Efficiently Searching Extreme Mechanical Properties via Boundless Objective-Free Exploration and Minimal First-Principles Calculations

Joshua Ojih,[1] Mohammed Al-Fahdi,[1] Alejandro David Rodriguez,[1] Kamal Choudhary,[2] and Ming Hu[1, *]

[1]Department of Mechanical Engineering, University of South Carolina, Columbia, SC 29208, USA

[2]Materials Science and Engineering Division, National Institute of Standards and Technology, Gaithersburg, MD, 20899, USA

Abstract

Despite the machine learning (ML) methods have been largely used recently, the predicted materials properties usually cannot exceed the range of original training data. We deployed a boundless objective-free exploration approach to combine traditional ML and density functional theory (DFT) in searching extreme material properties. This combination not only improves the efficiency for screening large-scale materials with minimal DFT inquiry, but also yields properties beyond original training range. We use Stein novelty to recommend outliers and then verify using DFT. Validated data are then added into the training dataset for next-round iteration. We test the loop of training-recommendation-validation in mechanical property space. By screening 85,707 crystal structures, we identify 21 ultrahigh hardness structures and 11 negative Poisson's ratio structures. The algorithm is very promising for future materials discovery that can push materials properties to the limit with minimal DFT calculations on only ~1% of the structures in the screening pool.

* Author to whom all correspondence should be addressed. E-Mail: hu@sc.edu (M.H.)

Introduction

The term extreme is defined as something farthest or of highest degree, which in terms of mechanical properties of materials imply unusual properties such as ultrahigh hardness[1] and extremely negative Poisson's ratio[2]. In the past decade, material scientists have been favorably using high throughput screening for structure property prediction with high accuracy in searching for promising materials[3]. However, high throughput screening prediction at the quantum level (first principles) is, although highly accurate, less efficient, and hence time consuming and computationally expensive[3,4]. In contrast, prediction at the classical level (such as classical molecular dynamics) is highly efficient but less accurate since they usually scale linearly with the number of atoms[5,6]. Because of the computational cost of density functional theory (DFT) and the less accuracy of classical potential, an intuitive idea is to bridge the gap between DFT-level accuracy and classical-level efficiency. ML methods offer the possibility of bridging this gap[7], and the application of ML has already help in speeding the process for material discovery[8].

ML methods have been extensively used for materials properties prediction over the past decade, because ML models can be trained to have high efficiency and accuracy close to DFT[9,10]. Generally speaking, the accuracy of a ML model depends on the effective input representation of the crystal structures, since the atomic positions are not suitable for direct input representation because they are not rotationally and translationally invariant[11]. Such input representation is known as descriptors or features. The idea behind the use of ML methods for structure properties prediction is to analyze and map the relationship between the properties of materials and their characteristics by extracting information from existing data without knowing any explicit knowledge on how to draw conclusion from those data[12]. With given data, ML algorithms learn the rules and relationship that underlie a dataset by assessing the data and build a model to make prediction[13]. For example, ML models have been used for the prediction of mechanical properties of metal alloy[14,15], band gap of crystals[16,17], the formation energies of crystals[18–20], melting temperature of binary inorganic compounds[21].

Though ML is highly efficient, it has some limitation which reduces its accuracy in predicting properties. Such limitations include, but are not limited to, measurement error[22], lack of generality and precision, reliance on high-quality data[23], inability to determine high level concept[24], prone to artifact[25], good in interpolation but poor in extrapolation[21,26]. Another critical drawback for ML methods is the lack of laws, understanding, and knowledge from their use because ML methods are treated as black box[6]. More importantly, the predicted materials properties of almost all existing ML models usually cannot exceed the range of the original training data. This means that, the trained ML models are usually good at predicting material properties within the original training data pool, the so-called interpolation prediction, while they can seldom predict material properties outside of training dataset, i.e., the extrapolation ability is poor. However, many previous studies have proved that most of extraordinary structures reside in the sparse area of the huge material space. To ensure building extrapolative machine learning materials property prediction model in the sparse area, it is critical to develop some advanced ML models to identify the promising candidates whose properties may exceed the range of the training data.

In this study, we implemented boundless objective-free exploration (BLOX) algorithm[27] for extreme mechanical property search. We use 3 different pairs of mechanical properties as the property space for the search, namely, bulk modulus vs. shear modulus, shear modulus vs. hardness, and Pugh's ratio vs. Poisson's ratio. The mechanical properties of a material are those properties that involve a reaction or behavior to an external or applied loading, and it is the characteristic that indicates the variation taking place in the material. The mechanical properties of a material characterize the reaction of the material to external loadings. Mechanical properties can be used to determine how a material would behave in each application and they are helpful in material selection process. They can also be used to estimate the lifetime of a material. In BLOX implementation, a ML model, namely Random Forest (RF) algorithm, is built to predict the properties of materials for which current data on calculated properties is available. In searching the property space, the BLOX algorithm searches outside the boundary to capture properties of materials that lie at the edge of the boundary. This can be made possible by using the Stein novelty (SN)

scores to recommend potential materials with tendency of being outside the boundary, i.e., different from original training data. The SN scores measures a deviation between the observed properties and the predicted properties by using Stein discrepancy[28]. After thoroughly screening of the 85,707 crystal structures from Materials Project[29] database, we found 30 structures with ultrahigh bulk and shear moduli, 21 structures with ultrahigh shear modulus and hardness, and 11 structures with negative Poisson's ratio. We compare our result with traditional ML methods such as crystal graph convolution neural network (CGCNN)[9], RF, Lasso Regression, and Ridge regression[30].

Results and Discussion

The result of this study is described in four major subsections based on the material property space searched.

A. High Bulk Modulus and Shear Modulus

Superhard materials are defined as materials with hardness exceeding 40 GPa[31] and they are of great importance because of their industrial applications such as abrasives, polishing, disc brakes, proactive coating, and cutting tools[32]. Diamond and related carbon nanostructures have been known to be at the very top of the hardest materials to date, with Vickers hardness in the range of 70 – 150 GPa[33]. However, diamond has several limitations for massive industrial applications such as high cost and oxidizing at temperatures above 800 °C[34]. A superhard material usually possesses a high bulk modulus (K) and shear modulus (G) and does not deform plastically. The shear modulus relates to strain response of a body to shear or torsional stress, and it involves change of shape without change of volume, while bulk modulus is related to the strain response of a body to hydrostatic stress which involves change in volume without change in shape[35]. Inspired by the mechanical properties of diamond[36], such as ultrahigh bulk and shear moduli, our first goal is to search the structures with high bulk and shear moduli which has tendency to be superhard materials.

The scatter plot of observed mechanical properties and BLOX prediction selected from top SN scores for the first round is shown in Fig. 1a, while Fig. 1b shows the first round of DFT calculation of the recommended structures by BLOX algorithm in comparison with traditional ML methods (CGCNN, RF, Lasso Regression, and Ridge Regression). The initial set of elastic moduli data for property prediction model was obtained from the JARVIS-DFT database[37,38], which consists for more than 65,000 materials and more than 15,000 elastic modulus data along with several other properties of materials. From Fig. 1a, we can clearly see that the BLOX algorithm recommended lots of materials that are out-of-trend from the initial observed materials. The materials with high SN score mean they will have higher chance to be out-of-trend, as shown by the color coding in Fig. 1a. The initial observed materials were randomly chosen from original pool of 10,192 structures downloaded from the Materials Project database[39] and their mechanical properties were calculated by DFT. With recommendation by BLOX, we continue to verify the material properties of these materials with DFT calculations and then found some materials that have extremely high and low bulk modulus and shear modulus. It is worth noting that, since our target is to find materials with extremely high mechanical strength, we then cleaned our data by removing any materials that have bulk modulus below 130 GPa. This step is necessary because BLOX algorithm searches for out-of-trend materials in all directions in the bulk modulus vs. shear modulus space and this will lead to some materials with low bulk and shear moduli being recommended as well. In other words, we guide the BLOX algorithm to search in the direction we are interested. It is worth pointing out that, the threshold value of 130 GPa was chosen based on the empirical experience, which is about half of the maximum value of bulk modulus in the original training data. From Fig. 1b, we compare the mechanical properties of recommended materials between ML models and DFT calculations. We built CGCNN, RF, Lasso regression, and Ridge regression models using the initial observed data and then used these ML models to predict the mechanical properties of recommended structures. By comparing prediction by ML models to the DFT results of BLOX recommended structures, Fig. 1b provides direct evidence that the traditional ML models could not push the material properties to the limit, even if the exact same recommended materials were tested, which is one of the main drawbacks for many existing ML models as

we pointed out earlier. This is understandable considering that traditional ML models are trained to be capable of predicting materials properties within the original range of training data, while they can hardly predict properties outside.

Once we got the DFT results for recommended structures, we added these data into original observed data and then use the expanded observe dataset to train a ML model, and then BLOX will recommend next round promising structures based on SN scores (see details in "Methods" section). We continued this loop in searching for ultrahigh bulk and shear moduli, by running BLOX for four rounds, and in each round, we added the previous materials with high mechanical properties verified by DFT calculations to the next round. In Fig. 2a, we observed that more and more materials in each round were identified to push the material property to the limit with DFT validation. From Fig. 2b, we can see that there are no significant changes in bulk modulus in our search, but in Fig. 2c, we observed that after the first iteration, there is an increase of about 60% in our maximum shear modulus as compared to our initial training data, since adding the verified DFT data to the initial training data improves BLOX recommendation, hence a need for further iteration. We observed there was no significant changes between third and fourth iteration, hence we stop the iteration. The stopping criteria depend largely on the specific material properties we are investigating. After four rounds we found 30 structures in total with ultrahigh bulk and shear moduli. It is interesting to notice that some identified structures even have almost doubled shear modulus as compared to the original observed data. To quantify the difference in the material properties between the BLOX recommended and DFT validated values and ML model predictions, we calculated the distance between the outlier of CGCNN and the real values by DFT calculations that are higher than the CGCNN predictions for each round as shown in Fig. 3a, by using the formula for distance between two points given below

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (1)$$

where x_2, y_2 are the bulk modulus and shear modulus, respectively, of materials higher than CGCNN prediction and x_1, y_1 is the bulk modulus and shear modulus of the highest outlier of CGCNN prediction. Fig. 3a shows the maximum and average distance between CGCNN prediction and DFT calculations for the recommended materials. We observed that as the number of BLOX rounds increases, the distance between the CGCNN prediction and DFT calculation decreases. This is understandable considering that more and more material properties that are outside of original training data were added into the next training process, i.e., the property range that CGCNN model can predict will also expand. To better illustrate our ideal, we added CGCNN prediction and DFT calculations for the 1st, 2nd, 3rd, and 4th round as observed in Fig. 3b – e for bulk modulus vs. shear modulus. The blue symbol denotes the outlier for CGCNN prediction that was used for calculating the distances compared to real DFT values. The traditional ML models do not continuously improve as we add a few hundred recommended observed data to the initial 2,000 observed data. This can be seen from Fig. 4 where we plot the mean absolute error (MAE) for CGCNN, Lasso regression, and Ridge regression for bulk and shear moduli prediction for different BLOX rounds. For most ML models the MAE for model prediction does not change noticeably. For Ridge regression model, we even found that the MAE increases with more data added into the training. There are several reasons responsible for this observation: (1) the total number of added data is still not significant as compared to the size of original observed data (2,000), roughly estimated as 5 – 10%; (2) a considerably large portion of added data is still in the range of original observed data, which has already been well trained in previous rounds, and thus those data actually do not provide any information or contribute too much in the next training process; (3) the ML models can predict very well on the subset of training data, and with larger dataset there is increase in variability and the model might come across data not well considered in our training.

B. Ultrahigh Hardness

Our recent high-throughput study on ultrahard carbon allotropes illustrates that the hardness has a strong positive correlation with the shear modulus[31,40]. Using the shear modulus and hardness property space, we

were able to find some materials with ultrahigh hardness. Here we are comparing the correlation between hardness vs. shear modulus and hardness vs. bulk modulus. Previous study has shown that, all materials with high shear modulus would normally have high hardness, but not all materials with high bulk modulus would have high hardness[1]. Therefore, shear modulus provides a better correlation with hardness than bulk modulus[41]. That is the reason we chose to search for materials in the hardness vs. shear modulus space.

In Fig. 5, we observed that more materials in each round were pushed to the limit with DFT validation. After four rounds we found 21 structures in total with ultrahigh hardness and shear modulus. Some identified structures were found to have Vickers hardness greater than 70 GPa, which is close to that of diamond or carbon allotropes. We also quantify the difference in the material properties between DFT validated values and CGCNN prediction, by calculating the distance between the outlier of CGCNN and the DFT calculated values higher than the CGCNN prediction for each rounds using Eqs (1). The results are shown in Fig. 6a. Here, we are calculating the distance of all DFT values that are greater than the highest value predicted by CGCNN model. As shown in Fig. 6b – e, for each round, an outlier predicted by CGCNN model was chosen (the blue symbol in the figure). Then, we calculated the distance of all DFT values that are higher than this outlier using Eqs. (1). We observed that as the number of BLOX rounds increases, the distance between the CGCNN prediction and DFT calculations decreases, which is the same phenomenon as found before for bulk and shear moduli space (see Fig. 3b – e). However, once again, we found that the distance for CGCNN prediction, representing the recommended material properties relative to the original range, cannot continuously decrease with BLOX rounds increasing. This means that CGCNN model cannot easily be trained to predict material properties in the rare or boundary region.

C. Negative Poisson's Ratio

Poisson's ratio is defined as the ratio of lateral strain in solid over the longitudinal strain measured in a simple tension experiment[42]. Most solid materials have positive Poisson's ratio, but a small portion of

solid materials have negative Poisson's ratio, which are known as auxetic materials[43,44]. The materials with negative Poisson's ratio have exceptional properties such as high energy absorption, high fracture resistance, difficult to deform under shear loading, enhanced toughness, and resistance to indentation. We use the Pugh's ratio (defined as the ratio between the shear modulus and the bulk modulus to distinguish the ductile/brittle behavior of material[45,46]) vs. Poisson's ratio for this search. We applied BLOX algorithm to explore in the negative direction to find structures with negative Poisson's ratio. In Fig. 7, we present the Poisson's ratio vs. Pugh's ratio for different BLOX rounds. In total we found 11 structures with negative Poisson's ratio from 85,707 crystal structures taking from Materials Project database. In contrast, there are only 2 materials in the original ~2,000 observed dataset that have negative Poisson's ratio. Our results indicate that the original Materials Project database does not include many materials with negative Poisson's ratio.

D. Data Driven Insight into Mechanical Properties

Once we obtain lots of high accuracy DFT data recommended by BLOX, we are now in the position to do further study to deeply understand the mechanisms of these outliers. The Pearson correlation matrix, as shown in Fig. 8, gives an insight on how much each property correlates with each other[47]. In principle, the mechanical behavior of a material depends on their interatomic bonding, which can then be further traced back to the electronic cloud such as charge density and spatial distribution. That is the reason we show the correlation between elastic properties and local potential (LOCPOT) and electron localization function (ELF) values. A Pearson correlation matrix relates two parameters to each other, and the values is between -1 and 1. A negative value of -1 shows a perfectly inverse correlation between the two parameters, while a positive value of 1 shows a perfectly positive correlation. A value of 0 shows no correlation. A value close to 0 indicate a weak direct correlation while a value close to -1 or 1 indicate a strong inverse or a strong direct correlation, respectively. Fig. 8 shows that there is a strong inverse correlation between Poisson's ratio and Pugh's ratio, and a strong positive correlation between bulk and shear modulus, between bulk and Young's modulus, and between shear and Young's modulus. There is

pretty strong negative and positive correlation between the mechanical properties and the mean values of LOCPOT and ELF, respectively.

To analyze the mechanism for ultrahigh hardness materials, we established the correlation between electron work function (EWF), interatomic bonding, and hardness. Generally, the mechanical behavior of materials depend on their interatomic bonding strength, which is a basic feature of most known superhard materials like diamond having strong covalent bonding, governed by the behavior of the electrons[48]. EWF is the minimum energy required to move electron inside a material at the Fermi level to its surface without kinetic energy[49]. It is determined by its composition and charge redistribution on its surface caused by dipole layer[50], and it reflects the electronic behavior of metals and atomic interaction[51]. Previous studies have demonstrated that, for hard materials, their hardness is mainly governed by the interatomic bonding strength through correlation with the EWF[48], i.e., the higher the EWF, the higher the hardness of the material. We also study the ELF which is the measure of electron localization in atomic and molecular system. The ELF in Fig. 9, reflects the probability of finding an electron in the system and the LOCPOT for the top two superhard structures recommended by BLOX and two materials, namely $Be_4C_8N_4$ (mp-1189451) and $C_{12}N_8$ (mp-1188347), that have never been published in literature to the best of our knowledge. From Fig. 9a, c, we can see that the ELF plot strongly illustrates the presence of electrons and strong covalent bonding existing between the elements of the materials. The Fig. 9b, d show the presence of more electrons and hence an increase in EWF. Fig. 9e – h also shows the presence of strong covalent bonds between the atoms of the structures. Thus, all plots agree with the ultrahigh hardness exhibited by the four materials as confirmed by DFT calculations.

Our results have demonstrated that BLOX algorithm can be effectively used to accelerate material discovery. Despite that some materials recommended by BLOX have been previously reported, it still shows that our screening results are accurate. One of such materials is BC_7[52,53] with symmetry $P\bar{4}m2$ (space group number: 115) and hardness of 75.2 GPa. In our DFT calculations we found that BC_7 has hardness of 71.7 GPa, which is very close to previous study. Finally, we found six superhard structures

which have never been reported in literature to the best of our knowledges. All these six structures have negative average local potential, indicating a strong average atomic attractive interaction in the unit cell. These structures with relevant structural information and hardness are reported in Table 1. To further confirm the thermodynamic stability of these structures, Fig. 10 shows the phonon dispersions of selected structures along high symmetry points in the Brillouin zone. No negative frequencies were found for the structures, indicating these structures are thermodynamically stable.

Before closing, we would like to discuss some important points about the ML+BLOX algorithms:

(1) The stopping criterion: Indeed, it is hard to recognize or quantify a stopping criterion by a mathematic formula. In practice, we stop iteration when there is no significant addition of structures coming out from validated DFT calculation from the previous round. This is an empirical and intuitive method. We would like to point out that, when and where we should stop BLOX loops would certainly depend on the specific material properties we are investigating. It also depends on the current region that the existing materials have already reached. Although it is almost impossible to theoretically or mathematically prove the upper or lower limit of material properties now, our study of coupling BLOX algorithms and DFT calculations paves the way to accelerating material discovery by identifying the out-of-trend materials using the boundless objective-free exploration approach. Our results demonstrate that the boundless objective-free exploration algorithm is very promising for future materials discovery that can push the materials properties to the limit with acceptable and achievable/realizable DFT calculations.

(2) The BLOX algorithm could be coupled with any traditional ML regressors. The reason for using RF+BLOX in our study is, in previous study conducted by Terayama et al[27] they compared different ML+BLOX of which RF+BLOX gave the best result. We simply follow that recipe in our current work. Systematic cross-checking and comparison of the performance among different combinations of ML+BLOX will be the focus of our future work.

(3) Certainly, it is hard to give out a mathematic formula for the maximum theoretical limit for material properties, including mechanical properties studied here. In many times people want to find the materials with enhanced properties. As we have presented, the ML+BLOX algorithm has great potential to accelerate such discovery, i.e., it allows to identify materials with the mechanical property close to or even going beyond the current boundary of the mechanical properties. Using this approach, we have identified 30 structures with ultrahigh bulk and shear moduli, 21 superhard structures with ultrahigh hardness, and 11 structures with negative Poisson's ratio from 85,707 crystal structures taking from the well-known Materials Project database. It is also worth noting that, the final findings also depend on the material pool to be screened. Here, we used the Materials Project database with 85,707 crystal structures. We believe that, if the same method is applied to even larger database, such as OQMD database that currently has around 1 million structures, more structures with extreme mechanical properties will be identified quickly.

(4) The transferability of the method to other material property: We believe that the same procedure can be straightforwardly carried out to find other material properties like lattice thermal conductivity, Grüneisen parameter, heat capacity, superconductivity, etc. In particular, the BLOX method is believed to be very suitable for finding extreme material properties that are hard to calculate by direct DFT. We would like to emphasize here again that, the overall performance of the BLOX algorithm depends on two major factors: (i) A well-defined to-be-pushed material property vs. dependent variable(s): according to our experience, the stronger correlation or relationship for such definition there is, the easier the BLOX algorithm can identify the trend and then recommend the outliers. (ii) More accurate descriptor(s) for the ML regressor model: more accurate descriptor(s) will result in more accurate prediction of target properties of to-be-screened materials (unchecked data), which will facilitate the BLOX algorithm to pinpoint the outliers more efficiently and accurately, so that in each iteration structures that are outside of the previous boundary of material property will be identified. In this way, the material property will be gradually pushed to the limit as search iteration goes on.

(5) Last but not least, the ML+BLOX algorithms used herein have helped us identify some structures in existing database, but their properties have not been previously explored yet. Such algorithms can be also coupled with the state-of-the-art crystal structure prediction methods or packages, such as Universal Structure Predictor (USPEX)[54–56], Crystal structure Analysis by Particle Swarm Optimization (CALYPSO)[57–59], to discover completely structures that are not included in existing materials databases with desired or extreme material properties.

Methods

A. Training Data and DFT Calculations

We used the classical force-field inspired descriptors (CFID)[38,60] to transform our crystal structure to ML input. We used boundless objective-free exploration (BLOX)[27], coupled with RF ML algorithm to screen 85,707 crystal structures downloaded from Materials Project[39] database. We split our 85,707 crystal structures into 10 different jobs and run them in parallel. For each job BLOX recommended 25 promising structures ranked by the SN score, making a total of 250 recommended candidates in each round. We then performed DFT calculations using the plane-wave basis projector augmented wave (PAW) method[61], within the Perdew-Burke-Ernzerhof exchange-correlation functional[62], as implemented in the VASP package[63–65]. The cutoff energy is set to be 500 eV for the recommended crystal structures to calculate mechanical properties. The energy and force criteria for the DFT calculation of elastic constants were 10^{-6} eV and 10^{-4} eV/Å, respectively. DFT calculation was conducted for validating the recommended structures by BLOX because the ML model prediction was not accurate enough due to transferability issue and limited number of training data. We performed crystal graph convolutional neural networks (CGCNN), Lasso regression, and Ridge regression for the recommended structures and compare with DFT. The phonon dispersions of selected structures were calculated by the finite displacement method using PHONOPY package[66] with harmonic second-order force constants calculated by VASP.

B. ML Workflow

Traditional ML model was trained and used for structure property prediction to see if we can rely on the ML model to find extreme mechanical properties, and this will in turn reduce the cost of DFT computation. The traditional ML models (RF, Ridge, and Lasso regression) were used in this study as implemented in the scikit-learn[67]. RF is a ML technique, proposed by Breiman in 2001[68] for classification and regression problems, through the ensembles of different decision trees. The RF regressor produces an estimation by averaging the prediction of many individual trees fitted on randomly resampled sets of training data. Three-fold cross validation was used for model fitting and hyperparameter optimization. We set the maximum number of trees to 100, 80% of observed data was used for training and 20% for testing. Ridge regression was originally proposed by Hoerl and Kennard in 1970[30,69] and used for analyzing data which are affected by multicollinearity, whereas Lasso regression was put forward by Tibshirani in 1996[70] for parameter estimation and variable selection simultaneously in regression analysis. Lasso and Ridge regression are both regularized methods that significantly reduces the intricacy of the models such as the number or absolute size of the sum of all coefficients in the model[71]. Lasso regression minimizes the absolute sum of the coefficients (L1 regularization), and Ridge regression minimizes the squared sum of the coefficients (L2 regularization). They aim to regularize complex models by introducing penalty factors and they are great at reducing overfitting. Three-fold cross validation was also used for model fitting and hyperparameter optimization, 80% of observed data was used for training and 20% for testing, the maximum alpha (α) value was set to 10.

The CGCNN model combines the descriptors and learning model into one inseparable step, i.e., the model learns material properties directly from the connection of atoms in the crystal[9]. The CGCNN framework has been demonstrated to represent periodic crystal that provides material property prediction with DFT accuracy[4,9]. Here, the crystal structures are represented by a crystal graph that encodes both atomic information and bonding interaction between atoms, and then build a convolutional neural network on top of the graph to automatically extract representations that are optimum for predicting

targets properties. The atomic properties are represented by nodes and encoded in the feature vector (v_i). For each atom, neighbors are first search within a 6 Å radius, and are considered as connected when the share a Voronoi face[72] with the center atom and have interatomic distance lower than the sum of the Cordero covalent bond length[73] of 0.25 Å. Crystal graphs do not form optimum representation by themselves; however, they are improved by using convolutional layers. After each convolutional layer, the features vectors gradually contain more information on the surrounding environment due to the concatenation between atom and bond features vectors[6]. The convolution function by Xie et al[9] consists of

$$v_i^{(t+1)} = v_i^{(t)} + \sum_{j,k} \sigma\left(z_{(i,j)_k}^{(t)} W_f^{(t)} + b_f^{(t)}\right) \odot g\left(z_{(i,j)_k}^{(t)} W_s^{(t)} + b_s^{(t)}\right) \quad (2)$$

where $W_f^{(t)}$, $W_s^{(t)}$, and $b_i^{(t)}$ are the convolution weight matrix, self-weight matrix and bias of the t^{th} layer respectively, g is the activation function for introducing nonlinear coupling between layers, σ denotes the sigmoid function, ⊙ denotes element-wise multiplication, and $z_{(i,j)_k}^{(t)}$ is the concatenation of the neighbor vectors. After *R* convolutions, a pooling layer reduces the spatial dimensions of convolutional neural network, the pooling layer operates on all feature vectors. For simplicity, a normalization summation is used as the pooling function. For optimization, backpropagation, and stochastic gradient descent (SGD) were used to update the weights with DFT calculated data. Here, we train the CGCNN model using the observed data, with 60% for training, 20% for testing, and 20% for validation. We then use our model to predict the properties for the unchecked 85,707 data. We add the DFT validated values to the observed data for the next round. We also compared the distance between the outlier of the CGCNN prediction with DFT calculations (see Fig. 3, 6).

C. Structural Descriptors

Machine learning techniques have shown great prospect for screening and discovery of crystals structures because of their high efficiency in predicting material properties as compared to high demanding DFT calculations. However, in ML based approach, controlling the performance to enhance its accuracy is based on how compound/crystalline structures are represented in dataset[74]. Transforming the input data into a suitable representation of atoms for ML is a necessary step as it will reduce the amount of required training data and help increase the accuracy[11].The transformation of input data is called descriptors/features extraction or engineering. Selecting a good descriptor is a very important step for ML model training, because a good descriptor can explain a target property well and this leads to a robust prediction model of a target property[74]. Combining descriptors with ML methods leads to model capable of accurately predicting structure properties. Chemical descriptors based on elemental properties have been successfully applied for various computational discovery[75], nonetheless, this is not suitable for modelling crystal structures with the same composition since they ignore structural information[76]. In this study, we use CFID[38,76] as descriptors, because the descriptors cover a wide range of crystal structures, and they are able to consider a combined form of elemental and structural representation[74,77]. The combined descriptors have been applied not only to crystalline systems but also to molecular systems[74]. Elemental representations we used in this study include atomic number, atomic mass, period, and group in the period table, first ionization energy, second ionization energy, electron affinity, Pauling electronegativity, Allen electronegativity, van der Walls radius, covalent radius, atomic radius, melting and boiling point, density, molar volume, heat of fusion, heat of vaporization, thermal conductivity, and specific heat. These elemental descriptors help capture essential information about compounds. Structural representations include simple coordination number, Voronoi polyhedron of central atom, angular distribution function, radial distribution function, bond-orientational order parameter[78], and angular Fourier series[79]. The CFID consists of 1,557 descriptors in total for each crystal structure: 438 average chemical, 4 simulation box size, 378 radial charge distribution, 100 radial distribution, 179 angle distribution up to the first neighbor, 179 angle distribution up to the second neighbor, 179 dihedral angle up to the first neighbor, and 100 nearest neighbor descriptors.

D. Computational Workflow and BLOX Algorithm

The schematic of our computational workflow performed in this study is illustrated in Fig. 11. The initial preparation is to randomly select 2,000 materials from database and calculate their mechanical properties by DFT. These 2,000 DFT data was served as observed data to initiate the whole process. We transformed these structures into ML input using CFID as descriptors. We also trained a RF model and used the model to predict structures from Materials Project database (unchecked data). After that, the search was performed by repeating the following steps. Step 1: construct a property prediction model; Step 2: recommend promising candidates ranked by the SN score based on kernel-based Stein discrepancy; Step 3: evaluate recommended candidates by DFT[27] and add DFT data into training dataset for next round ML training. In step 1, RF model is built as a property prediction model on the already evaluated materials and their property data. Structures with high SN score were recommended for evaluation as potential candidates. SN score for each unchecked materials intuitively measures a deviation between the observed property and the predicted property[28] as given in the equation below

$$SN(V \cup \{p\}) = SD(V) - SD(V \cup \{p\}) \quad (3)$$

where $SD(V)$ is the Stein discrepancy for the evaluated data (observed data), p is predicted point by ML, and $\cup$ is union operator in set theory. We select the candidates with top SN scores. The SN score is based on Stein discrepancy, which can boundlessly evaluate a distance between any two distributions in any dimensional space.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Code availability

The BLOX and CGCNN code used for this research are open source and free to download on GitHub.

Acknowledgements

Research reported in this publication was supported in part by the NSF (award number 1905775, 2030128, 2110033) and SC EPSCoR/IDeA Program under NSF OIA-1655740 via SC EPSCoR/IDeA SAN program (20-SA05).

Author Contributions

M.H. conceived the idea of the project. J.O. implemented the BLOX model and trained all ML models. M.A.F. conducted Pearson's correlation analysis and contributed to the results explanation. A.D.R wrote some codes to analyze results. K.C. helped in obtaining the DFT data and J.O. performed DFT calculations for the BLOX recommended structures. J.O. wrote the draft of the manuscript. M.H. and K.C. revised the manuscript. All authors contributed to manuscript preparation.

Competing Interests

The authors declare no competing financial or non-financial interests.

References

1. Mansouri Tehrani, A. *et al.* Machine Learning Directed Search for Ultraincompressible, Superhard Materials. *J. Am. Chem. Soc.* **140**, 9844–9853 (2018).

2. Sabouni-Zawadzka, A. Al. Extreme mechanical properties of regular tensegrity unit cells in 3D lattice metamaterials. *Materials (Basel).* **13**, 1–17 (2020).

3. Chibani, S. & Coudert, F. X. Machine learning approaches for the prediction of materials properties. *APL Mater.* **8**, (2020). https://doi.org/10.1063/5.0018384.

4. Noh, J., Gu, G. H., Kim, S. & Jung, Y. Uncertainty-Quantified Hybrid Machine Learning/Density Functional Theory High Throughput Screening Method for Crystals. *J. Chem. Inf. Model.* **60**, 1996–2003 (2020).

5. MacKerell, A. D. *et al.* All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**, 3586–3616 (1998).

6. Schmidt, J., Marques, M. R. G., Botti, S. & Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *npj Comput. Mater.* **5**, (2019).

7. Schütt, K. T. *et al.* How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Phys. Rev. B - Condens. Matter Mater. Phys.* **89**, 1–5 (2014).

8. Callaghan, S. Toward machine learning-enhanced high-throughput experimentation for chemistry. *Patterns* **2**, 100221 (2021).

9. Xie, T. & Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **120**, 145301 (2018).

10. Seko, A., Hayashi, H., Tsuda, K., Chaput, L. & Tanaka, I. Prediction of Low-Thermal-Conductivity Compounds with First-Principles Anharmonic Lattice-Dynamics Calculations and

Bayesian Optimization. *Phys. Rev. Lett.* **205901**, 1–5 (2015).

11. Himanen, L. *et al.* DScribe: Library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **247**, 106949 (2020).

12. Lu, L. *et al.* Extraction of mechanical properties of materials through deep learning from instrumented indentation. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 7052–7062 (2020).

13. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).

14. Chatterjee, S., Murugananth, M. & Bhadeshia, H. K. D. H. δ TRIP steel. *Mater. Sci. Technol.* **23**, 819–827 (2007).

15. Bhadeshia, H. K. D. H., Dimitriu, R. C., Forsik, S., Pak, J. H. & Ryu, J. H. Performance of neural networks in materials science. *Mater. Sci. Technol.* **25**, 504–510 (2009).

16. Pilania, G. *et al.* Machine learning bandgaps of double perovskites. *Sci. Rep.* **6**, 1–10 (2016).

17. Dey, P. *et al.* Informatics-aided bandgap engineering for solar materials. *Comput. Mater. Sci.* **83**, 185–195 (2014).

18. 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**, 1–5 (2015).

19. Meredig, B. *et al.* Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B - Condens. Matter Mater. Phys.* **89**, 1–7 (2014).

20. Curtarolo, S., Morgan, D., Persson, K., Rodgers, J. & Ceder, G. Predicting crystal structures with data mining of quantum calculations. *Phys. Rev. Lett.* **91**, 1–4 (2003).

21. Seko, A., Maekawa, T., Tsuda, K. & Tanaka, I. Machine learning with systematic density-functional theory calculations: Application to melting temperatures of single- and binary-

component solids. *Phys. Rev. B - Condens. Matter Mater. Phys.* **89**, 1–9 (2014).

22. Fan, J., Han, F. & Liu, H. Challenges of Big Data analysis. *Natl. Sci. Rev.* **1**, 293–314 (2014).

23. Keith, J. A. *et al.* Combining Machine Learning and Computational Chemistry for Predictive Insights into Chemical Systems. *Chem. Rev.* **121**, 9816–9872 (2021).

24. Bietti, A. & Mairal, J. On the inductive bias of neural tangent kernels. *Adv. Neural Inf. Process. Syst.* **32**, (2019).

25. Lapuschkin, S. *et al.* Unmasking Clever Hans predictors and assessing what machines really learn. *Nat. Commun.* **10**, 1–8 (2019).

26. Pun, G. P. P., Batra, R., Ramprasad, R. & Mishin, Y. Physically informed artificial neural networks for atomistic modeling of materials. *Nat. Commun.* **10**, 1–10 (2019).

27. Terayama, K. *et al.* Pushing property limits in materials discovery: Via boundless objective-free exploration. *Chem. Sci.* **11**, 5959–5968 (2020).

28. Liu, Q., Lee, J. D. & Jordan, M. I. A Kernelized Stein Discrepancy for Goodness-of-fit Tests and Model Evaluation. *Proceedings of the 33rd International Conference on Machine Learning*, (2016).

29. Jain, A. *et al.* Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 1–11 (2013).

30. Hoerl, R. W. Ridge Regression: A Historical Context. *Technometrics* **62**, 420–425 (2020).

31. Mohammed, Al-Fahdi; Tao, Ouyang; and Ming, H. High-Throughput Computation of Novel Ternary B-C-N Structures and Carbon Allotropes with Electronic-Level Insights into Superhard Materials from Machine Learning. *J. Mater. Chem. A* **9**, 27596–27614 (2021).

32. Chung, H. Y., Weinberger, M. B., Yang, J. M., Tolbert, S. H. & Kaner, R. B. Correlation between

hardness and elastic moduli of the ultraincompressible transition metal diborides RuB2, OsB2, and ReB2. *Appl. Phys. Lett.* **92**, 2008–2010 (2008).

33. Synthetic, D. Superhard material. Wikipedia. 1–15 (2021).

34. John, P., Polwart, N., Troupe, C. E. & Wilson, J. I. B. The oxidation of (100) textured diamond. **11**, 861–866 (2002).

35. Phani, K. K. & Sanyal, D. The relations between the shear modulus, the bulk modulus and Young's modulus for porous isotropic ceramic materials. *Mater. Sci. Eng. A* **490**, 305–312 (2008).

36. Chen, X.-Q., Niu, H., Li, D. & Li, Y. Intrinsic Correlation between Hardness and Elasticity in Polycrystalline Materials and Bulk Metallic Glasses. **2**, (2011).

37. Choudhary, K., Cheon, G., Reed, E. & Tavazza, F. Elastic properties of bulk and low-dimensional materials using van der Waals density functional. *Phys. Rev. B* **98**, 1–39 (2018).

38. Choudhary, K. *et al.* The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *npj Comput. Mater.* **6**, (2020).

39. Jain, A. *et al.* Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, (2013).

40. Mohammed, Al-Fahdi; Alejandro, Rodriguez; Tao, Ouyang; and Ming, H. High-Throughput Computation of New Carbon Allotropes with Diverse Hybridization and Ultrahigh Hardness. *Crystals*. 1-15 (2021) doi:10.3390/cryst11070783.

41. Levine, B. J. B., Tolbert, S. H. & Kaner, R. B. Advancements in the Search for Superhard Ultra-Incompressible Metal Borides. 3519–3533 (2009) doi:10.1002/adfm.200901257.

42. Lakes, R. Advances in negative poisson's ratio materials. *Adv. Mater.* **5**, 293–296 (1993).

43. Choudhary, K., Cheon, G., Reed, E. & Tavazza, F. Elastic properties of bulk and low-dimensional

materials using van der Waals density functional. *Phys. Rev. B* **98**, 1–12 (2018).

44. Dagdelen, J., Montoya, J., De Jong, M. & Persson, K. Computational prediction of new auxetic materials. *Nat. Commun.* **8**, 1–8 (2017).

45. Wen, Y., Wang, L., Liu, H. & Song, L. Ab Initio Study of the Elastic and Mechical Properties of B19TiAl. *Crystals*. 1–11 (2017) doi:10.3390/cryst7020039.

46. Liu, Z. T. Y., Zhou, X., Gall, D. & Khare, S. V. First-principles investigation of the structural , mechanical and electronic properties of the NbO-structured 3d , 4d and 5d transition metal nitrides First-principles investigation of the structural , mechanical and electronic properties of the NbO-struc. *Comput. Mater. Sci.* **84**, 365–373 (2014).

47. Adler, J. & Parmryd, I. Quantifying Colocalization by Correlation : The Pearson Correlation Coefficient is Superior to the Mander ' s Overlap Coefficient. (2010) doi:10.1002/cyto.a.20896.

48. Hua, G. & Li, D. The Correlation Between the Electron Work Function and Yeild Strength of Metals. *Phys. Status Solidi B* **1520**, 1517–1520 (2012).

49. Lang, N. D. & Kohn, W. Theory of Metal Surfaces: Charge Density and Surface Energy. *Phys. Rev. B* **1**, 12 (1970).

50. Leung, T. C., Kao, C. L., Su, W. S., Feng, Y. J. & Chan, C. T. Relationship between surface dipole, work function and charge transfer: to an established rule. *Phys. Rev. B*. 1–6 (2003) doi:10.1103/PhysRevB.68.195408.

51. Lu, H. *et al.* Electron work function – a promising guiding parameter for material design. *Nat. Publ. Gr.* 1–11 (2016) doi:10.1038/srep24366.

52. Xing, M., Li, B., Yu, Z. & Chen, Q. Elastic Anisotropic and Thermodynamic Properties of Two BC 7 Phases. **132**, 1340–1346 (2017).

53. Liu, H., Li, Q., Zhu, L. & Ma, Y. Superhard polymorphs of diamond-like BC 7. *Solid State Commun.* **151**, 716–719 (2011).

54. Glass, C. W., Oganov, A. R. & Hansen, N. USPEX-Evolutionary crystal structure prediction. *Comput. Phys. Commun.* **175**, 713–720 (2006).

55. Oganov, A. R., Lyakhov, A. O. & Valle, M. How evolutionary crystal structure prediction works-and why. *Acc. Chem. Res.* **44**, 227–237 (2011).

56. Lyakhov, A. O., Oganov, A. R., Stokes, H. T. & Zhu, Q. New developments in evolutionary structure prediction algorithm USPEX. *Comput. Phys. Commun.* **184**, 1172–1182 (2013).

57. Wang, Y., Lv, J., Zhu, L. & Ma, Y. CALYPSO: A method for crystal structure prediction. *Comput. Phys. Commun.* **183**, 2063–2070 (2012).

58. Yin, K. *et al.* An automated predictor for identifying transition states in solids. *npj Comput. Mater.* **6**, 1–10 (2020).

59. Su, C. *et al.* Construction of crystal structure prototype database: Methods and applications. *J. Phys. Condens. Matter* **29**, (2017).

60. Choudhary, K., Decost, B. & Tavazza, F. Machine learning with force-field-inspired descriptors for materials: Fast screening and mapping energy landscape. *Phys. Rev. Mater.* **2**, 1–8 (2018).

61. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).

62. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).

63. Kresse, G. & Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* **47**, 558–561 (1993).

64. Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B*

- *Condens. Matter Mater. Phys.* **59**, 1758–1775 (1999).

65. Vargas-Hernández, R. A. Bayesian Optimization for Calibrating and Selecting Hybrid-Density Functional Models. *J. Phys. Chem. A* **124**, 4053–4061 (2020).

66. Togo, A. & Tanaka, I. First principles phonon calculations in materials science. *Scr. Mater.* **108**, 1–5 (2015).

67. Barupal, D. K. & Fiehn, O. Generating the blood exposome database using a comprehensive text mining and database fusion approach. *Environ. Health Perspect.* **127**, 2825–2830 (2019).

68. Jin, Z. *et al.* RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **12343 LNCS**, 503–515 (2020).

69. Hoerl, A. E. & Kennard, R. W. American Society for Quality Ridge Regression : Biased Estimation for Nonorthogonal Problems American Society for Quality Stable URL : http://www.jstor.org/stable/1267351 Linked references are available on JSTOR for this article : Ridge Regression : Biase. **12**, 55–67 (1970).

70. Shrinkage, R. Regression Shrinkage and Selection via the Lasso Author (s): Robert Tibshirani Source : Journal of the Royal Statistical Society . Series B (Methodological), Vol . 58 , No . 1 (1996), Published by : Wiley for the Royal Statistical Society Stable URL. **58**, 267–288 (2016).

71. Muthukrishnan, R. & Rohini, R. LASSO: A feature selection technique in predictive modeling for machine learning. *2016 IEEE Int. Conf. Adv. Comput. Appl. ICACA 2016* 18–20 (2017) doi:10.1109/ICACA.2016.7887916.

72. Blatov, V. A. Voronoi-Dirichlet polyhedra in crystal chemistry: Theory and applications. *Crystallogr. Rev.* **10**, 249–318 (2004).

73. Cordero, B. *et al.* Covalent radii revisited. *J. Chem. Soc. Dalt. Trans.* 2832–2838 (2008)

doi:10.1039/b801115j.

74. Seko, A., Hayashi, H., Nakayama, K., Takahashi, A. & Tanaka, I. Representation of compounds for machine-learning prediction of physical properties. *Phys. Rev. B* **95**, 1–11 (2017).

75. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2**, 1–7 (2016).

76. Choudhary, K., Decost, B. & Tavazza, F. Machine learning with force-field-inspired descriptors for materials: Fast screening and mapping energy landscape. *Phys. Rev. Mater.* **2**, (2018).

77. Tanaka, I. Nanoinformatics. *Nanoinformatics,* 1-298 (2018). doi:10.1007/978-981-10-7617-6.

78. Steinhardt, P. J., Nelson, D. R. & Ronchetti, M. Bond-orientational order in liquids and glasses. *Phys. Rev. B* **28**, 784–805 (1983).

79. Bart, A. P., Kondor, Risi & Csanyi, G. On representing chemical environments. *Phys. Rev. B* **184115**, 1–16 (2013).

Table 1. Structures identified by BLOX recommendation with corresponding ultrahigh hardness.

Material ID	Formula	Average LOCPOT	Energy above hull (eV/atom)	Space group	Number of atoms in primitive cell	Hardness (GPa)
mp-1188347	$C_{12}N_8$	-12.15	0.264	$P\bar{4}3m$ (215)	20	58.43
mp-1102681	C_8N_{16}	-12.98	0.569	$I\bar{4}2d$ (122)	12	50.99
mp-1077595	C_4N_8	-12.85	0.707	$Cmc2_1$ (36)	6	48.78
mp-1105655	$C_{12}N_8$	-12.12	0.264	$Pm\bar{3}m$ (221)	20	48.61
mp-1189451	$Be_4C_4N_8$	-12.34	0	$Pna2_1$ (33)	16	47.74
mp-9410	$C_{12}N_{16}$	-12.62	0.298	P31c (159)	28	47.32

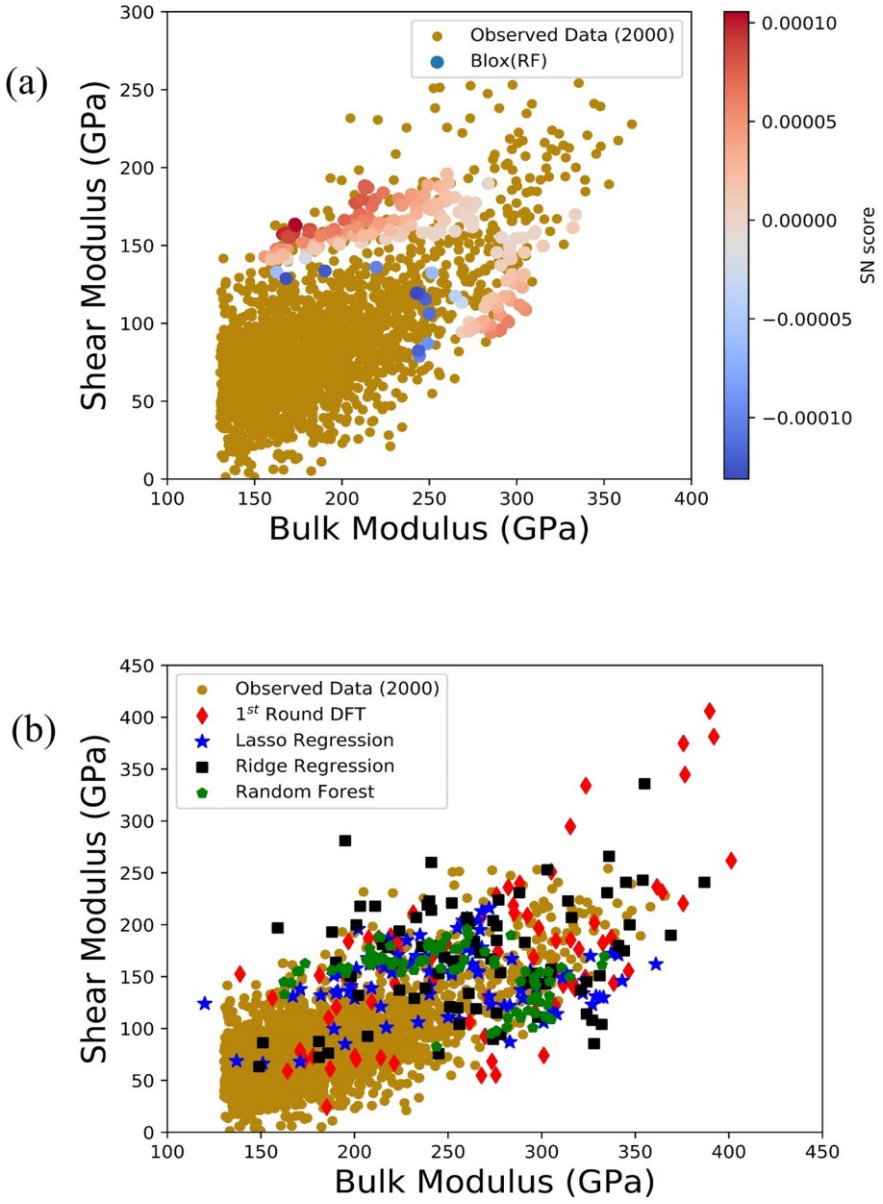

Fig. 1. Performance of BLOX algorithm in searching ultrahigh shear and bulk moduli materials. (a) Observed data and BLOX prediction for selected structure with highest Stein novelty (SN) score. 2,000 structures were used as input to ML in BLOX to train a model which predicts the unchecked data. The SN score is a measure of discrepancy between the predicted properties of unchecked data and properties of observed data. Candidates with high SN scores are recommended for DFT calculations. (b) Comparison of CGCNN, RF, Ridge regression, Lasso regression, and DFT calculation for structures recommended in the 1st round. The ML models could not push the properties to the outside of the original dataset (observed data).

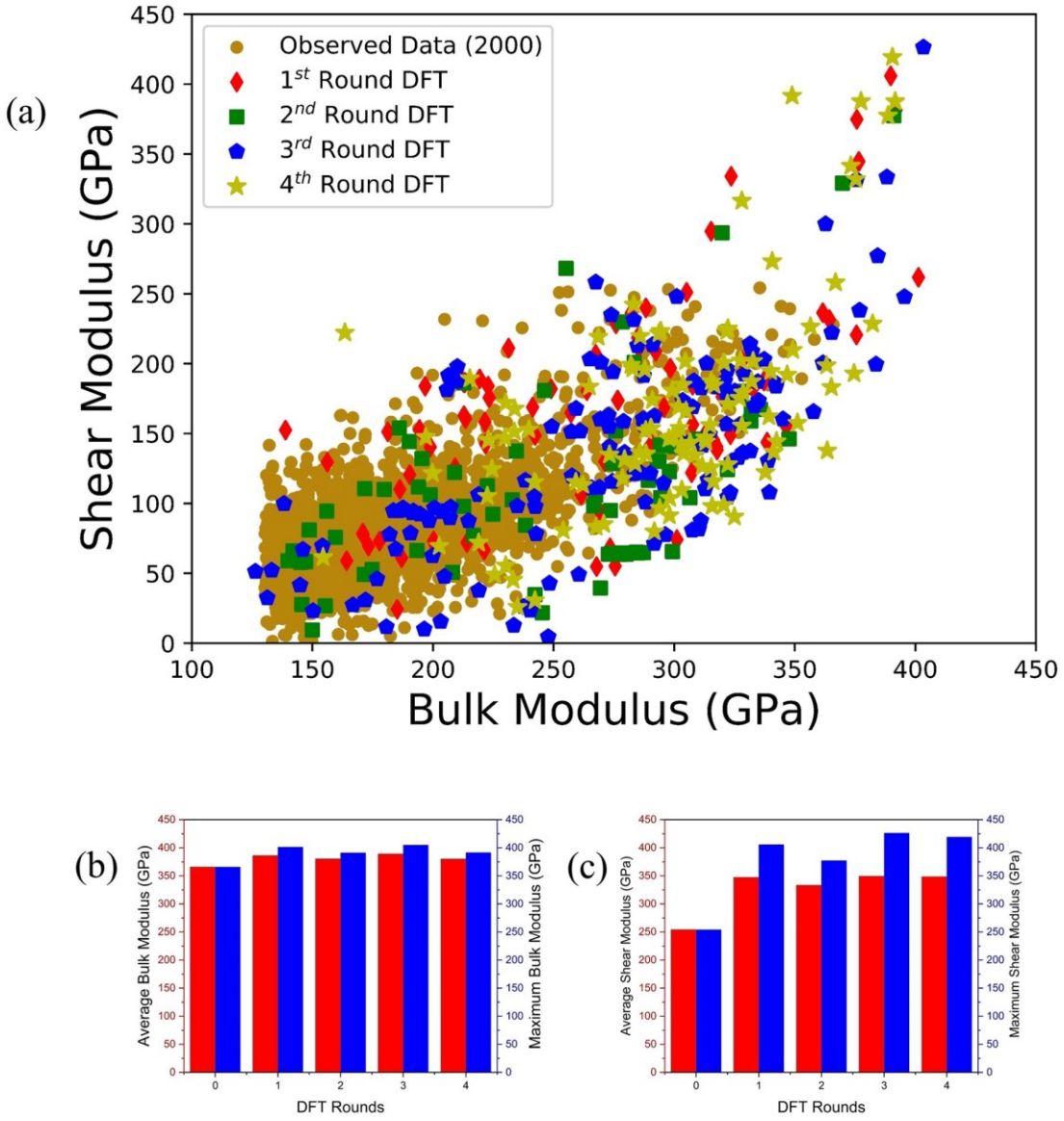

Fig. 2. Training procedure of BLOX algorithm. (a) Comparison of observed data and all rounds of DFT calculations recommended by BLOX. (b) – (c) Bar chart showing average (left y-axis) and maximum (right y-axis) value of bulk modulus and shear modulus respectively for each round of DFT calculations. "0" DFT round means the original training data.

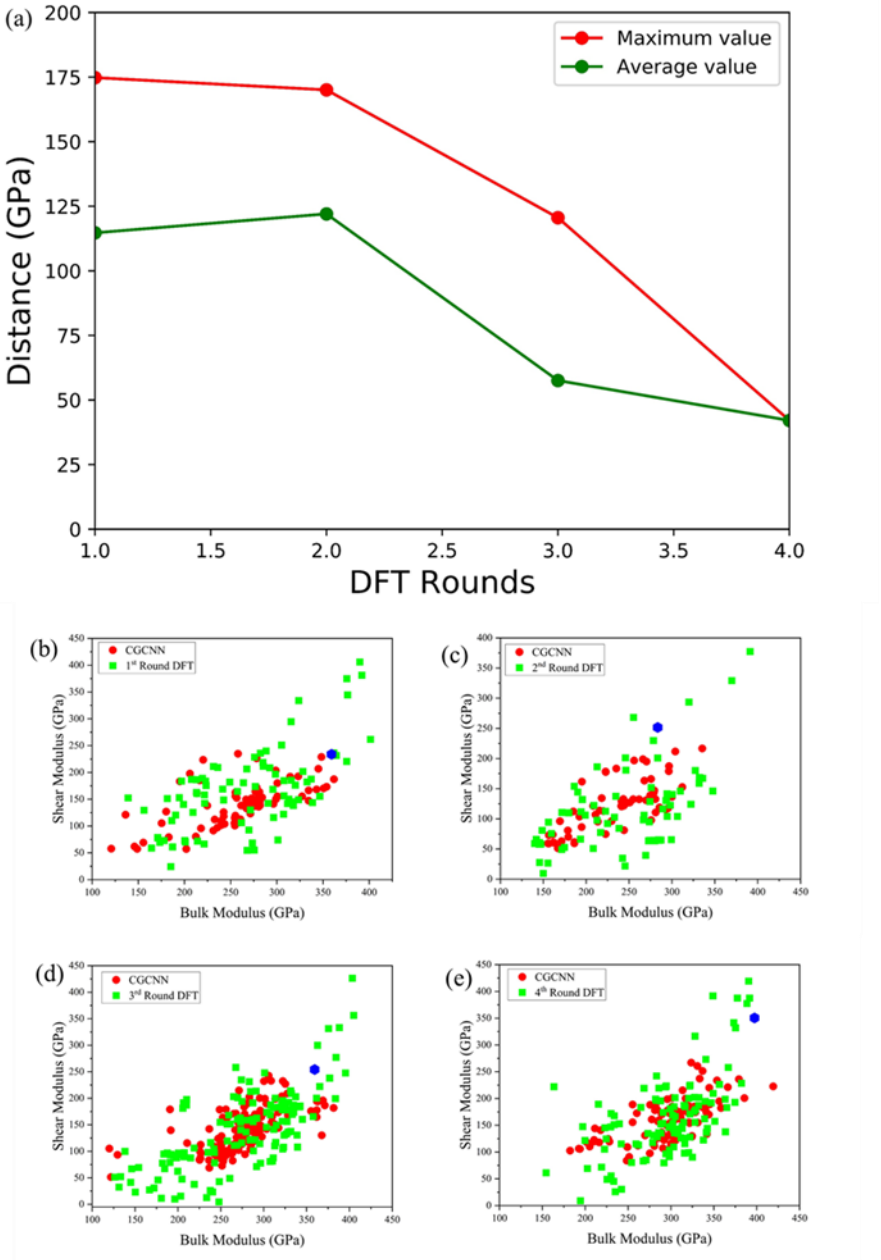

Fig. 3. Evaluation of prediction of CGCNN model for shear vs. bulk modulus. (a) Maximum and average distance between outlier of CGCNN prediction and DFT calculation for bulk and shear moduli. For each round we measure the distance between the outlier of CGCNN prediction (blue symbols in bottom panels) and all DFT values with bulk and shear moduli higher than CGCNN prediction. (b) – (e) Comparison between CGCNN prediction and DFT values for the 1st, 2nd, 3rd, and 4th round, respectively. The blue symbol denotes the outlier structure from CGCNN prediction that is used for calculating the distances to real DFT values.

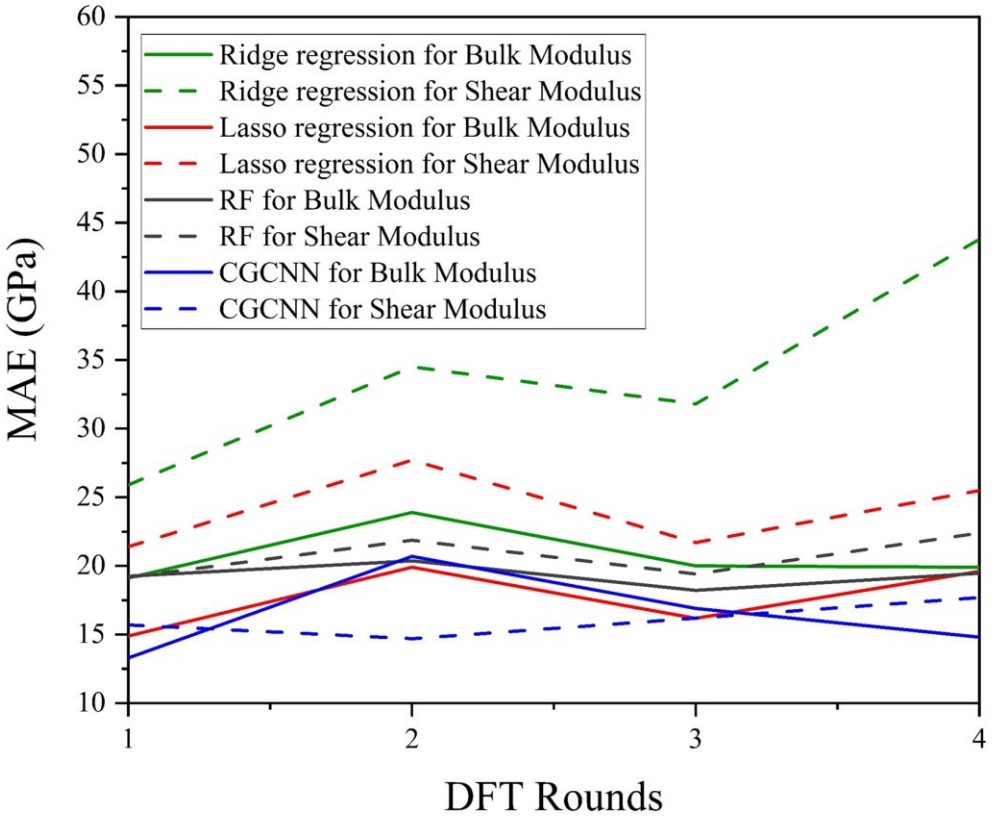

Fig. 4. Performance of traditional ML models. Mean absolute error (MAE) for Ridge regression, Lasso regression, Random Forest, and CGCNN models for bulk and shear moduli prediction at each DFT round.

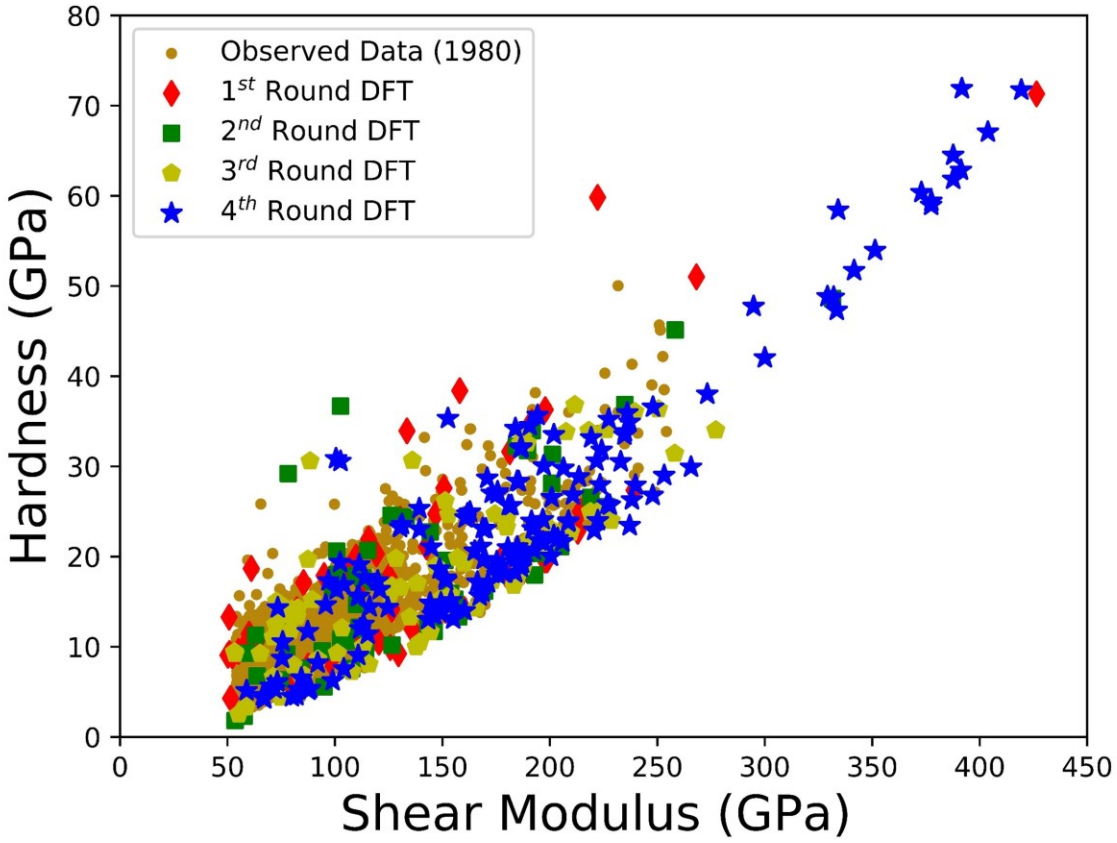

Fig. 5. Performance of BLOX algorithm in searching ultrahigh hardness materials. The hardness vs. shear modulus plot with comparison of observed data and all rounds of DFT calculations. After looping through all rounds of DFT, we were able to get 21 structures with ultrahigh shear modulus and ultrahigh hardness. The range of hardness and shear modulus is increased (expanded) by 40% and 70%, respectively.

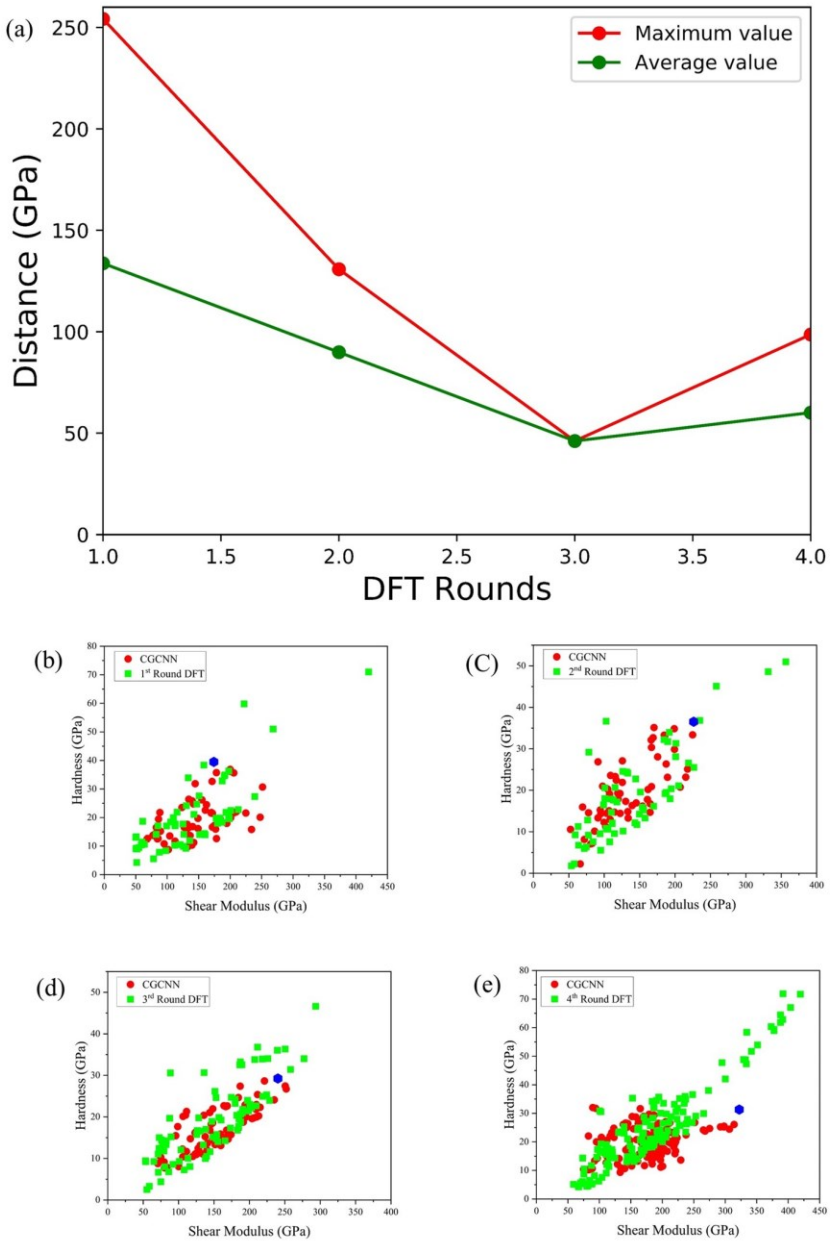

Fig. 6. Evaluation of prediction of CGCNN model for hardness vs. shear modulus. (a) Maximum and average distance between outlier of CGCNN prediction and DFT calculation for shear modulus and hardness. For each round we measure the distance between the outlier of CGCNN prediction (blue symbols in bottom panels) and all DFT values with shear modulus and hardness higher than CGCNN prediction. (b) – (e) Comparison between CGCNN prediction and DFT values for the 1st, 2nd, 3rd, and 4th round, respectively. The blue symbol denotes the outlier structure from CGCNN prediction that is used for calculating the distances to real DFT values.

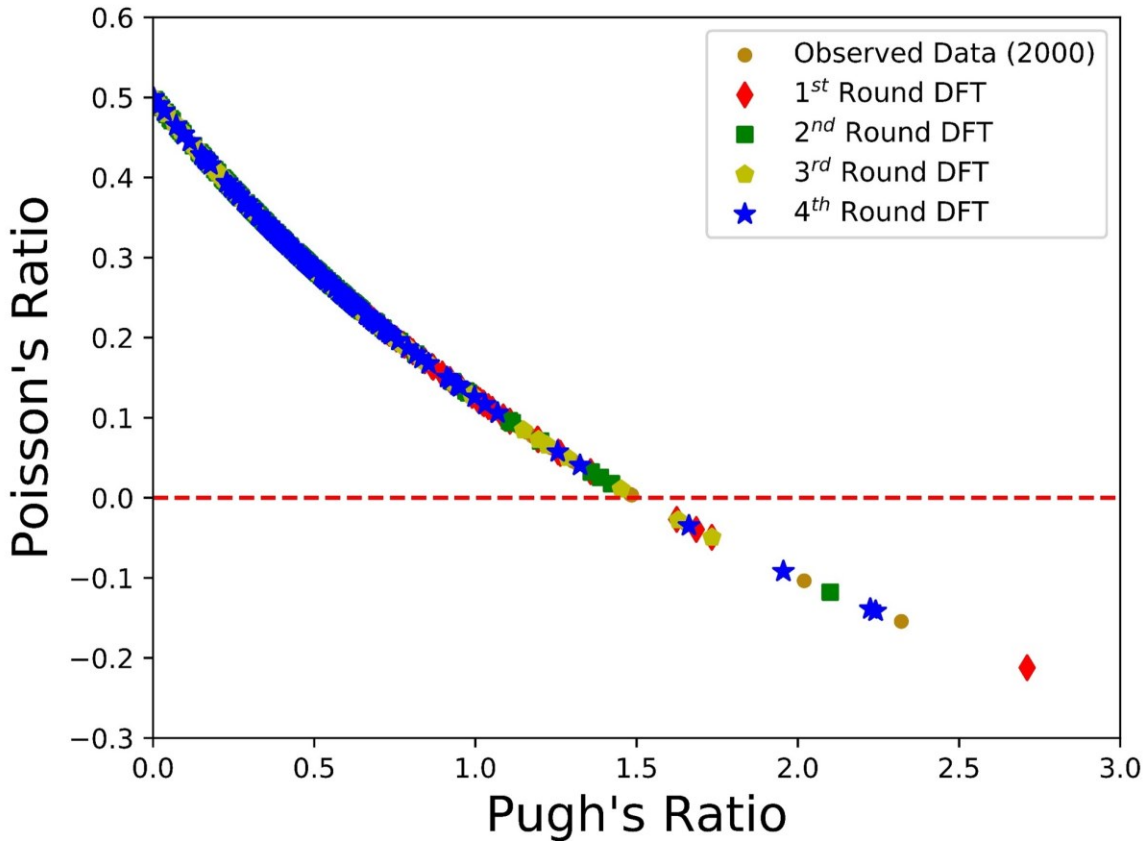

Fig. 7. Performance of BLOX algorithm in searching negative Poisson's ratio materials. Observed data and all rounds of DFT calculations recommended by BLOX for Poisson's ratio vs. Pugh's ratio. After a few loops, we were able to find 11 structures with negative Poisson's ratio. The dashed line represents zero Poisson's ratio and is the guide for eyes.

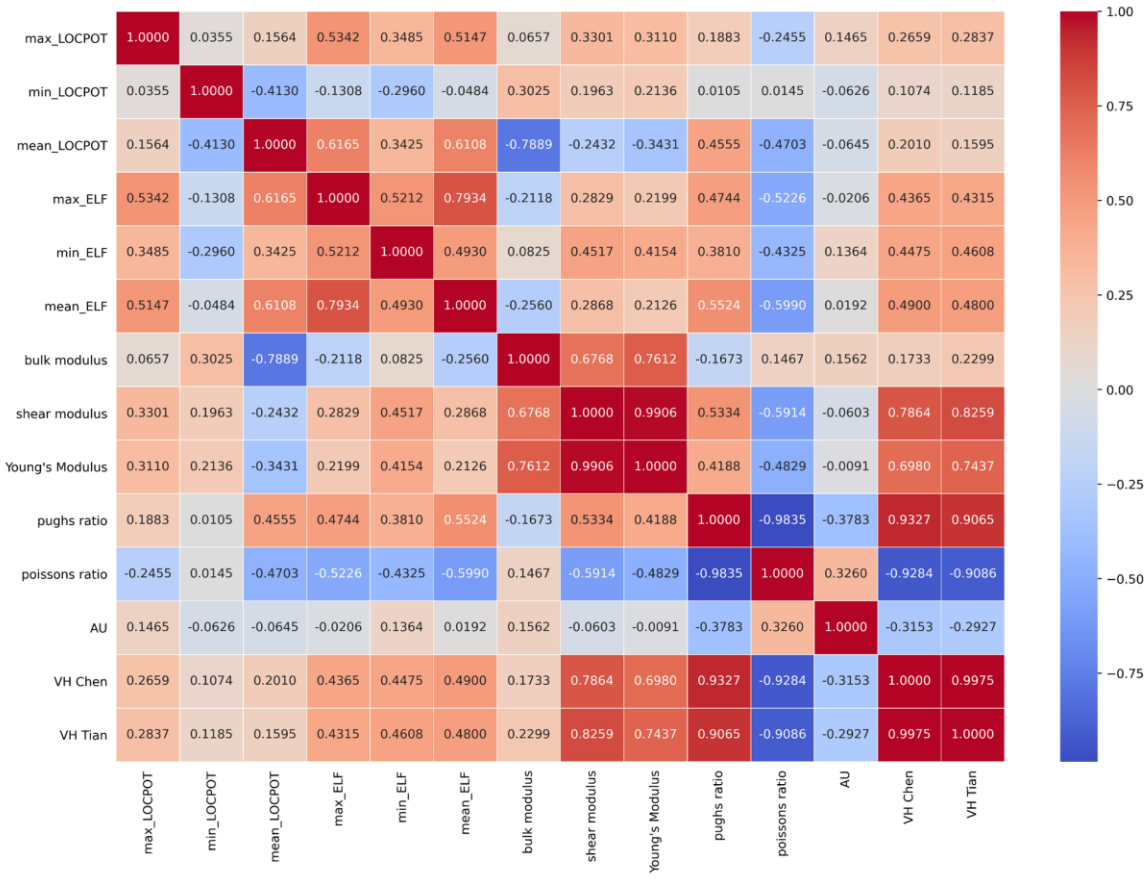

Fig. 8. Material descriptor analysis. Pearson correlation matrix between maximum local potential, minimum local potential, mean local potential, maximum electron localization function, minimum electron localization function, mean electron localization function, bulk modulus, shear modulus, elastic modulus Pugh's ratio, and Poisson's ratio.

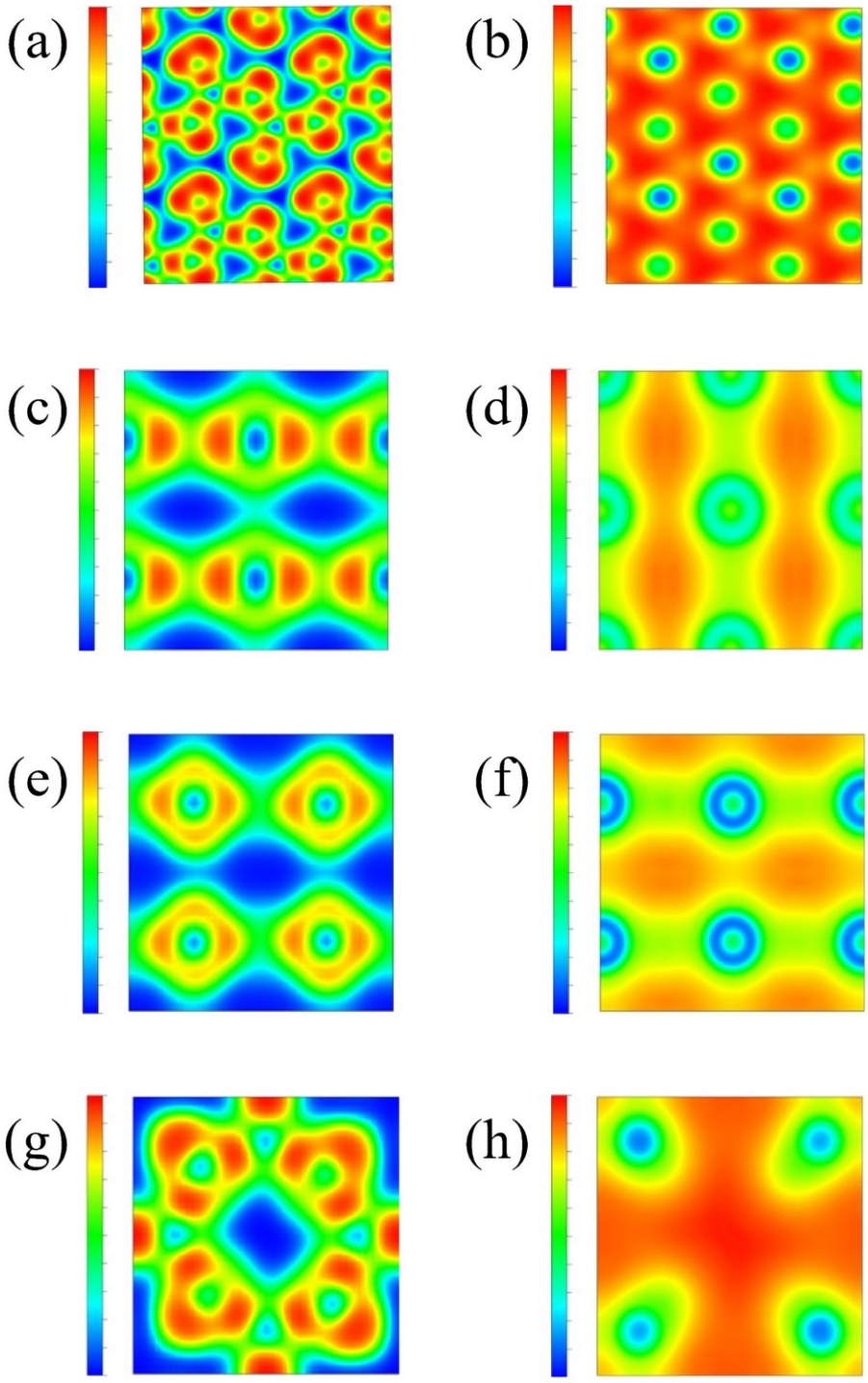

Fig. 9. Electronic level insight into ultrahigh hardness. The ELF and LOCPOT plot of the structure $B_4C_8N_4$ (mp-1079201) (a, b), BC_7 (mp-1078935) (c, d) recommended by BLOX showing the presence of more electrons and covalent bonding. (e-h) The ELF and LOCPOT plot of two identified structures $Be_4C_8N_4$ (mp-1189451) and $C_{12}N_8$ (mp-1188347) showing the presence of strong covalent bonding.

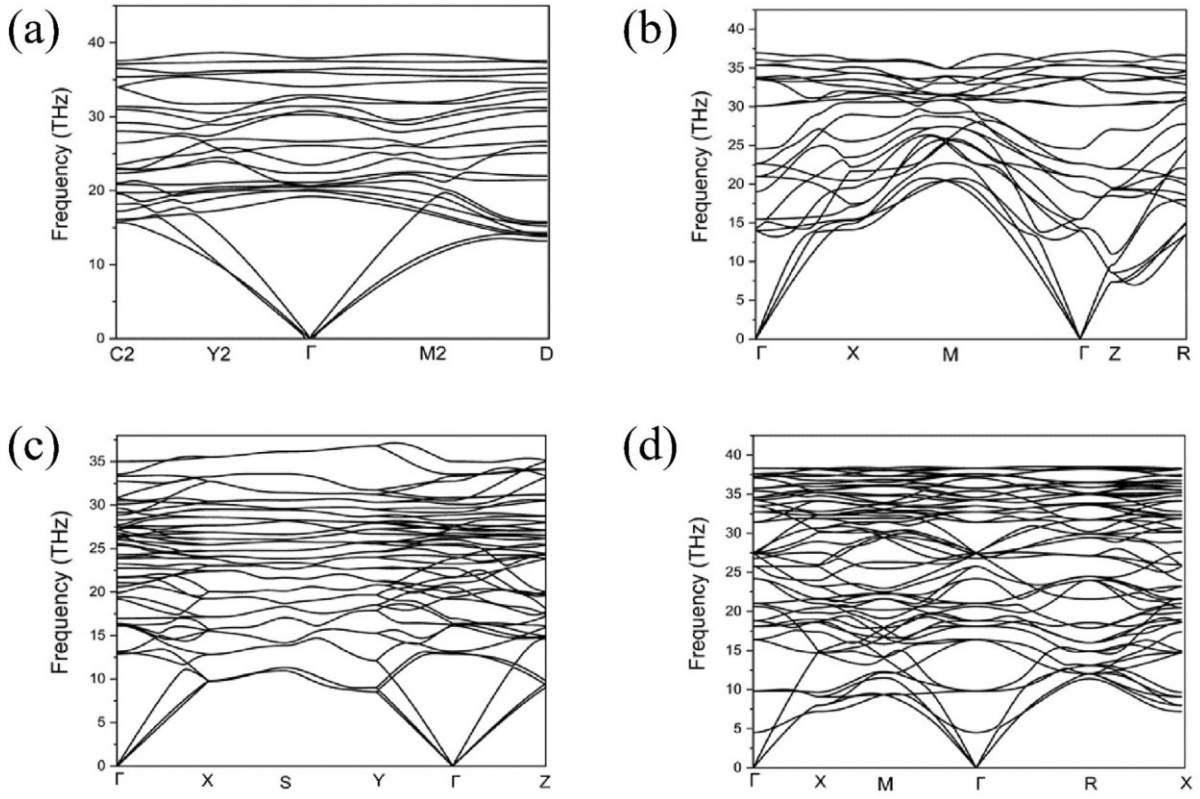

Fig. 10. Thermodynamic stability analysis of ultrahard materials. Phonon dispersions of (a) $B_4C_8N_4$, (b) BC_7, (c) $Be_4C_8N_4$, and (d) $C_{12}N_8$ along high symmetry paths in the Brillouin zone. There is no negative frequency in phonon dispersions, indicating these structures are thermodynamically stable.

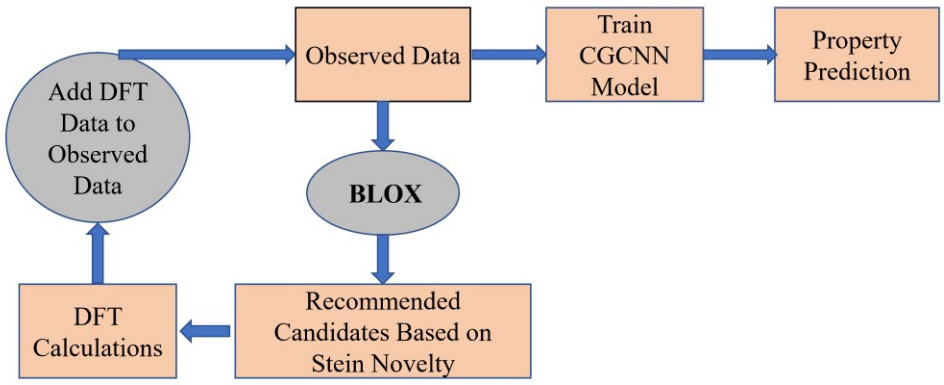

Fig. 11. Schematic of workflow of BLOX algorithm. The loop of DFT/BLOX/recommendations were performed at least 4 rounds until there is no significant amount of interested material properties recommended by BLOX algorithm.