Journal of Computational Science 82 (2024) 102402

Contents lists available at ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

Material hardness descriptor derived by symbolic regression

Christian Tantardini [a,b,c,*], Hayk A. Zakaryan [d], Zhong-Kang Han [e], Tariq Altalhi [f], Sergey V. Levchenko [e], Alexander G. Kvashnin [g,*], Boris I. Yakobson [b,f,*]

[a] Hylleraas center, Department of Chemistry, UiT The Arctic University of Norway, P.O. Box 6050 Langnes, Tromsø, N-9037, Norway
[b] Department of Materials Science and Nanoengineering, Rice University, 6100 Main St, Houston, 77005, TX, United States of America
[c] Institute of Solid State Chemistry and Mechanochemistry SB RAS, Kutateladze 18, Novosibirsk, 630128, Russian Federation
[d] Yerevan State University, 1 Alex Manoogian St., Yerevan, 0025, Armenia
[e] Zhejiang University, 866 Yuhangtang Rd, Hangzhou, 310027, China
[f] Chemistry Department, Taif University, P.O. Box 11099, Taif, 21944, Saudi Arabia
[g] Skolkovo Institute of Science and Technology, Bolshoy Boulevard 30, bld. 1, Moscow, 121205, Russian Federation

ARTICLE INFO

Keywords:
Hardness
SISSO
Machine learning
Symbolic regression
Superhard materials

ABSTRACT

Hardness is a materials' property with implications in several industrial fields, including oil and gas, manufacturing, and others. However, the relationship between this macroscale property and atomic (i.e., microscale) properties is unknown and in the last decade several models have unsuccessfully tried to correlate them in a wide range of chemical space. The understanding of such relationship is of fundamental importance for discovery of harder materials with specific characteristics to be employed in a wide range of fields. In this work, we have found a physical descriptor for Vickers hardness using a symbolic-regression artificial-intelligence approach based on compressed sensing. SISSO (Sure Independence Screening plus Sparsifying Operator) is an artificial-intelligence algorithm used for discovering simple and interpretable predictive models. It performs feature selection from up to billions of candidates obtained from several primary features by applying a set of mathematical operators. The resulting sparse SISSO model accurately describes the target property (i.e., Vickers hardness) with minimal complexity. We have considered the experimental values of hardness for binary, ternary, and quaternary transition-metal borides, carbides, nitrides, carbonitrides, carboborides, and boronitrides of 61 materials, on which the fitting was performed.. The found descriptor is a non-linear function of the microscopic properties, with the most significant contribution being from a combination of Voigt-averaged bulk modulus, Poisson's ratio, and Reuss-averaged shear modulus. Results of high-throughput screening of 635 candidate materials using the found descriptor suggest the enhancement of material's hardness through mixing with harder yet metastable structures (e.g., metastable VN, TaN, ReN_2, Cr_3N_4, and ZrB_6 all exhibit high hardness).

1. Introduction

Hardness is a mechanical property of materials important for several industrial applications. In particular, hardness is measured and used as a parameter determining the type of application itself for materials in construction or manufacturing, e.g., cutting, drilling, or grinding [1–3]. Over the years, various scales of hardness have been proposed. The Vickers scale is considered universal because it spans both macro- and micro-scales and is independent of the size of the indenter [4]. Thus, Vickers hardness is commonly measured in various applications to determine whether a material is superhard, typically 40 GPa or higher hardness [5,6]. In some applications these materials are required to fulfil additional requirements. For example, they need to preserve their hardness at high pressure and temperature, be non-toxic, and so on. Therefore, searching for hard and superhard materials with different chemical compositions remains an important challenge. The best candidates for superhard materials are borides, carbides, and nitrides of metals [7,8].

In order to find materials with high hardness among many candidates, one can synthesize and test all of them one by one. However, this is obviously a very inefficient approach. Alternatively, one can find a correlation between hardness and features (or their mathematical combinations) that are easy to evaluate. This correlation can then be used to quickly explore the chemical space of candidate materials (see also discussion and Fig.3 in Penev et al. [9]). Such a combination of features is called descriptor.

* Corresponding authors.
E-mail addresses: christiantantardini@ymail.com (C. Tantardini), A.Kvashnin@Skoltech.ru (A.G. Kvashnin), biy@rice.edu (B.I. Yakobson).

https://doi.org/10.1016/j.jocs.2024.102402
Received 30 March 2023; Received in revised form 31 July 2024; Accepted 31 July 2024
Available online 6 August 2024

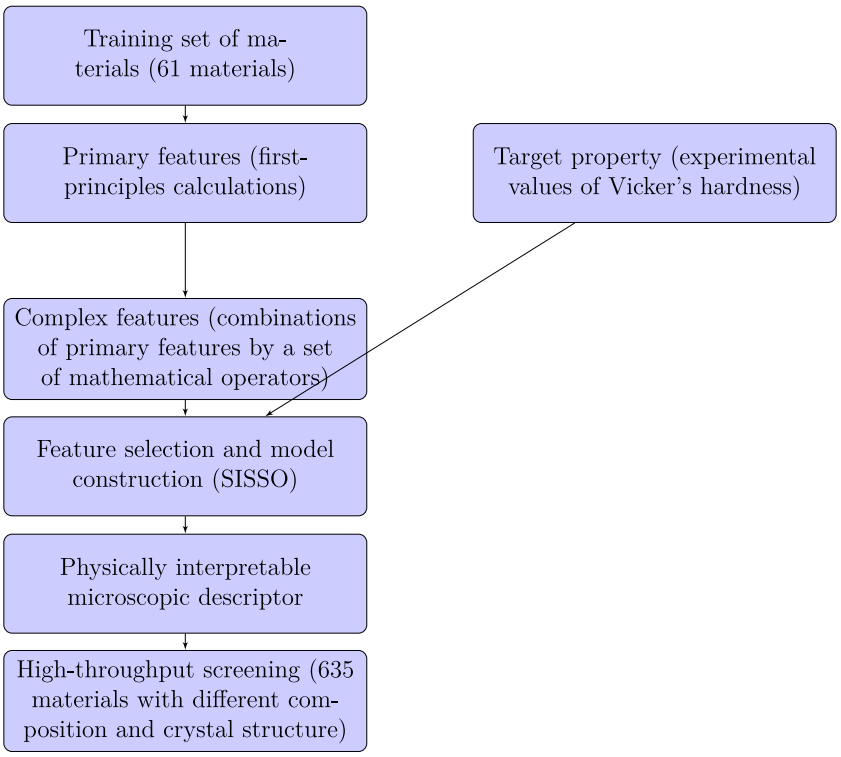

Fig. 1. Schematic workflow of the SISSO method, illustrating the steps from the training set and primary features to the final high-throughput screening.

Both macroscopic properties (fluidity, elastic stiffness, ductility, strength, crack resistance and viscosity) and microscopic properties (from atomistic simulations) can be important constituents of the descriptor [4]. Which features are most important or sufficient is unknown, but it is known that none of the single features tested so far is a good descriptor. On the other hand, a set of physically relevant features can be relatively easily obtained from atomistic simulations. Therefore, it is attractive to explore whether a mathematical combination of such features (perhaps non-linear) correlates with Vickers hardness. In the last decade several semi-empirical models were developed that use elastic properties as input to predict the hardness [10–12]. Low accuracy of such models points out that other features must be included.

Recently, Podryabinkin et al. [13] proposed an alternative approach to calculate the hardness, which utilizes first-principles calculations and machine learning potentials that were actively learned on local atomic environments in order to explicitly model the process of nanoindentation. Although this method is highly accurate, it is too computationally expensive for a high-throughput screening. Thus, there is a need for a physical, easily computable descriptor of hardness that can be used for a high-throughput search for superhard materials in the vast chemical space. Such descriptor is found here using sure independence screening plus sparsifying operator (SISSO) [14–19]. In the past, neural networks have been used several times to describe Vickers hardness [20–23]. However, the considered chemical and structural spaces were quite limited. Moreover, no microscopic information was used in these studies. Also, due to the nature of such machine learning methods, all physical meaning is lost, providing no insight into the structural features that could be explored to create new, harder materials.

2. Methods

The workflow that we have used and showed in Fig. 1 is described below. The initial step is to define the set of primary features of materials. The primary features are readily available or obtainable properties that may have a physical relation to the target property (i.e. Vicker's hardness) of 61 compounds for which the experimental value is known. We have included all features that can be obtained from atomistic simulations easier than hardness itself and are physically related to bond strength, bond breaking, and restructuring [4]. For example, first ionization energy is related to chemical bond hardness [24]. Bulk and shear moduli are used in the well-known Chen's model [11]. Other features are mechanical properties (Young's modulus, Poisson's ratio, and so on) that are related to hardness, but this relationship is non-trivial [4]. We use SISSO to both select most relevant features (or rather combination of features) from the initial set and find a (possibly non-linear) relationship between these features and hardness, as described below. The accuracy of the model and in particular its predictive power (accuracy of predicting hardness for materials not included in the training set) indicate whether the initial set of primary features is sufficient for predicting hardness.

Primary features are combined using a set of mathematical operators to form a large number (up to tens of billions) of complex features. Each complex feature is a (generally) non-linear formula including one, two, or more primary features, depending on requested complexity level (more complex combinations lead to larger feature spaces). These complex features are used in SISSO as a basis in materials space: the target property is expressed as a linear combination of complex features, but each complex feature may be a non-linear function of primary features. Primary features themselves are also included in the basis. Using complex features is a crucial step, because in general primary features may not correlate well with the target property, while their combination may represent a physical relation that describes target property very well. For example, there is no physical reason why either electron affinity or ionization potential should correlate to electronic excitation energy, while the difference between them is directly physically related to it.

Here, we have produced an initial set with 20 primary features, see Table 1. This set is used to generate 1260 candidate features (complex plus primary). The list of primary features includes the properties such as radii of the atoms in the compound, density, bulk and shear moduli, as well as the elasticity tensor components, elastic anisotropy, Poisson's ratio, Young's modulus, and more (see Table 1).

Table 1
Primary features used for construction of the descriptor.

Name	Units	Abbreviation
Density	g/cm^3	D
Voigt averaging of bulk modulus B_V	GPa	B_V
Reuss averaging of bulk modulus B_R	GPa	B_R
Voigt–Reuss–Hill averaging of bulk modulus B_{VRH}	GPa	B_{VRH}
Voigt averaging of shear modulus G_V	GPa	G_V
Reuss averaging of shear modulus G_R	GPa	G_R
Voigt–Reuss–Hill averaging of shear modulus G_{VRH}	GPa	G_{VRH}
Young's modulus	GPa	Y
Fraction		Fr
Elastic anisotropy		el
Poisson's ratio		σ
Maximum atomic radius	Å	R_X
Minimum atomic radius	Å	R_N
Weighted atomic radius	Å	R_W
Maximum atomic weight	a.m.u.	A_X
Minimum atomic weight	a.m.u.	A_N
Weighted atomic weight	a.m.u.	A_W
Maximum first ionization energy	eV	I_X
Minimum first ionization energy	eV	I_N
Weighted first ionization energy	eV	I_W

All of the primary features utilized in this study were obtained from either the literature [25] or from the Materials Project database [26]. The primary features were combined using the following set of operators [14–19]:

$$\hat{H} \equiv \{+, -, *, /, ^{-1}, ^2, ^3, \sqrt{}, \sqrt[3]{}, \exp, \log, |-|\}[\phi_1, \phi_2], \tag{1}$$

where ϕ_1 and ϕ_2 are primary features (in case of a unary operator, only one feature ϕ_1 is considered). The set of operators is applied recursively to generate features of increasing complexity. Complexity level zero (Φ_0) contains only primary features. Φ_1 contains Φ_0, features obtained by applying unary operators to all primary features, and binary combinations of primary features. Φ_2 contains Φ_1 plus all new features obtained by unary and binary operations on Φ_1. In this work, Φ_2 is the highest considered level of complexity.

SISSO is used to both *select* the most important complex (or primary) features and find the model for the target property. Thus, SISSO automatically, as explained below, finds the most important primary features and the physically interpretable descriptor from data. The number of selected features (descriptor dimension) depends on the required accuracy of training data fitting: the larger it is, the better is the fitting. However, larger number of complex features will eventually result in overfitting, which leads to worsening prediction accuracy for new materials not included in the training set. The optimal number of complex features is determined by cross-validation. In this work we have used 10-fold cross-validation (CV10). This consisted in randomly subdividing the data set in ten subsets and progressively using nine subsets for training the SISSO model and one subset for verification of the model. The prediction (CV10) error is then evaluated as an average model error for the 10 verification subsets. As the number of complex features in the model (which is an input parameter) increases, the fitting error first decreases, but eventually starts to increase, indicating overfitting. The dimension that yields minimum CV10 error is then used to find the best SISSO model using all training data.

3. Results and discussion

3.1. Development of the descriptor

We have collected a set of 635 compounds, selected from the Materials Project database [26]. Materials without reliable experimental data for Vickers hardness (i.e., target property) were eliminated from the dataset, as those that were deemed unstable according to DFT calculations from aforementioned database. This led to a total of 61 compounds for our training dataset, containing both hard materials (borides, carbides, nitrides, etc.) and comparatively soft ionic crystals and oxides (NaCl, Al_2O_3, etc.). In order to access the values of such primary features and the properties for the training data sets, a request can be made via the GitHub (see link in the Appendix A). The dataset for the target property (hardness) was created using information gathered from Zhang et al. [27]. The best found descriptors with different dimension (from 1D to 6D), see Eqs. (A1)–(A6) are presented in Appendix A. The CV10 error as a function of dimension is shown in Fig. 2a. As we can see, CV10 error increases from dimension larger than two, while the fitting root-mean square error (RMSE) reduces monotonically as the dimension of the descriptor increases (red curve in Fig. 2a). CV10 is expected to increase in case of overfitting, which happens when the number of parameters, in this case the dimension of descriptor, is so high that the model learns random details and noise in the dataset, making it unable to correctly predict the property of unexplored materials. Thus, according to CV10, the obtained optimal descriptor dimension is two. This 2D descriptor bears a relatively complex analytical form:

$$H^{SISSO}_{predicted} = 0.147 \cdot \frac{B_V}{\sigma \sqrt[3]{G_R}} - 1.136 \cdot \frac{B_R \log R_X}{A_W} - 5.679 \tag{2}$$

where B_V, B_R are the values of bulk modulus calculated using Voigt and Reuss averaging methods, [28,29] respectively, while G_R is the shear modulus calculated using Reuss averaging method, σ is a Poisson's ratio, A_W is the average atomic mass of the compound, and R_X is the maximum atomic radius of the species in the compound. The computed atomic radius and mass data were obtained from the Python library for materials analysis, Pymatgen [30].

The correlation between the optimal SISSO model and experimental hardness values is shown in Fig. 2b. The error distribution for hardness prediction using the optimal model with the 2D descriptor is shown as the inset to Fig. 2b. We attained a relatively low fitting RMSE of 4.28 GPa, as well as CV10 RMSE of 5.48 GPa for 2D model, with a maximum absolute error (MaxAE) of 10.1 GPa on the training set. The average relative error with respect to experimental hardness is 1.99%, which is small enough for a fast screening of promising candidates and lowest compared to what obtained with different previous models as Teter [10], Chen [11] and Mazhnik [12] that are respectively 21 %, 14 % and 16 %. The small CV10 error also indicates that the chosen set of primary features contains the important physical quantities that are necessary for describing the hardness. Although including more advanced and therefore computationally more expensive primary features, e.g. surface or defect formation energies, may significantly improve the model, the reasonable predictive power of the current model indicates that these more advanced features can be approximately expressed through the employed primary features.

3.2. Analysis of the impact of dimension of descriptor

Furthermore, to better understand the impact that each component of the two-dimensional descriptor has on the outcome, we calculated an importance score IS for each term in Eq. (2) towards the total error of our model. This involved eliminating one component of the descriptor at a time and re-fitting the model with the remaining component. The resulting one-dimensional derivative models are formulated as follows:

$$H_1 = a_1 \cdot \frac{B_R \log R_X}{A_W} + b_1 \tag{3}$$

and

$$H_2 = a_2 \cdot \frac{B_V}{\sigma \sqrt[3]{G_R}} + b_2 \tag{4}$$

Coefficients a_1, b_1, and a_2, b_2 were fitted separately for H_1 and H_2 by minimizing RMSE, and are equal to $a_1 = 15.384$, $b_1 = 0$, and $a_2 = 0.1485$, $b_2 = -7.2$.

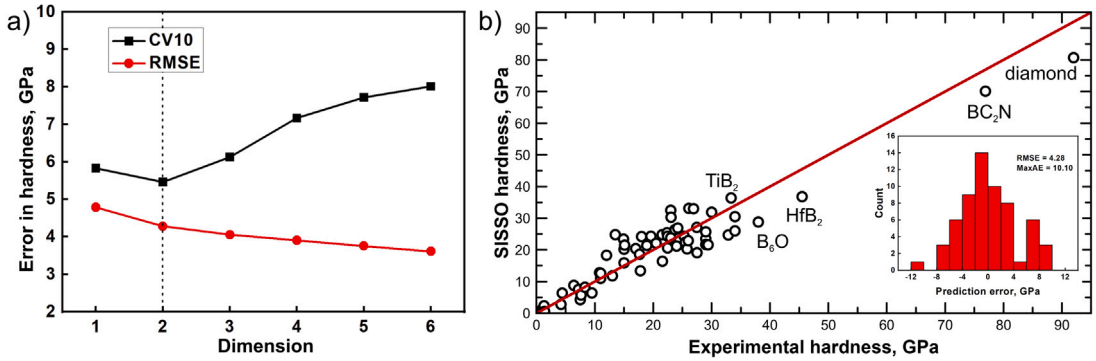

Fig. 2. (a) Root-mean square error (RMSE) for the SISSO model and the average RMSE of CV10. Dashed vertical line denotes the optimal descriptor dimension. (b) The correlation between predicted hardness by 2D SISSO descriptor and experimental values of 61 compounds. The inset shows the distribution of RMSE and maximum absolute error (MaxAE) for the prediction of hardness 2D SISSO descriptor.

The IS is calculated by using the RMSE and MaxAE values for H_1 and H_2, respectively, for our dataset as follows:

$$IS_i^{\mathrm{RMSE}} = 1 - \frac{\mathrm{RMSE}(H_{predicted})}{\mathrm{RMSE}(H_i)} \tag{5}$$

$$IS_i^{\mathrm{MaxAE}} = 1 - \frac{\mathrm{MaxAE}(H_{predicted})}{\mathrm{MaxAE}(H_i)} \tag{6}$$

Calculated importance scores based on RMSE and MaxAE are respectively 0.49 and 0.52 for IS_1, and 0.07 and 0.06 for IS_2. Thus, the first descriptor component in Eq. (2) plays a more significant role in hardness according to our model. However, including both descriptor components in the SISSO model reduces the RMSE and MaxAE errors by 6%–7%. The RMSE on the shared dataset is 5.2 GPa for H_1 and 9.3 GPa for H_2. The use of both H_1 and H_2 results in a lower error value of 4.28 GPa, highlighting the importance of the 2D descriptor in comparison to the 1D counterpart.

The obtained 2D model was used to perform high-throughput screening of hard and superhard materials belonging to binary, ternary, and quaternary transition metal borides, carbides, and nitrides. The required crystal structures of experimentally known and hypothetical structures were extracted using the Materials Project database [26]. In total, 635 structures were gathered for the selected classes of materials. For each structure, we have extracted the necessary properties for the developed model, including bulk and shear moduli, Poisson's ratio, and the averaged atomic mass of each compound. The maximum radius of the atoms in the compound was determined using the Pymatgen library [30].

To analyse the collected data, we have constructed the correlation plot displaying the relationships between SISSO Vickers hardness, bulk modulus, Poisson's ratio, and shear modulus for 635 inorganic compounds. This excludes diamond, borocarbides, carbonitrides, and layered compounds, as shown in Fig. 3a. The colour scale of the points indicated the energy above the convex hull to represent the (meta)stability of each compound. A clear trend in increasing hardness with higher B_v/σ values is visible. There are outliers which show high hardness and quite low shear modulus, together with a low B_v/σ value which contradicts the general trend. These outliers correspond to metastable structures (see red and green points in Fig. 3a). Despite the denominator of Eq. (2) containing G_R, the correlation between B_V and G_R (Fig. B.1 in Appendix B) results in an overall increase in hardness as shear modulus increases. The correlation between B_V and G_R for stable structures is further enhanced (see

Fig. B.1 in Appendix B). Moreover, compounds conforming to the general trend exhibit the Pugh's ratio (i.e., G/B) ranging from 0.5 to 0.8, as shown in Figs. B.1 and B.2 in Appendix B, which is indicative of how brittle the material becomes or not. This supports non-linear relationship between hardness and other properties, emphasising the significance of accounting for this non-linearity to identify hard materials.

In Fig. 3a, well-known hard and superhard compounds for a total of 635 are identified as reference points. This aids in understanding the location of other compounds in relation to them. The highest values of hardness belong to boride and carbide compounds (see Fig. 3b, c).

Among selected borides (Fig. 3b), ZrB_6 (mp-1001788), a metastable compound located 0.4 eV/atom above the convex hull (according to data from the Materials Project) and with a predicted SISSO hardness of 46 GPa, can be highlighted. ZrB_6 has a crystal structure similar to that of calcium hexaboride, consisting of a 3D boron cage, which contributes to its high bulk modulus and hardness. The influence of the boron cage on the mechanical and elastic properties of borides has previously been demonstrated for hafnium borides [31]. It should be noted that, while such a crystal type is typical for borides of rare-earth elements, it is an unusual metastable structure for transition metals, with an extremely low Reuss-averaged shear modulus of 2 GPa and a low B_v/σ of 500 GPa (the Poisson's ratio is 0.39). However, this discovery suggests a promising way to enhance the hardness of rare-earth borides by incorporating transition metals as substitutes within the crystal structure. Also, high hardness is predicted for well-known superhard compounds, namely TiB_2, ReB_2, HfB_2, and CrB_4 (see Fig. 3b).

Among the carbides, cubic polymorphic modification of tungsten carbide (WC) with $F\bar{4}3m$ space group (see Fig. 3c) has the highest hardness of 46 GPa. WC (mp-1008635) has a zincblende structure where each tungsten atom forms corner-sharing WC_4 tetrahedra with four equivalent carbon atoms. The structure has a bulk modulus of 249 GPa and a shear modulus of 3 GPa, resulting in a very high Poisson's ratio of 0.48. Despite its high hardness, this structure is deemed unstable with an energy of formation 0.67 eV/atom above the convex hull (according to data from the Materials Project). The well-known hexagonal modification of WC has an SISSO hardness of 35 GPa with bulk and shear moduli equal to 387 and 276 GPa respectively. Predicted values are in good agreement with experimental data and those obtained by other models [32]. Hexagonal WC has the highest B_v/σ ratio compared to the other considered carbides at a value of 1842 GPa. Two other structures with comparable mechanical characteristics to WC are CrC (mp-1018050) and MoC (mp-2305) as shown in Fig. 3c. Both of these have a hexagonal $P\bar{6}m2$ space group, the same as in hexagonal WC. Each metal atom in the structure forms bonds with six equivalent carbons to create a mixture of distorted face, edge, and corner-sharing MeC_6 pentagonal pyramids. It is predicted that the SISSO hardness of both CrC and MoC is approximately 30 GPa. The bulk modulus of both structures is roughly 350 GPa, whereas the shear modulus is about 240 GPa. CrC proves metastable with an energy of formation 80 meV/atom above the convex hull, whereas MoC is stable and holds a calculated energy of formation of only 1 meV/atom above the convex hull.

The hardest found compounds among nitrides are VN, TaN, and ReN_2, see Fig. 3d. VN (mp-1002105) belongs to the $Pm\bar{3}m$ space group

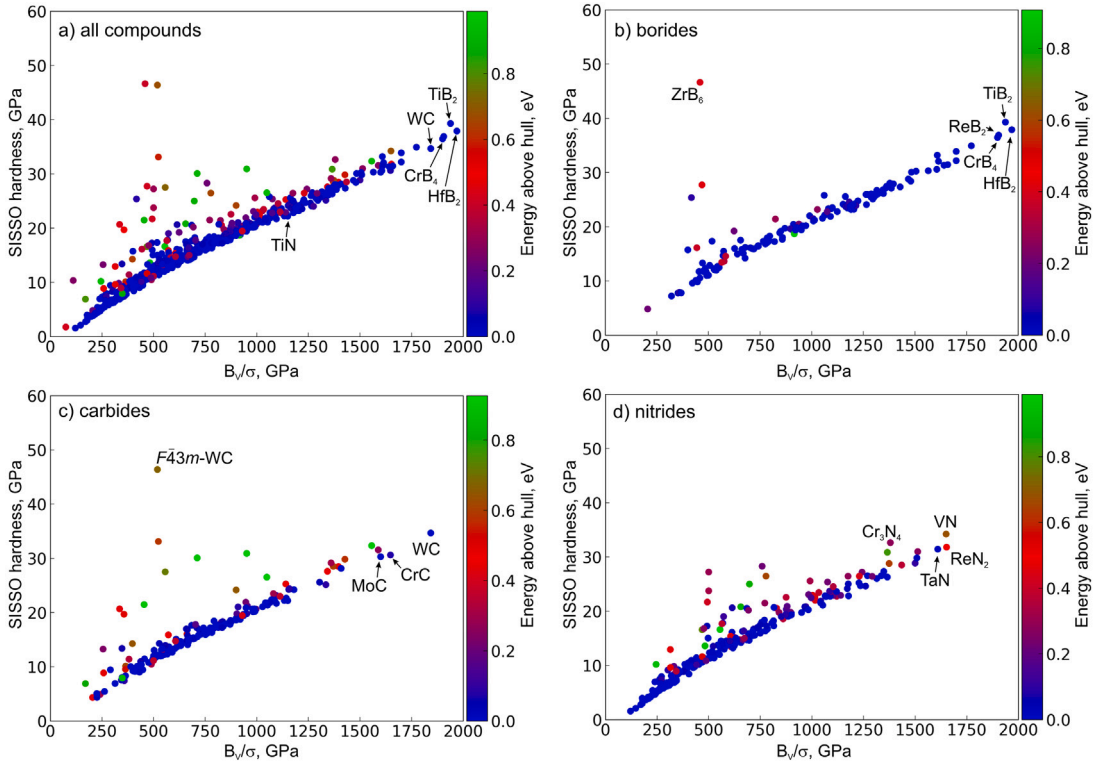

Fig. 3. (a) The SISSO H_V model predictions are plotted against B_v/σ for considered 635 inorganic compounds. Specific classes of materials are also shown, including (b) borides, (c) carbides, and (d) nitrides. Colorbar shows the energy of formation above the convex hull denoting stability of each structure.

and is located 0.68 eV/atom above the convex hull. The SISSO model predicts its hardness to be 34 GPa with $B_v/\sigma = 1650$ GPa (while Poisson's ratio is 0.16). TaN (mp-1009831), which has a SISSO hardness of 31 GPa is isostructural to well-known WC structure and belongs to the $P\bar{3}m2$ space groups. It has a Poisson's ratio of 0.21 and $B_v/\sigma = 1610$ GPa, as shown in Fig. 3d. Renium dinitride (mp-1019055) is located 0.49 eV/atom above the convex hull and predicted to have a hardness of 32 GPa with $B_v/\sigma = 1650$ GPa.

Another interesting nitride material is Cr_3N_4 (mp-1014460), see Fig. 3d. The material has a $Pm\bar{3}m$ space group and can be depicted as a rocksalt structure with a missing atom in the 4*a* Wyckoff position which results in fractional composition. Its predicted SISSO hardness is 33 GPa, and it has a low Poisson's ratio of 0.1, leading to a high B_v/σ value of 1380 GPa.

Furthermore, we compare our SISSO hardness model with other machine learning and empirical models, we have used Teter's [10], Chen's [11], Mazhnik-Oganov's [12], and XGBoost [33] models to predict the hardness of structures in the created dataset. The Fig. 4 portrays their correlations with the SISSO model for stable structures lied on the convex hull. The colour scale indicated variations between the SISSO and the considered reference model. Our model yields a good agreement of predicted hardness values with the Teter model, as shown in Fig. 4a. The greatest difference between predictions was 12 GPa for hexagonal $NaBPt_3$ (mp-28614), and the next greatest was 20 GPa for zincblende FeN (mp-6988). The largest deviation between the SISSO model and Chen's model was found to be 15.5 GPa for $NaBPt_3$ (see Fig. 4b). The SISSO hardness of this compound is 21.2 GPa, whereas Chen's hardness is only around 5 GPa. This significant variation could be attributed to the highly anisotropic structure of $NaBPt_3$, resulting in a difference of 33 GPa between Reuss- and Voigt-averaged shear moduli according to the Materials Project. In our model, we use Reuss averaging, resulting in higher hardness than Chen's model, which uses the Voigt–Reuss–Hill averaged shear modulus. The latter is lower compared to the Reuss averaged value for $NaBPt_3$. Predictions of our model align well with the recent Mazhnik-Oganov model [12] as shown in Fig. 4c, except for $NaBPt_3$ and FeN, where the differences are similar to Teter's model.

The use of machine-learning XGBoost model for predicting hardness was innovative and highly efficient [33]. We trained the same XGBoost model as was used in Ref. [33] on our training set, and predicted hardness for all the considered compounds. First, we performed the 10-fold cross-validation using the same techniques and dataset as for SISSO training, that is, we divided the dataset into 10 subsets and trained the XGBoost model using 9 of those subsets. The CV10 error was calculated as the mean value of the test RMSE acquired for each of the ten subsets, and equated to 7.8 GPa, which is approximately twice as high as the CV10 error for SISSO. The distribution of errors for the XGBoost model for CV10 is shown in the Appendix B (Fig. B.3). The correlations between the XGBoost model and the SISSO model is shown in Fig. 4d. Numerous structures have a hardness disparity ranging from 12 to 17 GPa. Most of these structures comprise rare-earth metal carbides, specifically Y_2C (mp-1334), Sc_4C_3 (mp-15661), Y_4C_5 (mp-9459), Y_2ReC_2 (mp-21003). Such significant differences in the hardness predicted by the XGBoost and SISSO models for our compounds can be attributed to the fitting hyperparameters of the XGBoost algorithm, which need to be redefined before training on the new training set. When considering only transition-metal borides, carbides, and nitrides, much lower differences can be obtained between XGBoost and SISSO, as XGBoost more accurately describes these classes of compounds (see Fig. B.4 in Appendix B).

Our findings demonstrate that SISSO identified a physical significance of the B_v/σ ratio for hardness, enabling one to quickly estimate the hardness of a compound across a diverse range of chemical compositions and crystal structures. A greater B_v/σ ratio corresponds to heightened hardness.

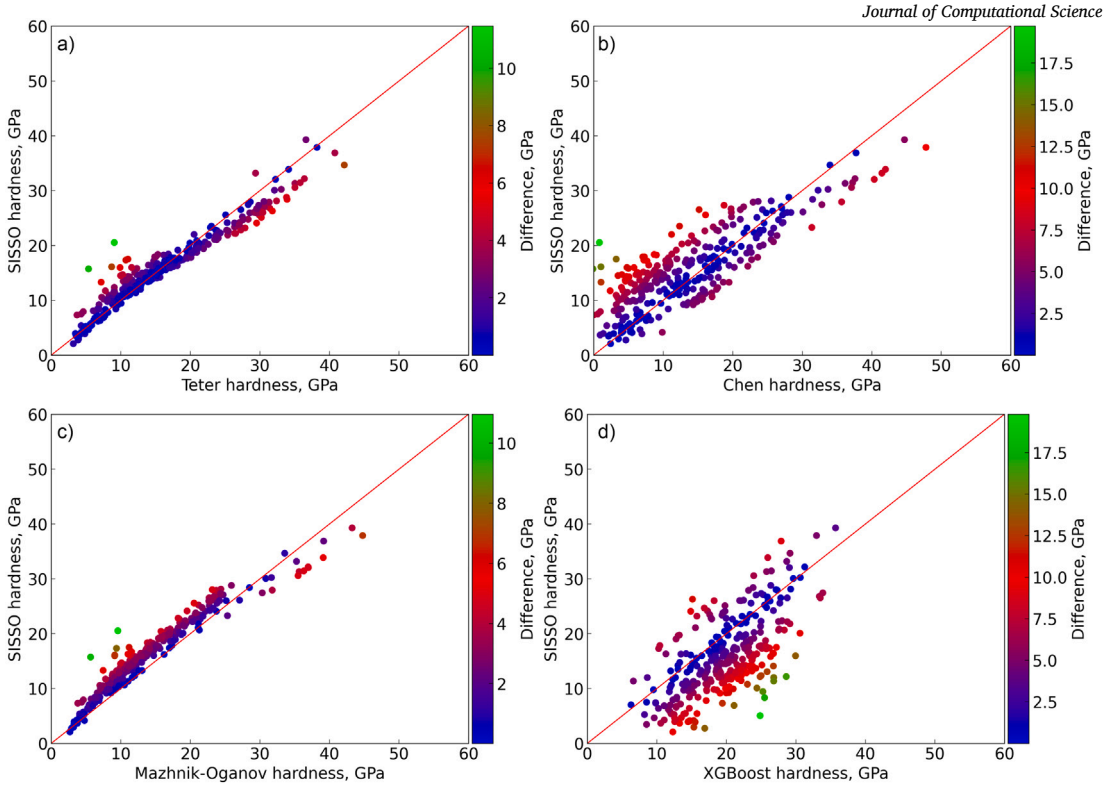

Fig. 4. Correlations between SISSO hardness and (a) Teter [10], (b) Chen [11], (c) Mazhnik-Oganov [12], (d) XGBoost [33] models for considered stable structures. Colorbar shows the difference between two sets of data.

4. Conclusion

In conclusion, our study successfully identified a physical descriptor for Vickers hardness through a novel application of the SISSO artificial-intelligence algorithm. By leveraging a symbolic-regression approach based on compressed sensing, we have derived a non-linear function correlating microscopic properties to macroscale hardness for a wide range of materials. The key contributors to this descriptor are the Voigt-averaged bulk modulus, Poisson's ratio, and Reuss-averaged shear modulus, reflecting the intricate relationship between these properties and material hardness. The model, validated against experimental values for a diverse set of transition-metal compounds, demonstrates significant predictive power. High-throughput screening of 635 candidate materials using this descriptor reveals promising pathways for enhancing material hardness, particularly through the incorporation of harder, metastable structures such as VN, TaN, ReN_2, Cr_3N_4, and ZrB_6.

Broadly, these findings underscore the transformative potential of artificial intelligence in materials science. By bridging the gap between atomic-level properties and macroscale characteristics, such approaches can accelerate the discovery and design of advanced materials with tailored properties, driving innovation across various industrial sectors. This research not only provides a robust tool for predicting material hardness, but also highlights the importance of interdisciplinary methodologies in solving complex materials challenges.

CRediT authorship contribution statement

Christian Tantardini: Writing – review & editing, Writing – original draft, Investigation, Formal analysis. **Hayk A. Zakaryan:** Data curation, Investigation. **Zhong-Kang Han:** Software, Investigation, Validation. **Tariq Altalhi:** Formal analysis. **Sergey V. Levchenko:** Methodology, Software, Writing – original draft. **Alexander G. Kvashnin:** Writing – review & editing, Writing – original draft, Conceptualization, Supervision. **Boris I. Yakobson:** Writing – review & editing, Formal analysis, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data is available via the link provided in Supporting Information.

Acknowledgements

The research was carried out within the state assignment to ISSCM SB RAS (project No. 121032500059-4). Ch.T was supported by the Norwegian Research Council through a Centre of Excellence grant (Hylleraas Centre 262695), a FRIPRO grant (ReMRChem 324590) and from NOTUR – The Norwegian Metacenter for Computational Science through grant of computer time (nn14654k). T.A. and B.I.Y. acknowledge the Taif University Research Support Project (TURSPHC2024/1, Saudi Arabia).

Appendix A. Predicted descriptors

All the data about datasets are available via the github link by request: github.com/AlexanderKvashnin/SISSO_hardness.git. There is a list of predicted descriptors by SISSO used for calculations the RMSE and CV10 in Fig. 1a.

$$H^{1D} = 0.182 \cdot \frac{B_R}{\sigma \sqrt[3]{Y}} - 6.191 \tag{A.1}$$

$$H^{2D} = 0.147 \cdot \frac{B_V}{\sigma \sqrt[3]{G_R}} - 1.136 \cdot \frac{B_R \log R_X}{A_W} - 5.679 \tag{A.2}$$

$$H^{3D} = 0.659 \cdot \frac{B_R}{\sigma \sqrt[3]{Y}} - 1.405 \cdot \frac{G_V}{A_W} \cdot \log R_X$$

$$- 0.042 \cdot \frac{Fr}{R_N \log el} - 12.221 \tag{A.3}$$

$$H^{4D} = 0.677 \cdot \frac{B_R}{\sigma \sqrt[3]{Y}} - 0.133 \cdot \frac{Y}{D} \cdot \log R_X + 0.041 \cdot \frac{Fr}{R_N \log el} - 13.228 \cdot \frac{I_W}{I_X \sqrt{R_W}} - 1.471 \tag{A.4}$$

$$H^{5D} = 0.155 \cdot \frac{B_R}{\sigma \sqrt[3]{G_V}} - 0.353 \cdot \frac{G_V}{D} \cdot \log R_X + 0.054 \cdot \frac{Fr}{R_W \log el} - 1027 \cdot \frac{|B_V - G_R|}{\exp A_N} + 3.190 \cdot \frac{R_W}{el|B_R - G_V|} - 5.873 \tag{A.5}$$

$$H^{6D} = 0.177 \cdot \frac{B_R}{\sigma \sqrt[3]{G_V}} - 41.972 \cdot \frac{\log R_X}{A_W} \cdot \sigma + 0.046 \cdot \frac{G_R}{R_N \log el} - 1175 \cdot \frac{|B_R - G_R|}{\exp A_N} + 0.047 \cdot \frac{D^3}{|B_V - G_V|} - 0.963 \cdot \frac{A_X}{A_W} \cdot \sqrt{A_N} + 3.815 \tag{A.6}$$

Appendix B. Additional data

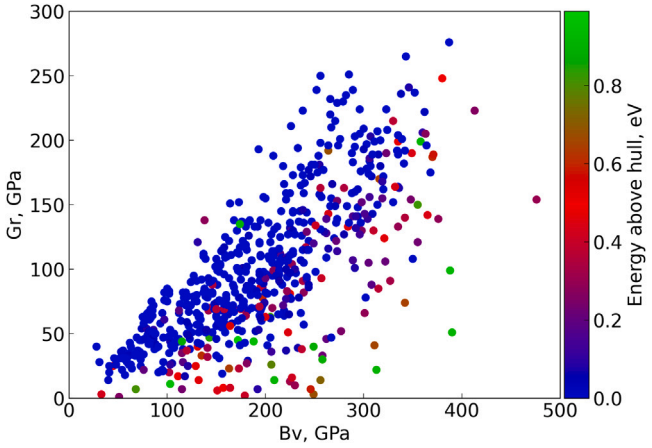

Fig. B.1. Correlation between Voigt-averaged bulk modulus and Reuss-averaged shear modulus of stable and metastable structures among borides, carbides, and nitrides. Colorbar shows the energy of formation above the convex hull denoting stability of each structure.

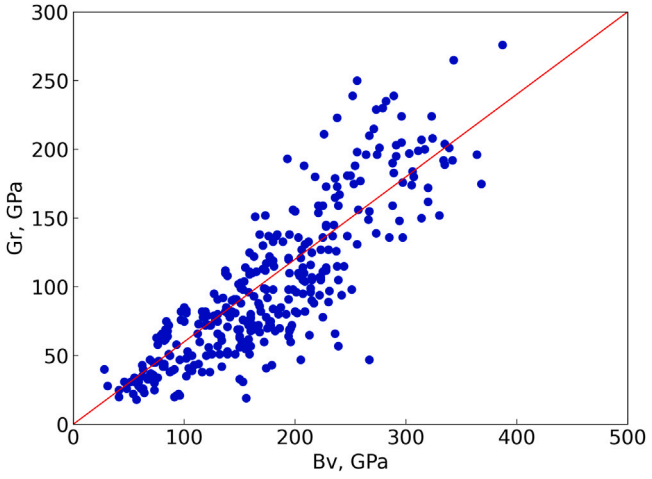

Fig. B.2. Correlation between Voigt-averaged bulk modulus and Reuss-averaged shear modulus of only stable structures among borides, carbides, and nitrides.

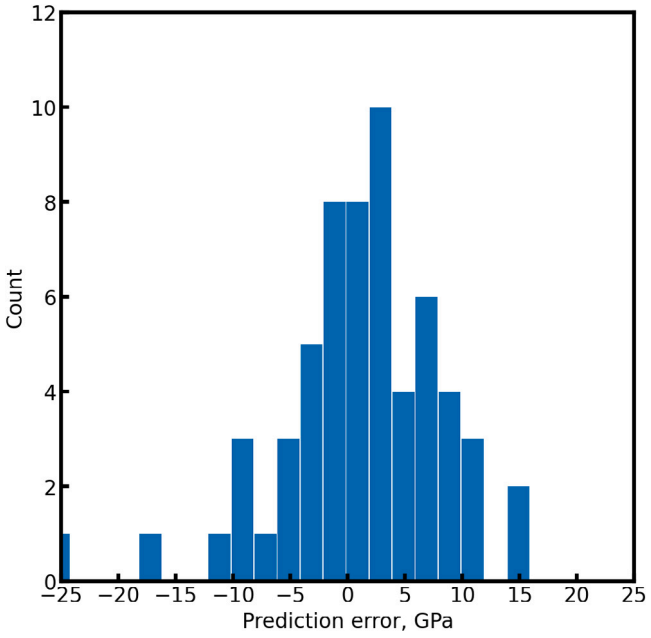

Fig. B.3. Distribution of CV10 errors for XGBoost model. Maximum absolute error is 25.6 GPa, RMSE is 7.8 GPa.

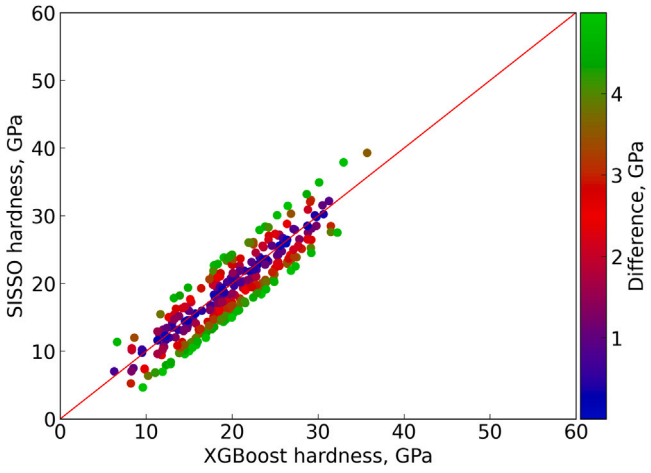

Fig. B.4. Correlation between SISSO hardness and XGBoost [33] model for considered stable carbides, borides and nitrides only. Colorbar shows the difference between two sets of data.

References

[1] V. Kanyanta, Hard, superhard and ultrahard materials: An overview, in: V. Kanyanta (Ed.), Microstructure-Property Correlations for Hard, Superhard, and Ultrahard Materials, Springer International Publishing, Cham, 2016, pp. 1–23.

[2] M. Kasonde, V. Kanyanta, Future of superhard material design, processing and manufacturing, in: V. Kanyanta (Ed.), Microstructure-Property Correlations for Hard, Superhard, and Ultrahard Materials, Springer International Publishing, Cham, 2016, pp. 211–239.

[3] J. Haines, J. Léger, G. Bocquillon, Synthesis and design of superhard materials, Annu. Rev. Mater. Res. 31 (2001) 1–23.

[4] E. Broitman, Indentation hardness measurements at macro-, micro-, and nanoscale: A critical overview, Tribol. Lett. 65 (1) (2016) 23.

[5] V.L. Solozhenko, E. Gregoryanz, Synthesis of superhard materials, Mater. Today 8 (11) (2005) 44–51.
[6] R.B. Kaner, J.J. Gilman, S.H. Tolbert, Designing superhard materials, Science 308 (5726) (2005) 1268–1269.
[7] V.L. Solozhenko, O.O. Kurakevych, D. Andrault, Y. Le Godec, M. Mezouar, Ultimate metastable solubility of boron in diamond: Synthesis of superhard diamondlike BC_5, Phys. Rev. Lett. 102 (1) (2009) 015506.
[8] V.L. Solozhenko, S.N. Dub, N.V. Novikov, Mechanical properties of cubic BC_2N, a new superhard phase, Diam. Relat. Mater. 10 (12) (2001) 2228–2231.
[9] E.S. Penev, N. Marzari, B.I. Yakobson, Theoretical prediction of two-dimensional materials, behavior, and properties, ACS Nano 15 (4) (2021) 5959–5976.
[10] D.M. Teter, Computational alchemy: The search for new superhard materials, MRS Bull. 23 (1) (1998) 22–27.
[11] X.-Q. Chen, H. Niu, D. Li, Y. Li, Modeling hardness of polycrystalline materials and bulk metallic glasses, Intermetallics 19 (9) (2011) 1275–1281.
[12] E. Mazhnik, A.R. Oganov, A model of hardness and fracture toughness of solids, J. Appl. Phys. 126 (12) (2019) 125109.
[13] E.V. Podryabinkin, A.G. Kvashnin, M. Asgarpour, I.I. Maslenikov, D.A. Ovsyannikov, P.B. Sorokin, M.Y. Popov, A.V. Shapeev, Nanohardness from first principles with active learning on atomic environments, J. Chem. Theory Comput. 18 (2) (2022) 1109–1121.
[14] R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, L.M. Ghiringhelli, SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates, Phys. Rev. Mater. 2 (8) (2018) 083802.
[15] Z.-K. Han, D. Sarker, R. Ouyang, A. Mazheika, Y. Gao, S.V. Levchenko, Single-atom alloy catalysts designed by first-principles calculations and artificial intelligence, Nature Commun. 12 (1) (2021) 1833.
[16] R. Ouyang, E. Ahmetcik, C. Carbogno, M. Scheffler, L.M. Ghiringhelli, Simultaneous learning of several materials properties from incomplete databases with multi-task SISSO, J. Phys. Mater. 2 (2) (2019) 024002.
[17] T.A. Purcell, M. Scheffler, L.M. Ghiringhelli, Recent advances in the SISSO method and their implementation in the sisso++ code, J. Chem. Phys. 159 (11) (2023).
[18] Y. Xu, Q. Qian, I-SISSO: Mutual information-based improved sure independent screening and sparsifying operator algorithm, Eng. Appl. Artif. Intell. 116 (2022) 105442.
[19] A. Wei, H. Ye, Z. Guo, J. Xiong, SISSO-assisted prediction and design of mechanical properties of porous graphene with a uniform nanopore array, Nanoscale Adv. 4 (5) (2022) 1455–1463.
[20] H. PourAsiabi, H. PourAsiabi, Z. AmirZadeh, M. BabaZadeh, Development a multi-layer perceptron artificial neural network model to estimate the Vickers hardness of Mn–Ni–Cu–Mo austempered ductile iron, Mater. Des. 35 (2012) 782–789.
[21] J. Sembiring, A. Amanov, Y. Pyun, Artificial neural network-based prediction model of residual stress and hardness of nickel-based alloys for UNSM parameters optimization, Mater. Today Commun. 25 (2020) 101391.
[22] A.F. Abd El-Rehim, H.Y. Zahran, D.M. Habashy, H.M. Al-Masoud, Simulation and prediction of the vickers hardness of AZ91 magnesium alloy using artificial neural network model, Crystals 10 (4) (2020) 290.
[23] W. Vermeulen, P. Van der Wolk, A. De Weijer, S. Van Der Zwaag, Prediction of Jominy hardness profiles of steels using artificial neural networks, J. Mater. Eng. Perform. 5 (1996) 57–63.
[24] R. Shankar, K. Senthilkumar, P. Kolandaivel, Calculation of ionization potential and chemical hardness: A comparative study of different methods, Int. J. Quantum Chem. 109 (4) (2009) 764–771.
[25] J.A. Dean, Lange's Handbook of Chemistry, McGraw-Hill, New York, N.Y., 1999, OCLC: 473528388.
[26] A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.a. Persson, The materials project: A materials genome approach to accelerating materials innovation, APL Mater. 1 (1) (2013) 011002.
[27] Z. Zhang, J. Brgoch, Determining temperature-dependent vickers hardness with machine learning, J. Phys. Chem. Lett. 12 (29) (2021) 6760–6766.
[28] O.L. Anderson, A simplified method for calculating the debye temperature from elastic constants, J. Phys. Chem. Solids 24 (7) (1963) 909–917.
[29] R. Hill, The elastic behaviour of a crystalline aggregate, Proc. Phys. Soc. A 65 (5) (1952) 349.
[30] S.P. Ong, W. Davidson Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V.L. Chevrier, K.A. Persson, G. Ceder, Python materials genomics (pymatgen): A robust, open-source python library for materials analysis, Comput. Mater. Sci. 68 (2013) 314–319.
[31] C. Xie, Q. Zhang, H.A. Zakaryan, H. Wan, N. Liu, A.G. Kvashnin, A.R. Oganov, Stable and hard hafnium borides: A first-principles study, J. Appl. Phys. 125 (20) (2019) 205109.
[32] A.G. Kvashnin, Z. Allahyari, A.R. Oganov, Computational discovery of hard and superhard materials, J. Appl. Phys. 126 (4) (2019) 040901.
[33] Z. Zhang, A. Mansouri Tehrani, A.O. Oliynyk, B. Day, J. Brgoch, Finding the next superhard material through ensemble learning, Adv. Mater. 33 (5) (2021) 2005112.

Dr. Christian Tantardini earned his Ph.D. in Materials Science & Engineering from the Skolkovo Institute of Science and Technology in 2020, under the supervision of Prof. Artem Oganov. His thesis focused on quantum chemical methods to understand chemical bonding. Notably, he reformulated Pauling's thermochemical electronegativity scale, overcoming its drawbacks, and published his findings in *Nature Communications* in 2021. This work has garnered over 200 citations, demonstrating its significant impact in chemistry, materials science, and physics. After his Ph.D., he joined Prof. Xavier Gonze's group at Skolkovo and at the same time working at the Institute of Solid State and Mechanochemistry SB RAS in Novosibirsk, contributing to projects like Abinit and PseudoDojo by developing specific norm-conserving pseudopotentials to study actinides under high pressure. Currently, he is a postdoc at UiT The Arctic University of Norway, working on numerical resolution approaches based on multiwavelets, and a visiting scientist at Rice University, focusing on first-principles and machine learning approaches for materials science with a specific interest in magnetic materials.

Dr. Hayk A. Zakaryan earned his Ph.D. in 2017 from Yerevan State University, where his research focused on the atomic-scale modeling of gas adsorption mechanisms on semiconductor surfaces. Post-Ph.D., he delved into groundbreaking work involving the discovery of novel superhard materials through the application of evolutionary algorithms and the design of innovative battery materials. In 2022, the Laboratory of Computational Materials Science was established, with Dr. Zakaryan as its head. Under his guidance, the laboratory expanded its scope to encompass not only superhard and battery materials but also dedicated efforts to explore 2D materials for catalytic applications.

Prof. Dr. Zhong-Kang Han is currently a researcher and doctoral supervisor in the "Hundred Talents Plan" of the School of Materials Science and Engineering at Zhejiang University. He obtained his Ph.D. from the University of Chinese Academy of Sciences in 2018 and then conducted postdoctoral research at the Fritz-Haber-Institute of the Max Planck Society in Germany and the Skolkovo Research Center in Russia. In April 2022, he joined the Institute of Materials Physics, School of Materials Science and Engineering at Zhejiang University. His research focuses on the intelligent design of advanced materials through multiscale computational simulation combined with machine learning, studying the evolution rules of material surface and interface structures, structural phase diagrams, and phase transition processes under real reaction conditions. He has published more than 40 papers in renowned journals such as *Physical Review Letters*, *Nature Communications*, *Angew. Chem.*, and *Advanced Functional Materials*. His research combines first-principles calculations and interpretable machine learning to conduct high-throughput screening and rational design of advanced materials, revealing key factors affecting material performance. He adds advanced machine learning algorithms to the cluster expansion model to study material structure phase diagrams and phase change processes under real reaction conditions, further revealing the intrinsic relationship between material structure and performance. He is also building a small-scale but practical material database for a specific material system.

Prof. Dr. Tariq Altalhi is an Associated Professor in the Department of Chemistry at Taif University, Saudi Arabia. He received his Ph.D. from the University of Adelaide, Australia, in 2014, earning the Dean's Commendation for Doctoral Thesis Excellence. His research interests lie in developing advanced chemistry-based solutions for solid and liquid municipal waste management, both organic and inorganic. He has worked on transforming solid organic

waste into valuable nanomaterials and economic nanostructures, such as converting plastic bags into carbon nanotubes during his Ph.D. studies and turning fly ash into efficient adsorbent materials. Another area of interest is natural extracts and their application in generating value-added products, such as nanomaterials and essences. Through his work as an independent researcher, he has developed strong management and mentoring skills, leading a group of multidisciplinary researchers in fields including chemistry, materials science, biology, and pharmaceutical science. His publications demonstrate that he has established a wide network of national and international researchers who are leaders in their respective fields. Additionally, he has formed key contacts with major industries in the Kingdom of Saudi Arabia.

Prof. Dr. Sergey V. Levchenko obtained his M.Sc. from the Moscow Institute of Physics and Technology and his Ph.D. from the University of Southern California, LA, USA, in 2005. After a postdoc period at the University of Pennsylvania, PA, USA, he was a group leader at the Fritz Haber Institute of the Max Planck Society in Berlin, Germany. Since 2018, he has been an associate professor at Skoltech, Moscow, Russia. Levchenko's main areas of interest include point defect formation in bulk and at surfaces of crystals under realistic temperature, pressure, and doping conditions, adsorption and chemical reactions at surfaces, heterogeneous catalysis for methane and carbon dioxide conversion, electrocatalysis for hydrogen production, and the development of symbolic-regression and data-mining artificial intelligence approaches for designing catalytic and other functional materials.

Prof. Dr. Alexander G. Kvashnin received a master's degree in applied mathematics and physics from the Moscow Institute of Physics and Technology in 2012 and then became a Ph.D. student at the same university. He also worked as a visiting scientist at Rice University in the USA in 2011 and 2013. Since 2015, he has been a research scientist at the Skolkovo Institute of Science and Technology. In 2016, Alexander obtained his Ph.D. in Condensed Matter Physics from the National University of Science and Technology MISIS. In 2021, he successfully defended his habilitation thesis devoted to high-temperature superconductivity and received the habilitation degree in Condensed Matter Physics. Since 2023, Alexander holds a position of a Full Professor at Skoltech, leading the Industry-Oriented Computational Discovery lab. In 2023, Alexander was listed among the world's top 2 % scientists based on their impact, as compiled by Elsevier.

Prof. Dr. Boris I. Yakobson is the Karl F. Hasselmann Chair in Engineering. Dr. Yakobson holds a joint appointment between the Department of Materials Science and Nano-Engineering and the Department of Chemistry. In 2008, Yakobson received a Nano 50 Award from the science magazine Nanotech Briefs for his innovation in nanotechnology, and in 2009, the Department of Energy R&D Award. He received his Ph.D. in 1982 from the Russian Academy of Sciences. Dr. Yakobson is an editorial board member of the Journal of Nanoparticle Research and a member of the American Physical Society and the Electrochemical Society. Karl F. Hasselmann was a pioneer in offshore oil and gas exploration and a philanthropist. Hasselmann was a close friend of Malcolm Lovett Sr., Chairman of the Rice Board of Trustees from 1967–1972. Yakobson's research interests are in the theory and modeling of the structure, kinetics, and properties of materials derived from macroscopic and fundamental molecular interactions. He has done groundbreaking work on the physical properties of nanotubes, particularly their electromechanics, and recently with graphene and graphane.