Finding the Next Superhard Material through Ensemble Learning

Ziyan Zhang,[1] Aria Mansouri Tehrani,[1] Anton O. Oliynyk,[2]
Blake Day,[1] Jakoah Brgoch[1*]

[1]Department of Chemistry, University of Houston,
Houston, TX 77204, USA
[2]Department of Chemistry and Biochemistry, Manhattan College,
Riverdale, NY 10471, USA

*To whom correspondence should be addressed; E-mail: jbrgoch@uh.edu.

We report an ensemble machine-learning method capable of finding new superhard materials by directly predicting the load-dependent Vickers hardness based only on the chemical composition. A total of 1062 experimentally measured load-dependent Vickers hardness data were extracted from the literature and used to train a supervised machine-learning algorithm utilizing boosting, achieving excellent accuracy (R^2 = 0.97). This new model was then tested by synthesizing and measuring the load-dependent hardness of several unreported disilicides as well as analyzing the predicted hardness of several classic superhard materials. The trained ensemble method was then employed to screen for superhard materials by examining more than 66,000 compounds in crystal structure databases, which showed that only 68 known materials surpass the superhard threshold. The hardness model was then combined with our data-driven phase diagram generation tool to expand the limited num-

ber of reported compounds. Eleven ternary borocarbide phase spaces were studied, and more than ten thermodynamically favorable compositions with superhard potential were identified, proving this ensemble model's ability to find previously unknown superhard materials.

Introduction

Superhard materials are essential in applications ranging from manufacturing to energy production. They also have substantial use in the aerospace, military, and even health care industries. (*1*) Finding new superhard materials that have a Vickers hardness (H_V) greater than 40 GPa has traditionally been guided by empirical design rules derived from classically known materials like diamond, *c*-BN, and more recently ReB_2, among others. (*2, 3*) It is widely accepted that a three-dimensional network of short covalent bonds and a high valence electron density, will likely lead to enhanced hardness. (*4–6*) However, these rules are all qualitative. Researchers have also considered quantitative approaches using computational methods like molecular dynamics (MD) simulations or density functional theory (DFT) for identifying and understanding superhard materials. (*7–11*) Unfortunately, large scale MD simulations are computationally expensive and generally impractical for multicomponent systems. (*12*) DFT is also computationally expensive and cannot directly calculate hardness, although there has been some success using DFT calculated proxies to estimate a material's hardness based on the elastic moduli. (*13–16*) For example, finding proportionality coefficients that relate H_V to different combinations of the elastic moduli have produced several semi-empirical hardness models with varying accuracy. (*17*) Nevertheless, constructing simple mathematical models remains insufficient to distinguish the multidimensional relationship between chemical composition, crystal structure, microstructure, and hardness. More recently, machine-learning methods, which can capture such complex connections, have been created to identify new superhard materials based

on the elastic moduli. (*18*) The speed of machine learning allows for rapid materials screening, but this method still relies on computationally-derived indirect proxies of hardness that are subject to misinterpretation.

These approaches are all generally able to roughly approximate some version of a material's H_V as a single value. However, they cannot predict a material's real response to plastic deformation that yields hardness. (*19*) A material's Vickers hardness varies with the load applied to the indenter tip, which is referred to as indentation size effect. (*20*) In every measurement, the hardness decreases asymptotically as the load increases. This observation is unexpected because hardness should be independent of the applied load. The origin and the mechanism of the indentation size effect is not well understood, although possible explanations include measurement error because of the small indentation imprint, sizeable elastic recovery, or changes in the microstructure. (*21, 22*) As a result, even if a method can predict a material's hardness at a single load, it is not likely to capture load-dependent response. This remains a significant barrier in superhard materials design.

In this study, we overcome this challenge by constructing a machine-learning model capable of directly predicting load-dependent hardness based only on chemical composition. The number of unique data available in the literature is relatively limited (only $\approx$1000 examples); thus, ensemble learning algorithms are employed to train the model. Ensemble learning methods are effective at dealing with small and sparse data-sets. (*23*) Generally, these algorithms work by running a base learner multiple times, and form a vote out of the resulting hypotheses. (*24*) The outcome is an improved model with reduced variance and bias. Here, a Random Forest (RF) machine-learning algorithm was employed as the starting point. (*25*) Boosting ensemble algorithms, including Gradient Boosting (GB) trees (*26*) and XGBoost (XGB) (*27*), were then demonstrated to considerably improve the model. The predictive power was subsequently validated using two different hold-out test sets. Eight unmeasured metal disilicides were synthe-

sized as a phase pure materials. Their Vickers hardness was then measured at different applied loads. Moreover, a customized hold-out test set containing several classic superhard materials was created. The results both showed a remarkable reproduction of the Vickers hardness, including capturing the load-dependent hardness curves for all compounds. Once the ensemble algorithms were trained and validated, this composition-only featurized model was used to screen ≈66,000 known compounds in Pearson's Crystal Data (PCD) set. Only 68 compounds (0.1%) are suggested to be superhard, and many of these phases are derivatives of already reported high hardness materials. The sparse number of possible superhard materials suggests it is improbable to find entirely new superhard materials using this strategy. We address this inadequacy by merging this hardness model with our recently developed formation energy and convex hull prediction tool to identify more than ten previously unreported superhard compounds. This work proves that using ensemble learning to predict load-dependent hardness can potentially provide the next big step in the search for new superhard materials.

Methods

Construction of the Ensemble Learning Models The training set was composed of 1062 experimentally measured Vickers hardness ($H_{V,\text{exp}}$) data extracted from literature. Regression models were constructed to predict the Vickers hardness using a random forest (RF) algorithm ($H_{V,\text{RF}}$), gradient boosting (GB) trees ($H_{V,\text{GB}}$), and XGBoost (XGB) algorithms ($H_{V,\text{XGB}}$). The data were represented by a feature-set containing 35 distinct compositional descriptors and four mathematical expressions, including the difference, the average, the maximum value, and the minimum value. In total, the initial feature-set contained 140 compositional features. Additionally, the applied load from each hardness measurement was included as a feature resulting in 141 total features. The full list of features is provided in Table S1. Recursive feature elimination (RFE) (*28*) was then used to reduce the 141 descriptors by pruning the least important

feature and recursively considering a smaller set of features for each model. The two metrics used to evaluate the model performance were the coefficient of determination (R^2) and the mean squared error (MSE), shown in Figure S1. The optimal features for each algorithm are provided in Tables S2, S3, and S4, respectively. Hyperparameter settings were adjusted for each model constructed using a 10-fold cross-validated grid search where exhaustive evaluations of all parameter combinations were performed. The searching space includes the maximum depth of trees in the range of [3, 4, 5, 6, 7, 8], the learning rate in the range of [0.01, 0.03, 0.05, 0.07, 0.10, 0.15, 0.20], the subsample ratio of columns when constructing each tree in the range of [0.6, 0.7, 0.8, 0.9, 1] and the subsample ratio of the training instances in the range of [0.6, 0.7, 0.8, 0.9, 1]. RF and GB were used through the Scikit-learn python implementation (*29*) while XGB was also used within python environment. (*27*) All codes, training data, and prediction sets associated with this work are provided in the open-source Github repository at https://github.com/BrgochGroup.

Synthesis of MSi$_2$ for Model Validation Pellets of the nominal compositions MSi$_2$ (M = Cr, V, Nb, Ta, Mo, W, La, and Ce) were prepared by starting from the high-purity metals: chromium powder (99.995%, Alfa Aesar), vanadium powder (99.995%, Alfa Aesar), niobium pieces (99.995%, Alfa Aesar), molybdenum powder (99.995%, Alfa Aesar), tungsten powder (99.995% Alfa Aesar), La metal (99.9%, HEFA Rare Earth Canada Co. Ltd., Canada), Ce metal (99.9% HEFA Rare Earth Canada C . Ltd., Canada), and silicon powder (99.999%, Alfa Aesar, USA). The powders were weighed out in a 1:2 ratio and homogenized by manually blending the samples using an agate mortar and pestle. The mixtures were then pressed into 6 mm pellets using a Carver hydraulic press with an applied pressure of 1.5 metric tons. These pellets were placed on a copper hearth in an arc melter (CenTorr Vacuum Industries, model 5SA) along with an oxygen getter (titanium metal). The chamber was sealed and evacuated under vacuum for 1

minute, followed by filling with argon; this process was repeated at least three times. The Ti was melted first to ensure any residual oxygen was removed, and then samples were arc-melted using a forward current of typically 30 A to 70 A for 1 to 2 min. The buttons were flipped and re-melted at least two times to ensure homogeneity. The weight loss after arc-melting was $<1\%$.

The products were split in two, with half ground into a fine powder using a CerCo Diamonite mortar and pestle for analysis by powder X-ray diffraction. The X-ray diffractograms were collected using a PANalytical X'Pert powder diffractometer equipped with Cu Kα radiation (λ = 1.54184 Å). The diffractograms were all analyzed by Le Bail refinement performed with the General Structure Analysis System (GSAS) software and the EXPGUI interface. The data and associated refinement details for each sample are provided in Figure S2 and Table S5. The other half of each sample was mounted in an epoxy resin and polished to a mirror surface using SiC polishing plates (600–1200 grit) followed by 7 μm, 3 μm, and 1 μm diamond paste. The Vickers hardness was measured by making ten indentations on each sample using microindentation (LECO AMH55, LM810AT) with applied loads of 0.49 N (0.049 kgf), 0.98 N (0.098 kgf), 2.94 N (0.294 kgf), 4.9 N (0.49 kgf) and averaging the results.

Result and Discussion

Feature Development and Ensemble Learning Model Construction Constructing a machine learning model to predict load-dependent hardness first requires gathering a sufficient amount of training data. A total of 1062 experimentally measured Vickers hardness values were manually extracted from literature, plus the applied load and chemical composition. The data distribution of the 1062 training samples is provided in Figure S3, along with the element population count among the training set. The training set is composed mostly of binary and ternary systems and covers H_V values spanning from low-load ($<$10 GPa) to high-load ($>$60

GPa) with ≈80% of the data falling between 0 GPa and 40 GPa. It also covers a variety of elements, except the alkali metals, halogens, and the noble gases, which are not frequently encountered in structural materials. The data were randomly split into ten separate training and test sets with a 9:1 ratio. All model statistics were determined by averaging the different seeded models. The initial hardness predictions were made using an RF algorithm with the 40 features identified by RFE (Table S2).

As shown in Figure 1a, the predicted $H_{V,\mathrm{RF}}$ for 106 compounds in the test set (10% of the entire training set) reproduces the experimentally measured hardness ($H_{V,\mathrm{exp}}$) reasonably well. The average R^2 was 0.90, whereas MSE shows a more significant deviation of 18.6 GPa. Although RF can capture the general trend and the lower hardness ($<$15 GPa) values, RF misses the high hardness predictions resulting in an overall inferior model not particularly capable of predicting superhard materials.

This model's limited capability is likely because there is sparse hardness data available in the literature, especially in the superhard region. Ensemble learning was thus investigated as a way to assist in the prediction of load-dependent hardness. Numerous ensemble learning methods have been developed with the most common being bagging (bootstrap aggregating) (*30*) and boosting. (*31*) Bagging works by sampling uniformly with replacement from the original dataset whereas boosting sequentially adds one weak leaner at a time to the ensemble focusing more on the data points poorly handled by the previous model. (*32*) Only recently have ensemble learning methods been used in materials science. (*33*)

Applying GB (Figure 1b) and XGB (Figure 1c) to predict Vickers hardness both show a dramatic improvement in their ability to reproduce compared to RF. The $H_{V,\mathrm{GB}}$ model delivers an excellent R^2 (0.96) and much a more agreeable MSE (6.6 GPa). The changes are particularly noticeable $>$20 GPa where excellent agreement between $H_{V,\mathrm{exp}}$ and $H_{V,\mathrm{GB}}$ is obtained. Switching from GB to XGB further improves the model with $H_{V,\mathrm{XGB}}$ having an R^2 = 0.97 and MSE

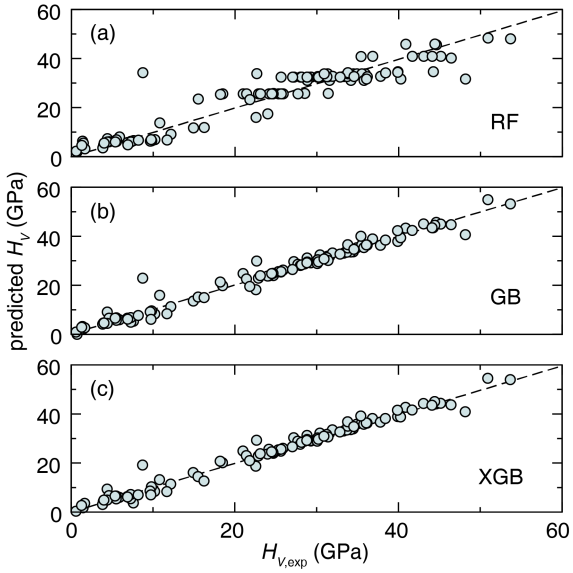

Figure 1: The experimentally measured hardness ($H_{V,\mathrm{exp}}$) for a representative test set is plotted against the (a) random forest (RF) model ($H_{V,\mathrm{RF}}$), (b) gradient boosting (GB) tree model ($H_{V,\mathrm{GB}}$), and the (c) XGBoost (XGB) model ($H_{V,\mathrm{XGB}}$). The ideal fit (1:1) is shown as the dashed line.

of only 5.7 GPa. The systematic optimization of XGB provides a significant improvement in terms of the model performance compared to conventional gradient boosting, and XGB is also faster to train . (*27*) Given the superb performance of XGB, this method was used for all future hardness predictions.

Validating Vickers Hardness and the Load-Response Curve Predictions Model validation requires a carefully controlled systematic study of relevant compounds. Metal disilicides, $M\mathrm{Si}_2$ (M = Cr, V, Nb, Ta, Mo, W, La, and Ce), have broad industrial applications as coating materials and can be used as high hardness materials. (*34*) Surprisingly, their load-dependent Vickers hardness is not widely published. (*35–38*) Given the diversity of compositions that make up these important structural materials, they are an ideal original validation set. Therefore, eight samples were synthesized as pure-phase products based on their refined powder X-ray diffractograms (Figure S2). The Vickers hardness of each compound was then experimentally obtained

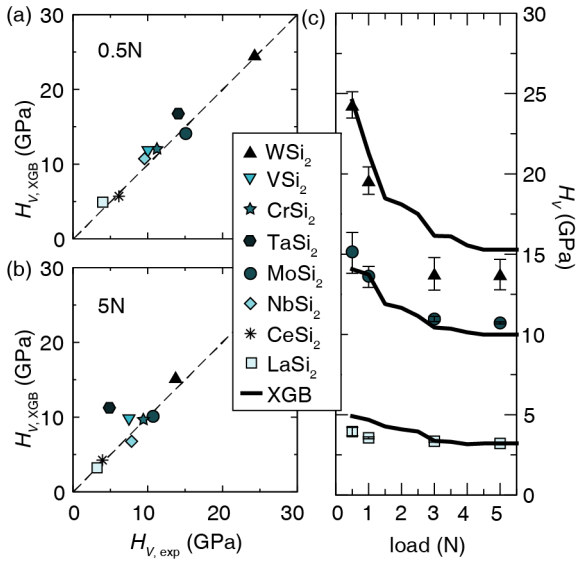

Figure 2: The machine learning predicted hardness ($H_{V,\mathrm{XGB}}$) against the experimentally measured hardness ($H_{V,\mathrm{exp}}$) of eight disilicides at (a) low applied load (0.5 N) and (b) high applied load (5 N). (c) The load-dependent hardness is plotted for three selected disilicides along with the predicted $H_{V,\mathrm{XGB}}$. All load-dependent hardness data are available in Figure S4. The symbols represent the different disilicides.

by microindentation. The load-dependent $H_{V,\mathrm{XGB}}$ was also predicted from 0.5 N to 5 N.

The $H_{V,\mathrm{XGB}}$ is shown for the eight compounds at low (0.5 N) load (Figure 2a) and at high (5 N) load (Figure 2b). These are two of the commonly reported applied loads in the literature. The $H_{V,\mathrm{XGB}}$ of all disilicides showed a striking reproduction of the $H_{V,\mathrm{exp}}$. The low load $H_{V,\mathrm{exp}}$ measured for the samples spans from ≈5 GPa ($LaSi_2$) to ≈25 GPa (WSi_2). The XGB model captures each compound's hardness across this entire range, with all of the predictions falling near the 1:1 line. The R^2 = 0.95 for the 0.5 N applied load and an MSE = 1.8 GPa, supporting this model's superior predictive power. The higher applied load (5 N) also shows the $H_{V,\mathrm{XGB}}$ of all disilicides all shift to lower values, which is observed experimentally. There is a bit more scatter in these predictions, presumably because the training set contains fewer high load data (Figure S3c). The model still quantitatively captures Vickers hardness and, more importantly, the trend of increasing the applied load producing a lower measured hardness.

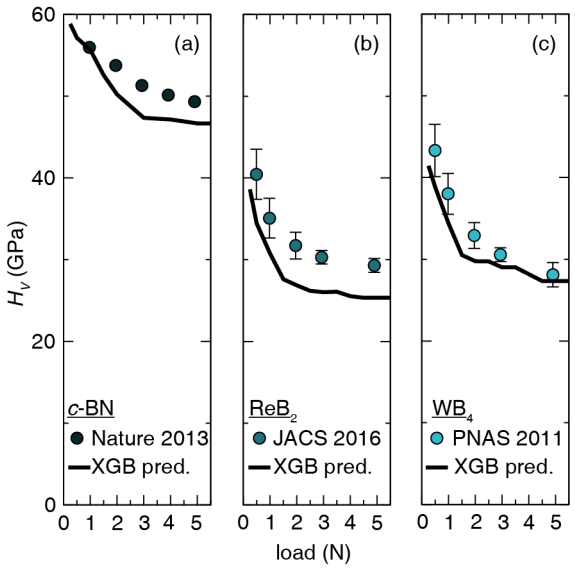

Figure 3: The hold-out test prediction of load-dependent hardness shows excellent agreemnt with the experimentally measured hardness for (a) cubic-BN (*39*) (b) WB_4 (*40*) (c) ReB_2 (*41*).

The full load-dependent hardness curves were also predicted for the disilicides to further demonstrate the power of the ensemble learning. Figure 2c shows that WSi_2, $MoSi_2$ and $LaSi_2$ all achieve a striking agreement between $H_{V,\text{exp}}$ and $H_{V,\text{XGB}}$ at every load. The predicted load-dependent hardness curves for the full set of disilicides are provided in Figure S4. The model can quantitatively predict the hardness and reliably reproduce the different curves of the hardness as a function of the applied load. These results prove that our XGB model can estimate the Vickers hardness based solely on chemical composition, and it can also reproduce the details and shape of load-dependent hardness measurements.

The XGB model can unquestionably predict the hardness of the disilicides, which have low to moderate hardness. A customized hold-out test of superhard materials was also created to ensure our model's ability to find superhard materials. Three well known materials with an $H_{V,\text{exp}} \geq 40$ GPa, cubic-BN, (*39*), ReB_2, (*41, 42*), and WB_4, (*40*) were each removed from the training set, and the model was retrained. This new model was then used to predict the load-dependent hardness for each compound. These data were then compared to $H_{V,\text{exp}}$ published in

the literature, plotted in Figure 3. The ensemble learning model brilliantly captures the 0.5 N hardness as well as the hardness loss with increasing load. Indeed, the predicted load-dependent hardness for *c*-BN, ReB_2, and WB_4 all achieved an excellent quantitative agreement with their reported values. Increasing the load causes a decrease in the measured hardness, captured by the $H_{V,\mathrm{XGB}}$ model, with only a slight underestimation of the hardness predicted for all three compounds. Most importantly, the model can reproduce the hardness of compounds containing only light elements as well as transition metal borides. This is notable because these materials have rather different chemistry, yet XGB can predict both with high accuracy suggesting remarkable transferability of the model.

Evaluating Known Materials for Superhard Response This model showed high statistical accuracy and reliability for quickly generating $H_{V,\mathrm{XGB}}$. Several disilicides and well-known high hardness compounds were then used for further validation. Given the successful implementation of the approach, our model was consequently employed to predict the hardness of inorganic compounds contained in Pearson's Crystal Data (PCD) set. It is essential to remember that the training set only covered a limited number of elements frequently encountered in structural materials (shown in Figure S4), and the model's extrapolation power to elements not present in the training set can not be guaranteed. Therefore, the prediction set was restricted to compositions containing only elements present in the training set, resulting in the analysis of 66440 compounds.

Predicting the load-hardness curves for such a large number of compounds allows high-level screening and can provide essential guidance before experimental synthesis. Of course, the maximum hardness at low load is not the only vital property for structural materials. The high load hardness is also essential for many applications. Ideally, materials should be minimally influenced by the indentation size effect and shown a minimal change between the low-load

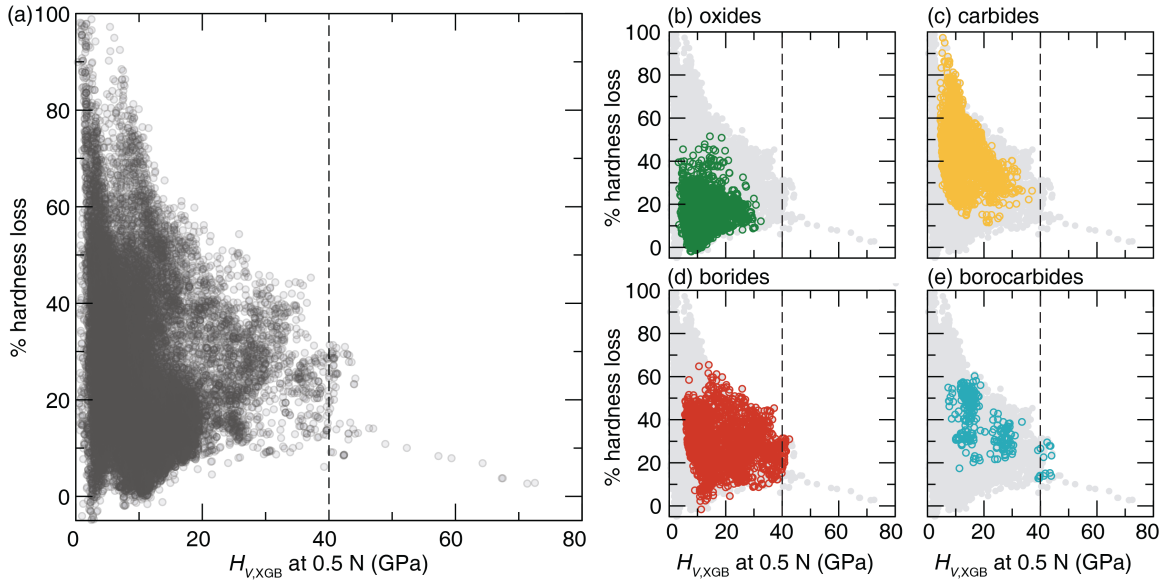

Figure 4: (a) The machine learning $H_{V,\mathrm{XGB}}$ model at 0.5 N is plotted against the % hardness loss when load increases from 0.5 N to 5 N for 66440 inorganic compounds in PCD. Darker regions indicate data overlap. Specific classes of traditional structural materials are shown, including (b) oxides, (c) carbides, (d) borides, and (e) borocarbides.

hardness and the high-load hardness, which we describe here as a hardness loss percentage following: % hardness loss= $((H_{V,0.5N} - H_{V,5N}) / H_{V,0.5N)}) \times 100\%$. Thus, the Vickers hardness was predicted at 0.5 N and 5.0 N for the compounds obtained from PCD, and these data were used to calculate the % hardness loss.

The overall $H_{V,\mathrm{XGB}}$ at low load (0.5 N) of the PCD prediction set was then plotted against the % hardness loss, shown in Figure 4a. The darker regions on this plot represent a higher density of compounds, and the vertical dashed line shows the superhard cut-off (40 GPa). Ideal superhard materials should fall in the bottom right region of this plot, indicating a high low load hardness and a minimal loss of hardness with increasing load. From this plot, it is clear that several outstanding compositions occur well above 40 GPa and have a minimal ($<$15%) drop in their hardness. Analyzing these points reveals these phases contain only light main group elements with diamond-type structures such as borocarbonitrides with slightly varied stoichiometry and

boron suboxide analogs. This outcome is somewhat expected given the superior superhard materials are diamond and *c*-BN. A larger number of compositions containing metals combined with the light main group elements are present closer to the superhard threshold. Nevertheless, only 68 compounds (0.1%) from our plot predicted to be superhard ($H_{V,\mathrm{XGB}} \geq 40$ GPa). All compositions located in the superhard region are listed in Table S6. Superhard materials are indeed exceedingly scarce. It is also noteworthy that even for moderately hard materials in the range of 25 GPa to 35 GPa, some compounds still have a high loss (up to 50%), which is not desirable. Almost all of the compounds (≈98%) analyzed from PCD are predicted to have a hardness lower than 25 GPa.

Decomposing the results into the compositional metadata reveal several trends that may prove critical for designing superhard materials. Plotting each compound-type shows the distribution of traditional structural materials: borides, carbides, oxides, and borocarbides. For example, Figure 4b shows that oxides tend only to have a moderate hardness (≤30 GPa), but when the applied load increases, they tend to have a small loss. Thus, these materials would be valuable for applications where moderate hardness is required, and a range of applied loads will be experienced. Carbides (Figure 4c) show similar hardness to oxides, although they have a hardness loss that spans the entire range, meaning the material choice for each application is delicate. Borides, shown in Figure 4d, are a crucial group containing several superhard materials such as ReB_2 and WB_4. The plot shows that borides have a higher hardness than most materials, as expected, including several compounds borides that surpass 40 GPa at 0.5 N. Many of these compounds are boron-rich phases, including $Ir_{0.02}B$, $Re_{0.01}B$, $Ti_{0.05}B$, $Mn_{0.05}B$, and $Sc_{1.61}B_{103}$, among numerous others. To the best of our knowledge, these phases have not been studied as superhard materials. There is still a clear opportunity to find new superhard borides. Finally, borocarbides, shown in Figure 4e, appear to have several promising compounds that have a $H_{V,\mathrm{XGB}}$ exceeding 40 GPa and a drop of only ≈30%, which makes these compounds

competitive with borides. Examining the composition of the top candidates in the borocarbide group revealed that all of these promising compounds contain early transition metal (*TM*) like Y, Ti, and V along with boron and carbon. This result is particularly exciting because the metals are more earth-abundant and undoubtedly cheaper to prepare than the 5*d* transition metal borides. Thus, borocarbides are of great interest and have significant potential to warrant further investigation as superhard materials.

Discovering Novel Superhard Borocarbides by Screening Phase Diagrams Screening large crystallographic databases using machine learning is one of the most common methods currently used for "materials discovery." Although this approach allows a rapid assessment of known compounds, this process does not tend to yield any transformative materials in most instances. Fortunately, the machine learning approach developed here is based only on composition and the applied load. This allows the model to make general predictions of hardness for any chemical composition without knowing the crystal structure. Merging this idea with our recently developed convex-hull phase diagram analysis (*43*), we can identify regions of composition space where new, unreported compounds with a hardness $>$ 40 GPa are likely to reside.

Owing to the promise of superhard response in borocarbides combined with the limited number of reported phases, we investigated eleven *TM*-B-C ternary phase diagrams where the *TM* is an early 3*d*, 4*d*, or 5*d* transition metal from groups III–VI on the periodic table. The *TM*-B-C phase diagrams were constructed by first creating a composition grid that contains 253 compositions in each ternary system. The hardness of each composition was then predicted using the XGB model (at 0.5 N). The resulting predictions for selected diagrams are shown in Figure 5, where the hardness range is plotted as a contour map only for $H_{V,\mathrm{XGB}} \geq$40 GPa. All of the ternary diagrams created are provided in Figure S5. Analyzing these plots shows that

the hardest compositions in each ternary plot occur for the boron and carbon-rich compositions. Most of these compositions may be superhard, at least at 0.5 N. The hardest phases are predicted to form when the carbon content is quite high with a maximum $H_{V,\mathrm{XGB}}$ at 0.5 N of 64 GPa. Moving toward the boron corner of the composition plot indicates these phases remain superhard, which is in agreement with the results from analyzing the PCD and previously developed empirical rules. The addition of a higher metal concentration reduces the hardness while a metal content above ≈25 mol.% falls below 40 GPa. Regardless, the combinations of boron and carbon with select early transition metals certainly show outstanding promise as high hardness materials.

These diagrams indicate the regions where the compositions may yield superhard behavior; however, there is no guarantee that any unknown compounds exist in the high hardness areas. Thus, to provide information on the thermodynamic stability of compounds on these ternary phase diagrams, the formation energy (E_f) and the associated convex hull (E_{hull} = 0 eV) was also determined using machine learning. Our previously built phase diagram model has proven extremely useful for the discovery of intermetallic materials. (*43*) Applying the method here produces multiple compounds on the convex hull, shown by the black squares on each plot. These represent the most energetically favorable compositions that should be observed at equilibrium (at 0 K). The high-temperature synthetic routes generally used for the synthesis of superhard materials, such as arc melting, also allows the acquisition of near-equilibrium phases. Thus, compositions with a $0 < E_{hull} < 50$ meV (580 K) were also identified.

The reliability of both machine learning models ($H_{V,\mathrm{XGB}}$ and E_f) was first confirmed by placing any experimentally reported compounds in Figure 5, plotted as the red circles. There are three reported Sc-B-C compounds in the PCD superhard region ($ScB_{13}C$, ScB_2C_2 and $Sc_3B_{51}C_{0.75}$) and these are also predicted to fall within 50 meV of the convex hull. Additionally, one Ti-B-C phase ($Ti_{0.93}B_{24}C$), plotted in Figure 5b, and one Y-B-C phase ($Y_{0.6}B_{14}C_{0.6}$),

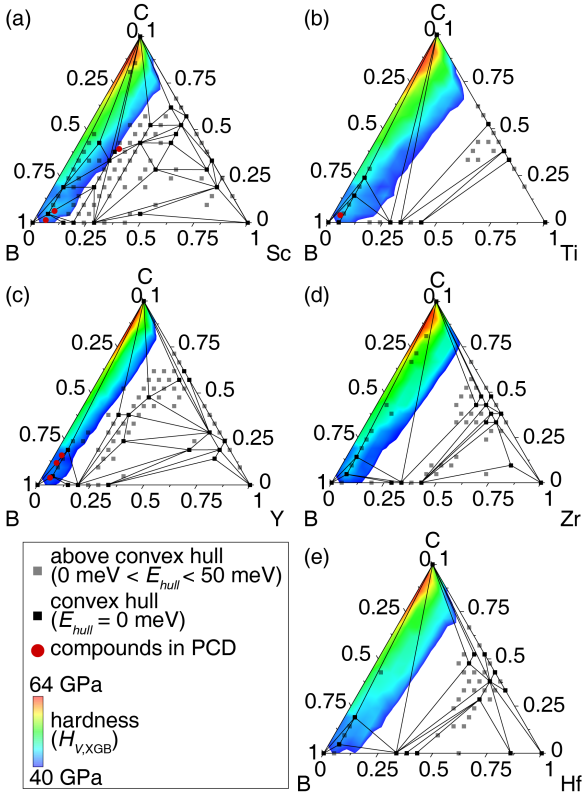

Figure 5: Possible superhard compounds can be identified by using machine learning to identify compositions on convex hull (black squares), within 50 meV of the convex hull (gray squares). Compounds experimentally reported in the PCD are shown by the red circles and the predicted hardness ($H_{V,\mathrm{XGB}}$) is the contour plot. The composition spaces examined include (a) Sc-B-C, (b) Ti-V-C, (c) Y-B-C, (d) Zr-B-C, and (e) Hf-B-C.

plotted in Figure 5c, were also found to occur on the convex hull, while two additional Y-B-C phases, $YB_{28.5}C_4$ and $YB_{24}C_{4.8}$, fall 57 meV and 64 meV, respectively, above the hull. They also have a $H_{V,\mathrm{XGB}}$ above 40 GPa. The observation that all of these compositions are on (or near) the convex hull supports the reliability of applying our convex hull analysis and hardness model in combination to find new superhard materials.

Analyzing the plots for unknown compositions that fall on (or near) the convex hull (the square points) shows plenty of opportunities to find new superhard materials. In Figure 5a, there are two possible Sc-B-C superhard phases, $Sc_2B_{10}C_9$ and $Sc_4B_9C_8$, that both fall on the convex hull and have a hardness of $>$40 GPa at 0.5 N. Similarly, there are multiple titanium-containing compositions (Figure 5b), including $TiB_{17}C_8$ and $TiB_{15}C_5$ that reside on the convex hull and show probable superhard behavior. The high hardness composition identified in the Zr-B-C phase is ZrB_5C_{15} with a $H_{V,\mathrm{XGB}}$ = 53 GPa at 0.5 N (Figure 5d). This composition is only 12 meV above the convex hull and thus could feasibly be synthesized. Further analysis of Figure 5 (and Figure S5) showcases many other reasonable superhard materials as well as regions on the phase diagram that could yield a myriad of structural properties. The application of machine learning and formation energy in tandem decidedly provides to be a vital approach for the rapid identification of novel superhard materials.

Conclusion

Hardness is an essential mechanical property necessary for a myriad of modern applications. The development of superhard materials has historically relied on experimentally derived empirical rules or indirect computational proxies. This has mostly restricted the ability to quickly identify new superhard materials to known structure types or derivative compositions. Here, we presented a new ensemble learning model that can directly predict Vickers hardness, including the anomalous load-dependent hardness with quantitative accuracy. The model was

validated by studying eight metal disilicides as well as creating a customized hold-out test set of superhard materials. Both provided impressive certainty in obtaining the hardness at all loads for these materials. The model was then employed to predict the hardness of 66440 compositions in Pearson's crystal data set, which suggested possible superhard properties in only 68 previously unstudied materials. Moreover, analyzing these data indicated that transition metal borocarbides are an underexplored yet promising space to find new superhard materials. Further investigating the phase diagrams by predicting the formation energies along with the hardness for numerous ternary borocarbide phase spaces suggested at least ten brand new superhard materials are waiting to be synthesized. This method of directed discovery is poised to modernize the search for new superhard materials benefiting from the efficient, scalable, and transferable nature of machine learning.

Acknowledgments

The authors thank Roy Arrieta and Jacob Hickey for the assistance with this project. The authors gratefully acknowledge the generous financial support provided by the University of Houston Division of Research through a High Priority Area Research Seed Grant. Additional support was provided by the Welch Foundation (E-1981) and the Texas Center for Superconductivity at the University of Houston (T_CSUH).

References

1. R. B. Kaner, J. J. Gilman, S. H. Tolbert, *Science* **308**, 1268 (2005).

2. H.-Y. Chung, *et al.*, *Science* **316**, 436 (2007).

3. R. Wentorf, R. C. DeVries, F. Bundy, *Science* **208**, 873 (1980).

4. J. Haines, J. Leger, G. Bocquillon, *Annu. Rev. Mater. Res.* **31**, 1 (2001).

5. M. T. Yeung, R. Mohammadi, R. B. Kaner, *Annu. Rev. Mater. Res.* **46**, 465 (2016).

6. J. B. Levine, S. H. Tolbert, R. B. Kaner, *Adv. Func. Mater.* **19**, 3519 (2009).

7. A. Tiwari, S. Natarajan, *Applied Nanoindentation in Advanced Materials* (Wiley Online Library, 2017).

8. Q. Gu, G. Krauss, W. Steurer, *Adv. Mater.* **20**, 3620 (2008).

9. A. G. Kvashnin, Z. Allahyari, A. R. Oganov, *J. Appl. Phys.* **126**, 040901 (2019).

10. P. Avery, *et al.*, *npj Comp. Mater.* **5**, 1 (2019).

11. A. M. Tehrani, J. Brgoch, *J. Solid State Chem.* **271**, 47 (2019).

12. P. Walsh, *et al.*, *Appl. Phys. Lett.* **82**, 118 (2003).

13. X. Jiang, J. Zhao, X. Jiang, *Comput. Mater. Sci.* **50**, 2287 (2011).

14. X. Jiang, J. Zhao, A. Wu, Y. Bai, X. Jiang, *J. Phys. Condens. Matter* **22**, 315503 (2010).

15. N. Miao, B. Sa, J. Zhou, Z. Sun, *Comput. Mater. Sci.* **50**, 1559 (2011).

16. X. Q. Chen, H. Niu, D. Li, Y. Li, *Intermetallics* **19**, 1275 (2011).

17. A. T. Mansouri, L. Ghadbeigi, J. Brgoch, T. D. Sparks, *Integr. Mater. Manuf. Innov.* **6**, 1 (2017).

18. A. T. Mansouri, *et al.*, *J. Am. Chem. Soc.* **140**, 9844 (2018).

19. L. Lu, *et al.*, *Proc. Natl. Acad. Sci.* **117**, 7052 (2020).

20. G. M. Pharr, E. G. Herbert, Y. Gao, *Annu. Rev. Mater. Res.* **40**, 271 (2010).

21. E. Broitman, *Tribo. Lett.* **65**, 1 (2017).

22. A. Iost, R. Bigot, *J. Mater. Sci.* **31**, 3573 (1996).

23. C. Zhang, Y. Ma, *Ensemble machine learning: methods and applications* (Springer, 2012).

24. T. G. Dietterich, *The handbook of brain theory and neural networks* **2**, 110 (2002).

25. A. Liaw, M. Wiener, *R news* **2**, 18 (2002).

26. J. H. Friedman, *Ann. Stat.* pp. 1189–1232 (2001).

27. T. Chen, C. Guestrin, *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (2016), pp. 785–794.

28. I. Guyon, J. Weston, S. Barnhill, V. Vapnik, *Mach. Learn.* **46**, 389 (2002).

29. F. Pedregosa, *et al.*, *J. Mach. Learn. Res.* **12**, 2825 (2011).

30. L. Breiman, *Mach. Learn.* **24**, 123 (1996).

31. Y. Freund, R. E. Schapire, *ICML* (Citeseer, 1996), vol. 96, pp. 148–156.

32. Y. Freund, R. E. Schapire, *European conference on computational learning theory* (Springer, 1995), pp. 23–37.

33. D. Xue, *et al.*, *Nat. Commun.* **7**, 1 (2016).

34. M. E. Schlesinger, *Chem. Rev.* **90**, 607 (1990).

35. T. C. Murthy, *et al.*, *J. Refract. Hard. Met.* **28**, 529 (2010).

36. Y. Pan, P. Mao, H. Jiang, Y. Wan, W. Guan, *Ceram. Int.* **43**, 5274 (2017).

37. G. Schultes, M. Schmitt, D. Goettel, O. Freitag-Weber, *Sen. Actuators A.* **126**, 287 (2006).

38. D. G. Morris, M. Leboeuf, M. Morris, *Mater. Sci. Eng. A.* **251**, 262 (1998).

39. Y. Tian, *et al.*, *Nature* **493**, 385 (2013).

40. R. Mohammadi, *et al.*, *Proc. Natl. Acad. Sci.* **108**, 10958 (2011).

41. A. T. Lech, *et al.*, *J. Am. Chem. Soc.* **138**, 14398 (2016).

42. J. Qin, *et al.*, *Adv. Mater.* **20**, 4780 (2008).

43. S. Lotfi, *et al.*, *Matter* p. in press (2020).

Supplementary materials

Figures S1 to S5

Table S1 to S6

Supplementary Materials for
Finding the Next Superhard Material through Ensemble Learning

Ziyan Zhang,[1] Aria Mansouri Tehrani,[1] Anton O. Oliynyk,[2]
Blake Day,[1] Jakoah Brgoch[1*]

[1]Department of Chemistry, University of Houston,
Houston, TX 77204, USA
[2]Department of Chemistry and Biochemistry, Manhattan College,
Riverdale, NY 10471, USA

[*]To whom correspondence should be addressed; E-mail: jbrgoch@uh.edu.

This PDF file includes:

Figures S1 to S5

Table S1 to S6

Table S1: Complete descriptor set for predicting the load dependent hardness

descriptor number	compositional variables
1−4	Atomic number
5−8	Atomic weight
9−12	Period
13−16	Group
17−20	Families
21−24	Mendeleev number
25−28	Atomic radii
29−32	Covalend radii
33−36	Zunger radii sum
37−40	Ionic radii
41−44	Crystal radii
45−48	Pauling EN
49−52	Martynov-Batsanov EN
53−56	Gordy EN
57−60	Mulliken EN
61−64	Allred-Rockow EN
65−68	Metallic valence
69−72	Number of valence electrons
73−76	Gilman number of valence electron
77−80	Valence s
81−84	Valence p
85−88	Valence d
89−92	Number of outer shell electrons
93−96	First ionization potential
97−100	Polarizability
101−104	Melting point
105−108	Boiling point
109−112	Density
113−116	Specific heat
117−120	Heat of fusion
121−124	Heat of vaporization
125−128	Thermal conductivity
129−132	Heat atomization
133−136	Cohesive energy
137−140	Electron affinity
141	Load

Table S2: The optimal feature set for the random forest (RF) model determined by recursive feature elimination.

descriptor number	compositional variables
1	Avg. Families
2	Avg. Mendeleev Number
3	Avg. Atomic radii
4	Avg. Covalent radii
5	Avg. Zunger radii sum
6	Avg. Ionic radii
7	Avg. Crystal radii
8	Avg. Pauling EN
9	Avg. Martynov-Batsanov EN
10	Avg. Gordy EN
11	Avg. Mulliken EN
12	Avg. Allred-Rockow EN
13	Avg. Number of valence electrons
14	Avg. Gilman number of valence electrons
15	Avg. Valence d
16	Avg. First ionization potential
17	Avg. Polarizability
18	Avg. Melting point
19	Avg. Density
20	Avg. Heat of fusion
21	Avg. Heat of vaporization
22	Avg. Thermal conductivity
23	Avg. Heat atomization
24	Avg. Cohesive energy
25	Avg. Electron affinity
26	Diff. Covalend radii
27	Diff. Zunger radii sum
28	Diff. Boiling point
29	Diff. Density
30	Diff. Heat of vaporization
31	Max. Covalent radii
32	Max. Melting point
33	Max. Specific heat
34	Max. Heat of vaporization
35	Min. Mendeleev number
36	Min. Mulliken EN
37	Min. First ionization potential
38	Min. Heat of fusion
39	Min. Thermal conductivity
40	Load

Table S3: The optimal feature set for the gradient boosting (GB) trees model determined by recursive feature elimination.

descriptor number	compositional variables
1	Avg. Mendeleev number
2	Avg. Covalent radii
3	Avg. Zunger radii sum
4	Avg. Ionic radii
5	Avg. Crystal radii
6	Avg. Martynov-Batsanov EN
7	Avg. Gordy EN
8	Avg. Mulliken EN
9	Avg. Gilman number of valence electrons
10	Avg. Valence d
11	Avg. First ionization potential
12	Avg. Polarizability
13	Avg. Melting point
14	Avg. Specific heat
15	Avg. Heat of fusion
16	Avg. Heat of vaporization
17	Avg. Thermal conductivity
18	Avg. Heat atomization
19	Avg. Cohesive energy
20	Avg. Electron affinity
21	Diff. Covalent radii
22	Diff. Martynov-Batsanov EN
23	Diff. Gordy EN
24	Diff. Heat of vaporization
25	Max. Melting point
26	Max. Specific heat
27	Min. Crystal radii
28	Min. First ionization potential
29	Min. Heat of fusion
30	Load

Table S4: The optimal feature set for the XGBoost (XGB) model determined by recursive feature elimination.

descriptor number	compositional variables
1	Avg. Families
2	Avg. Atomic Radii
3	Avg. Covalend Radii
4	Avg. Zunger radii sum
5	Avg. Ionic radii
6	Avg. Crystal radii
7	Avg. Gordy EN
8	Avg. Mulliken EN
9	Avg. Number of valence electrons
10	Avg. Valence d
11	Avg. Number of outer shell electrons
12	Avg. Polarizability
13	Avg. Melting point
14	Avg. Specific heat
15	Avg. Heat of fusion
16	Avg. Heat atomization
17	Avg. Cohesive energy
18	Diff. Mendeleev number
19	Diff. Covalent radii
20	Diff. Martynov-Batsanov EN
21	Diff. Gordy EN
22	Diff. Mulliken EN
23	Diff. Gilman number of valence electron
24	Diff. Heat of fusion
25	Diff. Heat of vaporization
26	Max. Atomic radii
27	Max. Valence d
28	Max. Melting point
29	Max. Specific heat
30	Max. Heat of fusion
31	Max. Thermal conductivity
32	Min. Mendeleev number
33	Min. Crystal radii
34	Min. Martynov-Batsanov EN
35	Min. Gordy EN
36	Min. Mulliken EN
37	Min. First ionization potential
38	Min. Melting point
39	Min. Heat of fusion
40	Load

Table S5: Refinement data of the disilicides MSi$_2$ (M = Mo, Ta, V, Nb, La, Ce, Cr and W).

sample	$MoSi_2$	$TaSi_2$	VSi_2	$NbSi_2$
space group	*I*4/*mmm*	$P6_222$	$P6_222$	$P6_222$
a (Å)	3.20693	4.78260	4.57376	4.79882
b (Å)	3.20693	4.78260	4.57376	4.79882
c (Å)	7.85024	6.56704	6.3753	6.5939
α (°)	90	90	90	90
β (°)	90	90	90	90
γ (°)	90	120	120	120
V (Å^3)	80.7	130.1	115.5	131.5
molar mass (g/mol)	152.11	237.12	107.11	149.08
T (K)	296	296	296	296
radiation type; λ (Å)	Cu Kα; 1.54184	Cu Kα; 1.54184	Cu Kα; 1.54184	Cu Kα; 1.54184
2θ (°)	10.00-80.00	10.00-80.00	10.00-80.00	10.00-80.00
refinement	Pawley	Pawley	Pawley	Pawley
number of data points	4117	4117	4117	4117
sample	$LaSi_2$	$CeSi_2$	$CrSi_2$	WSi_2
space group	$I4_1/amd$	*Imma*	$P6_222$	*I*4/*mmm*
a (Å)	4.3089	4.1994	4.424055	3.198119
b (Å)	4.3089	4.1947	4.424055	3.198119
c (Å)	13.8611	13.9419	6.358241	7.797669
α (°)	90	90	90	90
β (°)	90	90	90	90
γ (°)	90	90	120	90
V (Å^3)	80.7	130.1	115.5	131.5
molar mass (g/mol)	257.4	245.6	322.47	480.04
T (K)	296	296	296	296
radiation type; λ (Å)	Cu Kα; 1.54184	Cu Kα; 1.54184	Cu Kα; 1.54184	Cu Kα; 1.54184
2θ (°)	10.00-80.00	10.00-80.00	20.00-90.00	15.00-90.00
refinement	Pawley	Pawley	LeBail	LeBail
number of data points	4117	4117	6538	6538

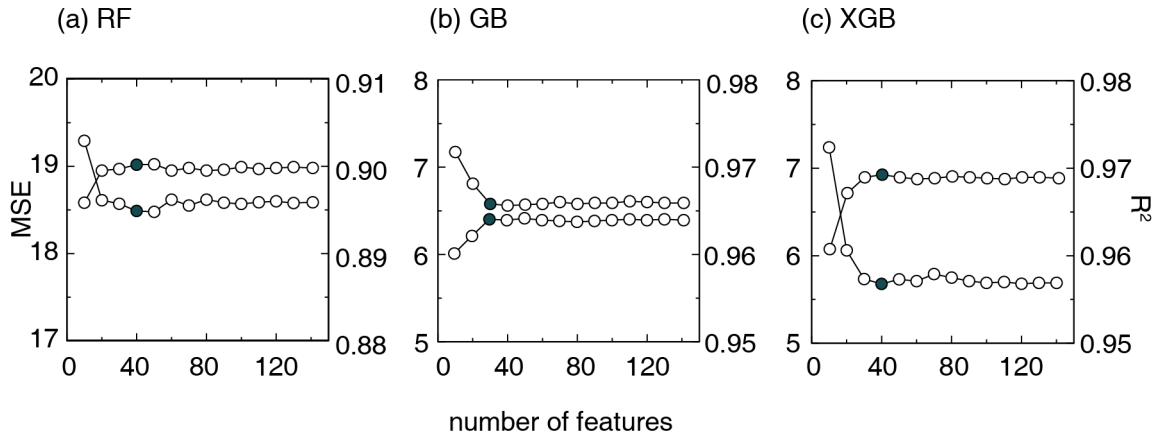

Figure S1: The recursive feature elimination (RFE) results for (a) random forest (RF) model, (b) gradient boosting (GB) trees model and (c) XGBoost (XGB) model, where filled circles represent the optimal number of features for each model.

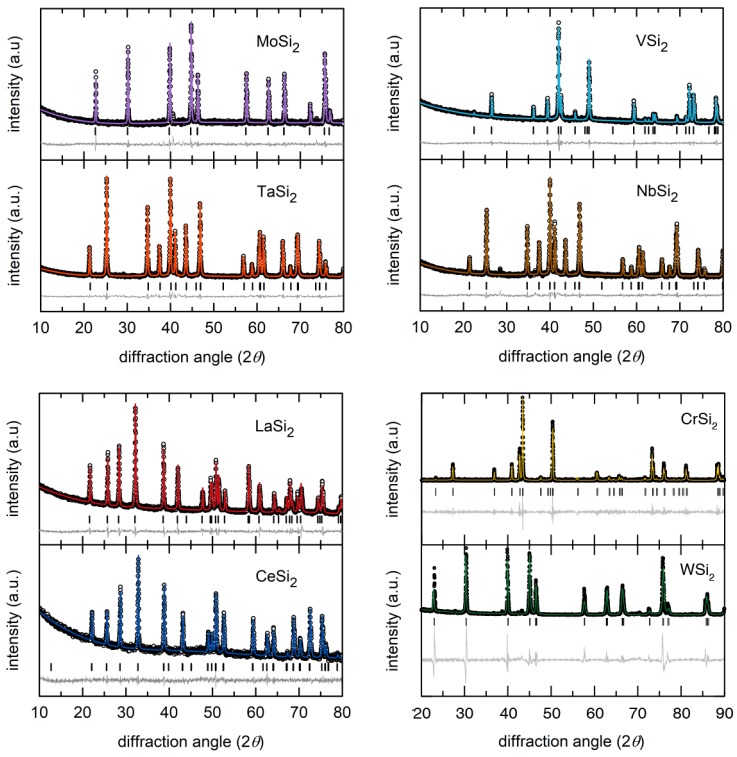

Figure S2: Refinement of disilicides MSi_2 (M = Mo, Ta, V, Nb, La, Ce, Cr and W) X-ray powder diffraction data. The observed data are black circles, the refinements are colored for each sample, and the differences are grey lines.

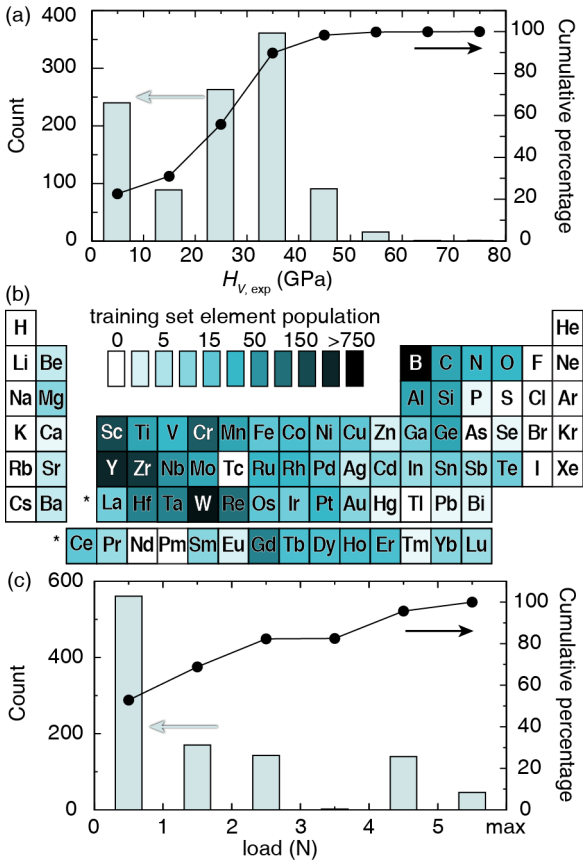

Figure S3: (a) Data distribution of 1062 training samples that were extracted from literature. The bars represent the total counts in each bin and the curve represent the cumulative percentage. (b) The element frequency of 1062 training samples. (c) Distribution of applied loads containing in the 1062 training samples extracted from literature. The bars represent the total counts in each bin and the curve represent the cumulative percentage.

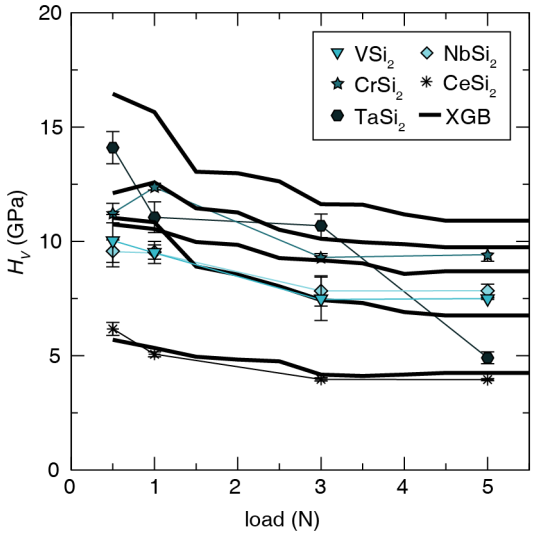

Figure S4: The XGB predicted load-dependent hardness for MSi_2 (M = Cr, V, Nb, Ta and Ce).

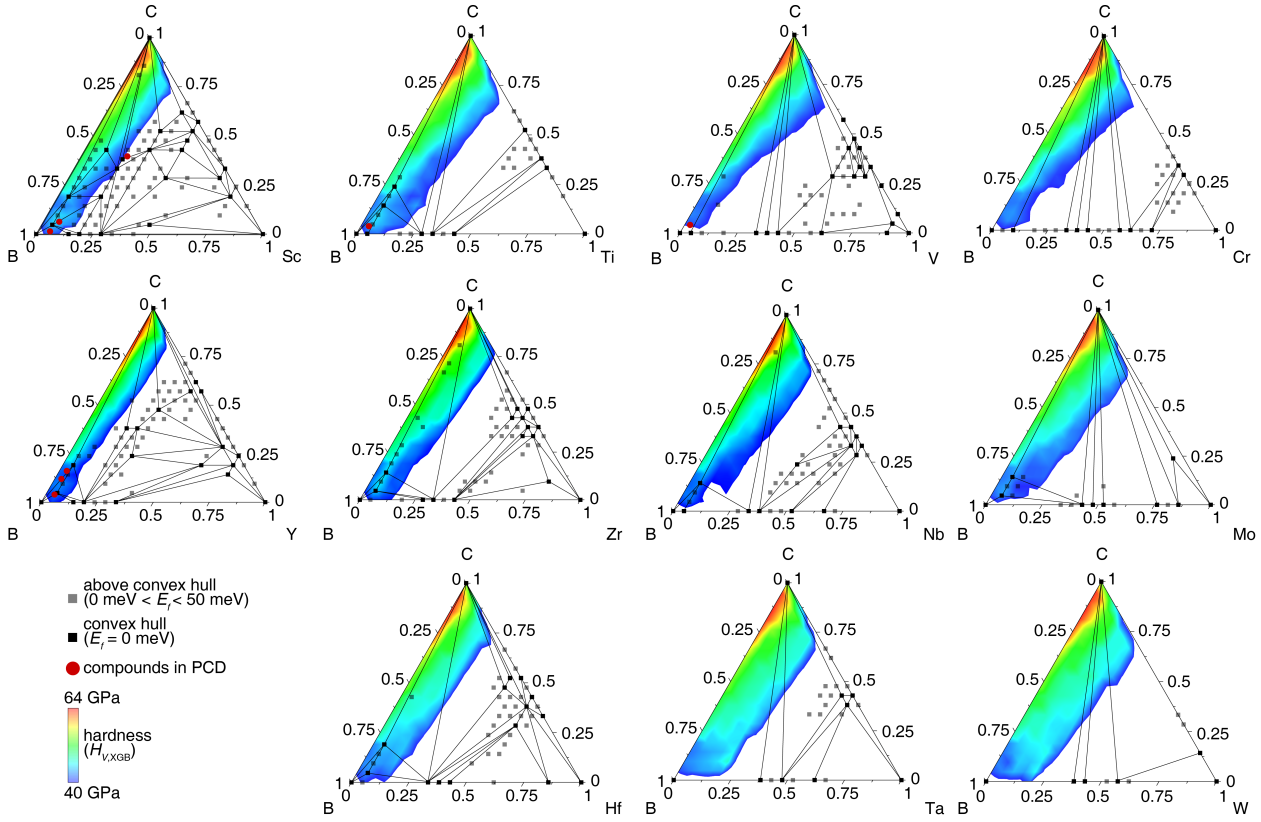

Figure S5: The predicted hardness phase diagrams for eleven TM-B-C ternary plots.

Table S6: The predicted superhard materials ($>$ 40GPa) in PCD, along with their % hardness loss when load increases to 5N.

Composition	$H_{V,XGB}$ at 0.5N	% hardness loss
$B_3C_{10}N_3$	72.56	2.80
$B_{0.57}C_{0.86}N_{0.57}$	71.32	2.65
$B_{0.67}C_{0.66}N_{0.67}$	67.47	3.73
$B_{0.67}C_{0.67}N_{0.66}$	67.31	3.74
$B_{0.86}C_{0.41}N_{0.73}$	64.29	6.87
$B_{0.918}C_{0.139}N_{0.943}$	59.36	7.91
$B_{1.1}N_{0.9}$	57.31	8.34
$B_{5.5}O$	52.31	10.03
$B_6O_{0.86}$	48.91	10.84
$B_{13}C_{1.33}O_{0.67}$	47.02	14.06
$B_6O_{0.79}$	46.36	14.12
$B_6O_{0.83}$	44.59	13.10
$B_6O_{0.84}$	44.15	13.00
$Ti_{0.93}B_{24}C$	43.95	23.29
$YB_{24}C_{4.8}$	43.85	13.85
$ScB_{13}C$	43.57	27.78
$YB_{28.5}C_4$	43.38	15.29
$Sc_{0.1}B_{12.3}C_{0.58}Si_{0.1}$	43.23	29.22
$Y_{0.6}B_{14}C_{0.6}$	43.18	27.84
$B_{25}N$	43.10	13.01
$B_{25.03}N_{0.94}$	43.09	13.01
$Al_{0.74}B_{24}C_4N_{1.02}$	43.09	11.77
$W_{1.2}Ni_{0.6}B_9$	42.53	30.96
$B_{13.72}C_{1.52}$	42.42	8.63
$B_{13.5}C_{1.5}$	42.42	8.63
$Sc_3B_{51}C_{0.75}$	42.36	29.48
$B_{13.6}C_{1.6}$	42.35	8.43
$B_{13.74}C_{1.51}$	42.34	8.57
$B_{13.75}C_{1.5}$	42.29	8.51
BeB_2C_2	41.72	15.04
$Y_{1.1}B_{66.4}$	41.37	25.40
$Y_{1.01}B_{66}$	41.32	25.43
$W_{1.2}Pt_{0.6}B_9$	41.20	30.76
$Al_{0.3}B_{13.3}C_{1.3}$	41.18	13.86
$Sc_{0.05}B$	41.17	29.38
$YB_{65.86}$	41.16	25.21
$Y_{1.15}B_{66}$	41.14	25.61

$V_{0.645}B_{24}C$	41.11	22.59
$Ir_{0.02}B$	41.08	25.38
$YB_{41.2}Si_{1.42}$	40.89	28.65
$Sc_{1.32}B_{105.2}$	40.82	20.00
$Au_{0.01}B$	40.81	24.57
$YB_{41}Si_{1.25}$	40.80	28.61
$Y_{0.55}B_{14}$	40.78	28.35
$Sc_{3.68}B_{101.78}$	40.75	27.15
$YB_{41}Si_{1.2}$	40.74	28.66
$Re_{0.01}B$	40.74	21.35
ScB_{15}	40.71	29.95
$Ti_{0.05}B$	40.71	22.97
$MgB_{12}Si_2$	40.70	27.10
$Sc_{1.61}B_{103}$	40.69	25.39
$W_{0.9}Pd_{0.9}B_9$	40.67	31.32
$YNi_{0.06}B_{41}Si_{1.3}$	40.67	28.57
$Mg_2B_{24}C$	40.54	29.67
$YRh_{0.02}B_{41.1}Si_{1.1}$	40.51	28.73
$Sc_2Cu_{0.77}B_{45}$	40.46	29.58
$W_{0.6}Rh_{1.2}B_9$	40.36	29.21
$B_{24.97}C_{0.91}$	40.31	9.14
$Mn_{0.78}B_{105.9}$	40.20	18.17
$Ti_{0.84}B_{25}$	40.19	19.85
$V_{0.09}Re_{0.91}B_2$	40.14	28.71
$Ta_{0.15}Re_{0.85}B_2$	40.13	31.37
$Mn_{0.05}B$	40.06	24.45
$Mn_{4.48}B_{103}$	40.04	24.87
$Mg_3B_{36}Si_9C$	40.02	25.60
ScB_2C_2	40.02	26.61
$Hf_{0.01}B$	40.01	19.64