WTe_2 has the largest unit cell of all the Group VI TMDCs and thus was chosen as a template to avoid strain on the layered systems. One unit cell of WTe_2 has two layers, so in order to produce a three-layered structure, 1.5 unit cells of WTe_2, containing nine atoms, were merged into one supercell. Periodic boundary conditions were applied in all directions. Ten angstrom of vacuum was added in the *z*-direction (normal to the structure surface) to prevent multiple images of the structure from interacting. All structure files were generated automatically and then uploaded to the MP database[32] using the pymatgen[40] python library (see Software Availability section for where to access this code). After submission, all calculations, job scheduling, parallelization, and workflow

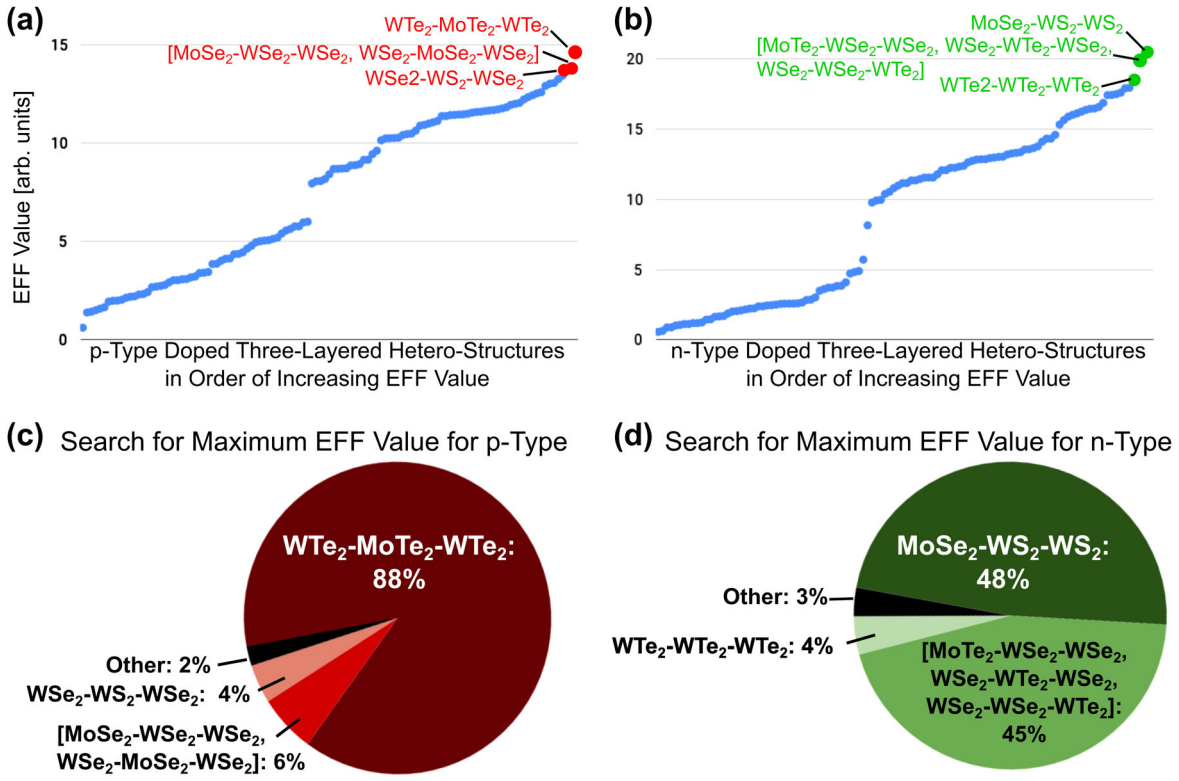

Fig. 5 **a**, and **b** Thermoelectric EFF values for all p-type-doped **a** and n-type-doped **b** three-layered hetero-structures, where hetero-structures on the *x*-axis are ordered by increasing EFF value. Points highlighted in red and green denote hetero-structures returned most frequently as the optimal structure by a BO model searching for the p-type- and n-type-doped structures, respectively, with maximum EFF value. **c**, **d** Pie chart showing the distribution of optimal thermoelectric structures returned in 500 independent BO searches for the p-type-doped **c** and n-type-doped **d** materials with maximum EFF value

management were handled on supercomputers by MP's infrastructure. Electronic structure data were calculated based on DFT,[41,42] using a plane-wave basis set and the projector augmented wave method[43] as implemented in the Vienna Ab initio simulation package.[44–47] The exchange and correlation energies were approximated with the Perdew–Burke–Ernzerhof functional.[48] First, both the unit cell and the atoms were allowed to relax to the lowest energy configuration. Next, self-consistent field iterations were performed to obtain the single-electron wave functions, and the electronic band structure was calculated from the resulting eigenenergies of the wave functions. The wave functions were constructed from a sum over a plane wave basis set, which consists of plane waves with kinetic energies up to 520 eV. Once the electronic structures were computed, the MP database used the results as input to BoltzTraP[33] to compute transport properties which were used to compute the thermoelectric EFF.

Gaussian process regression

A Gaussian process (GP) is a collection of random variables, in which any finite number of these variables has a joint Gaussian distribution. GP's are used to describe a distribution over functions and are completely specified by their mean function, $m(x)$, and covariance function (or kernel) $k(x,x')$. Thus, a GP may be written as $f(x) \sim GP(m(x),k(x,x'))$. GP regression (GPR) is a non-parametric regression technique, which models a distribution of functions that are consistent with a given set of N training observations (x^N, y^N). In our case, x is the n-dimensional input feature vector for each structure and y is either the structure's band gap, CBM or VBM dispersion curve, or thermoelectric EFF curve. We take an n-dimensional squared exponential kernel as the covariance kernel, shown in Eq. 1:

$$k(x,x\prime) = \exp\left\{-\frac{\sum_{i=1}^{n}\left\|x_i - x_i'\right\|^2}{\sigma_i^2}\right\} i = 1, \ldots, n \quad (1)$$

Here, σ_i are hyper-parameters associated with each dimension of the feature vector and are estimated using the maximum likelihood estimate. After training the model, we use the model to make predictions on test data set. The interested reader should refer to the Data Availability section for where to access our code used to run GPR.

Bayesian optimization

BO is an optimization technique, which is generally used for hyper-parameter tuning of deep neural networks and optimization of black box functions. Pseudocode for the algorithm is shown in Table 1. For band gap optimizations, a total of 500 independent BO runs were carried out to gather statistics on how frequently the true optimal structure was found in each of the cases of searching for the maximum peak EFF value, the maximum band gap value, and the band gap closest to 1.1 eV. In each BO run, five structures were chosen at random to act as the initial training data set. From the initial training data set, a GPR model is built and the acquisition function is computed for the remaining uncalculated structures. Details of the acquisition function calculation are shown in Table 1. In band gap (EFF) optimization, one of the top four (five) optimal structures is found within 30 BO iterations in over 95% of the 500 runs. Distributions of how many iterations it took to find one of the top five optimal structures in each case are shown in Figures S6 and S7 in Supplementary Information. The interested reader should refer to the Data Availability section for where to access our code used to run BO.

DATA AVAILABILITY

Electronic structure data is found at https://magics.usc.edu/data/, which contains data for all layered TMDC hetero-structures used in this study. Further information about these materials and many more materials are available on the Materials Project database at https://materialsproject.org/. Machine learning code for GPR and BO

along with the training data set can be found at https://github.com/rajak7/Bayesian_Optimization_Material_design, along with codes for automatically generating structure files for *N*-layered hetero-structures and uploading them to the Materials Project database with using pymatgen library functions. Code for computing the EFF from BoltzTraP output data is found at http://faculty.missouri.edu/singhdj/transm.shtml.

ACKNOWLEDGEMENTS

This work was supported as part of the Computational Materials Sciences Program funded by the US Department of Energy, Office of Science, Basic Energy Sciences, under Award Number DE-SC0014607. Calculations were performed at the Center for High Performance Computing of the University of Southern California, as well as the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231.

AUTHOR CONTRIBUTIONS

L.B.O. and P.R. contributed equally to this work. A.N., R.K.K., P.V., and F.S. designed the research. L.B.O. and P.R. performed material property calculations of the hetero-structures as well as machine learning computation. D.J.S. and J.S. developed key data analysis code for thermoelectric calculations. K.P., M.A., and P.H. played a key role in calculating and collecting data on the Materials Project database. All participated in data analysis and writing the paper.

ADDITIONAL INFORMATION

Supplementary information accompanies the paper on the *npj Computational Materials* website (https://doi.org/10.1038/s41524-018-0129-0).

Competing interests: The authors declare no competing interests.

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

REFERENCES

1. Gupta, A., Sakthivel, T. & Seal, S. Recent development in 2D materials beyond graphene. *Prog. Mater. Sci.* **73**, 44–126 (2015).
2. McDonnell, S. J. & Wallace, R. M. Atomically-thin layered films for device applications based upon 2D TMDC materials. *Thin Solid Films* **616**, 482–501 (2016).
3. Congxin, X. & Jingbo, L. Recent advances in optoelectronic properties and applications of two-dimensional metal chalcogenides. *J. Semicond.* **37**, 051001 (2016).
4. Wang, Q. H., Kalantar-Zadeh, K., Kis, A., Coleman, J. N. & Strano, M. S. Electronics and optoelectronics of two-dimensional transition metal dichalcogenides. *Nat. Nanotechnol.* **7**, 699–712 (2012).
5. Chhowalla, M. et al. The chemistry of two-dimensional layered transition metal dichalcogenide nanosheets. *Nat. Chem.* **5**, 263–275 (2013).
6. Geim, A. K. & Grigorieva, I. V. Van der Waals heterostructures. *Nature* **499**, 419–425 (2013).
7. Wang, H., Yuan, H., Hong, S. S., Li, Y. & Cui, Y. Physical and chemical tuning of two-dimensional transition metal dichalcogenides. *Chem. Soc. Rev.* **44**, 2664–2680 (2015).
8. Fang, H. et al. Strong interlayer coupling in van der Waals heterostructures built from single-layer chalcogenides. *Proc. Natl. Acad. Sci. USA* **111**, 6198–6202 (2014).
9. Choi, J., Zhang, H. & Choi, J. H. Modulating optoelectronic properties of two-dimensional transition metal dichalcogenide semiconductors by photoinduced charge transfer. *ACS Nano* **10**, 1671–1680 (2016).
10. Furchi, M. M., Pospischil, A., Libisch, F., Burgdörfer, J. & Mueller, T. Photovoltaic effect in an electrically tunable van der Waals heterojunction. *Nano Lett.* **14**, 4785–4791 (2014).
11. Rajan, K. Materials informatics. *Mater. Today* **8**, 38–45 (2005).
12. LeSar, R. Materials informatics: an emerging technology for materials development. *Stat. Anal. Data Min.* **1**, 372–374 (2009).
13. Mueller, T., Kusne, A. G. & Ramprasad, R. Machine learning in materials science: recent progress and emerging applications. *Rev. Comput. Chem.* **29**, 186–273 (2016).
14. Lee, J., Seko, A., Shitara, K., Nakayama, K. & Tanaka, I. Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques. *Phys. Rev. B* **93**, 115104 (2016).
15. Gu, T., Lu, W., Bao, X. & Chen, N. Using support vector regression for the prediction of the band gap and melting point of binary and ternary compound semiconductors. *Solid State Sci.* **8**, 129–136 (2006).
16. Pilania, G., Gubernatis, J. E. & Lookman, T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Comput. Mater. Sci.* **129**, 156–163 (2017).
17. Mannodi-Kanakkithodi, A., Pilania, G., Huan, T. D., Lookman, T. & Ramprasad, R. Machine learning strategy for accelerated design of polymer dielectrics. *Sci. Rep.* **6**, 20952 (2016).
18. Kim, C., Pilania, G. & Ramprasad, R. From organized high-throughput data to phenomenological theory using machine learning: the example of dielectric breakdown. *Chem. Mater.* **28**, 1304–1311 (2016).
19. Kim, C., Pilania, G. & Ramprasad, R. Machine learning assisted predictions of intrinsic dielectric breakdown strength of ABX3 perovskites. *J. Phys. Chem. C* **120**, 14575–14580 (2016).
20. Cubuk, E. D. et al. Identifying structural flow defects in disordered solids using machine-learning methods. *Phys. Rev. Lett.* **114**, 108001 (2015).
21. Zhaochun, Z., Ruiwu, P. & Nianyi, C. Artificial neural network prediction of the band gap and melting point of binary and ternary compound semiconductors. *Mater. Sci. Eng. B* **54**, 149–152 (1998).
22. Cubuk, E. D., Malone, B. D., Onat, B., Waterland, A. & Kaxiras, E. Representations in neural network based empirical potentials. *J. Chem. Phys.* **147**, 024104 (2017).
23. Huan, T. D., Mannodi-Kanakkithodi, A. & Ramprasad, R. Accelerated materials property predictions and design using motif-based fingerprints. *Phys. Rev. B* **92**, 014106 (2015).
24. Yang, J. et al. On the tuning of electrical and thermal transport in thermoelectrics: an integrated theory–experiment perspective. *NPJ Comput. Mater.* **2**, 15015 (2016).
25. Hicks, L. & Dresselhaus, M. S. Effect of quantum-well structures on the thermoelectric figure of merit. *Phys. Rev. B* **47**, 12727 (1993).
26. Xing, G. et al. Electronic fitness function for screening semiconductors as thermoelectric materials. *Phys. Rev. Mater.* **1**, 065405 (2017).
27. Shockley, W. & Queisser, H. J. Detailed balance limit of efficiency of p–n junction solar cells. *J. Appl. Phys.* **32**, 510–519 (1961).
28. Mockus, J. Bayesian Approach to Global Optimization: Teory and Applications Vol. 37 (Springer, Dordrecht, 2012).
29. Snoek, J., Larochelle, H. & Adams, R. P. Practical Bayesian Optimization of Machine Learning Algorithms. *Proc. Neural Inform. Process. Syst*. **25**, 2951–2959 (2012).
30. Shahriari, B., Swersky, K., Wang, Z., Adams, R. P. & de Freitas, N. Taking the human out of the loop: a review of bayesian optimization. *Proc. IEEE* **104**, 148–175 (2016).
31. Kushner, H. J. A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. *J. Basic Eng.* **86**, 97–106 (1964).
32. Jain, A. et al. Commentary: The Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
33. Madsen, G. K. & Singh, D. J. BoltzTraP. A code for calculating band-structure dependent quantities. *Comput. Phys. Commun.* **175**, 67–71 (2006).
34. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *NPJ Comput. Mater.* **2**, 16028 (2016).
35. Pilania, G. et al. Machine learning bandgaps of double perovskites. *Sci. Rep.* **6**, 19375 (2016).
36. Dey, P. et al. Informatics-aided bandgap engineering for solar materials. *Comput. Mater. Sci.* **83**, 185–195 (2014).
37. Toyao, T. et al. Toward effective utilization of methane: machine learning prediction of adsorption energies on metal alloys. *J. Phys. Chem. C.* **122**, 8315–8326 (2018).
38. Yanaki, A. & Obolonchik, V. Preparation of transition-metal tellurides by means of hydrogen telluride. *Inorg. Mater.* **9**, 1855–1858 (1973).
39. Bergerhoff, G., Hundt, R., Sievers, R. & Brown, I. The inorganic crystal structure data base. *J. Chem. Inf. Comp. Sci.* **23**, 66–69 (1983).
40. 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).
41. Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **136**, B864 (1964).
42. Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, A1133 (1965).
43. Blöchl, P. E. Projector augmented-wave method. *Phy. Rev. B* **50**, 17953 (1994).
44. Kresse, G. & Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* **47**, 558 (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**, 15–50 (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. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758 (1999).
48. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865 (1996).