ARTICLE   OPEN

Predicting superhard materials via a machine learning informed evolutionary structure search

Patrick Avery [1], Xiaoyu Wang[1], Corey Oses [2], Eric Gossett[2], Davide M. Proserpio [3,4], Cormac Toher [2,5], Stefano Curtarolo [2,5] and Eva Zurek [1]

The computational prediction of superhard materials would enable the in silico design of compounds that could be used in a wide variety of technological applications. Herein, good agreement was found between experimental Vickers hardnesses, H_v, of a wide range of materials and those calculated by three macroscopic hardness models that employ the shear and/or bulk moduli obtained from: (i) first principles via AFLOW-AEL (AFLOW Automatic Elastic Library), and (ii) a machine learning (ML) model trained on materials within the AFLOW repository. Because H_v^{ML} values can be quickly estimated, they can be used in conjunction with an evolutionary search to predict stable, superhard materials. This methodology is implemented in the XTALOPT evolutionary algorithm. Each crystal is minimized to the nearest local minimum, and its Vickers hardness is computed via a linear relationship with the shear modulus discovered by Teter. Both the energy/enthalpy and $H_{v,Teter}^{ML}$ are employed to determine a structure's fitness. This implementation is applied towards the carbon system, and 43 new superhard phases are found. A topological analysis reveals that phases estimated to be slightly harder than diamond contain a substantial fraction of diamond and/or lonsdaleite.

npj Computational Materials (2019)5:89 ; https://doi.org/10.1038/s41524-019-0226-8

INTRODUCTION

Superhard materials are important in a wide variety of applications including cutting and polishing, or as abrasives and coatings. They typically contain light elements such as B, C, N, and O, which can form short and strong covalent bonds, and they may have complex potential energy surfaces (PES) with numerous low lying minima. Because the main atomic constituents of superhard materials all have similar masses, it is sometimes impossible to determine their crystal structures from X-ray diffraction patterns alone. As a result, first-principles calculations have been instrumental in uncovering the structures of a number of superhard phases including $R3m$-BC$_2$N,[1] the M-carbon phase formed upon cold compression of graphite,[2–4] $Pnnm$-CN,[5] and cubic BC$_3$.[6]

Many superhard materials are metastable at ambient pressure/temperature conditions, and the synthesis procedure can have an impact on the product that is made. Therefore, a synergistic feedback loop between theory and experiment is required for the rational design of superhard phases with specific properties. For example, recently an entropy-forming ability descriptor was developed and used to predict novel carbides containing five metals with a high hardness that were subsequently synthesized.[7] Another example is the development of machine learning (ML) models for the elastic moduli, which were subsequently employed to screen hundreds of thousands of compounds found in a crystal structure database. The most promising superhard ternaries, Re$_{0.5}$W$_{0.5}$C and ReWC$_{0.8}$, were subsequently synthesized and experiments confirmed they were superhard at low load.[8]

The a priori prediction of superhard materials is a global optimization problem that requires maximizing the computed hardness in the space of the low-lying local minima on the PES. Microscopic models employing quantities that can be readily obtained such as the geometry of the crystal, atomic connectivity, valence electron density, and electronegativity or ionicity of the constituent atoms[9–13] can be used to estimate a structure's hardness. Recently, the well-known crystal structure prediction (CSP) algorithms CALYPSO[12] and USPEX[13] have been extended towards the prediction of superhard phases. The structures generated in the course of a CSP search with these programs are typically optimized using first-principles calculations, and their hardnesses are calculated using modified versions of either the Šimunek-Vackář (SV)[10] or the Li et al.[11] methods, respectively. The main deficiency of the original Li and SV equations is that they predict unreasonably high hardness values for crystals with non-bonding interactions that are necessary for maintaining the three dimensional structure of the system, such as the van der Waals forces between different layers of graphene in graphite. The hardness models developed in CALYPSO and USPEX overcome these limitations by using chemical graph theory to determine which atoms are bonded to each other.

The hardness of a material can also be estimated given its elastic properties, such as the bulk and shear moduli.[14–16] One of the advantages of these macroscopic hardness models is that they do not depend explicitly upon quantities that are ambiguous or hard to define, such as the atomic radii, bond electronegativity, or the bond strength. However, because elastic properties are expensive to calculate from first-principles, they have so far not been coupled with CSP searches where the hardness of hundreds, if not thousands, of structures would need to be obtained within a

[1]Department of Chemistry, State University of New York at Buffalo, Buffalo, NY 14260-3000, USA; [2]Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC 27708, USA; [3]Dipartimento di Chimica, Università degli Studi di Milano, Via Golgi 19, 20133 Milano, Italy; [4]Samara Center for Theoretical Materials Science (SCTMS), Samara State Technical University, Molodogvardeyskaya St. 244, 443100 Samara, Russia and [5]Center for Autonomous Materials Design, Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC 27708, USA
Correspondence: Stefano Curtarolo (stefano@duke.edu) or Eva Zurek (ezurek@buffalo.edu)

single run. Models that require the computed elastic properties are typically only employed to determine the hardness of a handful of promising structures found within a search.

Large materials databases that contain many measured and calculated observables[17–22] have allowed for the advent of ML models that have demonstrated predictive power for numerous properties, for example superconducting critical temperatures,[23] electronic band gaps,[24,25] elastic properties,[24] and the melting temperatures of unary and binary component solids.[26] AFLOW (Automatic FLOW) is an automatic framework for high-throughput materials discovery[17] that includes a materials database with over 2 million entries.[27] Many properties have been calculated for the materials in the AFLOW database including vibrational properties with the Automatic Phonon Library (APL),[17] thermal properties with the Automatic GIBBS Library (AGL),[28] and thermomechanical properties such as the bulk and shear moduli with the Automatic Elasticity Library (AEL).[22,29] It is now possible to interact with the AFLOW database via a RESTful API[30] and the AFLUX materials search API.[31]

Herein, we illustrate that the Vickers hardnesses, H_v, of a wide variety of crystalline materials predicted by using a macroscopic hardness model in conjunction with ML-derived bulk and shear moduli obtained via the RESTful API[32] available on AFLOW are in excellent agreement with results obtained from first-principles calculations. Both are in good agreement with experiment. These developments make it possible to quickly calculate reasonable hardness values for a given crystal structure using ML-based elastic properties, and these hardness estimates can subsequently be employed to calculate an individual's fitness in a CSP algorithm designed for the prediction of superhard phases. This technique is implemented within the XTALOPT evolutionary algorithm (EA),[33,34] and is subsequently applied towards the carbon system to search for stable and superhard phases. Seventy-nine dynamically stable, low energy, distinct topologies with $H_v > 40$ GPa are found in our searches. Forty-three of the predicted structures have not been reported previously.

RESULTS AND DISCUSSION

Macroscopic hardness models coupled with machine learning

Because the chemical bonding within a crystalline lattice affects the bulk modulus, B, of the material, it has been proposed that B can be a good indicator of hardness.[35,36] While this is true for specific classes of materials, such as diamond-like semiconductors, it turns out that, although some exceptions exist, the shear modulus, G, is a much better predictor of hardness.[14,37,38] The linear correlation between hardness and G was originally noted by Teter in 1998,[14] and in 2011, Chen et al. used the geometric shape of the Vickers indenter to derive the correlation coefficient.[15] We refer to the resulting equation for estimating hardness,

$$H_{v,Teter} = 0.151G, \tag{1}$$

as the *Teter model* to distinguish it from another model developed by Chen and co-workers in the same manuscript. Chen et al. suggested that discrepancies between $H_{v,Teter}$ and the experimentally measured Vickers hardnesses resulted from neglecting the plastic deformations in Eq. (1). In order to better account for these effects, they proposed a new formula that employs the famous Pugh's modulus ratio, $k = G/B$, which correlates well with a material's brittleness. The parameters in the modified equation, which we refer to as the *Chen model*, were obtained by an empirical fit yielding

$$H_{v,Chen} = 2(k^2 G)^{0.585} - 3. \tag{2}$$

Equation (1) was found to work well for brittle materials with a large k. The Chen model, Eq. (2), yielded better estimates of the

hardness in most cases, with one exception being crystals with a low k where the bonding was primarily metallic, such as fcc Al.

In 2012 Tian and co-workers noted that the intercept term in Eq. (2) did not have a physical basis, and would yield negative values for some materials such as KI and KCl.[16] Therefore, they obtained a revised formula, which we refer to as the *Tian model*, via refitting the original function proposed by Chen et al. without the intercept term as

$$H_{v,Tian} = 0.92k^{1.137}G^{0.708}. \tag{3}$$

The Vickers hardnesses estimated with the formulae of Chen and Tian are in good agreement with experimental measurements for many systems, but they tend to overestimate the hardness of ionic compounds and materials for which H_v is less than ~5 GPa.

The two parameters in the above equations, G and B, have been computed for over 5000 unique materials in the AFLOW database via AEL.[27] We employed their Voigt-Reuss-Hill average values, G_{VRH}, and B_{VRH}, to estimate the Vickers hardnesses of 64 systems, including many of those studied by Chen[15] (data for the full set can be found in Section S1 in the Supplementary Information, SI; Table 1 lists results only for those systems considered in ref. [15]). This set of materials spanned a wide range of hardness values ($H_v^{Exp} = 0.2 - 96.0$ GPa), and included ionic and covalently bonded crystals, as well as intermetallics. It has been demonstrated that AFLOW-AEL calculates reasonable bulk and shear moduli,[29] so it should not be a surprise that the Vickers hardnesses obtained using them are in good agreement with experiment, and the predicted values reflect the variation in the results of the models themselves.

To analyze the agreement between the Vickers hardnesses computed via AFLOW-AEL and the three different models in more detail, a correlation analysis was performed, and the results are provided in Fig. 1, which also plots the theoretical vs. the experimental data. The Pearson and Spearman coefficients measure the linear correlation between two variables, and the monotonicity of the relation between them, respectively. The normalized root-mean-squared relative deviation, RMSrD, between the experimental and computed values was also found, as it can be useful for distinguishing between different methods that may have similar correlations with, but different deviations from, experiment. Whereas the Pearson and Spearman coefficients should ideally be equal to 1, a lower RMSrD is indicative of a better agreement between theory and experiment. All three models showed a high correlation, though the Pearson coefficient was somewhat closer to unity for the Chen and Tian models, as compared to the Teter model. The RMSrD is clearly lower for the Teter model than the other two for this data set when AEL moduli are used. However, the Teter model tends to underestimate the Vickers hardnesses of materials with $H_v > 40$ GPa. Many of the systems in the full data set are soft, with $H_v < 5$ GPa, and, as previously mentioned, the macroscopic models are not so good at predicting their hardness values. A correlation analysis was also carried out for the subset of materials studied by Chen,[15] see Table 1, and the results are provided in Section S2 in the SI. The correlation coefficients for this subset of systems are higher than for the full data set, and the RMSrD values are significantly lower, <1.

The overwhelming majority of structures generated in a CSP search will not be found within a repository such as AFLOW. Moreover, it would not be feasible to carry out the first-principles geometry optimizations (24 per structure, unless reducible by symmetry) that AFLOW-AEL requires to obtain the bulk and shear moduli for the hundreds, if not thousands, of structures generated within a single CSP search. Therefore, we wondered if it would be possible to take advantage of the power of data science by using bulk and shear moduli obtained from the ML model that was trained on the materials within the AFLOW repository instead?[24] This Property-Labeled Materials-Fragments model[24] uses atomic

Table 1. A comparison of the Vickers hardness, H_v, for various materials computed via the Teter (Eq. (1)), Chen (Eq. (2)), and Tian (Eq. (3)) models using the Voigt-Reuss-Hill (VRH) averages of the bulk, B, and shear, G, moduli obtained from the Automatic Elasticity Library (AEL) and via a Machine Learning (ML) model trained on the AFLOW database, with experiment. H_v, B, and G are given in units of GPa. The dataset used was taken from ref. [15], which also provides the experimental values. A full table with results for all of the 64 structures used in the correlation analyses and the plots in Fig. 1 is provided in Section S1 in the SI

Name	$H_v^{Exp,a}$	G_{VRH}^{AEL}	B_{VRH}^{AEL}	$H_{v,Teter}^{AEL}$	$H_{v,Chen}^{AEL}$	$H_{v,Tian}^{AEL}$	G_{VRH}^{ML}	B_{VRH}^{ML}	$H_{v,Teter}^{ML}$	$H_{v,Chen}^{ML}$	$H_{v,Tian}^{ML}$
Diamond	96.0	517.9	434.0	78.2	92.2	93.9	514.0	430.2	77.6	91.9	93.5
BC$_2$N	76.0	411.6	372.5	62.2	73.1	73.1	363.6	354.0	54.9	62.0	61.6
c-BN	66.0	380.3	372.4	57.4	63.2	63.2	386.8	332.7	58.4	74.9	74.2
β-SiC	34.0	186.6	212.4	28.2	33.6	32.2	163.8	201.5	24.7	28.0	26.9
SiO$_2$	33.0	201.2	270.1	30.4	28.5	28.1	187.2	237.8	28.3	29.3	28.5
ReB$_2$	30.1	270.0	331.8	40.8	38.6	38.3	257.9	333.4	38.9	35.1	35.0
WC	30.0	272.3	377.4	41.1	33.3	33.6	247.0	369.0	37.3	28.4	28.8
VC	29.0	215.6	295.7	32.6	29.0	28.8	164.1	274.8	24.8	18.6	18.9
ZrC	25.8	157.0	220.9	23.7	22.8	22.4	136.0	194.4	20.5	20.3	19.9
TiC	24.7	177.8	238.8	26.8	26.3	25.8	145.9	220.8	22.0	19.7	19.6
TiN	23.0	171.2	269.8	25.9	20.8	20.9	141.5	240.5	21.4	16.5	16.8
RuO$_2$	20.0	138.0	267.6	20.8	13.5	14.2	125.6	231.5	19.0	13.5	14.1
NbC	18.0	201.8	298.7	30.5	25.2	25.2	157.8	266.0	23.8	18.0	18.3
AlN	18.0	122.2	195.1	18.5	16.2	16.2	126.7	191.2	19.1	18.0	17.8
NbN	17.0	129.1	305.9	19.5	9.5	10.8	107.1	286.9	16.2	6.7	8.2
HfN	17.0	158.7	269.5	24.0	17.9	18.2	132.9	248.5	20.1	13.8	14.4
GaN	15.1	106.0	175.3	16.0	14.0	14.1	105.5	173.7	15.9	14.0	14.1
ZrO$_2$	13.0	102.0	233.5	15.4	8.4	9.5	101.8	194.8	15.4	11.0	11.6
Si	12.0	62.5	89.1	9.4	11.8	11.5	51.2	81.3	7.7	8.6	8.8
GaP	9.5	51.3	78.8	7.7	9.1	9.2	37.5	74.0	5.7	4.5	5.5
AlP	9.4	46.4	82.9	7.0	6.6	7.2	45.7	82.6	6.9	6.4	7.0
InN	9.0	54.9	124.4	8.3	5.0	6.2	50.0	115.5	7.5	4.4	5.7
Ge	8.8	46.2	61.5	7.0	10.5	10.0	45.5	60.3	6.9	10.4	10.0
GaAs	7.5	40.8	62.7	6.2	7.6	7.8	34.0	58.3	5.1	5.4	6.1
Y$_2$O$_3$	7.5	62.3	137.9	9.4	5.9	6.9	60.7	123.8	9.2	6.6	7.5
InP	5.4	31.6	60.4	4.8	4.1	5.1	25.7	58.5	3.9	2.1	3.6
AlAs	5.0	39.3	67.4	5.9	6.1	6.7	36.3	64.6	5.5	5.3	6.1
GaSb	4.5	29.6	47.0	4.5	5.4	6.0	27.2	44.4	4.1	4.8	5.5
AlSb	4.0	28.5	49.4	4.3	4.5	5.3	26.5	47.9	4.0	3.8	4.8
InAs	3.8	26.2	63.6	4.0	1.8	3.4	26.1	59.4	3.9	2.1	3.6
InSb	2.2	20.1	38.1	3.0	2.5	3.7	17.4	36.5	2.6	1.4	3.0
ZnS	1.8	33.9	71.2	5.1	3.6	4.8	30.9	68.8	4.7	2.8	4.2
ZnSe	1.4	27.5	58.2	4.2	2.8	4.1	24.9	56.7	3.8	2.0	3.5
ZnTe	1.0	22.1	43.8	3.3	2.5	3.8	18.4	37.3	2.8	1.8	3.2

^{a}Experimental data are taken from ref. [15]

distance and a Voronoi tessellation to characterize the connectivity of a crystal structure: Voronoi cells that are centered on each atom are constructed and two atoms are considered connected if they are within a bonding distance threshold and their Voronoi cells share a face. The connected atoms form a graph, and sections of the graph define the "materials fragments". The nodes of the graph are decorated with elemental properties of the atomic species at that site, such as electronegativity, ionization energy, and atomic number. Combinations of these elemental properties, along with other properties of the crystal such as lattice parameters and space group, form the feature vector representing the material. The ML model is based on gradient boosted decision trees,[39] and is trained on elastic properties calculated using AEL. This model was able to predict the G_{VRH} and B_{VRH} of over 85% of the systems tested to within 20 GPa of the values calculated via AEL, with the ML moduli typically being smaller. The deviation was most pronounced for systems with a large G_{VRH}, and significant differences for B_{VRH} were found for graphite and two theoretical high-pressure B-N phases.

While it would be preferable to train the ML model directly to predict hardness values, this would require training directly on experimental data, since hardness cannot be calculated directly from first-principles.[15,16,37] However, because experimental hardness data is not readily available in the quantity and quality required to train ML models (reported experimental values can vary by more than 10%,[37] as discussed in Section S4 in the SI), models based on calculated elastic moduli must be used. For example, a recent combined theoretical/experimental ML directed search for superhard materials trained an ML model to predict G and B, which were assumed to be good proxies for hardness, and no attempts to estimate hardness were made.[8]
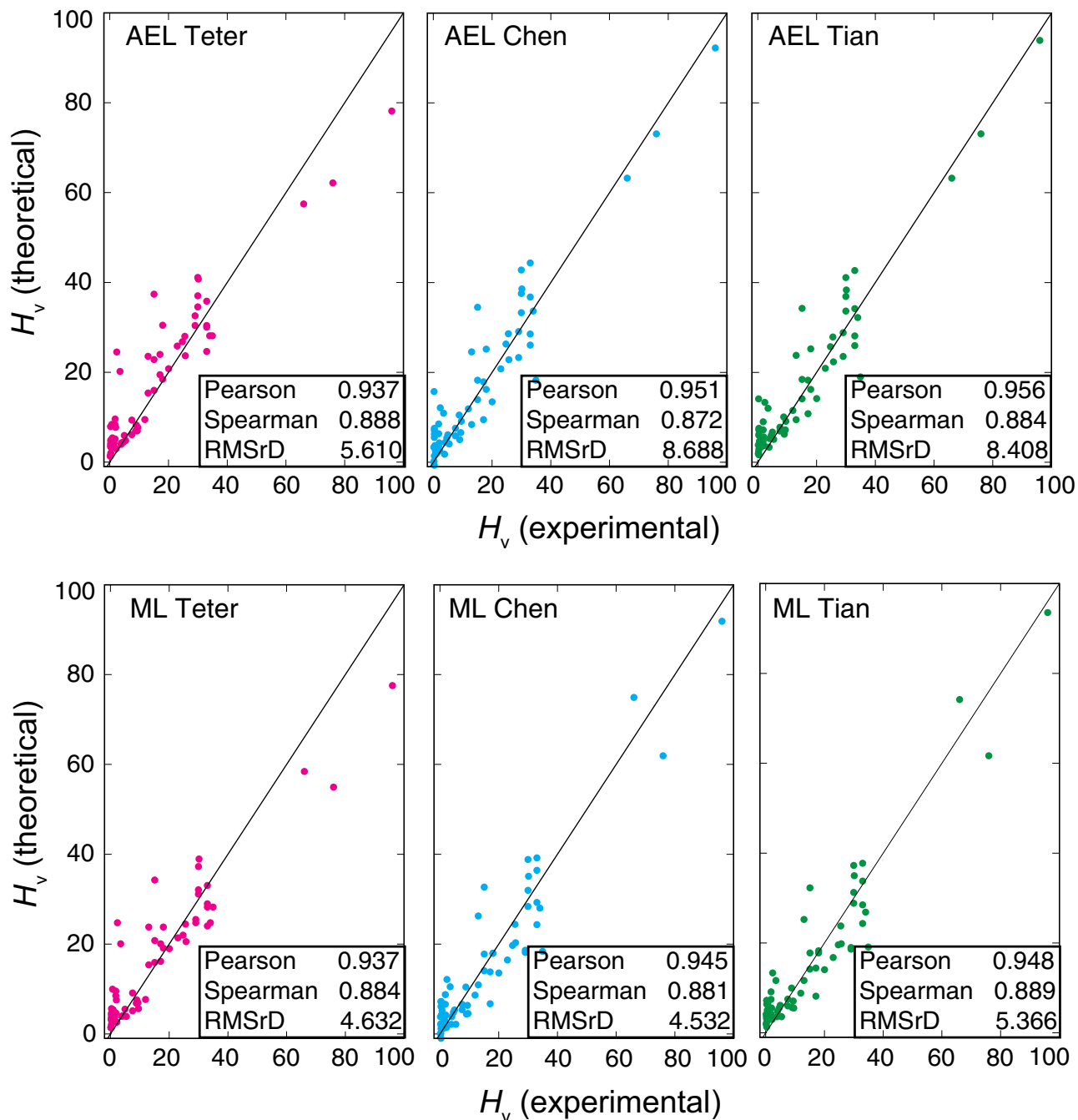
Fig. 1 A comparison of the Vickers hardness, H_v in GPa, for 64 materials (see the SI) computed via the Teter (Eq. (1)), Chen (Eq. (2)), and Tian (Eq. (3)) models using the Voigt-Reuss-Hill averages of the bulk and shear moduli obtained from the Automatic Elasticity Library (AEL) and via a Machine Learning (ML) model trained on the AFLOW database, with experiment. The line represents a perfect correlation. The Pearson and Spearman coefficients between the experimental and theoretical data are provided, as is the root-mean-squared relative deviation, RMSrD

Unsurprisingly, Table 1 shows that the AFLOW-ML model also typically underestimated the bulk and shear moduli of the structures considered herein, and it tended to do a better job in predicting B_{VRH} as opposed to G_{VRH}. This suggests that the ML-derived Vickers hardnesses will be smaller than those obtained using AEL for all three models. Comparison of the hardness values obtained with ML vs. AEL in Table 1 and Fig. 1 shows that in general these expectations hold, with the exception of c-BN and AlN for all three models, and SiO_2, ZrO_2, and Y_2O_3 for the Chen and Tian models. The correlation coefficients for the ML and AEL results shown in Fig. 1 are similar, but the RMSrD values

are somewhat smaller for the results obtained using the ML moduli.

Our findings suggest that the ML moduli can be used to obtain reasonable estimates of the hardness of a given material quickly, making it possible to employ them to determine the fitness of an individual in an EA based structure search. But which one of the macroscopic models should be used? During our EA searches on the elemental carbon system we found that in some instances the ML-derived B_{VRH} were extremely small, and when they were used in the denominator in Eqs. (2) and (3), unrealistically hard values were obtained, with several systems being predicted to be much

harder than diamond itself. Only the Teter model was able to predict that these crystals had a very low Vickers hardness. For the 64 systems in our test set, the Pearson and Spearman correlation coefficients between the Vickers hardness values calculated using the Teter model and AEL vs. ML were 0.993 and 0.980, respectively, and the RMSrD was 0.156. Moreover, the Chen and Tian models are theoretically less satisfying than the Teter model, since they are empirically fitted and result in an incorrect dimension. Thus, the Teter model, Eq. (1), was chosen for the hardness evaluations since it can differentiate between the hard crystals that are kept in the gene pool during the evolutionary search, as described in the following section, and those that are soft.

The XTALOPT evolutionary algorithm

EAs employ concepts from biological evolution to find an optimal solution for problems that have many degrees of freedom. When applied towards a priori CSP, EAs search for the lattice parameters and atomic coordinates that minimize or maximize a computed quantity. Because EAs are typically concerned with finding the global minimum (along with important local minima) on the PES, they attempt to minimize the computed energy/enthalpy.[40] An EA starts by generating a set of random structures (structures chosen by the user, also known as seeds, may also be employed) that are relaxed to the nearest local minima by an external program. The fitness of each optimized individual is calculated and used to determine the probability that it will be chosen for procreation. Child structures are created either via a two-parent breeding operation, or by a mutation of a single parent. Further details about the XTALOPT EA, and its subprograms can be found in refs. [33,34,41–43] (https://xtalopt.github.io/).

The original implementation of the open source EA for CSP, XTALOPT,[33] uses roulette wheel selection where the probability p_i that a structure with energy/enthalpy E_i is chosen for procreation is equal to its fitness. The probability is calculated as:

$$p_i = N\left(1 - \frac{E_i - E_{\min}}{E_{\max} - E_{\min}}\right), \qquad (4)$$

where $E_{\min}$ and $E_{\max}$ are the lowest and highest energies/enthalpies in the breeding pool, respectively, and N is a normalization constant chosen to ensure that $\sum p_i = 1$. The lower the energy/enthalpy of an individual, the more likely it is to be selected as a parent for the subsequent generation.

In the algorithms designed to predict superhard materials that are implemented within CALYPSO[12] and USPEX,[13] the hardness, as calculated via a microscopic model, is employed to determine an individual's fitness. Using roulette wheel selection, and maximizing the Vickers hardness instead of minimizing the energy/enthalpy, Eq. (4) becomes

$$p_i = N\left(1 - \frac{H_{v,\max} - H_{v,i}}{H_{v,\max} - H_{v,\min}}\right). \qquad (5)$$

The higher the hardness of an individual, $H_{v,i}$, the more likely it is to be selected for breeding.

Because we are interested in predicting the structures of superhard materials that could potentially be synthesized, it is desirable that they correspond to low lying local minima. Therefore, the new fitness function implemented within XTALOPT combines Eqs. (4) and (5) to favor the selection of structures that are both low in enthalpy and high in hardness via:

$$p_i = N\left[1 - w\left(\frac{H_{v,\max} - H_{v,i}}{H_{v,\max} - H_{v,\min}}\right) - (1 - w)\left(\frac{E_i - E_{\min}}{E_{\max} - E_{\min}}\right)\right], \qquad (6)$$

where the weight, w, is a fractional number between 0 and 1. Herein, we employed a $w = 0.5$, and the Vickers hardnesses used to determine p_i were computed via the Teter model, Eq. (1), using the ML values of the shear modulus.

Most EAs do not keep every optimized structure within the breeding pool, but rather only a user-specified number of the fittest individuals. In the original version of XTALOPT, the lowest energy/enthalpy structures were placed in the breeding pool. In the current implementation, the fitness of each individual is computed prior to normalization using Eq. (6). The crystals with the highest p_i are placed in the breeding pool, and their normalized probabilities are computed and used.

Applications to carbon

Above we have shown that the Vickers hardness of a compound can be computed with reasonable accuracy via Teter's equation using the shear modulus obtained with a ML algorithm, and proposed a fitness function that can be used in conjunction within an evolutionary structure search to predict superhard materials. To determine how well this approach works, we performed an EA search on elemental carbon. This system was chosen because of the diversity of the crystalline carbon family that arises from the possibility of sp, sp^2, and sp^3 hybridized bonds, leading to an infinite number of structures within its chemical space. Moreover, the hardness of its family members ranges from that of soft layered graphite to diamond, the hardest known material on this planet, as well as many well-known superhard species including lonsdaleite, M-carbon, Z-carbon, F-carbon, among many others. The wide range of possible Vickers hardnesses and crystalline topologies within the carbon system creates a tremendous space for XTALOPT to explore, and we are able to evaluate its performance by observing its efficiency on finding pearls, the superhard species, from the ocean of carbon crystals. We expect our EA search will find many reported structures, illustrating that it works, and show that it is a predictive tool that can discover new topologies.

Figure 2 plots $H_{v,\text{Teter}}^{\text{ML}}$ for all of the 5624 optimized individuals vs. their energy. The plot is partitioned into four quadrants following the standard suggested by Zhang et al.[12] We define superhard structures to have a Vickers hardness larger than 40 GPa, and thermodynamically stable structures to be those whose 0 K energies are less than −8.80 eV/atom, which is within 0.30 eV/atom of diamond. For comparison, the energy of M-carbon, which has been synthesized under pressure,[2–4] was computed to be −8.93 eV/atom. The superhard species that could potentially be synthesized lie in the top left quadrant of the hardness-energy map, and soft stable ones are found in the bottom left. Superhard phases with comparably high energies are located in the top right
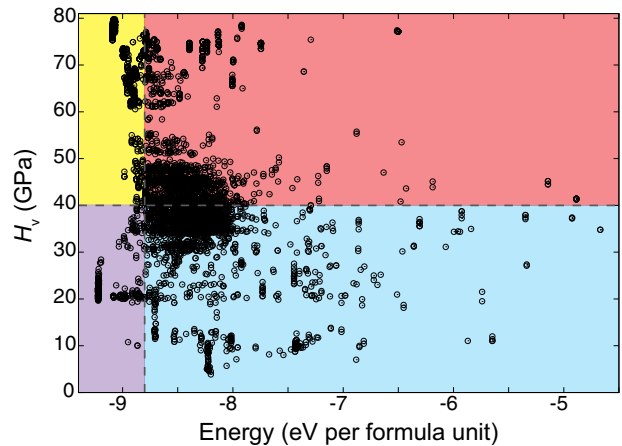

Fig. 2 The hardness vs. energy map of all carbon topologies predicted by the XTALOPT searches. The horizontal line corresponds to a $H_{v,\text{Teter}}^{\text{ML}} = 40$ GPa, and the vertical line to an energy of −8.80 eV/atom. The structures in the yellow quadrant are both stable and superhard

hand side of the plot, and unstable soft systems are at the bottom right. The energy of 1324 structures was less than −8.80 eV/atom, and from these 827 had a Vickers hardness greater than 40 GPa. Duplicates were removed from the yellow quadrant in Fig. 2, resulting in 89 distinct topologies, which were found by XTALOPT, that are both hard and thermodynamically favored. Phonon calculations showed that ten of these phases, all containing sp^2 carbons, were dynamically unstable. Phonon dispersion curves and phonon density of states plots for all structures, as well as thermodynamic properties for stable structures can be found in Section S3 in the SI. Therefore, in our searches, which employed cells containing 8, 12, 16, and 20 carbon atoms, 79 hard and stable individuals were found.

To determine if the predicted structures are already known, and avoid making false claims of novelty, the Samara Carbon Allotrope Database (SACADA) was employed (http://sacada.sctms.ru).[44] SACADA has collected the crystal structures and physical properties from both experiment and theory of more than 500 3D carbon allotropes. Among the hard species found via XTALOPT, 36 are already known, including diamond (#1, **dia**), lonsdaleite (#37, **lon**), a chiral-framework (#29, **unj**), M-carbon (#224, **cbn**), Z-carbon (#141, **sie**), F-carbon (#225, 4^4T35), bct4-carbon (#60, **crb**), and BC_8-carbon (#20, **gsi**). Some of these are also present in the AFLOW crystal prototype library.[45] The SACADA ID numbers and densities of the previously known phases are reported in Table 2, which also provides a comparison between our $H_{v,Teter}^{ML}$ results with the hardness values computed by others. The reason why the $H_{v,Teter}^{ML}$ for diamond given in Table 1 differs slightly from that in Table 2 is because in the first case the structure was taken as-is from the AFLOW database, and in the second case it was optimized again by us.

Figure 1 clearly illustrates that the Teter model tends to underestimate the Vickers hardnesses of superhard materials. Therefore, it is not surprising that almost all of our $H_{v,Teter}^{ML}$ values are lower than the Vickers hardness estimates of others. However, once a structure has been predicted by an EA, it is always possible to compute its hardness using a more accurate and computationally expensive hardness model. What is important is that for the unit cell sizes employed herein, our EA is able to find most of the hard stable carbon structures predicted earlier in refs. [12,13]. Plus, we find many new stable superhard phases not reported in SACADA, or known as of yet in the literature.

Table 2 shows that the computed $H_{v,Teter}^{ML}$ for the previously known individuals is typically 10–20 GPa lower than the values obtained using the microhardness models implemented within CALYPSO[12] and USPEX.[13] One exception is for the chiral-framework structure (#29, **unj**) originally predicted in ref. [46], which we find to be 28.5 GPa less hard than diamond. The aforementioned microhardness models computed diamond to be only 8–8.4 GPa harder than this $P6_522$ symmetry structure. It should be noted, however, that the chiral framework is significantly less dense than diamond ($\rho = 3.21$ g/cm³ vs. 3.50 g/cm³), and it would be quite remarkable if it were nearly as hard. The value of $H_{v,Teter}^{ML}$ obtained herein for this phase, 47.1 GPa, is close to Vickers hardnesses we calculate for phases with a similar density: bikitaite carbon, 6B, G158 and 3,4,4T154.

In addition to the previously predicted phases, our algorithm identified 53 new low-energy topologies that were superhard. Calculations of the phonons and elastic constants confirmed that 43 of these were dynamically and mechanically stable. A summary table, which lists the hardness, relative stability and topological properties of these phases is provided in Section S5 in the SI, and their fascinating structures, properties and electronic structures will be discussed in detail in a follow-up paper. Table 3 lists the ML and AEL values of their bulk and shear moduli, as well as the corresponding Vickers hardnesses computed with the Teter model. Their energy relative to diamond, density, and pressure above which their enthalpies become lower than that of graphite

Table 2. Stable superhard structures discovered using the XTALOPT evolutionary algorithm that have been predicted before. Either the well-known names, or those used in the first paper reporting the structure, are provided, along with the SACADA ID and topology. The Vickers hardness (GPa) as calculated via the Teter model using the ML shear modulus, $H_{v,Teter}^{ML}$, or computed in other studies, H_v^{others}, as well as the density, ρ, in g/cm³ of the optimized structures is also given

Names	SACADA ID (Topology)	$H_{v,Teter}^{ML}$	H_v^{others}	ρ
Diamond	1 (**dia**)	75.6	89.5[b], 87.3[b], 95.4[b], 97.5[b], 89.7[c]	3.50
bc8	20 (**gsi**)	74.4	88.6[b]	3.56
Chiral-framework	29 (**unj**)	47.1	81.5[b], 90.8[b], 81.3[c]	3.21
Lonsdaleite	37 (**lon**)	79.3	89.1[b], 89.1[c], 97.3[d]	3.49
bct-C_4	60 (**crb**)	65.6	84.4[b], 82.0[b], 92.9[b], 84.0[c]	3.31
Y-carbon	81 (**cag**)	62.8		3.31
T12	107 (**cdp**)	68.2	85.1[b]	3.35
4H-diamond	111 (**cfc**)	78.3	98.1[d]	3.50
Bikitaite carbon	124 (**bik**)	48.4		3.23
Z-carbon	141 (**sie**)	71.8	85.1[b], 84.4[b], 95.1[b]	3.40
6B	179 (NSI)	48.8		3.25
$C2/m$-16	186 (4^3T85)	68.6	84.9[b]	3.35
12R-diamond	209 (SiC12)	77.0		3.50
oc16	210 ($3,4^3$T72)	52.3	42.9[d]	3.27
W-carbon	218 (**cnw**)	70.4	85.3[b], 83.1[b], 93.8[b]	3.35
M-carbon	224 (**cbn**)	68.0	85.0[b], 82.7[b], 93.5[b], 84.3[c]	3.34
F-carbon	225 (4^4T35)	70.4	84.7[b]	3.33
$P2_1/m$	245 (4^5T27)	70.8	85.4[b]	3.36
12D	256 (4^6T15)	72.1		3.40
oP20	260 (4^6T16)	75.4	91.4[d]	3.42
Z-carbon-3	262 (4^6T7)	72.1	92.6[d]	3.40
mS32	291 (4^8T16)	72.6	90.8[d]	3.41
Z-carbon-2	293 (4^8T5)	75.6	92.4[d]	3.42
mP16	294 (4^8T15)	75.8	91.0[d]	3.41
16D	336 (4^{16}T3)	75.9		3.41
16B	337 (4^{16}T2)	72.7		3.42
G6	358 (4^6T28)	66.8		3.33
G21	365 (4^5T49)	69.3		3.33
G120	420 (4^4T85)	68.7		3.36
G153	447 (**cfe**)	77.8		3.50
G158	451 (mbc-4,4-lmma)	47.1		3.26
G178	470 (4^8T44)	78.3		3.48
G225	511 (4^8T49)	72.6		3.41
3,4,4T154[a]	not published	48.0		3.19
8-layered SiC polytype[a]	not published	77.0		3.50
Deem hyp. Zeolite	8047698	54.5		3.30

[a]Structures are in the SACADA database but have not yet been published online
[b]Hardness values are from ref. [12] and references therein
[c]Hardness values are from ref. [13]
[d]Hardness values are from the SACADA database

Table 3. New stable superhard structures predicted by the XTALOPT evolutionary algorithm. E_{ref} is the energy relative to diamond (eV/atom). The Voigt-Reuss Hill (VRH) average of the bulk, B, and shear, G, moduli obtained from the Automatic Elastic Library (AEL) and via a Machine Learning (ML) model trained on the AFLOW database, and the Vickers hardness, H_v, as computed via the Teter model (Eq. (1)) are provided. The density, ρ, in g/cm^3, percent of diamond and lonsdaleite found in the structure, and pressure, P, above which the structures are computed to become more stable than graphite are listed. B, G, H_v, and P are given in units of GPa. The Pearson and Spearman coefficients, along with the RMSrD between the AEL and ML results are 0.858, 0.895, 0.115 for the Teter model, 0.811, 0.849, 0.201 for the Chen model and 0.818, 0.854, and 0.200 for the Tian model

Name	E_{ref}	B_{VRH}^{ML}	B_{VRH}^{AEL}	G_{VRH}^{ML}	G_{VRH}^{AEL}	$H_{v,Teter}^{ML}$	$H_{v,Teter}^{AEL}$	ρ	% **dia/lon**	P
$R\bar{3}m$-16	0.001	404.5	434.7	504.9	521.6	76.2	78.8	3.50	100	7.6
$R3m$-16	0.007	404.6	435.1	504.8	523.0	76.2	79.0	3.50	100	8.1
$P2/m$-12	0.133	421.6	414.6	494.2	467.1	74.6	70.5	3.38	31.6	20.3
$Imm2$-12	0.137	387.0	415.2	474.3	468.1	71.6	70.7	3.38	31.9	20.7
$C2/m$-20c	0.142	392.1	419.7	468.4	487.7	70.7	73.6	3.40	38.3	20.6
$Imma$-16	0.143	387.2	413.4	468.8	452.8	70.8	68.4	3.38	72.9	21.2
$P\bar{1}$-16a	0.149	399.0	413.3	447.9	461.3	67.6	69.7	3.36	0	22.4
$C2/m$-20d	0.160	380.5	406.4	424.3	453.0	64.1	68.4	3.32	38.0	24.4
$Pmm2$-20	0.165	401.1	406.9	424.0	454.3	64.0	68.6	3.32	38.3	25.0
$C2/m$-20a	0.176	386.5	409.1	447.7	440.2	67.6	66.5	3.35	19.0	25.3
$C2/m$-16a	0.179	378.5	406.1	439.1	419.8	66.3	63.4	3.33	11.9	26.1
$C2/m$-16b[a]	0.187	382.1	409.1	449.1	403.9	67.8	61.0	3.35	24.3	26.3
$C2/m$-20b	0.193	384.7	410.5	450.0	445.2	67.9	67.2	3.36	20.1	26.7
$C2/m$-12a	0.193	384.9	411.0	455.3	448.3	68.7	67.7	3.36	0	26.5
$C2/m$-16c	0.207	382.8	411.4	474.3	448.4	71.6	67.7	3.38	48.0	27.6
$C2/m$-12b	0.211	370.4	396.1	317.0	424.3	47.9	64.1	3.24	0	33.1
$P\bar{1}$-20	0.215	392.5	388.7	303.5	403.2	45.8	60.9	3.18	0	37.1
$Cmmm$-12a	0.216	367.7	389.6	300.0	397.3	45.3	60.0	3.18	0	36.9
$C2/m$-12c	0.226	370.4	391.7	314.3	392.6	47.5	59.3	3.23	0	35.9
$Cmmm$-20[a]	0.227	394.4	400.4	346.8	396.1	52.4	59.8	3.28	19.3	33.2
$C2/m$-16d	0.228	377.4	401.9	363.8	414.7	54.9	62.6	3.30	11.4	32.7
$P2_1/m$-16a	0.228	414.2	404.5	440.0	444.6	66.4	67.1	3.32	12.7	32.1
$Cmmm$-12b[a]	0.233	381.8	386.1	295.1	319.8	44.6	48.3	3.17	0	39.7
$P\bar{1}$-16c[a]	0.242	407.2	412.6	472.1	445.0	71.3	67.2	3.38	48.4	31.2
$Pmma$-16a	0.242	394.3	399.0	360.6	408.8	54.5	61.7	3.29	23.8	34.6
$Pmma$-12	0.242	396.4	391.3	330.1	401.6	49.8	60.6	3.22	0	38.8
$P\bar{1}$-16d	0.248	411.5	421.8	504.5	480.3	76.2	72.5	3.45	61.6	29.8
Pm-16	0.250	410.2	399.2	375.8	411.9	56.8	62.2	3.29	0	35.5
$P2_1$-16	0.255	404.0	408.9	451.5	448.1	68.2	67.7	3.35	0	33.8
$P\bar{1}$-16e[a]	0.258	408.5	412.5	473.3	449.1	71.5	67.8	3.38	49	33.0
$P\bar{1}$-16b[a]	0.259	410.3	412.6	479.2	448.9	72.4	67.8	3.38	48.7	33.3
$P2_12_12_1$-16	0.263	389.2	389.6	311.4	380.3	47.0	57.4	3.22	0	41.7
$P1$-12	0.267	386.9	386.7	305.5	366.1	46.1	55.3	3.17	0	46.5
$Pmma$-16b	0.274	397.0	393.0	343.6	392.6	51.9	59.3	3.26	46.7	40.4
$P2/c$-16	0.276	393.8	395.8	322.5	415.2	48.7	62.7	3.25	0	41.9
$P1$-16d	0.276	412.3	414.4	479.6	461.7	72.4	69.7	3.40	0	34.4
$P1$-16a[a]	0.278	403.8	406.0	447.0	409.9	67.5	61.9	3.35	43.4	36.5
$P\bar{1}$-12	0.281	417.0	424.5	500.4	492.7	75.6	74.4	3.47	33.7	32.3
$P1$-16b[a]	0.282	385.7	388.6	300.0	372.4	45.3	56.2	3.21	0	45.9
$P1$-16c	0.285	414.6	417.2	485.9	467.4	73.4	70.6	3.43	23.5	34.3
$P\bar{1}$-16f	0.286	410.1	411.8	476.6	451.9	72.0	68.2	3.40	36.7	35.5
$P2/m$-16	0.288	401.3	389.5	329.0	386.6	49.7	58.4	3.22	0	44.8
$P1$-16e	0.290	413.1	417.4	486.5	470.2	73.5	71.0	3.42	23.7	34.9

[a]Structures that contain sp^2 carbon atoms

are also given. The caption of Table 3 provides the correlation coefficients and RMSrD between the ML and AEL Vickers hardnesses computed for these new phases using the Teter, Chen and Tian equations. The best agreement between the ML and AEL values is obtained for the Teter model.

M-carbon (#224, **cbn**), whose $H_{v,Teter}^{ML}$ is computed to be 68.0 GPa, can be made by compressing graphite to

~17–19 GPa.[2,3] Thus, it might be possible to synthesize some of the newly predicted superhard phases under pressure by an appropriate choice of the carbon allotrope starting material and the synthesis route. Metastable carbon phases have been previously synthesized by compressing glassy carbon[47,48] or onion carbon,[49,50] for instance. To study this further, we computed the pressures above which the newly predicted superhard phases become more stable than graphite. The transition pressures for diamond-like structures (more on this below) were typically between 7 and 8 GPa, whereas other structures have transition pressures within 23 and 47 GPa. Generally speaking, the higher density phases were predicted to become enthalpically favored over graphite at lower pressures, and they had larger Vickers hardnesses than the low density phases, as expected.

Since both the ML and AEL shear moduli of lonsdaleite are larger than that of diamond, but the bulk moduli are about the same, lonsdaleite is predicted to have a slightly higher Vickers hardness by all three macroscopic hardness models. Lonsdaleite has, in fact, been previously predicted to have superior mechanical properties as compared to diamond,[51,52] but its experimental existence has been questioned,[53] although there are some recent claims of its synthesis both through static[47] and shock compression.[54]

$H_{v,Teter}^{ML}$ for a few of the predicted structures was computed to be at least as high as that of diamond. This includes the following from Table 2: 4H and 12R diamond, G153, and 8-layered SiC. These are all composed entirely of mixtures of lonsdaleite and diamond: they are members of the infinite family of **dia-lon** polytypes. The tiling approach, wherein a 3D periodic net is represented as a tiling of generalized polyhedra, has previously been employed to analyze the topology of sp^3 bonded and amorphous carbon allotropes.[55,56] Using this method to extract the content of diamond and lonsdaleite topological building blocks in each superhard allotrope found in our EA searches revealed two new **dia-lon** polytypes that were also, unsurprisingly, found to be harder than diamond ($R\bar{3}m$-16 and $R3m$-16, see Table 3).

The same analysis technique revealed that G178 (#470 4^8T44), and 16D (#336 $4^{16}T3$), illustrated in Fig. 3a, b, as well as $mP16$ (#294 4^8T16), shown in Fig. 2 in ref. [55], contained more than 47% of diamond/lonsdaleite tiles (see Section S5 in the SI). The newly discovered $P\bar{1}$-16d and $P\bar{1}$-12 superhard species that are shown in Fig. 3c, d contained 61.6 and 33.7% diamond/lonsdaleite, respectively, along with tiles found in $mC12$ (#147, 4^2T112) and $oC16$-I (#175, 4^3T84), respectively. The $mC12$ and $oC16$-I phases have previously been predicted to be superhard,[57] and herein their $H_{v,Teter}^{ML}$ values were computed to be somewhat lower than that of diamond, 72.6 and 73.5 GPa, respectively. Thus, the newly predicted species that are estimated to be at least as hard as diamond contain a mixture of tiles found in diamond, lonsdaleite, or other superhard phases.

The first superhard phase that did not contain diamond or lonsdaleite, $P1$-16d, is shown in Fig. 3e. It is very asymmetric and appears close to amorphous, or disordered, with 16 different atom types. Indeed, it has been shown that disorder in a material can enhance its hardness.[7] The topological analysis further revealed that systems with sp^2 hybridized carbon atoms that had an $H_{v,Teter}^{ML}$ > 70 GPa contained a high percentage of diamond and/or lonsdaleite, at least 40%. They are illustrated in Fig. 3f–h.

Finally, we mention some classic cases where the microhardness models have failed, and illustrate that even for these difficult systems our simple approach yields results that are sufficiently accurate to guide a CSP search. *T*-carbon (#33, **dia-a**) is a hypothetical carbon allotrope[58] whose prediction prompted scrutiny of the various computational descriptors of hardness. The Gao[9] and SV[10] microhardness models yielded Vickers hardnesses of 61.1 GPa and 40.5 GPa, respectfully, but it was unclear why such a porous structure with a highly anisotropic distribution of the sp^3-like C–C bonds should be superhard. The

macroscopic Chen model yielded results that were more in-line with physical reasoning, $H_{v,Chen}$ = 5.6 GPa.[59] The failure of the Gao and SV models to accurately predict the hardness was attributed to the fact that they both assume uniformly distributed bonds. Herein, we find $H_{v,Teter}^{ML}/H_{v,Teter}^{AEL}$ = 8.7/7.0 GPa for *T*-carbon, which is a reasonable estimate. Layered van der Waals bonded systems such as graphite have also proven to be challenging for the microhardness models. For example, the Li model[11] predicts a hardness of 57.4 GPa for graphite,[13] whereas experimental measurements yield <1 GPa.[60] The value we calculate for a graphite structure that was optimized with Grimme's D3 dispersion correction, 27.8 GPa, is too high for a quantitative estimate. However, the fitness of structures with such a small Vickers hardness would be quite low, and it is unlikely that any of them would be present in the pool in the final stages of an EA search. Herein, 62.5% of the structures whose energy was below −8.80 eV/atom had an $H_{v,Teter}^{ML}$ greater than 40 GPa, confirming the preferential exploration of the PES of hard structures, which can further be tuned by adjusting the weight in the relation used to determine fitness, Eq. (6).

In the past, a priori crystal structure prediction (CSP) algorithms designed to pinpoint novel superhard phases have estimated a structure's hardness using microscopic models. Macroscopic hardness models overcome these limitations, but they depend upon elastic properties that are expensive to calculate from first principles.

We have shown that the Vickers hardnesses, H_v, of a wide range of materials computed via the Teter, Chen and Tian macroscopic models along with the bulk and/or shear moduli calculated using AFLOW-AEL agree well with experiment. Moreover, the H_v calculated using the moduli obtained via a ML model that was trained on the AFLOW repository are in excellent agreement with the AFLOW-AEL results. Because the ML based values can be computed quickly given only a crystal's structure, it is possible to employ them to determine fitness in a CSP search. This methodology has been implemented in the XTALOPT evolutionary algorithm, which can now calculate an individual's probability for procreation using both the Vickers hardness and the energy/enthalpy, to accelerate the search for hard and stable species.

The method was applied towards the carbon system. In addition to finding 36 known phases, 43 dynamically stable superhard individuals that have not been reported previously were identified. Their structures were analyzed using a topological analysis, which showed that all but one of the newly predicted phases with $H_{v,Teter}^{ML}$ *gt*; 70 GPa were comprised of a substantial fraction of diamond and/or lonsdaleite. The other superhard phase was close to amorphous or disordered. We expect that in the future CSP algorithms will become increasingly coupled with ML derived values to predict materials with desired properties.

METHODS

All of the calculations were carried out using density functional theory as implemented within the Vienna ab initio Simulation Package (VASP)[61] coupled with the Perdew, Burke and Ernzerhof (PBE) functional,[62] and the projector augmented wave (PAW) method.[63]

Detailed computational settings for the AFLOW-AEL[29] calculations used to determine the elastic properties are described in ref. [64]. A set of 4 normal and shear deformations (2 compressive and 2 expansive) are applied in each independent direction (1 to 3 depending on the structure's symmetry) to the fully relaxed structure of a given compound. The ionic positions are optimized while keeping the cell size and shape fixed, and the stress tensors for each deformed structure are calculated. The resulting stress-strain data is fitted to obtain the elastic constants in the form of the symmetric 6 × 6 elastic stiffness tensor (Voigt notation). The bulk and shear moduli are calculated from elastic constants as described in ref. [22].

The EA searches were carried out using XTALOPT version r12[34] on unit cells containing 8, 12, 16, and 20 formula units per primitive cell, and a total of 5624 structures were generated. These cell sizes were chosen
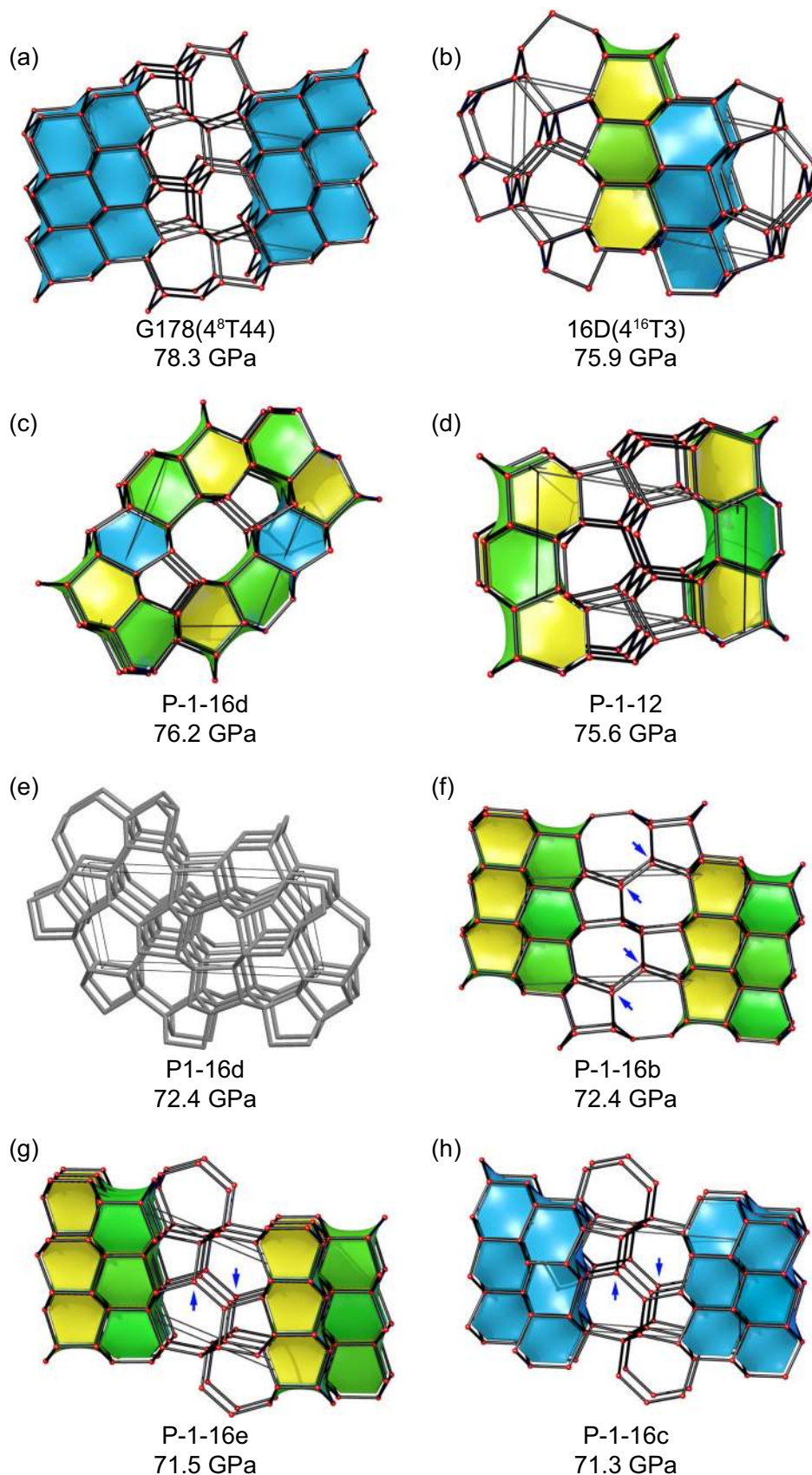
Fig. 3 a, b Previously known (see Table 2), and (**c**–**h**) newly predicted (see Table 3) superhard phases. The computed $H_{v,Teter}^{ML}$ are provided. The cages colored in blue are structurally related to diamond (**dia**), and the cages colored in yellow and green are structurally related to lonsdaleite (**lon**). The arrows in (**f**–**h**) point towards the sp^2 carbons

because they are able to predict phases with 1–6, 8, 10, 12, 16, and 20 atoms in the primitive cell, and because refs. [12,13] considered primitive cells with up to 24 carbon atoms. Tests showed that the likelihood of predicting superhard individuals increased when the initial volumes of the generated crystals were constrained to be about the same as that of diamond. The first generation was constructed using the RandSpg method,[43] and duplicates were removed from the breeding pool via the XtalComp algorithm.[42] The spacegroups of the predicted superhard phases were determined using AFLOW-SYM,[65] and their structures were analyzed via the tiling approach as implemented in the ToposPro package.[66]

The carbon $2s^2 2p^2$ electrons were treated as valence, and the core states were described using the PAW method with an energy cutoff of 500 eV in the structure searches, and 600 eV otherwise. Γ-centered Monkhorst-Pack k-meshes were employed where the number of divisions along each reciprocal lattice vector was chosen such that the product of this number with the real lattice constant was 25 Å for the final step in the EA searches, 40 Å for the phonon calculations, and 50 Å for the final optimizations and equation of state calculations. Crystals that satisfied the hardness and stability criteria described in the section "Applications to carbon" were optimized so that all forces were smaller than 10^{-5} eV/Å. Phonon calculations were carried out using AFLOW-APL (Automatic Phonon Library) in conjunction with VASP.[17] APL uses the finite displacement method to calculate the phonon dispersion. Supercells are generated with displaced atoms, and the interatomic forces are calculated using VASP. APL uses these forces to calculate the interatomic force constants and generate the dynamical matrix, which is diagonalized to obtain the phonon eigenmodes and dispersions.

DATA AVAILABILITY

The datasets generated during and/or analyzed during the current study are summarized in the Supplementary Information, and are available from the corresponding author on reasonable request.

AUTHOR CONTRIBUTIONS

E.Z. conceived the research. P.A. and C.T. tested the macroscopic hardness models. P. A. carried out the method development within the XᴛᴀʟOᴘᴛ code. P.A. and X.W. performed the evolutionary searches, used the AFLOW-ML API to predict hardness values with assistance from E.G., and analyzed the results. X.W. ran the phonon calculations to check dynamical stability, with assistance from C.O. D.M.P. led the topological analysis. E.Z. and S.C. supervised the study. All authors participated in discussing the results, and commented on the paper.

ADDITIONAL INFORMATION

Supplementary information accompanies the paper on the *npj Computational Materials* website (https://doi.org/10.1038/s41524-019-0226-8).

Competing interests: The authors declare no competing interests.

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

REFERENCES

1. Li, Q. et al. Rhombohedral superhard structure of BC_2N. *J. Appl. Phys.* **105**, 053514 (2009).
2. Mao, W. L. et al. Bonding changes in compressed superhard graphite. *Science* **302**, 425–427 (2003).
3. Wang, Y., Panzik, J. E., Kiefer, B. & Lee, K. M. Crystal structure of graphite under room-temperature compression and decompression. *Sci. Rep.* **2**, 520 (2012).
4. Li, Q. et al. Superhard monoclinic polymorph of carbon. *Phys. Rev. Lett.* **102**, 175506 (2009).
5. Stavrou, E. et al. Synthesis of ultra-incompressible sp^3-hybridized carbon nitride with 1:1 stoichiometry. *Chem. Mater.* **28**, 6925–6933 (2016).
6. Zhang, M. et al. Superhard BC_3 in cubic diamond structure. *Phys. Rev. Lett.* **114**, 015502 (2015).
7. Sarker, P. et al. High-entropy high-hardness metal carbides discovered by entropy descriptors. *Nat. Commun.* **9**, 4980 (2018).
8. Tehrani, A. M. et al. Machine learning directed search for ultraincompressible superhard materials. *J. Am. Chem. Soc.* **140**, 9844–9853 (2018).
9. Gao, F. et al. Hardness of covalent crystals. *Phys. Rev. Lett.* **91**, 015502 (2003).
10. Šimůnek, A. & Vackář, J. Hardness of covalent and ionic crystals: first-principle calculations. *Phys. Rev. Lett.* **96**, 085501 (2006).
11. Li, K., Wang, X., Zhang, F. & Xue, D. Electronegativity identification of novel superhard materials. *Phys. Rev. Lett.* **100**, 235504 (2008).
12. Zhang, X. et al. First-principles structural design of superhard materials. *J. Chem. Phys.* **138**, 114101 (2013).
13. Lyakhov, A. O. & Oganov, A. R. Evolutionary search for superhard materials: methodology and applications to forms of carbon and TiO_2. *Phys. Rev. B* **84**, 092103 (2011).
14. Teter, D. M. Computational alchemy: The search for new superhard materials. *MRS Bull.* **23**, 22–27 (1998).
15. Chen, X.-Q., Niu, H., Li, D. & Li, Y. Modeling hardness of polycrystalline materials and bulk metallic glasses. *Intermetallics* **19**, 1275–1281 (2011).
16. Tian, Y., Xu, B. & Zhao, Z. Microscopic theory of hardness and design of novel superhard crystals. *Int. J. Refract. Met. H.* **33**, 93–106 (2012).
17. Curtarolo, S. et al. AFLOW: an automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.* **58**, 218–226 (2012).
18. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD). *JOM* **65**, 1501–1509 (2013).
19. Landis, D. D. et al. The computational materials repository. *Comput. Sci. Eng.* **14**, 51–57 (2012).
20. Bergerhoff, G., Hundt, R., Sievers, R. & Brown, I. D. The inorganic crystal structure data base. *J. Chem. Inf. Comp. Sci.* **23**, 66–69 (1983).
21. Jain, A. et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
22. de Jong, M. et al. Charting the complete elastic properties of inorganic crystalline compounds. *Sci. Data* **2**, 150009 (2015).
23. Stanev, V. et al. Machine learning modeling of superconducting critical temperature. *npj Comput. Mater.* **4**, 29 (2018).
24. Isayev, O. et al. Universal fragment descriptors for predicting properties of inorganic crystals. *Nat. Commun.* **8**, 15679 (2017).
25. Pilania, G. et al. Machine learning bandgaps of double perovskites. *Sci. Rep.* **6**, 19375 (2016).
26. 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* **89**, 054303 (2014).
27. Curtarolo, S. et al. AFLOWLIB.ORG: a distributed materials properties repository from high-throughput ab initio calculations. *Comput. Mater. Sci.* **58**, 227–235 (2012).
28. Toher, C. et al. High-throughput computational screening of thermal conductivity, Debye temperature, and Grüneisen parameter using a quasiharmonic Debye model. *Phys. Rev. B* **90**, 174107 (2014).
29. Toher, C. et al. Combining the AFLOW GIBBS and elastic libraries to efficiently and robustly screen thermomechanical properties of solids. *Phys. Rev. Mater.* **1**, 015401 (2017).
30. Taylor, R. H. et al. A RESTful API for exchanging materials data in the AFLOWLIB. org consortium. *Comput. Mater. Sci.* **93**, 178–192 (2014).
31. Rose, F. et al. AFLUX: the LUX materials search API for the AFLOW data repositories. *Comput. Mater. Sci.* **137**, 362–370 (2017).
32. Gossett, E. et al. AFLOW-ML: a RESTful API for machine-learning predictions of materials properties. *Comput. Mater. Sci.* **152**, 134–145 (2018).
33. Lonie, D. C. & Zurek, E. XtalOpt: an open-source evolutionary algorithm for crystal structure prediction. *Comput. Phys. Commun.* **182**, 372–387 (2011).
34. Avery, P., Toher, C., Curtarolo, S. & Zurek, E. New version announcement: XtalOpt version r12: An open-source evolutionary algorithm for crystal structure prediction. *Comput. Phys. Commun.* **237**, 274–275 (2018).
35. Liu, A. Y. & Cohen, M. L. Prediction of new low compressibility solids. *Science* **245**, 841–842 (1989).
36. Sung, C.-M. & Sung, M. Carbon nitride and other speculative superhard materials. *Mater. Chem. Phys.* **43**, 1–18 (1996).
37. Gao, F. M. & Gao, L. H. Microscopic models of hardness. *J. Superhard Mater.* **32**, 148–166 (2010).
38. Brazhkin, V. V., Lyapin, A. G. & Hemley, R. J. Harder than diamond: dreams and reality. *Philos. Mag. A* **82**, 231–253 (2002).
39. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).

40. Zurek, E. Discovering new materials via a priori crystal structure prediction. In *Reviews in Computational Chemistry* (eds. Parrill, A. L. & Lipkowitz, K. B.) Vol. 29, 274–326 (John Wiley & Sons, Inc., Hoboken, New Jersey, 2016).

41. Avery, P., Falls, Z. & Zurek, E. XtalOpt version r11: an open-source evolutionary algorithm for crystal structure prediction. *Comput. Phys. Commun.* **222**, 418–419 (2018).

42. Lonie, D. C. & Zurek, E. Identifying duplicate crystal structures: XtalComp, an open–source solution. *Comput. Phys. Commun.* **183**, 690–697 (2012).

43. Avery, P. & Zurek, E. RandSpg: an open-source program for generating atomistic crystal structures with specific spacegroups. *Comput. Phys. Commun.* **213**, 208–216 (2017).

44. Hoffmann, R., Kabanov, A. A., Golov, A. A. & Proserpio, D. M. Homo citans and carbon allotropes: for an ethics of citation. *Angew. Chem. Int. Ed.* **55**, 10962–10976 (2016).

45. Mehl, M. J. et al. The AFLOW library of crystallographic prototypes: Part 1. *Comput. Mater. Sci.* **136**, S1–S828 (2017).

46. Pickard, C. J. & Needs, R. J. Hypothetical low-energy chiral framework structure of group 14 elements. *Phys. Rev. B* **81**, 014106 (2010).

47. Shiell, T. B. et al. Nanocrystalline hexagonal diamond formed from glassy carbon. *Sci. Rep.* **6**, 37232 (2016).

48. Hu, M. et al. Compressed glassy carbon: an ultrastrong and elastic inter-penetrating graphene network. *Sci. Adv.* **3**, e1603213 (2017).

49. Huang, Q. et al. Nanotwinned diamond with unprecedented hardness and stability. *Nature* **510**, 250–253 (2014).

50. Tao, Q. et al. Nanotwinned diamond synthesized from multicore carbon onion. *Carbon* **120**, 405–410 (2017).

51. Pan, Z., Sun, H., Zhang, Y. & Chen, C. Harder than diamond: superior indentation strength of wurtzite BN and lonsdaleite. *Phys. Rev. Lett.* **102**, 055503 (2009).

52. Qingkun, L., Yi, S., Zhiyuan, L. & Yu, Z. Lonsdaleite–a material stronger and stiffer than diamond. *Scr. Mater.* **65**, 229–232 (2011).

53. Németh, P. et al. Lonsdaleite is faulted and twinned cubic diamond and does not exist as a discrete material. *Nat. Commun.* **5**, 5447 (2014).

54. Turneaure, S. J., Sharma, S. M., Volz, T. J., Winey, J. M. & Gupta, Y. M. Transformation of shock-compressed graphite to hexagonal diamond in nanoseconds. *Sci. Adv.* **3**, eaao3561 (2017).

55. Baburin, I. A., Proserpio, D. M., Saleev, V. A. & Shipilova, A. V. From zeolite nets to sp^3 carbon allotropes: a topology-based multiscale theoretical study. *Phys. Chem. Chem. Phys.* **17**, 1332–1338 (2015).

56. Deringer, V. L., Csányi, G. & Proserpio, D. M. Extracting crystal chemistry from amorphous carbon structures. *ChemPhysChem* **18**, 873–877 (2017).

57. Selli, D., Baburin, I. A., Martoňák, R. & Leoni, S. Superhard sp^3 carbon allotropes with odd and even ring topologies. *Phys. Rev. B* **84**, 161411 (2011).

58. Sheng, X.-L., Yan, Q.-B., Ye, F., Zheng, Q.-R. & Su, G. T-carbon: a novel carbon allotrope. *Phys. Rev. Lett.* **106**, 155703 (2011).

59. Chen, X.-Q., Niu, H., Franchini, C., Li, D. & Li, Y. Hardness of *T*-carbon: density functional theory calculations. *Phys. Rev. B* **84**, 121405 (2011).

60. Patterson, J. R., Catledge, S. A., Vohra, Y. K., Akella, J. & Weir, S. T. Electrical and mechanical properties of C_{70} fullerene and graphite under high pressures studied using designer diamond anvils. *Phys. Rev. Lett.* **85**, 5364–5367 (2000).

61. Kresse, G. & Hafner, J. *Ab Initio* molecular dynamics for liquid metals. *Phys. Rev. B.* **47**, 558 (1993).

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

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

64. Calderon, C. E. et al. The AFLOW standard for high-throughput materials science calculations. *Comput. Mater. Sci.* **108**, 233–238 (2015).

65. Hicks, D. et al. AFLOW-SYM: platform for the complete, automatic and self-consistent symmetry analysis of crystals. *Acta Cryst. A* **74**, 184–203 (2018).

66. Blatov, V. A., Shevchenko, A. P. & Proserpio, D. M. Applied topological analysis of crystal structures with the program package ToposPro. *Cryst. Growth Des.* **14**, 3576–3586 (2014).